Examining the Performance of the Metropolis–Hastings Robbins–Monro Algorithm in the Estimation of Multilevel Multidimensional IRT Models

Applied Psychological Measurement

2017, Vol. 41(5) 323–337

© The Author(s) 2017

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0146621616688923

journals.sagepub.com/home/apm

Bozhidar M. Bashkov¹ and Christine E. DeMars²

Abstract

The purpose of this study was to examine the performance of the Metropolis–Hastings Robbins–Monro (MH-RM) algorithm in the estimation of multilevel multidimensional item response theory (ML-MIRT) models. The accuracy and efficiency of MH-RM in recovering item parameters, latent variances and covariances, as well as ability estimates within and between clusters (e.g., schools) were investigated in a simulation study, varying the number of dimensions, the intraclass correlation coefficient, the number of clusters, and cluster size, for a total of 24 conditions. Overall, MH-RM performed well in recovering the item, person, and group-level parameters of the model. Ratios of the empirical to analytical standard errors indicated that the analytical standard errors reported in flexMIRT were somewhat overestimated for the cluster-level ability estimates, a little too large for the person-level ability estimates, and essentially accurate for the other parameters. Limitations of the study, implications for educational measurement practice, and directions for future research are offered.

Keywords

estimation, multilevel models, multidimensional item response theory, simulation

Assessment practitioners usually design and administer tests that measure not one but several abilities or latent traits. Moreover, researchers are interested in the relationships among different domains. Nevertheless, each subtest is typically scored independently using a unidimensional item response theory (IRT) model. A serious drawback of this independent calibration approach is that it ignores the relationships among the domains during parameter estimation, which is likely to result in loss of information in the estimation of item parameters and person ability measures. Such loss of information is especially evident when the domains are highly correlated (which is often the case with cognitive assessments) and the number of items per

¹American Board of Internal Medicine, Philadelphia, PA, USA

²James Madison University, Harrisonburg, VA, USA

Corresponding Author:

Bozhidar M. Bashkov, American Board of Internal Medicine, 510 Walnut Street, Suite 1700, Philadelphia, PA 19106, USA.

Email: bbashkov@abim.org

domain is small, as it is in large testing programs such as National Assessment of Educational Progress (NAEP; Zhang, 2012).

An alternative approach to modeling multiple constructs is simultaneous estimation via multidimensional item response theory (MIRT), which extends the unidimensional model to include multiple latent traits. Here, the estimation of item parameters and ability estimates is aided by the auxiliary information contained in the correlations among the latent dimensions. With fewer items, several dimensions, and high correlations among the dimensions, this additional information can substantially increase the precision of parameter estimates (de la Torre & Patz, 2005; Wang, Chen, & Cheng, 2004). In addition, the correlations among dimensions are used in the prior if the person ability estimates are obtained via Bayesian methods (e.g., expected a posteriori [EAP]).

Another aspect of educational data is that they typically have a nested structure. That is, individual students are not simple random samples from the population to which one wishes to generalize, but instead they are nested or clustered within larger units (e.g., classrooms, schools). Due to this nesting, objects of measurement often share many characteristics with others within their unit. That is, lower level units within a higher level unit or cluster are typically more similar to one another than to lower level units in other clusters. This relatedness results in violation of the assumption of independent residuals in regression. Ignoring this violation results in underestimated standard errors (*SEs*) and inflated Type I error for inferential tests.

A family of models known as multilevel models were developed to accommodate and appropriately model data with hierarchical (i.e., nested) structure (Hox, 2010; Raudenbush & Bryk, 2002; Snijders & Bosker, 2010). Despite the increasing popularity of multilevel models in educational research, less attention has been given to the measurement of latent traits in nested data structures. In multilevel measurement models, the variance associated with item response patterns can be “decomposed” so that separate latent traits can be specified at the individual examinee level and at the cluster (e.g., classroom or school) level.

Estimating MIRT models can be extremely computationally challenging, and extending the MIRT model to multiple levels increases the challenges. The Metropolis–Hastings Robbins–Monro (MH-RM) is a promising algorithm for these complex models. The purpose of this study was to assess how well the MH-RM algorithm can estimate the parameters of 3-parameter logistic (3PL) multilevel MIRT models. Further details about multilevel measurement models and MH-RM are provided in the following sections.

Multilevel Measurement Models

Prior efforts in measuring latent traits in nested data structures have focused on modeling the measurement error in predictors (items) by specifying items as nested within examinees and then specifying a latent dependent variable in a structural measurement model, also known as a nonlinear multilevel model (e.g., Adams, Wilson, & Wu, 1997; Cheong & Raudenbush, 2000). The advantage of these models is that by adding covariates, the model can be extended to an explanatory model. The main disadvantage is that to actually model the dependency among examinees in the same school, a third level must be added, which makes the model computationally more complex, especially with random item discrimination parameters (Kamata, Bauer, & Miyazaki, 2008).

One alternative is to extend the measurement model to multiple levels, decomposing the variance in the latent trait to between-cluster and within-cluster levels. This is the approach taken by Höhler, Hartig, and Goldhammer (2010), although they specified only a 2-parameter logistic (2PL) model and were only interested in the latent covariance structure at different levels. To

account for the probability of correct response by examinees of low proficiency, we extended their model to a 3PL model.

