Article

Educational and Psychological Measurement
2019, Vol. 79(2) 272–287

DOI: 10.1177/0013164418777569
journals.sagepub.com/home/epm
SAGE

A Short Note on Obtaining Point Estimates of the IRT Ability Parameter With MCMC Estimation in Mplus: How Many Plausible Values Are Needed?

Yong Luo[1] and Dimiter M. Dimitrov[1,2]

Abstract

Plausible values can be used to either estimate population-level statistics or compute point estimates of latent variables. While it is well known that five plausible values are usually sufficient for accurate estimation of population-level statistics in large-scale surveys, the minimum number of plausible values needed to obtain accurate latent variable point estimates is unclear. This is especially relevant when an item response theory (IRT) model is estimated with MCMC (Markov chain Monte Carlo) methods in Mplus and point estimates of the IRT ability parameter are of interest, as Mplus only estimates the posterior distribution of each ability parameter. In order to obtain point estimates of the ability parameter, a number of plausible values can be drawn from the posterior distribution of each individual ability parameter and their mean (the posterior mean ability estimate) can be used as an individual ability point estimate. In this note, we conducted a simulation study to investigate how many plausible values were needed to obtain accurate posterior mean ability estimates. The results indicate that 20 is the minimum number of plausible values required to obtain point estimates of the IRT ability parameter that are comparable to marginal maximum likelihood estimation(MMLE)/expected a posteriori (EAP) estimates. A real dataset was used to demonstrate the comparison between MMLE/EAP point estimates and posterior mean ability estimates based on different number of plausible values.

[1]National Center for Assessment, Riyadh, Saudi Arabia
[2]George Mason University, Fairfax, VA, USA

Corresponding Author:
Yong Luo, National Center for Assessment, King Khalid Road, Riyadh 11537, Saudi Arabia.
Email: jackyluoyong@gmail.com

Keywords

plausible value, IRT ability estimation, MCMC, MMLE, EAP

The mathematical equivalency between the categorical confirmatory factor analysis (CCFA; Wirth & Edwards, 2007) and item response theory (IRT; Lord, 1980) is well documented in the psychometric literature (e.g., Kamata & Bauer, 2008; Muthén, 1984; Muthén & Christoffersson, 1981; Takane & de Leeuw, 1987). In CCFA, factor loadings and item thresholds correspond to item discrimination and item difficulty, respectively, and the factor score is often referred to as ability (or trait) in IRT (e.g., Hambleton & Swaminathan, 1985). As CCFA can be viewed as a special case of structural equation modeling (SEM; Bollen, 1989) without the structural model, it has been recommended for users of SEM and IRT to utilize the advantages of both SEM and IRT for the modeling of categorical data (e.g., Finney & DiStefano, 2013; Glockner-Rist & Hoijtink, 2003).

Traditionally, CCFA with categorical data are estimated with limited-information methods such as maximum likelihood estimation with robust standard errors (MLR) or weighted least squares methods (WLSM) that only require information in the two-way contingency table. In IRT, categorical data are usually estimated with full-information methods such as marginal maximum likelihood estimation (MMLE; Bock & Aitkin, 1981) or Markov chain Monte Carlo (MCMC) that utilize information in the whole response pattern (the full multi-way contingency table). Numerous studies (e.g., Bolt, 2005; Forero & Maydeu-Olivares, 2009; Knol & Berger, 1991; Luo, 2018a, 2018b; Paek, Cui, Öztürk Gübeş, & Yang, 2018) have investigated the feasibility of using limited information methods for IRT model estimation and found that they provide comparable estimation results to MMLE and MCMC. Likewise, SEM software programs have incorporated some full-information estimation methods as an option for the estimation of CCFA. For example, the latent variable modeling software Mplus (Muthén & Muthén, 1998-2012) added the MLR estimator, which implements MMLE estimation for categorical data (in version 3), and the Bayes estimator, which implements MCMC estimation (in version 6).

The addition of full information estimation methods in Mplus, along with limited information estimation methods such as the *weighted least squares adjusted by mean and variance* (WLSMV; B. Muthén, du Toit, & Spisic, 1997), makes Mplus increasingly popular as a viable software for estimation of various IRT models (e.g., Finch, 2010; Huggins-Manley & Algina, 2015). Under the one-factor CCFA model, Mplus produces estimates of factor loadings and item thresholds and provides their IRT parameterization based on CCFA-IRT relationships, using the estimators of interest here (MLR or WLSMV) with an appropriate (logit or probit) link function (Asparouhov & Muthén, 2015; see also the Note in Appendix A). For multidimensional IRT models (MIRT; Reckase, 2009), Mplus estimates of factor loadings and thresholds can be converted into their IRT counterparts using formulae provided in the literature (e.g., Finch, 2010; Luo, 2018a; McDonald, 1999).

For estimates of IRT ability parameters, Mplus produces factor scores that are equivalent to IRT ability estimates and can be directly used as such without further conversion under some restrictions on the mean and variance of the latent variable (factor, ability) (see the Note in Appendix A). In frequentist IRT, ability is often estimated as a single value used to represent the individual latent variable distribution. In Bayesian IRT via MCMC, the ability is often estimated as an empirical posterior latent variable distribution, which can also be summarized as a single value. The computation of such single values, called IRT ability point estimates, within the two (frequentist and Bayesian) IRT frameworks is discussed next.

