**Psychological Test and Assessment Modeling**, 2024.66:116-143
DOI: https://doi.org/10.2440/001-0015

# Parameter Recovery for the Four-Parameter Item Response Model: A Comparison of Marginal Maximum Likelihood and Markov Chain Monte Carlo Approaches

*Hoan Do*[1] *& Gordon P. Brooks*[2]

[1] Clinical Outcomes Solutions, Tucson AZ
[2] Department of Educational Studies, Patton College of Education, Ohio University

***Abstract***:
This study assessed the parameter recovery accuracy of Marginal Maximum Likelihood (MML) and two Markov Chain Monte Carlo (MCMC) methods, Gibbs and Hamiltonian Monte Carlo (HMC), under the four-parameter unidimensional binary item response function. Data were simulated under the mixed factorial design with sample size (1,000; 2,500; and 5,000 respondents) and latent trait distribution (normal and negatively skewed) as the between-subjects factors, and estimation method (MML, Gibbs, and HMC) as the within-subjects factor. Results indicated that in general, MML was more heavily impacted by latent trait skewness, but MML also improved its performance more strongly than MCMC when sample size increased. Two MCMC methods remained advantageous with lower root mean square errors (RMSE) of item parameter recovery across all conditions under investigation, but sample size increase brought a correspondingly narrower gap between MML and MCMC regardless of theta distributions. Gibbs and HMC provided nearly identical outcomes across all conditions, and no considerable difference between these two MCMC methods was detected. Sample size and latent trait distribution had little observable effect on trait score estimation by MCMC and Expected a Posteriori following MML (MML-EAP), which were essentially unbiased and had similar RMSE across all conditions. Discussions of the findings and model calibration issues are presented together with practical implications and future research recommendations.

*Keywords:* the four-parameter IRT model, Marginal Maximum Likelihood (MML), Markov chain Monte Carlo (MCMC), Gibbs sampling, Hamiltonian Monte Carlo (HMC)

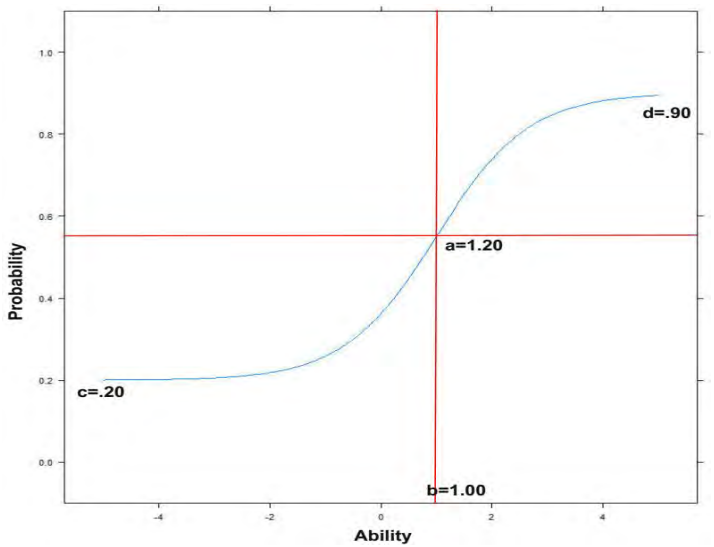
**Correspondence:**

Dr. Hoan Do, Ph.D., Research Scientist – Quantitative Sciences, Clinical Outcomes Solutions, Tucson AZ 85718, td898911@ohio.edu

The four-parameter item response model (4PM; Barton & Lord, 1981) expands on the three-parameter model with an upper asymptote less than one to capture the non-zero probability of "slip" among proficient test-takers. Under the four-parameter logistic function, the probability of the correct answer to an item is $P(x = 1 \mid \theta_i, a_j, b_j, c_j, d_j) = c_j + (d_j - c_j) \frac{e^{a_j(\theta_i - b_j)}}{1 + e^{a_j(\theta_i - b_j)}}$, in which $\theta_i$ is the latent trait score of examinee $i$, and $a_j$, $b_j$, $c_j$, $d_j$ are discrimination, difficulty, lower asymptote, and upper asymptote parameters of item $j$, respectively. Figure 1 below illustrates the characteristic curve of a hypothetical test item modeled with the 4PM. This example item is characterized by four parameters: difficulty parameter $b$ = 1.00, discrimination parameter $a$ = 1.20, lower asymptote parameter $c$ = .20, and upper asymptote parameter $d$ = .90. The upper asymptote implies the chance of endorsing the wrong answer to this item, even for very high ability examinees, equals 1- $d$ (i.e., chance of slipping equals .10 or 10 %).

Figure 1

*Characteristic Curve of a Hypothetical Item Modeled with the 4PM*




In recent years, the 4PM has rekindled research interest and found its utility in a variety of fields, including computerized adaptive testing (CAT; Liao et al., 2012; Rulison & Loken, 2009; Yen, Ho, Liao, & Chen, 2012; Yen, Ho, Laio, et al., 2012), cognitive appraisal such as financial literacy, mathematics, reading, and physics testing (Barnard-Brak et al., 2018; Culpepper, 2017; Sideridis et al., 2016; Walstad & Rebeck, 2017), health behavior assessment (Culpepper, 2016; Loken & Rulison, 2010), food security (Gregory, 2019), psychopathology and personality assessment (Feuerstahler & Waller, 2014; Waller & Feuerstahler, 2017; Waller & Reise, 2010), genetics

research (Tavares et al., 2004), and potential applications in other areas (Myszkowski & Storme, 2017; Primi et al., 2018; Storme et al., 2019). The interpretations of the d parameter and its usefulness vary across disciplines. For example, Gregory (2019) used the upper asymptote less than one to model the underreporting of food insecurity in adults and children; and in genetics research, Tavares (2004) proposed that the *d* parameter below one allows the possibility that a gene is not active in persons with a high predisposition to a disease. In educational assessment, general conclusions on the values of the 4PM involve measurement efficiency improvement, and reduced imprecision for scores among high-achieving students and among test-takers who make early mistakes in CAT (Culpepper, 2017; Rulison & Loken, 2009; Yen, Ho, Liao, & Chen, 2012).

As the 4PM is gaining a stronger foothold in item response theory (IRT) applications, theoretical discussions have also emerged. Sijtsma and Hemker (2000) showed that the 4PM shares the properties of stochastic ordering of latent trait based on unweighted sum of scores and non-invariant item ordering with the 2PM and 3PM. Ogasawara (2012) extended the previous work of Lord (1983) by deriving the asymptotic approximations of ability estimator, and Ogasawara (2017) discussed identified and unidentified cases of the fixed-effects 4PM. Tendeiro and Meijer (2012) recommended the 4PM to model test anxiety as a specific form of aberrant behavior in item responses, and Magis (2013) delineated the information function for the 4PM. More recent developments in the literature directly focused on examinations of the 4PM identifiability and estimation. Culpepper (2016) presented the full conditionals for the 4-parameter ogive model for Gibbs procedure and developed a package to estimate this model in R (R Core Team, 2020). Kern and Culpepper (2020) introduced the Dyad-4PM, in which items divided into groups of two (dyads) load on one latent binary attribute, and showed that this model is identified. Zhang et al. (2020) developed the Gibbs-slice sampling algorithm for estimation of the 4PM with two steps, one to update the upper and lower asymptote parameters with truncated beta distribution conjugation, and the other to slice sample for the item discrimination and difficulty parameters with different auxiliary variables. Meng et al. (2020) reformulated the 4PM as a mixture model and estimated it with marginalized maximum a posteriori via a newly developed Expectation-Maximization (EM) algorithm.

Like other IRT models, the 4PM is useful only when person and item latency is well understood in the form of accurately estimated parameters. In applied measurement with the 4PM, both Marginal Maximum Likelihood (MML) and Markov Chain Monte Carlo (MCMC) within the Bayesian framework have been used (Barnard-Brak et al., 2018; Culpepper, 2016, 2017; Feuerstahler & Waller, 2014; Waller & Reise, 2010; Walstad & Rebeck, 2017), and their statistical properties have been investigated individually (Culpepper, 2016; Feuerstahler & Waller, 2014; Loken & Rulison, 2010; Sheng, 2015; Waller & Feuerstahler, 2017). The coexistence of multiple estimation approaches logically motivates researchers to pose questions about how the available methods compare to one another under various measurement conditions. Extensive research comparing the properties of MML and MCMC estimations has been reported on a number of IRT models, such as Rasch model (Kim, 2001), two-parameter model

(Baker, 1998), three-parameter model (Béguin & Glas, 2001), two-parameter testlet model (Luo & Wolf, 2019), graded response model (Kieftenbeld & Natesan, 2012; Kuo & Sheng, 2016), and nominal response model (Wollack et al., 2002). A similar question about merits of different estimation methods for the 4PM awaits exploration. This study will fill this gap.