Specifically, the latent ability level of examinee j from cluster k on dimension g can be decomposed as $\theta_{gjk} = \theta_g + \theta_{gk}^B + \theta_{gjk}^W$, where θ_g is the grand mean on dimension g (which is typically constrained to 0 for identification purposes), $\theta_{gk}^B = \theta_{gk} - \theta_g$ is the deviation of cluster k 's mean from the grand mean, and $\theta_{gjk}^W = \theta_{gjk} - \theta_{gk}$ is examinee j 's deviation from cluster k 's mean on dimension g . Then the 3PL multilevel multidimensional item response theory (ML-MIRT) model becomes

$$P(U_{ijk} = 1 | \theta_k^B, \theta_{jk}^W, \mathbf{a}_i^B, \mathbf{a}_i^W, c_i, d_i) = c_i + (1 - c_i) \frac{e^{a_i^B \theta_k^{B'} + a_i^W \theta_{jk}^{W'} + d_i}}{1 + e^{a_i^B \theta_k^{B'} + a_i^W \theta_{jk}^{W'} + d_i}}.$$

Assuming simple structure (i.e., an item loads on a single theoretical dimension, which amounts to one between- and one within-cluster dimension) and fixing item discrimination parameters to be equivalent for the same item across levels, the model simplifies to

$$P(U_{ijk} = 1 | \theta_k^B, \theta_{jk}^W, a_i, c_i, d_i) = c_i + (1 - c_i) \frac{e^{a_i (\theta_k^{B'} + \theta_{jk}^{W'}) + d_i}}{1 + e^{a_i (\theta_k^{B'} + \theta_{jk}^{W'}) + d_i}}.$$

Note that the typical exponent term $a(\theta - b)$ is written in the slope/intercept form $a\theta + d$ by replacing $-ab$ with the scalar d (Reckase, 2009, Chapter 4). Now item difficulty can be interpreted as item easiness. Despite the simple structure, the dimensions with zero loadings still play a role in the estimation of parameters. Also of importance are the two sets of ability estimates from the model: a between-cluster ability estimate $\hat{\theta}_k^B$ associated with Level 2 and a within-cluster ability estimate $\hat{\theta}_{jk}^W$ associated with Level 1. Conceptually, these represent the school-level and the student-level ability estimates, both of which are direct estimates of the model. In summary, not only does the model allow for the proper accommodation of nesting, but it also allows for the estimation of more reliable school-level ability measures due to their direct estimation rather than simple averaging of individual examinee ability estimates. Estimating and reporting direct and more reliable school-level estimates of ability would be especially appealing to policy makers and education practitioners.

The “Curse of Dimensionality”

Despite the vast theoretical support for MIRT models, applications in research and especially in practice remain limited, even when researchers have large enough sample sizes. The main reason is the so-called “curse of dimensionality” (Bellman, 2003, p. ix). With respect to measurement, this term means that when estimating a MIRT model using maximum marginal likelihood (MML), the estimation of item parameters requires numerical integration over a large number of quadrature points, which makes the estimation process computationally intensive and often intractable (Asparouhov & Muthén, 2012; Cai, 2010a, 2010b). Several alternative estimation methods that do not require the use of quadrature have been proposed; however, none of them appear to offer an optimal solution (see Cai, 2010a).

The MH-RM Algorithm

Overview

Recently, a new algorithm, MH-RM, was developed to overcome the “curse of dimensionality.” MH-RM was first proposed by Cai (2008) and later extended to high-dimensional exploratory (Cai, 2010a) and confirmatory item factor analysis (Cai, 2010b). Unlike other

methods, which either rely on high-dimensional numerical integration or approximation of the entire multidimensional response surface, MH-RM makes use of Fisher's identity to employ stochastic imputation (i.e., data augmentation) via the Metropolis–Hastings sampler and then apply the stochastic approximation method of Robbins and Monro to approximate the observed data likelihood. Thus, MH-RM is able to avoid both numerical integration and approximation of the entire posterior distribution, which makes it a particularly useful estimator for data with a large number of items, many dimensions, or missing data (Cai, 2010b). MH-RM has been implemented in flexMIRT (Cai, 2013), IRTPRO (Cai, Thissen, & du Toit, 2011), and the “mirt” package (Chalmers, 2012) in R (R Core Team, 2013). For more details, see Cai and Thissen (2015), who provided a short illustration comparing the calculations involved in the MH-RM and MML algorithms.

Prior Research on the Functionality of MH-RM

Several studies have compared MH-RM with some of the most popular estimation techniques, and the results are promising. Specifically, the bias and root mean square error (RMSE) of MH-RM parameter estimates and marginal maximum likelihood expectation-maximization (MML-EM) estimates were nearly identical in studies using a two- or three-dimensional graded-response model (Cai, 2010a, 2010b). The two procedures also yielded similar graded-response parameter estimates for a five-dimensional model (Cai, 2010a) and an eight-dimensional model (Cai, 2010b) with real data. Similarly, the bias and RMSE of MH-RM parameter estimates and Markov chain Monte Carlo (MCMC) estimates were nearly identical in studies using a four-dimensional graded-response model (Cai, 2010a), and MH-RM was considerably faster than MML or MCMC. Yavuz and Hambleton (2016) found similar RMSE for MML, MH-RM, and MCMC for a three-dimensional 3PL model. MML was faster than the other two methods for three dimensions, but their five-dimensional model could not be run in MML due to lack of computing power.