In the frequentist paradigm, MMLE only estimates item parameters, whereas IRT ability estimates can be obtained via three approaches, namely, maximum likelihood estimation (MLE), expected a posteriori (EAP), and maximum a posteriori (MAP). As all three approaches use a single point to summarize the individual latent variable distribution, the resulting IRT ability estimates are all point estimates. It should be noted that MLE has the inherent flaw of being unable to provide estimates for examinees with zero or perfect scores, whereas EAP and MAP (hereafter referred to as MMLE/EAP and MMLE/MAP) introduce ancillary information about the latent distribution via a prior distribution and can provide ability estimates of such examinees. Also, MMLE/EAP computes the mean of the posterior distribution via quadrature points and uses this posterior mean as a point estimate, whereas MMLE/MAP computes the mode of the posterior distribution via an iterative method and uses this posterior mode as a point estimate.

In the Bayesian framework, the use of MCMC usually provides a large number of drawn values that form the empirical posterior distribution of individual ability. Also, the computation of the mean of the posterior distribution (referred to as MCMC/EAP hereafter) is straightforward by simply taking the average of the drawn values. This is not the case with the computation of MMLE/EAP and MMLE/MAP, which requires either an iterative method or quadrature points. Conceptually, the MCMC/EAP computation is not different from drawing a sample from the population and using the sample mean to infer the population mean. Consequently, the number of drawn values (sample size) affects the inferential power of the sample mean. Popular Bayesian software programs such as WinBUGS (Spiegelhalter, Thomas, Best, & Lunn, 2003) and Stan (Carpenter et al., 2017) automatically provide such MCMC/EAP point estimates based on the drawn values. Also, as the draws used for MCMC estimation are also utilized to compute point estimates of the latent variables, the sample size is usually not a concern due to the large number of draws (the number of iterations minus the burn-in iterations, multiplied by the number of parallel chains) used for computation of MCMC/EAP point estimates. For WinBUGS or other Bayesian software programs that implement the Metropolis-Hastings algorithm or the Gibbs sampler, thousands of iterations are usually required for MCMC estimation. As a result of this, MCMC/EAP estimates are also based on thousands of drawn values. For Stan, an emerging Bayesian software program that implements the efficient Hamiltonian Monte Carlo (HMC; Neal, 2011) algorithm, hundreds of iterations may suffice for

common IRT models. Also, the MCMC/EAP point estimates are based on hundreds of drawn values, a sample size large enough to ensure accuracy of MCMC/EAP estimates. For an illustration of the sampling efficiency of Stan for IRT models, the reader may refer to Luo and Jiao (2018). However, this is not the case with Mplus, where the sample size issue becomes relevant considering that the draws for MCMC estimation are not used for the computation of latent variable point estimates and the users are at liberty to specify the number of draws from the posterior distribution used for the computation of latent variable point estimates.

Compared to estimators like MLR and WLSMV, the Bayes estimator in Mplus offers some special advantages. For example, Mplus automatically implements the posterior predictive model checking (PPMC; Rubin, 1984) procedure for model check. In addition, the MCMC method implemented in the Bayes estimator has computational advantages when dealing with high-dimensional models and/or a large number of items, scenarios where estimators such as MLR and WLSMV may become infeasible due to either issues with dimensionality or difficulty with inverting an excessively large matrix. When the Bayes estimator is used, Mplus does not automatically provide MCMC/EAP point estimates as WinBUGS and Stan do. Instead, Mplus estimates the entire posterior distribution of each individual latent variable.

To obtain MCMC/EAP point estimates in Mplus, plausible values can be used. The idea of plausible values was originally developed for the analysis of National Assessment of Educational Progress data in 1983-1984 (Mislevy, 1991). Essentially, multiple scores are provided for each student to assess the measurement error associated with each individual, and as stated by Wu (2005), ''If measurement error is small, then multiple scores for an individual will be close together. If measurement error is large, then multiple scores for an individual will be far apart'' (p. 115). Formally, plausible values are random draws from the posterior distribution of individual score given one's response pattern and represent ''the likely distribution of a student's proficiency'' (von Davier, Gonzalez, & Mislelvy, 2009, p. 11).

Asparouhov and Muthén (2010) stated that the posterior mean factor score (MCMC/EAP point estimate) can be computed by using multiple plausible values drawn from the posterior distribution and demonstrated that MCMC/EAP estimates based on 500 plausible values have desirable psychometric properties over MMLE/EAP with small sample size and nonnormal latent distributions. Regarding the number of plausible values, they suggested that ''it is necessary to use many imputed values. For example 100 or 500 such values can yield a precise posterior distribution for a latent variable which can be used to compute the posterior mean factor score estimate'' (Asparouhov & Muthén, 2010, p. 2). Although it is intuitive that more plausible values can approximate the posterior distribution better, drawing plausible values in Mplus takes time and requires postprocessing by the researcher to compute statistics of interest such as MCMC/EAP point estimates. Also, researchers may face computer hardware constraints and time limitation when drawing plausible values. For example, when analyzing a dataset with half a million examinees, which is not uncommon in large-scale testing organizations, drawing 500 plausible values for

each examinee takes excessively long time and results in a matrix with half a million rows and 500 columns that may be too large for the computer memory to handle. Therefore, it is of practical interest to have guidelines on the minimum number of plausible values required to compute accurate MCMC/EAP point estimates.