The purpose of this investigation is to evaluate the quality of parameter recovery for the 4PM by MML with EM algorithm and two MCMC sampling mechanisms, Gibbs and Hamiltonian Monte Carlo (HMC), within a fully Bayesian framework under several measurement conditions. Specifically, answers to the following questions are sought:

Question 1: How accurate are MML, Gibbs and HMC estimations of person and item parameters for the 4PM across latent trait distributions?

Question 2: How accurate are MML, Gibbs and HMC estimations of person and item parameters for the 4PM across sample size levels?

## Methods

### Design

This simulation was designed as a mixed factorial study with two between-subjects factors and one within-subjects factor. Two between-subjects factors were latent distribution (two conditions: normality and negative skewness) and sample size (three levels: 1,000; 2,500; and 5,000 respondents). The within-subjects factor was estimation method with three types: MML, Gibbs, and HMC. This 2x3x3 mixed factorial design with two fully crossed factors resulted in 18 unique combinations of estimation features (18 cells). The treatment of estimation method as a within-subjects factor produced more accurate comparison because all three procedures were subject to identical random sampling variations in the data generation process. The number of items was fixed at 20, a rather common scale length in the 4PM research and applications (Culpepper, 2016; Sheng, 2015; Sideridis et al. 2016; Storme et al., 2019; Waller & Feuerstahler, 2017).

### Data Generation

Item parameters were generated under the assumption that tests are well-designed, and few parameters are beyond their typical and useful range: $a \sim N(1.2, 0.35^2)$ truncated at 0, $b \sim N(0,1^2)$, $c \sim N(0.15, 0.04^2)$ truncated at 0, and $d \sim N(0.88, 0.04^2)$ truncated at 1. The additional constraint $c < d$ was placed on asymptote parameters to further ensure item quality and appropriate representation of the 4PM item response function. Person parameters were generated to represent two scenarios. First, $\theta$ was

drawn from $N(0,1^2)$ under the assumption that the test-taker sample comes from a normally distributed population. Second, to represent latent traits when a relatively easy test is administered (Lord, 1955; Sass et al., 2008), beta(8.2, 2.8) was employed to construct a negatively skewed distribution with moderate skewness of -0.60. This skewness level is within the range of median skewness values for IRT scale scores across state-level tests reported by Ho and Yu (2015). Random θ draws were scaled to have mean of 0 and variance of 1 before being inputted in the item response generation process. To reduce sampling errors and ensure diverse response patterns simultaneously, this study adopted Feinberg and Rubright's (2016) strategy: in each cell, only one θ sample is drawn and fixed for all replications, but item parameters differ across replications. A random seed was set for each cell so that the person and item parameter values would vary across cells. The generated parameters were used to simulate response data under the 4PM with the *simdata* function in the mirt package (Chalmers, 2012) in R (R Core Team, 2020).

## Replications

In light of previous research into parameter recovery under the four-parameter item response function, and the time- and computer resource-consuming nature of MCMC (Feinberg & Rubright, 2016; Harwell et al., 1996), 50 replications were set for each cell in this simulation. With this design, a total of 300 unique data sets were generated and 900 sets of results were obtained from three estimation procedures, with 72,000 estimates for item parameters and 555,000 estimates for theta.

## Model Estimation

Item parameter estimation with MML was performed with the mirt package (Chalmers, 2012) in R (R Core Team, 2020) using 41 quadrature points to ensure little estimation bias (Cao et al., 2014; Kim & Lee, 2017; Kim & Moses, 2016; Seong, 1990; Sinharay & von Davier, 2005; van Rijn, 2014). Subsequent to item calibration, the person scoring phase was executed with Expected A Posteriori (EAP) using $N(0, 1.2^2)$ as θ prior.

Bayesian estimation via MCMC was performed with following priors: $a \sim LN(0, 0.2^2)$ where LN indicates the lognormal distribution, $b \sim N(0, 1.3^2)$, $c \sim$ beta(2, 10), $d \sim$ beta(10, 2), and $\theta \sim N(0, 1.2^2)$. For both Gibbs and HMC, four parallel Markov chains were run with 50,000 iterations per chain and the initial burn-in (warm-up in rstan) of 20,000 iterations was discarded. No chain thinning was performed due to its adverse effect on estimation accuracy (Link & Eaton, 2012). Gibbs sampling was executed with JAGS (Plummer, 2003) via the R interface package runjags (Denwood, 2016), and HMC was conducted using Stan via its R interface package rstan (Carpenter et al., 2017; Stan Development Team, 2018a, 2018b). MCMC modeling was run in

several office computers with Windows 10 and in the Ohio Supercomputer Center with Linux using batch mode.

## Model Assessment

For MML, the convergence threshold of 0.001 was set for all replications. To accommodate possible convergence failures in the 4PM estimation, only data sets which resulted in identified models and successful estimation convergence in MML were fed to Gibbs and HMC until a predetermined number of successful calibrations (i.e., replications) in each cell were reached. It should be noted that both technical and practical convergence in MML were required. Pilot simulations indicated that MML frequently failed to practically converge (i.e., produce sound IRT estimates) despite technical convergence, especially for small samples. Thus, additional criteria adopted from Waller & Feuerstahler (2017) to ensure reasonable parameter estimates ($\hat{a} \leq 3$, $-6 \leq \hat{b} \leq 6$, $\hat{c} < \hat{d}$) were imposed on selected data.

For MCMC, the potential scale reduction factor (PSRF) and its multivariate counterpart (MPSRF) were used to assess Markov chain convergence using the conservative threshold of 1.05 (Brooks & Gelman, 1998). The MPSRF was used for the joint posterior distribution of item parameters only, because person parameters tend to be accurately estimated and rarely have convergence issues (Sinharay, 2004).

## Outcome Analysis

Two common measures of the difference between true parameters and their estimates were used as outcome variables: bias, which represents the systematic discrepancy between the true and estimated parameter, and root mean square error (RMSE), which denotes total estimation error (Feinberg & Rubright, 2016; Mooney, 1997). Average bias and RMSE were calculated using the following equations:

For item parameters

$$\text{bias} = \frac{1}{R}\frac{1}{L}\sum_{r=1}^{R}\sum_{j=1}^{L}(\widehat{x_{jr}} - x_T)$$

$$\text{RMSE} = \frac{1}{R-1}\sum_{r=1}^{R}\sqrt{\frac{\sum_{j=1}^{L}(\widehat{x_{jr}} - x_T)^2}{L}}$$

For person parameters

$$\text{bias} = \frac{1}{R}\frac{1}{n}\sum_{r=1}^{R}\sum_{i=1}^{n}(\widehat{x_{ir}} - x_T)$$

$$\text{RMSE} = \frac{1}{R-1}\sum_{r=1}^{R}\sqrt{\frac{\sum_{i=1}^{n}(\widehat{x_{ir}} - x_T)^2}{n}}$$

in which $x_T$ is the true parameter value (either person or item parameter), $\widehat{x_{jr}}$ is the estimated parameter of item $j$ in replication $r$, $\widehat{x_{ir}}$ is the estimated parameter of person $i$ in replication $r$, $R$ is the number of replications, $L$ is scale length, and $n$ is the number of simulated examinees in each data set.

## Results

Both technical and practical convergence in MML were found to follow the same patterns of improving with increased sample size and deteriorating with the violation of θ normality assumption. Given technical convergence in mirt, MML successfully produced reasonable estimates less than 1 % of the time at $n$ = 1,000 for both latent distributions. At $n$ = 5,000, MML plausible estimation, conditional on technical convergence, jumped to 39 % under skewed θ and 73 % under normal θ. Subsequent to data filtering with MML, selected item response matrices were subject to calibrations with Gibbs and HMC. For both MCMC methods, all PSRF and MPSRF estimates were far below the conservative cut point of 1.05, which suggested Markov chains reached a stationary state and mixed thoroughly.