One study using seven-dimensional graded-response or 2PL data (Asparouhov & Muthén, 2012) found that confidence interval coverage of the loading estimates was significantly lower for MH-RM (54% for the polytomous data and 42% in the dichotomous case) than it was for other methods, including MCMC, although the parameter estimates were nearly unbiased. In summary, the MH-RM algorithm appears to be a promising tool in estimating MIRT models. Although prior research on its ability to recover item parameters is not entirely unanimous, the majority of simulation studies are quite favorable.

The existing research on MH-RM and its applications to operational data (e.g., Falk & Cai, 2016) show the potential of MH-RM to estimate a wide variety of models in practice. It is important to note, however, that though flexible, the algorithm is still fairly new, and much more research is needed to examine its performance when applied to various models. The algorithm has seldom been studied for 3PL data, although Chalmers (2015) found relatively unbiased estimates and reasonable *SEs* when recovering a multiple-group 3PL model. In addition, several studies have used MH-RM to examine various multilevel nonlinear models (Camilli & Fox, 2015; Kuhfeld, 2016; Sulis & Toland, 2017; Yang & Cai, 2014). Thus, multilevel models are another important area of research with respect to MH-RM in particular and as applied to educational measurement in general.

Study Purpose

The purpose of this study was to examine the performance of the MH-RM algorithm in the estimation of a 3PL ML-MIRT model under different conditions. The study conceptually

represented students nested within schools. Given the applications of MH-RM to single-level multidimensional and multilevel unidimensional data, it is important to know how accurate and efficient the MH-RM algorithm is in estimating these and more complex models (e.g., multilevel multidimensional measurement models). In particular, the study aimed to answer the following research questions:

Research Question 1: How well does the MH-RM algorithm recover the true parameter values?

Of particular interest was the bias and efficiency associated with item parameter estimates, Level 2 (between) variances and covariances, Level 1 (within) covariances, and ability estimates at both levels. In addition, it was of interest to examine the accuracy (i.e., lack of bias) of the *SEs* of item parameter estimates, latent trait variance and covariance estimates, and ability estimates.

Research Question 2: What conditions are optimal in obtaining accurate estimates of parameters?

Specifically, what combinations of number of dimensions, intraclass correlation coefficient (ICC), and sample size (i.e., number of clusters and cluster size) affect these estimates?

Method

Given no prior research on the performance of the MH-RM algorithm applied to 3PL ML-MIRT models, the focus of this study was on the most important aspects of such complex models. Specifically, the main challenge in estimating MIRT models has been the “curse of dimensionality.” Thus, examining whether and how well MH-RM can estimate 3PL ML-MIRT models was of primary interest. Similarly, analyzing nested data appropriately requires knowledge and understanding of several important characteristics of the data (e.g., the number of clusters and cluster size, the intraclass coefficient). As such, these characteristics were also considered in multilevel modeling. Finally, prior simulation and real-data research were used to inform the specific conditions to be investigated.

The current study varied the number of dimensions (three vs. five), the ICC level (.15, .25, and .35), the number of clusters (40 vs. 200), and cluster size (20 vs. 100). Crossing the levels of these four factors resulted in 2 dimension levels $\times$ 3 ICC levels $\times$ 2 numbers of clusters $\times$ 2 cluster sizes = 24 conditions total (see Appendix A in the supplemental materials).

Data Generation

The data for all conditions were generated via the “Simulation” mode in flexMIRT (Houts & Cai, 2013). The generating model was based on user-supplied item parameters and a latent variance–covariance matrix.

Item parameters. Given the number of dimensions (three or five), the hypothetical test length was set at 45 items, which allowed 15 items per dimension in the condition with three latent trait dimensions and nine items per dimension in the condition with five dimensions. The generating item parameters were held fixed across replications. In both the three- and five-dimensional models, three alternating values for the discrimination parameters were chosen ($a = 1, 1.5$, and 2). These values are on the logistic metric and correspond to about 0.59, 0.88, and 1.18 on the

normal metric. Normal priors were used with $\mu = 1.7$, $\sigma = 1$. The item difficulties were set at nine different values ($0, \pm 0.380, \pm 0.787, \pm 1.262$, and ± 1.922) determined by the inverse of the normal cumulative distribution with $\mu = 0$, $\sigma = 1.5$. These values are the familiar difficulty parameters in IRT. Before being supplied to flexMIRT, they were converted by $d_i = -b_i a_i$. The pseudoguessing parameters for all items were fixed to .20, assuming five-option multiple-choice items.