Although it is well known that five plausible values are usually sufficient for accurate estimation of population-level statistics in large-scale surveys (e.g., von Davier et al., 2009), one can expect that this guideline is not applicable to the case of MCMC/EAP computation as plausible values are used differently in the following two scenarios. First, for the estimation of population-level statistics, plausible values are usually used via multiple imputation techniques (Rubin, 1987). Second, when MCMC/EAP point estimates are of interest, plausible values are simply used to compute their arithmetic mean.

Based on the above discussion, the purpose of this study is to investigate the number of plausible values needed to obtain MCMC/EAP point estimates under common large-scale testing conditions (e.g., sufficiently large sample size and normally distributed latent variable) that are comparable to MMLE/EAP point estimates produced when the MLR estimator is used in Mplus. We do not consider the WLSMV estimator for CCFA in Mplus as it produces MAP estimates, which is conceptually less aligned with MCMC/EAP estimates and has been shown to be inferior to EAP estimates in many ways (Bock & Mislevy, 1982; Mislevy & Bock, 1997).

Method

Simulation Design

We used a small-scale simulation study to explore the number of plausible values needed to obtain accurate MCMC/EAP point estimates in the sense that they are comparable with MMLE/EAP point estimates. Specifically, we investigated MCMC/EAP estimates based on 5, 10, 20, 50, 100, 200, and 500 plausible values (abbreviated as PV5, PV10, PV20, PV50, PV100, PV200, and PV500, respectively). The number of five plausible values, which is recommended for estimations of population-level statistics, was chosen here as the minimum number of plausible values as we suspected that five plausible values were not enough to produce accurate MCMC/EAP point estimates. The number of 500 plausible values, used by Asparouhov and Muthén (2010), was chosen to be the maximum number of plausible values as we suspected that 500 plausible values might be an overkill when the goal is to obtain accurate MCMC/EAP point estimates. Due to the conceptual equivalence between drawing plausible values from a posterior distribution and drawing a sample from the population, we expected that when the initial number of plausible values was small, increasing it would result in considerably better MCMC/EAP point estimates. However, when the number of plausible values reached certain thresholds, increasing it would have only negligible returns in terms of improvement of estimation accuracy.

The IRT model of choice is the two-parameter logistic model (2PLM). This model was chosen for the purpose of illustration. More complex IRT models, such as

multidimensional or polytomous models, can also be used as long as model convergence is achieved. Drawing plausible values from the estimated posterior distribution is essentially the same regardless of the complexity of the IRT model. Under the 2PLM model, the probability of correct item response is

$$p_{ij}(u_{ij} = 1|\theta_i, a_j, b_j) = \frac{1}{1 + \exp(-a_j(\theta_i - b_j))} \quad (1)$$

where u_{ij} is the (1/0) score of examinee i on item j, a_j and b_j are the discrimination and difficulty parameters, respectively, of item j, and θ_i is the latent ability of examinee i.

We simulated a test of 40 dichotomously scored items taken by 1,000 examinees. This particular test length was chosen to mimic a multiple-choice question test of a medium length; the sample size of 1,000 was selected to ensure accurate estimation of item parameters with the 2PLM (Stone, 1992). The ability parameters were generated from a standard normal distribution; that is, $\theta \sim N(0, 1)$. For the item parameters listed in Table 1, the discrimination parameters were generated from a normal distribution $N(1, 0.04)$, and the difficulty parameters from a standard normal distribution $N(0,1)$. We generated 100 datasets based on Equation 1. For each dataset, we fit the 2PLM with both the MLR and Bayes estimators in Mplus. With the Bayes estimator, the default uninformative prior $N(0, 5)$ in Mplus was used for the factor loading and item threshold parameters. We specified Mplus to run four parallel chains with each containing a minimum of 5,000 iterations. In addition, we requested Mplus to draw from the estimated posterior distribution of each individual latent variable 500 plausible values, from which we used the first 5 values to compute MCMC/EAP point estimates for PV5, first 10 for PV10, first 20 for PV20, first 50 for PV50, first 100 for PV100, first 200 for PV200, and all 500 for PV500. The Mplus syntax codes for estimating the 2PL IRT model with the MLR estimator and the Bayes estimator are provided in Appendixes A and B, respectively.

Outcome Variables

We evaluate the quality of point estimates for the ability parameter in terms of the correlation between estimated and true values, Bias, standard error (*SE*), and *root mean square error* (RMSE). The Bias, *SE*, and RMSE statistics were chosen as they account for systematic error, random error, and total error of parameter recovery, respectively. These statistics are defined as follows:

$$Bias(\hat{\theta}) = \frac{\sum_1^R (\hat{\theta}_r - \theta)}{R}, \quad (2)$$

$$SE(\hat{\theta}) = \sqrt{\frac{\sum_1^R \left(\hat{\theta}_r - \bar{\hat{\theta}}\right)^2}{R}}, \quad (3)$$

Table 1. Item Parameter Values Used for Data Generation.