Bias and RMSE across conditions are displayed in Table 1, and boxplots in Figures 2-6 depict estimation bias to aid examinations. In general, MML took a more substantive impact of latent trait skewness but also absorbed the momentum from sample size increase to improve its performance more strongly than MCMC. Two MCMC methods remained advantageous with lower RMSE of item parameter recovery across all conditions under investigation, but sample size increase brought a correspondingly narrower gap between MML and MCMC regardless of latent trait condition. Gibbs and HMC provided nearly identical outcomes across all conditions, and no considerable difference between the two MCMC methods was detected.

Table 1

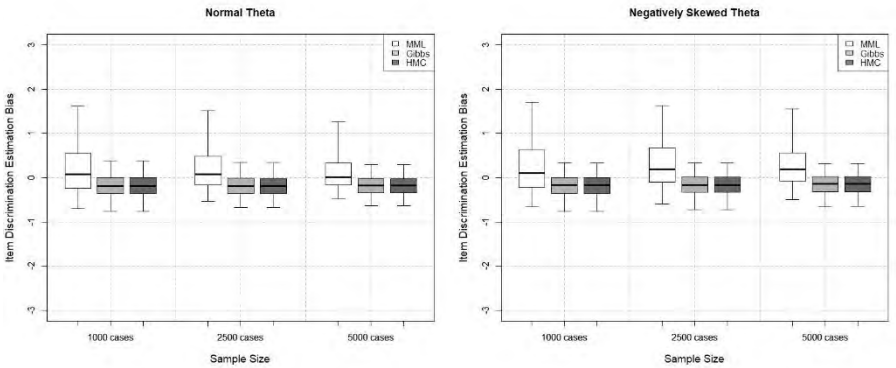
*Mean Bias and RMSE of Item and Person Parameter Estimations by MML, Gibbs and HMC across Sample Size Levels and Latent Trait Distributions*

| Parameter | Estimator | Bias | | | RMSE | | |
|---|---|---|---|---|---|---|---|
| | | *n* = 1,000 | *n* = 2,500 | *n* = 5,000 | *n* = 1,000 | *n* = 2,500 | *n* = 5,000 |
| *a* | MML | 0.2047 | 0.1971 | 0.1282 | 0.6323 | 0.5671 | 0.4707 |
| | | 0.2503 | 0.3172 | 0.2844 | 0.6733 | 0.6689 | 0.5959 |
| | Gibbs | -0.1886 | -0.1846 | -0.1837 | 0.3389 | 0.3150 | 0.2998 |
| | | -0.1783 | -0.1688 | -0.1516 | 0.3289 | 0.3157 | 0.2918 |
| | HMC | -0.1887 | -0.1846 | -0.1836 | 0.3389 | 0.3151 | 0.2999 |
| | | -0.1782 | -0.1687 | -0.1516 | 0.3289 | 0.3156 | 0.2918 |
| *b* | MML | -0.0573 | 0.0110 | 0.0026 | 0.5833 | 0.4244 | 0.3597 |
| | | -0.2486 | -0.2358 | -0.2558 | 0.6288 | 0.5614 | 0.4954 |
| | Gibbs | -0.0757 | -0.0116 | -0.0228 | 0.2985 | 0.2764 | 0.2704 |
| | | -0.1827 | -0.2254 | -0.2756 | 0.3421 | 0.3747 | 0.3963 |
| | HMC | -0.0756 | -0.0112 | -0.0225 | 0.2983 | 0.2765 | 0.2691 |
| | | -0.1828 | -0.2251 | -0.2756 | 0.3423 | 0.3746 | 0.3964 |
| *c* | MML | -0.0219 | -0.0004 | -0.0033 | 0.1204 | 0.1013 | 0.0927 |
| | | -0.0502 | -0.0250 | -0.0282 | 0.1204 | 0.1046 | 0.0871 |
| | Gibbs | 0.0007 | 0.0034 | 0.0047 | 0.0537 | 0.0515 | 0.0545 |
| | | -0.0105 | -0.0173 | -0.0225 | 0.0527 | 0.0546 | 0.0496 |
| | HMC | 0.0007 | 0.0035 | 0.0048 | 0.0537 | 0.0515 | 0.0544 |
| | | -0.0105 | -0.0172 | -0.0225 | 0.0527 | 0.0547 | 0.0497 |
| *d* | MML | -0.0022 | -0.0006 | 0.0012 | 0.1179 | 0.1003 | 0.0871 |
| | | -0.0346 | -0.0561 | -0.0648 | 0.1358 | 0.1267 | 0.1190 |
| | Gibbs | -0.0235 | -0.0165 | -0.0116 | 0.0642 | 0.0589 | 0.0541 |
| | | -0.0381 | -0.0420 | -0.0518 | 0.0728 | 0.0775 | 0.0843 |

*Note*. Outcomes under normal theta are displayed in unshaded areas. Outcomes under negatively skewed theta are shaded

Figure 2

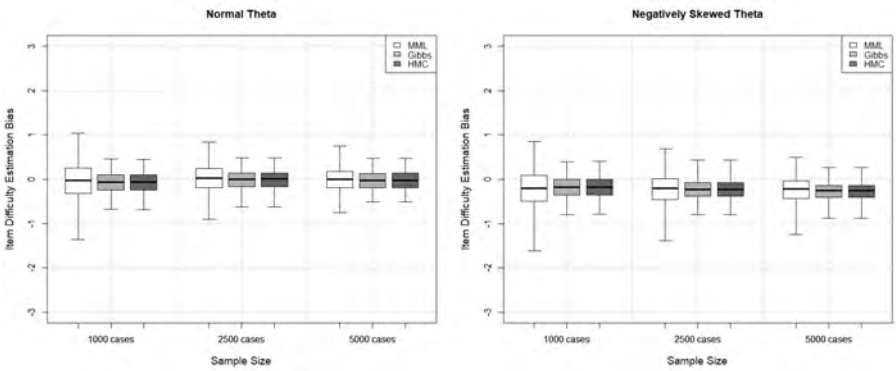
*Bias in Estimating the 4PM Item Discrimination by MML and MCMC across Sample Size and Latent Trait Conditions*




*Note*. Boxplot whiskers display 2.5 to 97.5 quantiles.

Figure 3

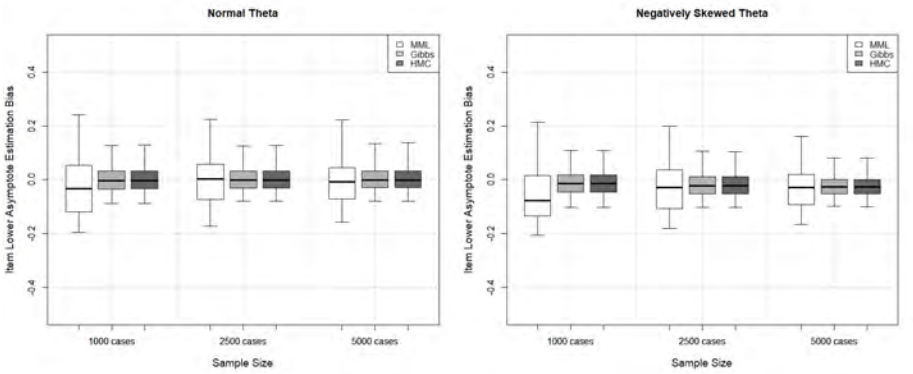
*Bias in Estimating the 4PM Item Difficulty by MML and MCMC across Sample Size and Latent Trait Conditions*




*Note*. Boxplot whiskers display 2.5 to 97.5 quantiles.