Latent variance–covariance matrix and ICC values. As the within-cluster variances were set to 1.0 for model identification purposes, the within-cluster covariances are on the correlation metric. Given the high correlations among dimensions found in educational data (Sinharay, 2010), both the within- and between-cluster correlations were set to values of .70, .80, and .90. The ICC values were specified by solving for σ_B^2 in the ratio of between to total variance $\rho = \sigma_B^2 / (\sigma_B^2 + \sigma_W^2)$, where $\sigma_W^2 = 1$.

Latent means and distributions. The abilities for each dimension in both levels were generated to be multivariate normal, with a mean of 0 and variance of 1.0 at Level 1 and variance at Level 2 at specific values to obtain the desired ICC levels of .15, .25, and .35.

Dependent Variables of Interest

Parameter recovery under the MH-RM algorithm was examined over 200 replications. Specifically, the accuracy and efficiency of item, group, and person-level parameters were assessed in terms of bias and sampling variability, respectively. The results were aggregated over parameters with the same generating value. For example, the bias and efficiency of all item discriminations with a generating value of 1 were aggregated across dimensions within the same condition. Given the inconclusive results of prior studies regarding MH-RM *SEs* described earlier, *SE* accuracy was also of interest. Bias and efficiency measures are defined below.

Bias. Bias is defined as the average difference between the estimated parameter and the generating parameter value. For a given parameter, β bias was computed as

$$\text{Bias}_\beta = \frac{\sum_{r=1}^R (\hat{\beta}_r - \beta)}{R},$$

where $\hat{\beta}_r$ is the parameter estimate from the r th replication, β is the true parameter value, and R is the total number of replications.

RMSE. The RMSE combines both bias and sampling variability of parameter estimates across replications. Specifically, RMSE was computed as

$$\text{RMSE}_\beta = \sqrt{\frac{\sum_{r=1}^R (\hat{\beta}_r - \beta)^2}{R}} = \sqrt{\text{Bias}_\beta^2 + SE_\beta^2},$$

where SE_β is the empirical standard error of the parameter (i.e., the standard deviation of the parameter estimates across replications), and all other elements are as defined above.

SE accuracy. The accuracy of *SEs* was examined by computing the average ratio of empirical to analytical *SE* for each parameter. Specifically, the empirical *SE* for a given parameter was

Table 1. Mean Bias, RMSE, and *SE* Accuracy for Item Parameters, Latent Variances and Covariances, and Ability Estimates.

Parameter	Bias (SD)	RMSE (SD)	Empirical <i>SE</i>	Analytical <i>SE</i>	<i>SE</i> ratio
Item difficulty	0.128 (0.058)	0.193 (0.109)	0.170	0.170	1.164
Item discrimination	0.013 (0.025)	0.123 (0.084)	0.148	0.145	1.044
Level 2 (between) variance	0.006 (0.007)	0.069 (0.038)	0.077	0.071	1.068
Level 2 (between) covariance/correlation	0.005 (0.005)	0.056 (0.034)	0.069	0.071	0.958
Level 1 (within) covariance/correlation	−0.020 (0.018)	0.030 (0.015)	0.020	0.026	0.831
Level 2 (between) ability	−0.075 (0.026)	0.211 (0.057)	0.154	0.189	0.927
Level 1 (within) ability	−0.007 (0.004)	0.498 (0.040)	0.408	0.534	0.778

Note. RMSE = root mean square error.

computed as the standard deviation of the parameter estimates across replications in a given condition. The analytical *SE* was output by flexMIRT for each replication. Then we computed the ratio of the empirical to analytical *SE* for each replication and averaged these ratio values by condition. The mean empirical and analytical *SE*s were also computed across conditions by taking the square root of the mean empirical and analytical error variances, respectively (see Reardon & Ho, 2015).

Results

Overall, item parameters (which were fixed to be equivalent across levels), latent variances and covariances, and abilities were reproduced well (Table 1). The EAP ability estimates at Level 2 (i.e., the cluster mean abilities) and Level 1 (i.e., individual deviations from the cluster mean abilities) were examined separately to distinguish bias at the cluster (e.g., school) level from bias at the individual (e.g., student) level. However, in practice, the ability estimates from the two levels would be summed to report individual students' ability estimates or scaled scores. Ratios of the empirical to analytical *SE*s (taking magnitude into account) indicated that the analytical *SE*s reported in flexMIRT were somewhat large for the Level 2 ability estimates, a little too large for the Level 1 ability estimates, and essentially accurate for the other parameters. Next, we focus on specific predictors of bias, efficiency, and *SE* accuracy based on linear regression. As a general rule, an effect had to be statistically significant, and, more importantly, explain at least 1% of the variance in the criterion, to be retained in the model.

Bias

Overall, all parameters were estimated accurately, except for item difficulty, which showed small positive bias, indicating that, on average, items were estimated to be easier than they actually were. Regression analysis revealed a positive relationship between generating *a* value and bias in item difficulty (see Figure 1). In addition, bias was consistently higher in the three-dimensional models than in the five-dimensional models. Finally, item difficulty bias appeared to be slightly lower with a larger number of small clusters (bottom left panel). Having the same overall sample size but made up of a small number of large clusters (top right panel) resulted in noticeably larger bias. Furthermore, the bias in item difficulty did

Figure 1. Item difficulty bias (y axis) as a function of generating a value (x axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

not appear to improve much by adding more large clusters (bottom right panel); in fact, it appeared to be more beneficial to have a large number of small clusters than the same large number of large clusters.