Item	*a*	*b*
1	1.333	−0.556
2	1.186	−0.783
3	1.053	1.357
4	0.833	0.231
5	1.258	0.172
6	0.973	1.200
7	0.853	−0.389
8	1.064	−0.268
9	0.883	−1.385
10	1.252	2.172
11	1.263	−0.524
12	0.927	0.601
13	0.943	−0.131
14	0.557	−0.666
15	0.672	0.860
16	0.783	−0.562
17	0.626	0.086
18	1.350	1.165
19	0.947	−0.032
20	1.400	0.852
21	0.748	0.305
22	1.318	1.310
23	1.284	2.369
24	0.842	0.735
25	1.494	1.668
26	0.926	−0.418
27	1.336	0.508
28	1.297	−1.560
29	1.164	1.398
30	1.075	0.189
31	1.017	0.226
32	0.890	0.951
33	1.076	1.805
34	1.278	−0.660
35	1.091	0.462
36	0.685	0.838
37	1.059	0.991
38	0.809	−1.283
39	1.051	−0.591
40	0.984	0.054

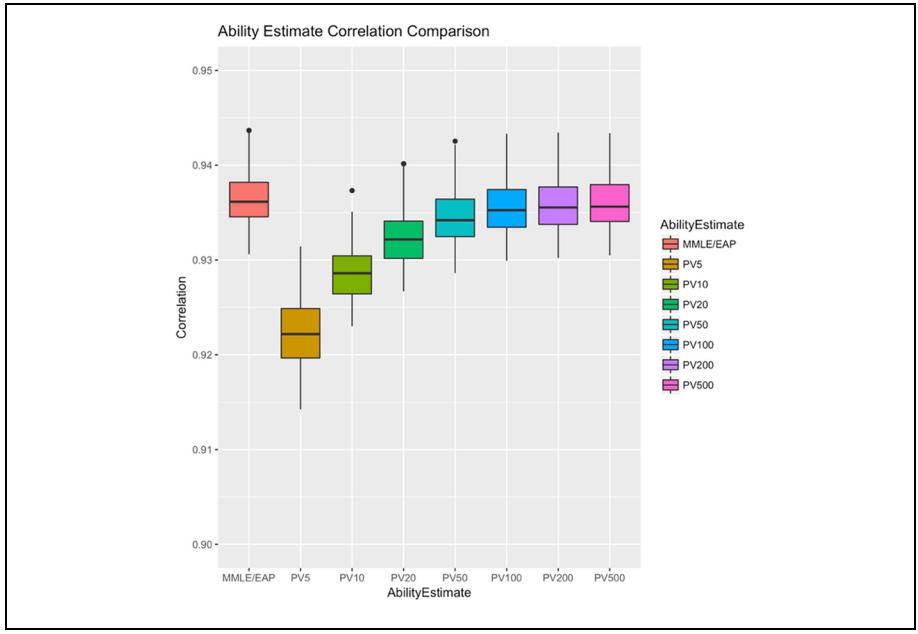

Figure 1. Comparison of ability estimates based on their correlation with generating values.

and

$$RMSE(\hat{\theta}) = \sqrt{\frac{\sum_1^R (\hat{\theta}_r - \theta)^2}{R}}, \quad (4)$$

where R is the number of replications (100 in the current study), θ is the generating ability parameter, $\hat{\theta}_r$ is the estimated ability parameter in the rth replication, $\bar{\hat{\theta}}$ is the mean estimated ability parameter averaged across the 100 replications. Cohen's f (hereafter abbreviated as f) was used as an effect size index to gauge the practical significance of the difference captured with the statistics in Equations 2, 3, and 4.

Results

The boxplot in Figure 1 provides a visual presentation of the correlation between the eight sets of ability estimates and the generating true ability values. As can be seen, the correlation between MCMC/EAP estimates based on plausible values and true values increases with an increase of the number of plausible values. However, the increase of the number of plausible values has diminishing returns regarding the increase of their correlation with the true values. With only five plausible values, the mean correlation between PV5 and the true values is 0.922, whereas the mean

Table 2. Bias, RMSE, and *SE* of Ability Estimates.