Figure 4

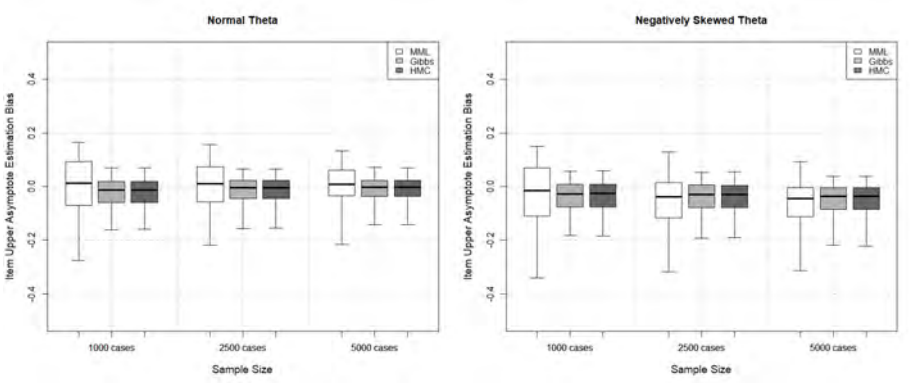
*Bias in Estimating the 4PM Item Lower Asymptote by MML and MCMC across Sample Size and Latent Trait Conditions*




*Note*. Boxplot whiskers display 2.5 to 97.5 quantiles.

Figure 5

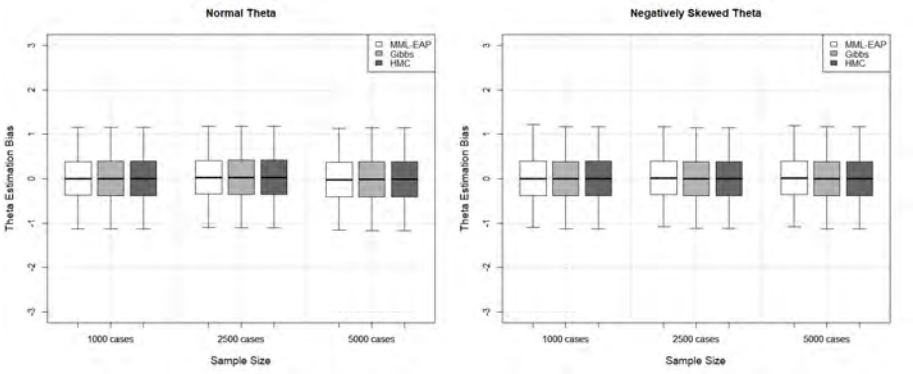
*Bias in Estimating the 4PM Item Upper Asymptote by MML and MCMC across Sample Size and Latent Trait Conditions*




*Note*. Boxplot whiskers display 2.5 to 97.5 quantiles.

Figure 6

*Bias in Estimating the 4PM Latent Trait Score by MML-EAP and MCMC across Sample Size and Latent Trait Conditions*




*Note*. Boxplot whiskers display 2.5 to 97.5 quantiles.

Specifically, when θs were generated from a normal distribution, MML and MCMC estimated the *b*, *c,* and *d* parameters with little mean bias, even at $n$ = 1,000. Estimates of the *a* parameter were positively biased for MML and negatively biased for MCMC, and mean bias by all methods was larger than 0.10 in absolute value even at $n$ = 5,000. While MML item parameter recovery had lower mean bias than Gibbs and HMC at $n$ = 5,000, two MCMC methods remained superior with smaller RMSE. Under normal θ, all methods consistently improved RMSE of item parameter recovery in conjunction with sample size increase, except for MCMC estimation of the *c* parameter which did not exhibit a clear trend.

When latent trait scores were skewed to the left, there was a concomitant deterioration in the quality of item parameter recovery by both MML and MCMC generally. Under skewed θ, MML had total errors of item parameter recovery diminished as more examinees took a test, yet sample size increase did not appear to benefit mean bias. Indeed, MML became increasingly negatively biased in its estimation of the *d* parameter as sample size increased, and mean biases of estimating other item parameters remained considerably large at $n$ = 5,000. For Gibbs and HMC, sample size increase under skewed θ benefited only mean bias of item slopes recovery while rendering their estimation of other item parameters more negatively biased. In addition, unlike MML, there was a slight increase in the *b* and *d* parameter estimation errors by two MCMC methods as more cases were drawn from a skewed θ distribution. Overall, MCMC still maintained its advantage over MML with lower total errors in recovering item parameters under skewed θ.

Sample size and latent trait distribution had little observable effect on person parameter recovery on average. Both MML-EAP and MCMC were essentially unbiased and

had similar RMSE of trait score estimation across all conditions. Spearman's correlations between true and estimated θs fluctuated between .82 and .84 for three methods across all conditions, indicating that MML-EPA and MCMC had similarly strong preservation of respondents' rankings despite changes in sample size and latent distribution.

## Follow-up Simulation 1

We followed up the main findings to explore how Gibbs and HMC would perform when MML estimation was technically successful in mirt but practically unsuccessful (i.e., when MML offers severely inflated/deflated item parameter estimates). To answer this question, additional simulations for $n$ = 1,000 under both normal and negatively skewed θ distributions were conducted. Simulated data sets were selected if they allowed technical convergence in mirt, and no additional constraints were imposed on MML estimates. Model configurations for MML and MCMC were identical to those described earlier. Results (Table 2) indicated that when MML technically converged but failed at practically sound estimates, both Gibbs and HMC continued to perform stably. No appreciable difference in MCMC estimations was found when compared to the findings in Table 1. On the contrary, in addition to the implausible item discrimination estimates (14 % of the $a$ parameter estimates were larger than 5.0), there was one case of anomalous item difficulty estimate of -84.00 by MML when the generated $b$ value was -3.25, which severely worsened mean bias and RMSE for its item difficulty recovery.

Table 2

*Calibrations by MML and MCMC at n = 1,000 when MML Only Converged Technically*

| Parameter | Estimator | Bias | RMSE |
|---|---|---|---|
| *a* | MML | 1.3126 | 2.8162 |
| | | 1.9651 | 3.9274 |
| | Gibbs | -0.1747 | 0.3341 |
| | | -0.1790 | 0.3352 |
| | HMC | -0.1748 | 0.3341 |
| | | -0.1789 | 0.3352 |
| *b* | MML | -0.0757 | 2.6412 |
| | | -0.2419 | 0.7321 |
| | Gibbs | -0.0458 | 0.3064 |
| | | -0.1900 | 0.3564 |
| | HMC | -0.0460 | 0.3064 |
| | | -0.1902 | 0.3561 |
| *c* | MML | 0.0176 | 0.1476 |
| | | 0.0058 | 0.1402 |
| | Gibbs | 0.0066 | 0.0575 |
| | | -0.0043 | 0.0528 |
| | HMC | 0.0066 | 0.0574 |
| | | -0.0043 | 0.0529 |
| *d* | MML | -0.0266 | 0.1413 |
| | | -0.0772 | 0.1728 |
| | Gibbs | -0.0288 | 0.0659 |
| | | -0.0381 | 0.0775 |
| | HMC | -0.0288 | 0.0659 |
| | | -0.0382 | 0.0776 |
| $\theta$ | MML | 0.0203 | 0.6136 |
| | | 0.0198 | 0.6021 |
| | Gibbs | 0.0215 | 0.5989 |
| | | 0.0039 | 0.5875 |
| | HMC | 0.0214 | 0.5989 |
| | | 0.0038 | 0.5875 |

*Note*. Outcomes under normal theta are displayed in unshaded areas. Outcomes under negatively skewed theta are shaded.

## Follow-up Simulation 2

Because MML showed signs of continuous improvement as more examinees participated in a test, a follow-up simulation was conducted with $n = 10{,}000$ under identical model configurations to examine how MML would perform at a larger sample size. Outcomes were averaged across 50 replications and are reported in Table 3. With 10,000 examinees and normal θ, MML estimation accuracy became acceptable. Mean bias went down below 0.10 for item slopes, and notable reduction in RMSE were observed for all item parameters, especially item discrimination and item difficulty.

Equally important, absurd item discrimination estimation happened only 8.4 % among generated data at $n$ = 10,000, as opposed to 27 % at $n$ = 5,000. However, the improvement in item parameter recovery at $n$ = 10,000 under negatively skewed θ was minor generally, and MML appeared to have reached a point of diminished returns when a very large sample alone no longer remedied the adverse impact of the non-normal latent trait to a desirable degree.

Table 3

*Mean Bias and RMSE of Item and Person Parameter Estimation by MML at n =10,000*

| Parameter | Bias | RMSE |
|---|---|---|
| *a* | 0.0956 | 0.3531 |
| | 0.3173 | 0.5511 |
| *b* | 0.0388 | 0.3082 |
| | -0.2632 | 0.4718 |
| *c* | 0.0113 | 0.0833 |
| | -0.0076 | 0.0752 |
| *d* | -0.0046 | 0.0742 |
| | -0.0842 | 0.1287 |
| *θ* | 0.0087 | 0.5786 |
| | 0.0259 | 0.5726 |

*Note*. Outcomes under normal theta are displayed in unshaded areas. Outcomes under negatively skewed theta are shaded.