Unlike Level 1 (within) ability estimates, which were unbiased on average (except for extreme generating θ values), Level 2 (between) abilities were slightly underestimated on average. As shown in Figure 2, bias in the Level 2 (between) ability estimates was primarily a function of the generating ability level (i.e., generating θ value). Specifically, lower generating abilities were overestimated, whereas higher generating abilities tended to be underestimated. This inward bias is to be expected with Bayes estimates, as they are usually “pulled” toward the mean. It is important to note, however, that for a large range of the proficiency continuum, bias in ability estimates was very small. As to the other factors, bias was larger for small clusters, and bias decreased as the ICC level increased. However, the effect of the ICC level was not so profound when cluster size was large.

RMSE

RMSE showed the same patterns as bias for item parameters and latent variances and covariances. RMSE for the Level 2 (between) ability estimates was small. However, RMSE was noticeably larger for the Level 1 (within) ability estimates, although they were unbiased on average, indicating a lot more measurement error at Level 1. Regression analyses revealed that for Level 2 (between) abilities, RMSE was largely a function of cluster size, whereas for Level 1 (within) abilities, RMSE was essentially a function of only the generating θ value. Figures B1 and B2 in the supplemental materials show these effects.

SE Accuracy

As described earlier, *SE* accuracy was assessed by computing the average ratio of the empirical to the analytical *SE* for each parameter. An important factor that could affect the examination of *SE* accuracy is the number of observations. In this case, *SE* ratios were computed based on the estimates and analytical *SEs* for each parameter in the 200 replications. As discussed earlier, more observations are desirable, to assess the trustworthiness of *SEs* more accurately. However, there was another issue, which led to an even smaller number of observations for some parameters. Specifically, the analytical error variances (i.e., the squared analytical *SEs*) of certain parameters and conditions were undefined in some replications. Because of this, the *SEs* for these parameters and conditions were treated as missing. Thus, the average analytical *SE* for a given parameter was based only on the replications with an error variance for that parameter. In the case of parameters with the same generating value, only those parameters with an undefined error variance were excluded from the computation of *SE* ratios and average analytical *SEs*.

The analytical error variances for item discrimination and item difficulty were undefined under only a few conditions, and in no more than 1% of the replications. Undefined analytical error variances were much more prevalent for the latent variances and covariances. In the three-dimensional models, undefined analytical error variances occurred in less than 1% of the Level 2 (between) variances and covariances, and in up to 25% of the Level 1 (within) covariances/correlations, usually in the conditions with the largest sample size (20,000). In the five-dimensional models, undefined analytical error variances occurred much more frequently: In up to 17% of the Level 2 (between) variances and covariances, and in up to 53% of the Level 1 (within) covariances/correlations, indicating the *SE* ratios for these parameters may not be trustworthy. There were no undefined *SEs* for the ability estimates at either level, so the accuracy of these *SEs* was examined more closely.

Recall that the ratios of the empirical to analytical *SEs* indicated that the analytical *SEs* reported in flexMIRT were somewhat large for the Level 2 ability estimates, and a little too large for the Level 1 ability estimates. As shown in Figure B3 in the supplemental materials, the *SE* ratio for Level 2 (between) ability estimates was primarily a function of generating θ value. Specifically, at extremely low and extremely high levels of proficiency, the *SE* ratio was too low (indicating the analytical *SEs* were too large), whereas for the remainder of the ability continuum, the *SE* ratio was close to 1 (indicating the analytical *SEs* were accurate). Moreover, this effect of generating θ value was moderated by the ICC level: the higher the ICC level, the more accurate the *SEs*.

Unlike Level 2, the *SE* ratio for Level 1 (within) ability estimates was noticeably lower than the desired level, on average. In other words, the analytical *SEs* for the Level 1 (within) ability estimates were too large across the entire proficiency continuum, and more so at the extremes. Figure B4 in the supplemental materials shows the effects of generating θ value and, to a lesser extent, number of dimensions on the *SE* ratio for the Level 1 (within) ability estimates. In addition to the expected effect of generating θ value, Figure B4 shows that *SEs* for Level 1 (within) ability estimates were slightly more accurate in the three-dimensional model.

Discussion

Summary

The purpose of this study was to examine the performance of the MH-RM algorithm in the estimation of a 3PL ML-MIRT model under different conditions.

Figure 2. Level 2 (between) ability estimate bias (y axis) as a function of rounded generating θ value (x axis), cluster size (top vs. bottom panels), and ICC level (columns of panels from left to right).

Note. ICC = intraclass correlation coefficient.

In terms of bias, item difficulty (i.e., “easiness”) was slightly overestimated on average, meaning that items were estimated to be easier than they actually were, and especially so when the items were highly discriminating, and the model had three dimensions.