Statistic	MMLE/EAP	PV5	PV10	PV20	PV50	PV100	PV200	PV500
Bias								
Mean	0.0001	0.0049	−0.0013	−0.0011	0.0005	−0.0030	−0.0045	−0.0025
SD	0.1339	0.2177	0.1921	0.1727	0.1581	0.1513	0.1511	0.1500
Min	−0.5466	−0.7818	−0.6176	−0.6338	−0.5781	−0.6137	−0.6196	−0.5938
Max	0.9134	1.1354	1.0772	1.0710	1.0132	0.9674	0.9779	0.9517
RMSE								
Mean	0.1228	0.1491	0.1374	0.1306	0.1265	0.1248	0.1243	0.1239
SD	0.0622	0.1037	0.0924	0.0857	0.0731	0.0717	0.0704	0.0690
Min	0.0633	0.0623	0.0633	0.0622	0.0616	0.0602	0.0603	0.0604
Max	0.9244	1.3922	1.2687	1.2456	1.1211	1.0317	1.0539	1.0044
SE								
Mean	0.1049	0.1017	0.1006	0.1008	0.1015	0.1019	0.1015	0.1015
SD	0.0171	0.0182	0.0174	0.0171	0.0171	0.0172	0.0171	0.0170
Min	0.0629	0.0622	0.0598	0.0598	0.0602	0.0600	0.0600	0.0600
Max	0.1691	0.1781	0.1727	0.1662	0.1693	0.1810	0.1740	0.1758

Note. RMSE = root mean square error; SE = standard error; MMLE = marginal maximum likelihood estimation; EAP = expected a posteriori; PV = plausible value.

correlation between MMLE/EAP estimates and the true values is 0.937. When the number of plausible values is greater than 50, the mean correlation with the true values remains approximately 0.936, which only differs from the correlation between MMLE/EAP and the true values at the third decimal place. Such an observation is consistent with our expectation that an increase of the number of plausible values results in a noticeable improvement of the point estimation accuracy, but only to a certain point.

The Bias, *SE*, and RMSE for the eight sets of ability estimates are provided with Table 2. For all eight methods, the absolute value of the mean Bias for estimation of the ability parameter is smaller than 0.005, a value small enough to be considered close to zero. Also, the mean Bias randomly fluctuates around zero with a different number of plausible values, suggesting that such MCMC/EAP estimates are essentially unbiased and their discrepancies with the true values are caused by sampling error. As shown in Table 2, the standard deviation of Bias, as well as the mean and standard deviation of the RMSE, tend to decrease slightly with the increase of number of plausible values

To further compare the estimation quality between the eight methods, we conducted analysis of variance (ANOVA) using Bias, *SE*, and RMSE as dependent variables, respectively. The ANOVA results with Bias as the dependent variable indicated that there were no significant differences in estimation biases between the eight methods, $F(7, 7{,}992) = 0.899$, $p = .506$. The ANOVA results with the RMSE as the dependent variable indicated that there were significant differences, $F(7, 7{,}992) = 13.111$, $p < .001$, $f = 0.105$. According to Cohen (1992), such an f value

indicates a small effect size. Subsequent Tukey post hoc tests revealed that (a) there were no significant differences between the RMSE of PV20, PV50, PV100, PV200, PV500, and MMLE/EAP, and (b) the RMSE of PV10 and PV20 was significantly lower than those of PV5. The ANOVA results with *SE* as the dependent variable indicated that there were significant differences, $F(7, 7{,}992) = 5.926$, $p < .001$, $f = 0.071$ (this effect size is practically negligible). The Tukey post hoc tests showed that there were no significant differences between the *SE* of all seven sets of MCMC/EAP estimates based on plausible values, whereas the *SE* of MMLE/EAP estimates were significantly greater than those of their MCMC/EAP counterparts.

Based on the above analyses, it can be stated that PV20 provides IRT ability estimates with Bias and RMSE comparable to those based on MMLE/EAP and significantly smaller *SE*. Therefore, we concluded that 20 plausible values are required to obtain point estimates as accurate as MMLE/EAP estimates.

An Illustration With Real Data

In this section, we illustrate with real data how MCMC/EAP point estimates of IRT abilities based on different numbers of plausible values compare with MMLE/EAP point estimates. The dataset used in the illustration was drawn from a test form of the verbal section of the General Aptitude Test (GAT-V), a high-stakes test used for college admission purpose in Saudi Arabia and other Middle-Eastern countries. GAT-V consists of 52 multiple-choice items distributed across four domains, namely, reading comprehension, sentence completion, verbal analogy, and synonymy. We randomly sampled 1,200 students from the examinees, and the subsequent analyses were based on a matrix of zeros and ones with a dimension of 1,200 by 52.

We fit the 2PLM with both the MLR and Bayes estimators in Mplus. The configuration for the Bayes estimator is the same as in the simulation section; that is, four parallel chains with a minimum of 5,000 iterations were specified and 500 plausible values were imputed. The iteration history in the Mplus output showed that the potential scale reduction factor (PSRF; Gelman & Rubin, 1992) values for all model parameters were smaller than 1.05 with less than 2,000 iterations, whereas with 5,000 iterations, the highest PSRF value was 1.022. We concluded that model convergence had been reached and proceeded to compare the IRT ability point estimates based on the eight methods.

As an illustration, Table 3 lists the IRT ability point estimates for the first 20 students with MMLE/EAP and MCMC/EAP based on different number of plausible values. As can be seen, except for PV5 and PV10, which have mean differences of 0.07 and 0.03 relative to MMLE/EAP, the other five sets of IRT ability point estimates based on plausible values are very similar to MMLE/EAP, with the mean difference being smaller than 0.01.

Figure 2 provides a visual presentation of how the correlation between the MCMC/EAP point estimates based on plausible values and the MMLE/EAP point estimates for all 1,200 students changes with the change of the number of plausible

Table 3. Ability Estimates (on the IRT Logit Scale) for the Real Data.