## Discussions

Careful inspections of the estimation results by MML and two MCMC methods across sample size levels and latent trait distributions gave rise to certain remarks. First, Lord (1986) noted, with mathematical proof, that given a useful prior density and the use of posterior mean as the point estimate of a parameter, Bayesian estimation would outperform MML in minimizing mean square error. This fact explains the consistent advantage of two MCMC methods with lower RMSE for item parameter recovery across all conditions. However, Bayesian estimation has an inherent trade-off between bias and RMSE in that securing minimal mean square error equals inflating bias (Lord, 1986). Our findings indicated that when θ normal assumption held, MML only lagged behind Gibbs and HMC in terms of mean bias at the lowest sample of 1,000 cases. At $n$ = 2,500, MML was less biased than MCMC in estimations of all item parameters except item slopes, and at $n$ = 5,000, MCMC was trailing MML in unbiasedness of recovering all item parameters.

It was also observed that as more respondents were available, MML made stronger progress than MCMC and narrowed the RMSE gap accordingly, regardless of the

shape of the parent θ distribution. These findings could be explained by the different consistency properties possessed by MML and MCMC estimates (Patz & Junker, 1999). It is well known that MML is based on the asymptotic theory, and MML item parameter estimates are consistent (i.e., approaching the true population parameters) as sample size increases and the number of items is constant (Ogasawara, 2012). MML enjoys the advantage of the separation between item calibration and person scoring, as more test-takers do not lead to an increase in the number of parameters to estimate. The consistency properties in MCMC, interestingly, depend on the way the Markov chain draws are used (Patz & Junker, 1999). When MCMC output is used to estimate both item (structural) and person (incidental) parameters, as is the case in this study, MCMC resembles Joint Maximum Likelihood in the lack of consistency because sample size leads to more parameters to estimate (Ogasawara, 2012; Patz & Junker, 1999).

Another interesting finding was under negatively skewed latent trait, larger sample sizes actually harmed MCMC estimation of the *b* and *d* parameters, albeit at a minor degree. It is important to note that compared to normal θ samples, skewed θ draws actually brought about two disadvantages to data calibrations. One was the violation of the underlying latent distribution assumption for MML, and for MCMC was the mismatch between the normal θ prior and its skewed parent distribution. The other detriment was the lack of high-ability respondents at the upper end of the latent scale. Specifically, the maximum possible θ value generated from the skewed distribution beta(8.2, 2.8) after z transformation is around 2.02. Examinations of the simulated data indicated that the largest θ drawn under the negatively skewed latent trait equaled 1.97 in the present study. The worsened performances in overall item calibrations by MML and MCMC under skewed theta could be attributed to the differences in these two factors. Within MCMC under skewed theta, however, the counterintuitive trend of increased RMSE of *b* and *d* estimations occurred when only sample size was designed to increase, and the mismatched prior was held constant (Table 1). A putative explanation for this phenomenon has to do with the interplay between the prior and data in Bayesian posterior reconstruction. Due to the nature of the skewed θ distribution mentioned above, sample size increase did not bring more respondents with latent trait scores larger than 2.02 (practically no larger than 1.97). The absence of larger θ values toward the right end of the latent continuum provided little information for accurate estimation of highly difficult items and particularly for the recovery of upper asymptote, which was shown to depend on the number of respondents with high latent trait scores (Culpepper, 2016). In the meantime, sample size increase led to the diminishing influence of priors overall. Thus, even as larger samples brought more information for item parameter recovery in general, the useful information associated with sample size increase under skewed θ was not as strong as its counterpart under normal θ to counter the declining effect of the *b* and *d* priors. Note that the rate of RMSE improvement slowed down for MML estimations of *b* and *d* under skewed θ as well, possibly because the lack of individuals at the upper tail of the θ continuum offset the overall value of sample size increase to some extent. The significant role of adequate information from data can also explain why little adverse impact of the negatively

skewed θ on the estimation of *c* was observed, presumably because the presence of more individuals in the lower tail of the latent scale (i.e., more respondents with low θ scores) considerably assisted accurate recovery of *c*. Examinations of the generated data revealed that there were 128 individuals with θ scores below -2.0 at $n$ = 5,000 under normal θ, whereas the corresponding number under negatively skewed θ was 190 examinees. At this largest sample size $n$ = 5,000, both MCMC and MML yielded lower RMSE in estimating *c* under skewed θ than under normal θ.

The low practical convergence rate in MML was notable in this study. Approximately one in every four samples failed to practically converge in MML under the most favorable conditions with $n$ = 5,000 and normal θ. At $n$ = 1,000, less than one percent of data sets generated were successfully calibrated with plausible MML estimates. Examinations of MML estimates revealed that out-of-bounds discrimination parameter estimates were the major culprit behind practical convergence failure in MML. Infinite estimates of item difficulty also appeared, albeit very rarely. Due to the infinite parameter estimation by MML even for less-parameterized models, suggestions to incorporate item priors (i.e., turn MML into Bayesian Modal Estimation [BME; Mislevy, 1986]) have been made and adopted in popular IRT software programs like BILOG and MULTILOG (Rupp, 2003). However, Waller and Feuerstahler's (2017) investigation of the 4PM estimation with BME indicated that the lack of practical model convergence was still prevalent with BME, especially at $n$ = 1,000. Our follow-up calibrations showed that Bayesian estimation via Gibbs and HMC can serve as more viable alternatives for IRT practitioners when MML and Bayesian analytic algorithm do not work well for the 4PM item parameter estimation.

A characteristic of the Markov chain simulation in this study was the use of informative priors for all item parameters, which was found to be vital for Markov chain convergence. In our pilot simulations when only the *d* parameter had the uniform prior U(.6, 1) and informative priors were used for other item parameters, rstan gave warnings about a large number of divergent transitions, which at times went up to 10,000. Essentially, divergent transitions mean that rstan is having difficulty sampling from the posterior region thoroughly (Stan Development Team, 2018b). When divergence happens in rstan, model calibration results from other MCMC mechanisms should be questioned as well because the Markov chains are unlikely to converge to the posterior distribution (B. Goodrich, personal communication, November 19, 2019). When more informative priors were imposed on all item parameters, the vast majority of calibrations were completed well and rstan gave warnings to only several to a couple of dozens replications, which is small compared to 120,000 iterations after burn-in (warm-up in rstan).

Given the significant role of informative priors for item parameters characterized with the 4PM, a relevant question is how analysts can obtain them ahead of actual model estimation. Lord (1986) commented that repeated administrations of parallel test forms to similar test-taker groups allow us to infer appropriate item and person parameter prior distributions. While parallel test forms and frequent administrations are not feasible in many situations, reasonable expectations of an IRT parameter value

range and types of distributions to effectively capture it are possible (Baker & Kim, 2004). Note also that when the *d* parameter is very close to one, little change in the *a*, *b*, and *c* parameter estimates is observed between the 4PM and the 3PM (see, e.g., Swist, 2015). Therefore, modeling results with traditional IRT models can also serve as useful starting points for the 4PM calibrations.

Of course, the challenge remains for *d* because the 4PM has found widespread applications very recently and preexisting research results are not always available to inform the *d* prior. However, recent 4PM applications in educational measurement (e.g., Culpepper, 2016; Sideridis et al., 2016; Walstad & Rebeck, 2017) appeared to support a common-sense approach to specifying an upper asymptote prior that allows most values to cluster around .80 to 1.00 and lower values to be less and less probable. For psychopathology data modeling as reported in Waller and Feuerstahler (2017) and Waller and Reise (2010), the *d* parameters in the Depression and Cynicism data fit this general anticipation well, whereas the Low Self-Esteem data had many *d* values below .70 and might require a closer collaboration between methodologists, psychometricians, and content experts, as suggested by König and van de Schoot (2018), to determine the amount of prior knowledge available and the appropriate statistical distribution to convey this knowledge. Substantive issues aside, using the U(0, 1) to indicate no prior information on the upper asymptote, as examined in previous research (e.g., Culpepper, 2016), is not entirely justified because it endorses an implicit assumption that values of .10 and .90 are equally probable. While uniform priors such as U(0, 1) and U(.6, 1) impose little belief on the probability of *d* within a certain range, our study showed that they are not always warranted and might be a statistical luxury researchers cannot afford due to the Markov chain convergence concern.