Unlike Level 1 (e.g., student) abilities, which were unbiased, Level 2 (e.g., school) ability estimates were underestimated. Although the reason for the direction of bias was not clear, bias behaved as expected. Specifically, low-generating abilities were overestimated, whereas high-generating abilities were underestimated. In addition, bias for Level 2 ability estimates was smaller when the clusters were larger, which also made sense—the bigger the clusters, the less biased the cluster-level abilities. Thus, if cluster means are of primary interest, larger clusters are desired. When interpreting the magnitude of bias for the abilities, it is important to consider that 95% of Level 1 units (e.g., students) had generating abilities between -1.96 and 1.96 , and 95% of Level 2 units (e.g., schools) had abilities within ± 0.822 for ICC = .15, within ± 1.131 for ICC = .25, and within ± 1.428 for ICC = .35. Within these ranges, bias was small, especially at Level 2. For example, in Figure 2, bias appeared greater when the ICC was lower, but this was largely because the -2 to 2 range included values that were relatively more extreme for the lower ICC levels. Thus, overall MH-RM recovered the vast majority of ability estimates quite well at both levels.

RMSE showed a pattern similar to bias for each of the parameters. This was not surprising, as RMSE combines bias and sampling variability. On average, RMSE was small for item difficulty and item discrimination, very small for the latent variances and covariances/correlations, not overly large for the Level 2 ability estimates, and noticeably larger for the Level 1 ability estimates.

Similar to bias, RMSE for item difficulty and item discrimination was smallest for items with low discrimination. For item parameters as well as the latent variances and covariances, the

largest total sample size (20,000) was associated with the smallest RMSE. For Level 2 variances and covariances, the mid-level sample size (4,000) showed a clear advantage of a larger number of small clusters over a smaller number of large clusters. Thus, although a large total sample size is desirable, it appears to be better to have more clusters, even if they are small. This finding aligns with the large body of research in the multilevel literature converging on the same conclusion (e.g., Maas & Hox, 2005; Snijders, 2005; Spybrook, 2008).

For the ability estimates, RMSE also behaved as expected. Specifically, both Level 2 and Level 1 ability estimates at the extremes of the proficiency continuum were associated with greater RMSE because there was less information in these ranges, whereas abilities near the middle had smaller RMSE. In addition, for Level 2 ability estimates, RMSE was smaller in larger clusters. Again, when Level 2 (e.g., school-level) ability estimates are of primary importance, the sampling design should include larger clusters (i.e., more students per school).

SE ratios were used to determine the accuracy of the analytical *SEs* produced by MH-RM in flexMIRT. It is important to note that the interpretation of *SE* ratios is dependent upon and limited by the availability of meaningful error variances. As mentioned above, the analytical error variances for the latent variances and covariances in a substantial number of replications were undefined, especially in the five-dimensional models. This problem stems from the limited amount of information in complex models; it is not an issue in simpler models. More specifically, in complex models such as the ones in this study, one is trying to estimate the error variances of unknown latent parameters (e.g., latent variances/covariances) with very little information about them. As a result, the algorithm encounters a nonpositive definite matrix, and the error variance becomes undefined (L. Cai, personal communication, April 19, 2015).

Because of this issue, the undefined error variances (and *SEs*) were treated as missing in the computation of *SE* ratios by which *SE* accuracy was evaluated. Therefore, one should interpret the *SE* ratios for the latent variances and covariances with great caution, as they were based on fewer replications, and even 200 replications might be considered too few in the evaluation of *SEs*. That said, the results revealed that, on average, the analytical *SEs* were somewhat large for the Level 2 ability estimates, a little too large for the Level 1 ability estimates, and essentially accurate for the other parameters.

Limitations

Despite the extensive scope of this article, the design and execution of the simulation study have several limitations that are worthy of consideration. First, it is important to acknowledge that in all models, item parameters were constrained to be of the same magnitude across levels. That is, a single item difficulty and item discrimination were estimated for each item for both levels of the measurement model. By applying this constraint, one is assuming that the items function the same way at the cluster level as they do at the individual level. This constraint was not mandatory. In fact, the model allows for the item parameters to differ across levels. In some disciplines (e.g., industrial/organizational psychology), there are constructs that have substantively different meanings and may necessitate the free estimation of item loadings at different levels (see Bliese & Jex, 2002). However, in educational measurement, it may be difficult to make an argument for freely estimating the item parameters at different levels, especially when one considers the interpretation of those parameters.

Perhaps a more serious limitation concerns the specification of multilevel measurement models. In this study, the Level 1 (within) variances were constrained to 1 for identification purposes, whereas the Level 2 (between) variances for each dimension were freely estimated. This is a rather strong assumption, and it implies that the within-cluster variance in all Level 2 units was the same. In practice, this translates into having the same variability in student ability across

schools, which may or may not be the case. In future work, researchers could revise the software algorithms to allow heterogeneous within-cluster variance.

Finally, the results of the study were based on 200 replications for each condition. Although this number of replications may be sufficient for the examination of bias, a much larger number of replications (e.g., 1,000 or more) is desirable for the proper assessment of *SE* accuracy. This is especially true for the latent variances and covariances whose error variances were often undefined in the five-dimensional models.