Examinee	MMLE/EAP	PV5	PV10	PV20	PV50	PV100	PV200	PV500
1	−0.606	−0.512	−0.622	−0.671	−0.619	−0.622	−0.610	−0.615
2	0.873	0.756	0.807	0.794	0.836	0.779	0.806	0.839
3	0.750	1.028	0.713	0.688	0.714	0.705	0.727	0.723
4	−0.829	−0.583	−0.729	−0.797	−0.785	−0.821	−0.814	−0.835
5	0.040	−0.014	0.008	0.056	0.116	0.088	0.044	0.037
6	−0.206	−0.145	−0.111	−0.239	−0.233	−0.177	−0.201	−0.225
7	−0.875	−0.645	−0.741	−0.793	−0.849	−0.830	−0.861	−0.868
8	0.447	0.527	0.495	0.445	0.462	0.471	0.463	0.457
9	−1.508	−1.361	−1.344	−1.413	−1.480	−1.465	−1.474	−1.462
10	−0.424	−0.368	−0.394	−0.557	−0.501	−0.453	−0.415	−0.417
11	−0.784	−0.806	−0.892	−0.813	−0.809	−0.814	−0.795	−0.805
12	0.041	0.054	0.087	0.126	0.059	0.049	0.068	0.061
13	0.496	0.568	0.509	0.560	0.496	0.498	0.477	0.457
14	0.249	0.333	0.322	0.277	0.230	0.280	0.294	0.276
15	0.408	0.538	0.614	0.532	0.482	0.444	0.456	0.441
16	−0.345	−0.446	−0.414	−0.364	−0.374	−0.363	−0.364	−0.343
17	−0.847	−0.681	−0.807	−0.831	−0.792	−0.817	−0.788	−0.807
18	0.112	0.242	0.052	0.153	0.146	0.129	0.138	0.115
19	1.114	1.112	1.175	1.105	1.158	1.086	1.095	1.089
20	−0.606	−0.512	−0.622	−0.671	−0.619	−0.622	−0.610	−0.615

Note. IRT = item response theory; MMLE = marginal maximum likelihood estimation; EAP = expected a posteriori; PV = plausible value.

values. Even with only five plausible values, the correlation between PV5 and MMLE/EAP is 0.9892 and it increases with more plausible values, but the increase is negligibly small. The correlation between MMLE/EAP and PV20, the MCMC/EAP estimates based on 20 plausible values, is as high as 0.9969.

Conclusion

The Bayes estimator in Mplus implements the Gibbs sampler and allows researchers and practitioners to estimate latent variable models such as IRT models with MCMC methods with relatively little programming efforts. When the Bayes estimator is used, Mplus automatically performs model check via PPMC, a feature that is especially attractive to those who are not familiar enough with general Bayesian programming software such as WinBUGS or Stan to program such procedures themselves. In addition, MCMC methods may be the only computationally feasible estimation methods with either multidimensional models, which cause dimensionality problems for MMLE estimation, or a large number of indicators that make limited information methods such as WLSMV too slow.

It should be noted that for latent variables such as the individual ability in IRT context, Mplus estimates the posterior distribution of each ability parameter but does