The MPSRF was highly useful in Markov chain convergence evaluation. In our trial simulations where the uninformative prior U(0, .6) was used for the *d* parameter, the PSRF quickly went down below 1.05 for all parameters in both HMC and Gibbs after 25,000 iterations, whereas the MPSRF remained well above the chosen threshold even when chain length was increased up to 100,000 iterations. The message conveyed by MPSRF was consistent with warnings of divergent transitions in rstan that MCMC algorithm was struggling to sample from the joint posterior distribution and the Markov chains did not mix well in a stationary state. Only when a more informative prior for the *d* parameter was employed, did MPSRF approach 1.0, which again resonated with the fact that no divergent transition warning was issued by rstan. These findings about the power of MPSRF corroborated Sinharay's (2004) report on its superiority to PSRF in monitoring Markov chain convergence.

## Estimator and Efficiency

MML was clearly faster than the simulation-based Gibbs and HMC at converging at a solution. Even with a sample of up to 10,000 cases, MML in mirt executed the calibration within matters of seconds and did not require a large computer memory.

Between the two MCMC candidates, HMC was faster than Gibbs overall. At $n = 1{,}000$ and normal θ, a core i7 desktop computer with 16 GB of memory handled a data set within 3.5 hours with HMC but spent up to 7.4 hours with Gibbs on average. HMC maintained its superior speed over Gibbs in computers with poorer computing powers, although its advantage appeared to shrink. In the supercomputer system with a Dell Intel Xeon E5-2680 v4 machine, both Gibbs and HMC took approximately 15 hours on average to calibrate one data set with 5,000 cases.

However, calibration time might not be the most or only reasonable criterion to inform the efficiency discussions because Bayesian MCMC provides richer information than MML. MCMC approximates the posterior distribution where distributional features can be summarized to answer research questions, in contrast with MML which offers only point estimates. Levy and Mislevy (2016) captured this difference with an interesting analogy that frequentist methods (e.g., MML) find the highest peak in a mountain range, but Bayesian methods aim to develop a panorama of the entire mountain range via the posterior distribution.

Following Carpenter et al.'s (2017) recommendation, mean effective sample size per second (ESS/s) was calculated to assess the efficiency of the two MCMC sampling algorithms when they were executed in the same computer station, and results are displayed in the Appendix Tables 1 and 2. In general, HMC produced more ESS per time unit than Gibbs across all parameter types, sample sizes and latent trait conditions, and both HMC and Gibbs sampled the person parameters far more efficiently than item parameters. For example, at $n = 1{,}000$ and normal θ, Gibbs obtained about 1,107 ESS/s for θ but less than 5 ESS/s for the discrimination parameter on average, while the corresponding numbers for HMC were approximately 11,253 and 125, respectively. The overall efficiency ratio between HMC and Gibbs ranged from 4.85 to 10.39 (i.e., HMC yielded about 5 to 10 times more ESS/s than Gibbs). Although HMC was found to sample the posterior space more efficiently than Gibbs in the same computing environment, we urge the reader to take caution in the generalization of ESS/s by the two MCMC methods because a mixture of computers with varied computing powers was used in this study. It is also worth pointing out that while HMC generally offers an efficiency advantage, rstan installation and execution are more complex than runjags. In our study, rstan occasionally terminated mid-way due to computer memory limits and required random seed adjustments to improve the poor progress of some Markov chains.

## Practical Implications

The analytical results in the present inquiry prompted several recommendations for applied researchers utilizing the 4PM. Similar to Waller and Feuerstahler's (2017) report on BME-EAP, this study suggested that if the overarching purpose of model calibration is recovery of trait scores, both MCMC and MML-EAP could be employed for similarly accurate θ estimates with samples of as few as 1,000 respondents. If

accurate recovery of item parameters is also targeted, which many studies undoubtedly aim for, issues of latent trait distribution and sample size must be considered to inform the choice of estimation methods. Based on the second follow-up simulation, it is recommended that MML should be employed when sample size reaches 10,000 cases to ensure acceptable accuracy in item parameter recovery. While obvious advantages of MML include high speed calibrations and the lack of the need to specify item priors, which is convenient due to the rather novel applications of the 4PM in many settings, our simulations showed that even at $n$ = 10,000 cases, practical model convergence with MML still failed about 8 % of the time. In such circumstances, MCMC appears to be a viable alternative for the 4PM estimation. Either Gibbs or HMC can be selected to estimate the 4PM item parameters with 1,000 examinees, providing that useful information is available in the form of informative priors. HMC might be preferred among two MCMC approaches due to its comparable parameter recovery accuracy but higher efficiency and superior built-in mechanism in rstan to detect non-convergence.

When there are reasons to believe departure from θ normality is present, such as test scores obtained from academically at-risk or gifted students, or non-normal latent trait population distribution, neither MML nor MCMC offer an optimal solution. In this study, only negatively skewed latent trait was explored, and MCMC still remained the better option with lower RMSE of item parameter recovery than MML. However, it was clear that for all methods, simply adding more data of the same type does not represent the silver bullet for improved accuracy. When data provide little information to accurately estimate the upper asymptote and difficulty parameters of the 4PM, one strategy to consider is to collect more data to fill the empty locations in the upper tail of the latent trait spectrum to aid accurate recovery of these parameters. Because accurate item parameter recovery in IRT depends on a sample which is both large and heterogenous (Hambleton & Jones, 1993), both quantity and quality of data matter. After all, our model is only as good as the data we feed into it.

Common Markov chain length and burn in are necessary in simulation studies because the total number of data sets and parameters to handle is large and it is impractical to have the number of iterations adapted to each peculiar estimation scenario (Wollack et al., 2002). In reality, applied researchers typically examine fewer data sets and parameters. Therefore, it is recommended that the Markov chain length and burn-in segment be tailored to individual response data sets and measurement conditions. Moreover, given that MCMC simulates random samples from the multivariate (joint) posterior distribution of all IRT parameters, the necessity to diagnose Markov chain convergence to the multivariate posterior distribution is self-evident. Because there have been complaints about the performance of PSRF in detecting the lack of MCMC convergence with the 4PM (Waller & Reise, 2010), researchers wishing to explore modeling data with the 4PM via MCMC might want to add MPSRF to their frequently used convergence diagnostic toolbox. In addition to numerical means such as MPSRF and PSRF, visual means like trace plots to investigate the stability and mixing of parallel chains are more convenient in actual measurement practice and are highly recommended. After all, MCMC convergence to the posterior region forever remains a

black box and one can never be too certain about it. Therefore, regarding techniques to evaluate MCMC convergence, one should be content to have more and willing to use more, not less.

## Limitations

We would like to acknowledge the limitations of the current study. First, scale length, an important factor for accurate person parameter recovery, was fixed at 20 items. Past research has demonstrated the influence of longer scales on the accuracy of recovering θ for various IRT models, including the 4PM (e.g., Kieftenbeld & Natesan, 2012; Loken & Rulison, 2010). In general, longer tests are likely to bring greater benefits to the person parameter recovery accuracy, all else being equal. Although scales with 20 items or even fewer are quite common for unidimensional IRT models, including the 4PM research and applications (e.g., Culpepper, 2016; Gregory, 2019; Sheng, 2015; Sideridis et al. 2016), 20-item tests are fairly short tests, and this test length limits the generalizability of the study's findings to scales with substantially more items. Second, only a moderately skewed latent distribution was examined as a representative of latent trait nonnormality within the study's design. While this level of skewness (-0.60) has been found to be present in real world test score distributions (Ho & Yu, 2015; Lord, 1955) and its impact is worth exploring, positive skewness and more extremely negative skewness are far from uncommon in practice and can be expected to bring different levels of estimation errors to item parameter recovery and possibly person parameter estimation as well. Additionally, the parameter values selected for data generation reflected a typical educational cognitive assessment scenario rather than other domains with higher *b* parameters like in psychopathology (Feuerstahler & Waller, 2014; Waller & Feuerstahler, 2017; Waller & Reise, 2010). Furthermore, an important feature of the MCMC methods in this study was the use of informative priors, particularly for *c* and *d*, as opposed to previous research by Culpepper (2016) where uniform priors were used for the asymptote parameters. Therefore, the findings on MML and MCMC performances might hold true for the measurement conditions examined in this study and should not be overgeneralized to other scenarios outside the confines of the study design.