MH-RM as an Estimator of Multilevel Measurement Models

Overall, the MH-RM algorithm performed well in the estimation of the three- and five-dimensional multilevel measurement models examined in this study. Although our findings are not directly comparable with prior research, a crude comparison of the results from this study and the research on the functionality of MH-RM compared with other estimation methods revealed that the results obtained here generally agree with the findings in published research. For example, the RMSE for item loadings reported in Cai (2010a, 2010b) is similar to the average RMSE for item discrimination reported here. Similarly, there was little bias for the item parameters and latent variances and covariances found here and in prior research. Thus, as far as item parameters and latent variances and covariances are concerned, the current study provided further support for MH-RM as a promising solution to the “curse of dimensionality” prohibiting the estimation of high-dimensional measurement models.

However, the purpose of modeling item response data is not only to obtain item parameters and examine the variances and covariances across dimensions but also to produce ability estimates. Thus, it is important to know how accurate those estimates are, and whether their *SEs* can be trusted. No known study has examined the accuracy of ability estimates produced by MH-RM with the 3PL ML-MIRT model. Therefore, the current study is the first to shed some light in this area. In terms of bias, the authors found that Level 2 ability estimates were slightly biased, whereas Level 1 ability estimates were unbiased. However, the bias was not overly large, and within level, it behaved in expected ways. Thus, the results supported MH-RM as a viable estimator on this front as well. What is concerning, on the contrary, is that the *SEs* for Level 2 and especially for Level 1 ability estimates were too large. Again, the examination of *SE* accuracy in the current study was somewhat limited; therefore, this finding warrants further investigation with more replications.

The bottom line is that MH-RM appears to be a viable option in the estimation of multilevel measurement models with as many as five dimensions on each level, as evidenced by the larger part of the results examined here. More importantly, this study unlocks the potential for application and future research of MH-RM in ML-MIRT models.

Implications for Practice

The 3PL ML-MIRT model presented in this article has enormous potential for educational measurement practice. K-12 education practitioners continuously demand more diagnostic feedback from assessments for accountability, such as the tests covering the Common Core State Standards, to also help diagnose and address students' specific needs. One way of meeting this demand is the implementation of multidimensional models, where one can model multiple subdomains within a subject area simultaneously. Importantly, specifying a model with multiple (typically highly correlated) dimensions can help reduce the number of items per dimension needed to achieve a certain level of precision compared with unidimensional models.

The other feature of the 3PL ML-MIRT model that can be extremely beneficial in practice is the measurement of proficiency at multiple levels. That is, the model not only properly accommodates the hierarchical structure of the data due to nesting (e.g., students nested within schools), but it also produces estimates of ability at the individual (e.g., student) and the cluster (e.g., school) levels. Thus, the 3PL ML-MIRT model allows for the estimation of more reliable cluster-level ability measures than those that would be obtained by simply averaging the individual ability estimates within clusters. This is because in the 3PL ML-MIRT model, the cluster means are estimated directly using information from all other schools in the model. In educational measurement, these school-level estimates would be particularly useful, especially within a school district or a state, where policy decisions are often based on aggregate school achievement metrics. As such, the models discussed here have direct implications for practice in that policy decisions will be made on the basis of more dependable scores. This, in turn, also increases the validity of inferences based on school-level estimates of achievement.

The applications of the 3PL ML-MIRT model are not limited to the school level. In practice, the 3PL ML-MIRT model could be applied with two-stage sampling designs frequently used in international achievement testing programs as well as with census data encountered in K-12 state assessments for accountability. Overall, the results presented in this study provided substantial support for the use of the model in practice. However, given the limited research on the performance of MH-RM as an estimator of the 3PL ML-MIRT model, it is up to education practitioners and policy makers to decide when to apply the model and for what purposes, depending on the questions at hand. As a reminder, the MH-RM algorithm is remarkably fast, especially when used with a powerful computer (e.g., four or more logical processors), which is especially desirable if the total sample size is large. Thus, one could always estimate a 3PL ML-MIRT model and compare the results across different models (e.g., a single-level MIRT or several multilevel unidimensional models). One could then assess the advantages and disadvantages of the models in light of the data, the research questions, and the simulation results described here.

Future Research

There are many opportunities for further research on the functionality of MH-RM in general. In this study, the accuracy and efficiency of MH-RM were specifically examined as applied to multilevel multidimensional models under various conditions, and although the scope of the study was extensive, there are many more avenues for further research. Below we point out to several directions for future research.

The research design of the current study could be enhanced in several ways. One modification of the design could involve the specification of different correlations among the dimensions at different levels. For example, it could be that in reality the dimensions are much more highly correlated at the school or classroom level than they are at the student level, controlling for cluster membership (e.g., Höhler et al., 2010). It would be interesting to explore whether and how specifying different correlations at different levels affects the results in terms of bias, RMSE, and possibly *SEs* at different levels.