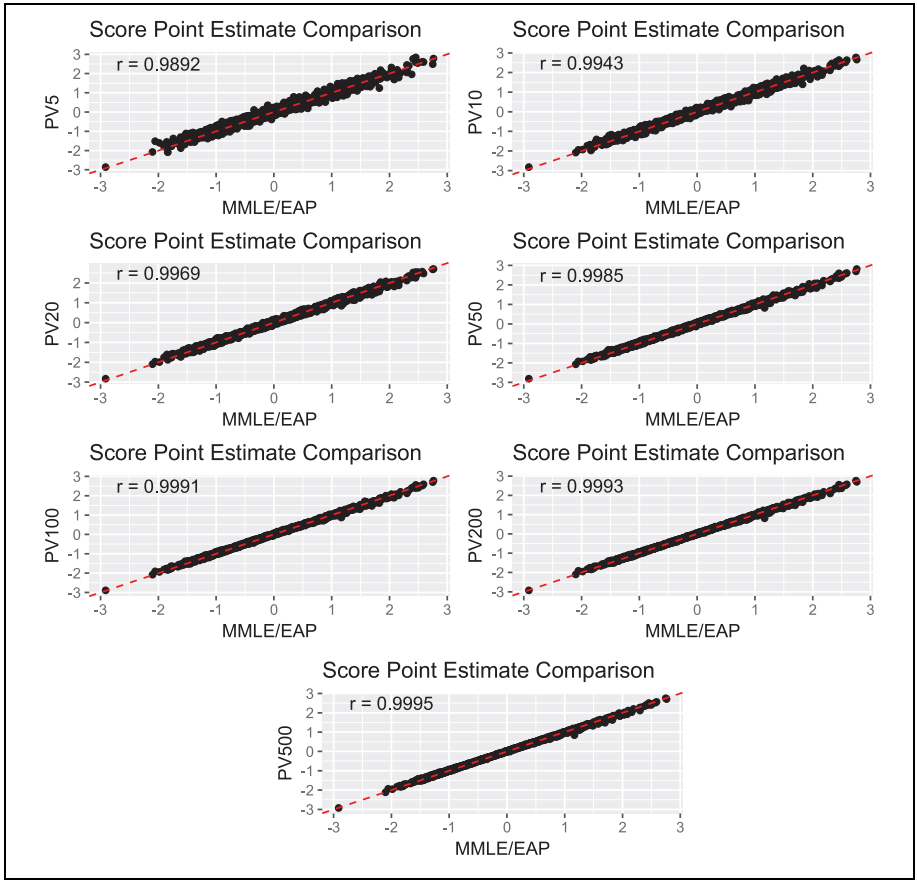

Figure 2. Comparison of point estimates of ability based on real data.

not provide MCMC/EAP point estimates. Instead, such estimates can be obtained by taking the average of a certain number of plausible values (specified by the user) drawn from the posterior distribution. However, it was unclear how many plausible values were needed to obtain accurate MCMC/EAP point estimates. In this note, we investigated the minimum number of plausible values required to obtain MCMC/EAP point estimates that are comparable to those based on MMLE/EAP, which is a common method for the estimation of IRT ability.

We found that MCMC/EAP estimates based on more plausible values had higher correlation coefficients with the generating values. This makes intuitive sense as more plausible values result in a larger sample size, and thus, higher precision of the sample mean as a population mean estimator. As is dictated by the central limit theorem, increasing the sample size results in diminishing returns in the precision of the

sample mean estimator (e.g., Pyrczak, 2003). Consequently, when the sample size reaches a certain level, its increase only produces negligible returns in precision, which is exactly what was observed in the simulation study. Specifically, increasing the number of plausible values from 5 to 10 resulted in the biggest increase (greater than 0.05) in the correlation between MCMC/EAP point estimates and the true values, which only changed at the third decimal place when the number of plausible values was greater than 50.

We also found that estimation biases based on different number of plausible values are not significantly different, which is hardly surprising. Indeed, as noted earlier, drawing plausible values from the posterior distribution is analogous to drawing a sample from the population. On the other side, it is known that the sample mean is an unbiased estimator of the population mean regardless of the population distribution. Therefore, the shape of the posterior distribution does not make a difference regarding the generalizability of the current finding; nor does the test length, as different test lengths only result in posterior distributions with varying standard deviations for individual latent ability, and the sample mean remains an unbiased estimator of the population mean. Although the mean of Bias does not exhibit a discernable pattern with an increase of the number of plausible values, the standard deviation of Bias was found to decrease with more plausible values. This observation also makes sense as more plausible values (i.e., a larger sample size) always result in smaller sampling errors, which in turn cause the mean and standard deviation of RMSE to drop.

Regarding the research question of how many plausible values can produce MCMC/EAP estimates comparable to MMLE/EAP estimates, the simulation study results indicated that MCMC/EAP point estimates based on five plausible values produce comparable estimation Bias, significantly smaller *SE*, and significantly greater RMSE. At least 20 plausible values are needed to reduce the RMSE to a level comparable to MMLE/EAP. As also shown in the simulation study, MCMC/EAP point estimates based on at least 20 plausible values produce comparable estimation Bias and RMSE and significantly smaller *SE* than their MMLE/EAP counterparts. Therefore, if researchers or practitioners are interested in obtaining accurate MCMC/EAP point estimates of IRT ability parameters with the Bayes estimator in Mplus, but do not want to draw a large number of plausible values due to practical constraints, they can use twenty plausible values.

Appendix A

Mplus Syntax for CCFA-Based Estimation of IRT Item Parameters Under the 2PL Model Using the MLR Estimator

```
TITLE:    CCFA-based estimation of IRT item parameters under the 2PL model
DATA:     FILE IS data.txt;
VARIABLE: NAMES ARE u1-u40;
          CATEGORICAL ARE u1-u40;
ANALYSIS: ESTIMATOR = MLR;
```

```
MODEL:      f BY u1-u40*;
             f@1;     !factor variance set equal to 1
            [f@0];    !factor mean set equal to 0
SAVEDATA:   FILE IS score_mlr.csv;
            SAVE = FSCORES;
```

Note. With the MLR estimator in the above syntax code, the logit link function is used in Mplus by default. With this, the following relationships between the CCFA item parameters (loading, λ_i, and threshold, τ_i) and the IRT item parameters under the 2PL model (discrimination, a_i, and difficulty, b_i) are in place (e.g., Asparouhov & Muthén, 2015):

$$a_i = \lambda_i \sqrt{\text{VAR}(f)} \text{ and } b_i = \frac{\tau_i - \lambda_i \bar{f}}{\lambda_i \sqrt{\text{VAR}(f)}}, \quad \text{(A.1)}$$

where $\bar{f}$ and VAR(f) are the mean (expected value) and the variance, respectively, of the latent factor f under the CCFA. However, in the Mplus syntax code, $\bar{f} = 0$, with using the command [f@0], and VAR(f) = 1, with using the command f@1, thus simplifying the equations in A.1 as follows:

$$a_i = \lambda_i \text{ and } b_i = \tau_i / \lambda_i. \quad \text{(A.2)}$$

The IRT estimates of item parameters provided in the Mplus output with the use of the above syntax code are produced with the use of the equations in A.2.