## Recommendations for Future Research

As is the case with other investigations, the current study provides answers to several specific questions while raising additional questions that await explorations to move the field forward. First, inquiries into how test length increase, different underlying item parameters, and latent distributional characteristics interact with estimation methods and influence the 4PM parameter recovery accuracy are warranted future directions. Second, it is important to understand how the prior specifications affect the MCMC estimation of the 4PM. In the present study, Gibbs and HMC simulations

were configured with informative and useful priors. Of course, prior informativeness and usefulness are matters of degree and can be adjusted by changing the parameters of the prior distributions. A sensitivity analysis to look at whether different levels of useful knowledge incorporated in the priors for the 4PM parameters have a considerable impact on the substantive conclusions regarding the merit of these two MCMC methods would be an interesting research idea to investigate. Third, the unsatisfactory performance of both MML and MCMC under skewed theta necessitates further research into more robust methods to handle the 4PM estimation under such a circumstance. The Bayesian non-parametric estimation approach (Paganin et al., 2023) represents another approach to dealing with parameter recovery in the context of non-normal θ. The literature on the use of Bayesian non-parametric method is still in its infant stage, but pioneering research has demonstrated that it brought more accurate item and person parameter recovery for Rasch IRT binary model than MML and MCMC, among other methods, in the presence of skewed latent trait while relaxing the normality assumption of person parameters (Finch & Edwards, 2016). The Bayesian non-parametric approach is an intriguing Bayesian inference development and a promising method for the 4PM estimation upon the departure from normal θ distribution. Future research efforts could examine this newer member in the Bayesian framework to model response data with the 4PM when θ normality cannot be reasonably assumed. Finally, it is imperative to provide formal proof of identifiability for the 4PM, as Culpepper (2016) also pointed out. IRT model identifiability is necessary for meaningful interpretations of its parameters, and future interests and useful applications of the 4PM might be hampered while waiting for such a significant piece of research to arrive.

## Acknowledgments

The authors would like to acknowledge the technical support of the Ohio SuperComputer Center.

## Declaration of Interest Statement

The authors declared no potential conflicts of interest with regard to the research, authorship, and publication of this study.

## References

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153–169. https://doi.org/10.1177/01466216980222005

Baker, F. B., & Kim, S-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). Dekker.

Barnard-Brak, L., Lan, W. Y., & Yang, Z. (2018). Differences in mathematics achievement according to opportunity to learn: A 4PM item response theory examination. *Studies in Educational Evaluation*, *56*, 1–7. https://doi.org/10.1016/j.stueduc.2017.11.002

Barton, M., & Lord, F. (1981). *An upper asymptote for the three-parameter logistic model*. (Research Report No. 81-20). Educational Testing Service.

Béguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*(4), 541–561. https://doi.org/10.1007/BF02296195

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, *7*(4), 434–455.

Cao, Y., Lu, R., & Tao, W. (2014). *Effect of item response theory (IRT) model selection on testlet-based test equating*. (Research Report No. 14-19). Educational Testing Service.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1), 1–32. http://dx.doi.org/10.18637/jss.v076.i01

Chalmers, R., P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. http://dx.doi.org/10.18637/jss.v048.i06

Culpepper, S. A. (2016). Revisiting the 4-parameter item response model: Bayesian estimation and application. *Psychometrika*, *81*(4), 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*(6), 706–725. https://doi.org/10.3102/1076998617705653

Denwood, M. J. (2016). Runjags: An R package providing interface utilities, model templates, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*, *71*(9), 1–25. https://doi.org/10.18637/jss.v071.i09

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Feuerstahler, L. M., & Waller, N. G. (2014). Abstract: Estimation of the 4-parameter model with marginal maximum likelihood. *Multivariate Behavioral Research*, *49*(3), 285–285. https://doi.org/10.1080/00273171.2014.912889

Finch, H., & Edwards, J. M. (2016). Rasch model parameter estimation in the presence of a nonnormal latent trait using a nonparametric Bayesian approach. *Educational and Psychological Measurement*, *76*(4), 662–684. https://doi.org/10.1177/0013164415608418

Gregory, C. A. (2019). Are we underestimating food insecurity? Partial identification with a Bayesian 4-parameter IRT model. *Journal of Classification*. https://doi.org/10.1007/s00357-019-09344-2

Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Harwell, M., Stone, C. A., Hsu, T-C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement*, *20*(2), 101–125. https://doi.org/10.1177/014662169602000201

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement*, *75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Kern, J. L., & Culpepper, S. A. (2020). A restricted four-parameter IRT model: The dyad four-parameter normal ogive (DYAD-4PNO) model. *Psychometrika*, *85*(3), 575–599. https://doi.org/10.1007/s11336-020-09716-3

Kieftenbeld, V., & Natesan, P. (2012). Recovery of graded response model parameters: A comparison of marginal maximum likelihood and Markov chain Monte Carlo. *Applied Psychological Measurement*, *36*(5), 399–419. https://doi.org/10.1177/0146621612446170

Kim, K. Y., & Lee, W. C. (2017). The impact of three factors on the recovery of item parameters for the three-parameter logistic model. *Applied Measurement in Education*, *30*(3), 228–242. https://doi.org/10.1080/08957347.2017.1316274

Kim, S., & Moses, T. (2016). *Investigating robustness of item response theory proficiency estimators to atypical response behaviors under two-stage multistage testing.* (Research Report No. 16–22). Educational Testing Service.

Kim, S-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement*, *25*(2), 163–176. https://doi.org/10.1177/01466210122031984

König, C., & van de Schoot, R. (2018). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, *70*(4), 486–509. https://doi.org/10.1080/00131911.2017.1350636

Kuo, T.-C., & Sheng, Y. (2016). A comparison of estimation methods for a multimimensional graded response IRT model. *Frontiers in Psychology*, *7*, Article 880. https://doi.org/10.3389/fpsyg.2016.00880

Liao, W.-W, Ho, R.-G., & Yen, Y.-C. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, *40*(10), 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Link, W. A., & Eaton, M. J. (2012). On thinning of chains in MCMC. *Methods in Ecology and Evolution*, *3*(1), 112–115. https://doi.org/10.1111/j.2041-210X.2011.00131.x

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. (1955). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement*, *15*, 383–389.

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

Lord, F. M. (1986). Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157–162.

Luo, Y., & Wolf, M. G. (2019). Item parameter recovery for the two-parameter testlet model with different estimation methods. *Psychological Test and Assessment Modeling*, *61*(1), 65–89.

Magis, D. (2013). A note on the item information function of the four-parameter logistic model. *Applied Psychological Measurement*, *37*(4), 304–315. https://doi.org/10.1177/0146621613475471

Meng, X., Xu, G., Zhang, J., & Tao, J. (2020). Marginalized maximum a posteriori estimation for the four-parameter logistic model under a mixture modelling framework. *British Journal of Mathematical and Statistical Psychology*,*73*(1), 51–82. https://doi.org/10.1111/bmsp.12185

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*(2), 177–195. https://doi.org/10.1007/BF02293979

Mooney, C. Z. (1997). *Monte Carlo simulation*. Sage.