Another way to build on the 3PL ML-MIRT model is to add predictors at Level 2 and Level 1. For example, several demographic/background variables at both the school and student levels could be added to the model to explain some of the variability in ability estimates. Now the model has two parts: a measurement part (such as 3PL ML-MIRT models examined in this study) and a structural part, which would provide regression coefficients and significance tests for the Level 2 and Level 1 predictors of the Level 2 (between) and Level 1 (within) variances. The main advantage of estimating both parts of the model in a unified framework (i.e., in a single hybrid model) is that the dependent variables in the structural part of the model are latent.

That is, unlike a traditional multilevel model in which one would model the ability estimates as observed dependent variables prone to measurement error, the hybrid model allows for more accurate estimates of the regression coefficients in the structural part of the model because measurement error in the ability estimates is taken into account. Although it would be interesting to examine such a model for research purposes, more research is needed to evaluate these models before they can be applied in practice.

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.

Supplemental Material

The online appendix is available at <http://journals.sagepub.com/doi/suppl/10.1177/0146621616688923>.

References

- Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, 22, 47-76.
- Asparouhov, T., & Muthén, B. (2012). *Comparison of computational methods for high dimensional item factor analysis* (Mplus Technical Report). Retrieved from <http://www.statmodel2.com/download/HighDimension.pdf>
- Bellman, R. (2003). *Dynamic programming*. Mineola, NY: Dover Publications.
- Bliese, P. D., & Jex, S. M. (2002). Incorporating a multilevel perspective into occupational stress research: Theoretical, methodological, and practical implications. *Journal of Occupational Health Psychology*, 7, 265-276.
- Cai, L. (2008). *A Metropolis-Hastings Robbins-Monro algorithm for maximum likelihood nonlinear latent structure analysis with a comprehensive measurement model* (Unpublished doctoral dissertation). Department of Psychology, University of North Carolina at Chapel Hill.
- Cai, L. (2010a). High-dimensional exploratory item factor analysis by a Metropolis-Hastings Robbins-Monro algorithm. *Psychometrika*, 75, 33-57.
- Cai, L. (2010b). Metropolis-Hastings Robbins-Monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35, 307-335.
- Cai, L. (2013). flexMIRT® Version 2.0: A Numerical Engine for Flexible Multilevel Multidimensional Item Analysis and Test Scoring [Computer software]. Chapel Hill, NC: Vector Psychometric Group.
- Cai, L., & Thissen, D. (2015). Modern approaches to parameter estimation in item response theory. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling* (pp. 41-59). New York, NY: Routledge.
- Cai, L., Thissen, D., & du Toit, S. (2011). IRTPRO for Windows [Computer software]. Chicago, IL: Scientific Software International.
- Camilli, G., & Fox, J.-P. (2015). An aggregate IRT procedure for exploratory factor analysis. *Journal of Educational and Behavioral Statistics*, 40, 377-401.
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.
- Chalmers, R. P. (2015). Extended mixed-effects item response models with the MH-RM algorithm. *Journal of Educational Measurement*, 52, 200-222.

- Cheong, Y. F., & Raudenbush, S. W. (2000). Measurement and structural models for children's problem behaviors. *Psychological Methods*, 5, 477-495.
- de la Torre, J., & Patz, R. J. (2005). Making the most of what we have: A practical application of multidimensional item response theory in test scoring. *Journal of Educational and Behavioral Statistics*, 30, 295-311.
- Falk, C. F., & Cai, L. (2016). A flexible full-information approach to the modeling of response styles. *Psychological Methods*, 21, 328-347.
- Höhler, J., Hartig, J., & Goldhammer, F. (2010). Modeling the multidimensional structure of students' foreign language competence within and between classrooms. *Psychological Test and Assessment Modeling*, 52, 323-340.
- Houts, C. R., & Cai, L. (2013). *flexMIRT® user's manual version 2.0: Flexible multilevel multidimensional item analysis and test scoring*. Chapel Hill, NC: Vector Psychometric Group.
- Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.
- Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345-388). Charlotte, NC: Information Age Publishing.
- Kuhfeld, M. R. (2016). *Multilevel item factor analysis and student perceptions of teacher effectiveness* (Unpublished doctoral dissertation). University of California—Los Angeles.
- Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 1, 86-92.
- Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.
- R Core Team. (2013). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Available from <http://www.R-project.org/>
- Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics*, 40, 158-189.
- Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.
- Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, 47, 150-174.
- Snijders, T. A. B. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1570-1573). Chichester, UK: John Wiley.
- Snijders, T. A. B., & Bosker, R. J. (2010). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, England: Sage.
- Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273-311). Charlotte, NC: Information Age Publishing.
- Sulis, I., & Toland, M. D. (2017). Introduction to multilevel item response theory analysis: Descriptive and explanatory models. *Journal of Early Adolescence*, 37, 85-128.
- Wang, W. C., Chen, P. H., & Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods*, 9, 116-136.
- Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis-Hastings Robbins-Monro algorithm. *Journal of Educational and Behavioral Statistics*, 39, 550-582.
- Yavuz, G., & Hambleton, R. K. (2016). Comparative analyses of MIRT models and software (BMIRT and flexMIRT). *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164416661220
- Zhang, J. (2012). Calibration of response data using MIRT models with simple and mixed structures. *Applied Psychological Measurement*, 36, 375-398.