Furthermore, with the MLR estimator and the logit link function in the Mplus syntax code, the following relationship takes place (Asparouhov & Muthén, 2015):

$$f = \bar{f} + \theta \sqrt{VAR(f)},$$

where θ is the IRT latent variable (ability), with $\theta \sim N(0, 1)$. Thus, with the restrictions $\bar{f} = 0$ and VAR(f) = 1 imposed for model identification, the factor scores, which are saved with using the last command in the above Mplus code, become identical to the IRT ability scores, θ.

Appendix B

Mplus Syntax for Estimation of IRT Item Parameters Under the 2PL Model With the Bayes Estimators

```
TITLE:  Estimation of IRT item parameters under the 2PL model with Bayes Estimators
DATA:       FILE IS data.txt;
VARIABLE:   NAMES ARE u1-u40;
            CATEGORICAL ARE u1-u40;
ANALYSIS:   ESTIMATOR = BAYES;
            PROCESSORS = 4;
            BITERATIONS = (5000);
              CHAINS = 4;
MODEL:      BY u1-u40*;
            f@1;    !factor variance set equal to 1
SAVEDATA:   FILE IS score_bayes.csv;
            SAVE=FSCORES(500);
OUTPUT:     STANDARDIZED TECH8;
```

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Asparouhov, T., & Muthén, B. (2010, August 21). *Plausible values for latent variables using Mplus*. Retrieved from https://www.statmodel.com/download/Plausible.pdf

Asparouhov, T., & Muthén, B. (2015, September 24). *IRT in Mplus* (Technical report). Retrieved from https://www.statmodel.com/irtanalysis.shtml

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Bolt, D. M. (2005). Limited and full-information IRT estimation. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary psychometrics* (pp. 27-71). Mahwah, NJ: Lawrence-Erlbaum.

Bollen, K. A. (1989). *Structural equations using latent variables*. New York: Wiley.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1). Retrieved from https://www.jstatsoft.org/article/view/v076i01

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

Finch, H. (2010). Item parameter estimation for the MIRT model: Bias and precision of confirmatory factor analysis–based models. *Applied Psychological Measurement*, *34*, 10-26.

Finney, S. J., & DiStefano, C. (2013). Non-normal and categorical data in structural equation modeling. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 439-492). Charlotte, NC: Information Age.

Forero, C. G., & Maydeu-Olivares, A. (2009). Estimation of IRT graded response models: limited versus full information methods. *Psychological Methods*, *14*, 275-299.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*, 544-565.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York, NY: Springer Science & Business Media.

Huggins-Manley, A. C., & Algina, J. (2015). The partial credit model and generalized partial credit model as constrained nominal response models, with applications in Mplus. *Structural Equation Modeling*, *22*, 308-318.

Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and item response theory models. *Structural Equation Modeling*, *15*, 136-153.

Knol, D. L., & Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research*, *26*, 457-477.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Luo, Y. (2018a). A short note on estimating the testlet model using different estimators in Mplus. *Educational and Psychological Measurement*, *78*, 517-529.

Luo, Y. (2018b, April). *Item parameter recovery of the 2PL testlet model with different estimation methods*. Paper presented at the annual meeting of the American Educational Research Association, New York, NY.

Luo, Y., & Jiao, H. (2018). Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, *78*, 384-408.

McDonald, R. P. (1999). *Test theory: A unified approach*. Mahwah, NJ: Lawrence Erlbaum.

Mislevy, R. (1991). Randomization based inference about examinees in the estimation of item parameters. *Psychometrika*, *56*, 177-196.

Mislevy, R. J., & Bock, R. D. (1997). BILOG 3: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.

Muthén, B., & Christoffersson, A. (1981). Simultaneous factor analysis of dichotomous variables in several groups. *Psychometrika*, *46*, 407-419.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). *Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes*. Unpublished manuscript.

Muthén, L., & Muthén, B. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, *2*, 113-162.

Paek, I., Cui, M., Öztürk Gübeş, N., & Yang, Y. (2017). Estimation of an IRT model by Mplus for dichotomously scored responses under different estimation methods. *Educational and Psychological Measurement*, *78*, 569-588. doi:10.1177/0013164417715738

Pyrczak, F. (2003). *Making sense of statistics: A conceptual overview* (3rd ed.). Los Angeles, CA: Pyrczak Publishing.

Reckase, M. (2009). *Multidimensional item response theory* (Vol. 150). New York: Springer.

Rubin, D. B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, *12*, 1151-1172.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: Wiley.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Lunn, D. (2003). WinBUGS 1.4* user manual [Computer program]. Cambridge, England: MRC Biostatistics Unit.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of MULTILOG. *Applied Psychological Measurement*, *16*(1), 1-16.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 9-36.

Wirth, R. J., & Edwards, M. C. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods*, *12*(1), 58-79.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation*, *31*(2-3), 114-128.