Myszkowski, N., & Storme, M. (2017). Measuring "good taste" with the Visual Aesthetic Sensitivity Test-Revised (VAST-R). *Personality and Individual Differences*, *117*, 91–100. http://dx.doi.org/10.1016/j.paid.2017.05.041

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

Ogasawara, H. (2017). Identified and unidentified cases of the fixed-effects 3- and 4-parameter models in item response theory. *Behaviormetrika*, *44*, 405–423. https://doi.org/10.1007/s41237-017-0032-x

Paganin, S., Paciorek, C. J., Wehrhahn, C., Rodríguez, A., Rabe-Hesketh, S., & de Valpine, P. (2023). Computational Strategies and Estimation Performance with Bayesian Semiparametric Item Response Theory Models. *Journal of Educational and Behavioral Statistics*, *48*(2), 147–188. https://doi.org/10.3102/10769986221136105

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*(2), 146–178. https://doi.org/10.3102/10769986024002146

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd international workshop on distributed statistical computing (DSC 2003)*. Vienna, Austria. Retrieved from https://www.r-project.org/conferences/DSC-2003/Proceedings/Plummer.pdf

Primi, R., Nakano, T. D. C., & Wechsler, S. M. (2018). Using four-parameter item response theory to model human figure drawings. *Avaliação Psicológica, 17*(4), 473–483. http://dx.doi.org/10.15689/ap.2018.1704.7.07

R Core Team. (2020). R: A language and environment for statistical computing [Computer software]. Vienna, Austria. Retrieved from http://www.R-project.org/

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*(2), 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*(4), 365–384. https://doi.org/10.1207/S15327574IJT0304_5

Sass, D. A., Schmitt, T. A., & Walker, C. M. (2008). Estimating non-normal latent trait distributions within item response theory using true and estimated item parameters. *Applied Measurement in Education*, *21*(1), 65–88. https://doi.org/10.1080/08957340701796415

Seong, T-J. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distributions. *Applied Psychological Measurement*, *14*(3), 299–311. https://doi.org/10.1177/014662169001400307

Sheng, Y. (2015). Bayesian estimation of the four-parameter IRT model using Gibbs sampling. *International Journal of Quantitative Research in Education*, *2*(3/4), 194–212. https://doi.org/10.1504/IJQRE.2015.071736

Sideridis, G. D., Tsaousis, I., & Al Harbi, K. (2016). The impact of non-attempted and dually-attempted items on person abilities using item response theory. *Frontiers in Psychology*, *7,* Article 1572. https://doi.org/10.3389/fpsyg.2016.01572

Sijtsma, K., & Hemker, B. T. (2000). A taxonomy of IRT models for ordering persons and items using simple sum scores. *Journal of Educational and Behavioral Statistics*, *25*(4), 391–415. https://doi.org/10.3102/10769986025004391

Sinharay, S. (2004). Experiences with Markov chain Monte Carlo convergence assessment in two psychometric examples. *Journal of Educational and Behavioral Statistics*, *29*(4), 461–488. https://doi.org/10.3102/10769986029004461

Sinharay, S., & von Davier, M. (2005). *Extension of the NAEP BGGROUP program to higher dimensions*. (Research Report No. 05–27). Educational Testing Service.

Stan Development Team. (2018a). *Stan user's guide* (Version 2.18). http://mc-stan.org/documentation/

Stan Development Team. (2018b). *Stan reference manual* (Version 2.18). http://mc-stan.org/documentation/

Storme, M., Myszkowski, N., Baron, S., & Bernard, D. (2019). Same test, better scores: Boosting the reliability of short online intelligence recruitment tests with nested logit item response theory models. *Journal of Intelligence*, *7*(3), Article 17. https://doi.org/10.3390/jintelligence7030017

Swist, K. (2015). Item analysis and evaluation using a four-parameter logistic model. *Edukacja*, *3*, 77–97.

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*(4), 679–685. http://dx.doi.org/10.1590/S1415-47572004000400033

Tendeiro, J. N., & Meijer, R. R. (2012). A CUSUM to detect person misfit: A discussion and some alternatives for existing procedures. *Applied Psychological Measurement*, *36*(5), 420–442. https://doi.org/10.1177/0146621612446305

van Rijn, P. W. (2014). Reliability of multistage tests using item response theory. In D. Yan, A. A. von Davier, & C. Lewis (Eds.), *Computerized multistage testing: Theory and applications* (pp. 251–264). Chapman and Hall/CRC.

Waller, N. G., & Feuerstahler, L. (2017). Bayesian modal estimation of the four-parameter item response model in real, realistic, and idealized data sets. *Multivariate Behavioral Research*, *52*(3), 350–370. https://doi.org/10.1080/00273171.2017.1292893

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard IRT models: Fitting the four-parameter model to the MMPI. In S. Embretson & J. S. Roberts (Eds.), *Measuring psychological constructs: Advances in model-based approaches* (pp. 147–173). American Psychological Association.

Walstad, W. B., & Rebeck, K. (2017). The *Test of Financial Literacy*: Development and measurement characteristics. *The Journal of Economic Education*, *48*(2), 113–122. https://doi.org/10.1080/00220485.2017.1285739

Wollack, J. A., Bolt, D. M., Cohen, A. S., & Lee, Y.-S. (2002). Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and Markov Chain Monte Carlo estimation. *Applied Psychological Measurement*, *26*(3), 339–352. https://doi.org/10.1177/0146621602026003007

Yen, Y.-C., Ho, R.-G., Liao, W.-W., & Chen, L.-J. (2012). Reducing the impact of inappropriate items on renewable computerized adaptive testing. *Educational Technology and Society*, *15*(2), 231–243.

Yen, Y.-C., Ho, R.-G., Laio, W.-W., Chen, L.-J., & Kuo, C.-C. (2012). An empirical evaluation of the slip correction in the four parameter logistic models with computerized adaptive testing. *Applied Psychological Measurement*, *36*(2), 75–87. https://doi.org/10.1177/0146621611432862

Zhang, J., Lu, J., Du, H., & Zhang, Z. (2020). Gibbs-slice sampling algorithm for estimating the four-parameter logistic model. *Frontiers in Psychology*, *11*, Article 2121. https://doi.org/10.3389/fpsyg.2020.02121

## Appendix

Table 1

*Effective Effect Size per Second by Gibbs and HMC*

| Parameter | Estimator | $n$ = 1,000 | $n$ = 2,500 | $n$ = 5,000 |
|---|---|---|---|---|
| a | Gibbs | 4.8318 | 0.3437 | 0.0387 |
| | | 5.7484 | 0.1623 | 0.0384 |
| | HMC | 125.6782 | 24.4941 | 1.0315 |
| | | 111.4864 | 5.6489 | 0.9828 |
| b | Gibbs | 1.7197 | 0.1654 | 0.0231 |
| | | 2.0527 | 0.0803 | 0.0244 |
| | HMC | 73.9134 | 16.6166 | 0.7705 |
| | | 64.4837 | 3.7612 | 0.7509 |
| c | Gibbs | 2.5047 | 0.2191 | 0.0299 |
| | | 3.3117 | 0.1275 | 0.0355 |
| | HMC | 89.5991 | 18.2892 | 0.8420 |
| | | 86.1628 | 5.1140 | 1.0195 |
| d | Gibbs | 2.5828 | 0.2530 | 0.0339 |
| | | 2.7914 | 0.1010 | 0.0293 |
| | HMC | 88.3873 | 19.8791 | 0.8887 |
| | | 71.3048 | 3.7451 | 0.6963 |
| θ | Gibbs | 1107.2935 | 434.6188 | 177.3456 |
| | | 1320.0575 | 204.3885 | 173.6008 |
| | HMC | 11253.3630 | 7041.0626 | 867.8881 |
| | | 9871.8391 | 1642.4100 | 839.5945 |
| Overall | Gibbs | 1118.9325 | 435.6000 | 177.4711 |
| | | 1333.9617 | 204.8595 | 173.7284 |
| | HMC | 11630.9411 | 7120.3417 | 871.4208 |
| | | 10205.2768 | 1660.6792 | 843.0440 |

*Note.* Outcomes under normal theta are displayed in unshaded areas. Outcomes under negatively skewed theta are shaded.

Table 2

*Ratio of HMC and Gibbs Sampling Efficiency*

| Parameter | *n* = 1,000 | *n* = 2,500 | *n* = 5,000 |
|---|---|---|---|
| a | 26.01 | 71.27 | 26.66 |
| | 19.39 | 34.80 | 25.57 |
| b | 42.98 | 100.47 | 33.36 |
| | 31.41 | 46.86 | 30.83 |
| c | 35.77 | 83.46 | 28.19 |
| | 26.02 | 40.11 | 28.74 |
| d | 34.22 | 78.57 | 26.21 |
| | 25.54 | 37.10 | 23.75 |
| θ | 10.16 | 16.20 | 4.89 |
| | 7.48 | 8.04 | 4.84 |
| Overall | 10.39 | 16.35 | 4.91 |
| | 7.65 | 8.11 | 4.85 |

*Note*. Ratio was calculated as $HMC_{ESS/s}$ : $Gibbs_{ESS/s}$. Outcomes under normal theta are displayed in unshaded areas. Outcomes under negatively skewed theta are shaded.