Spring 2015

Examining the performance of the Metropolis-Hastings Robbins-Monro algorithm in the estimation of multilevel multidimensional IRT models

Bozhidar M. Bashkov
James Madison University

Follow this and additional works at: https://commons.lib.jmu.edu/diss201019

Part of the Educational Assessment, Evaluation, and Research Commons, Multivariate Analysis Commons, Quantitative Psychology Commons, Statistical Methodology Commons, and the Statistical Models Commons

Approved and recommended for acceptance as a dissertation in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Special committee directing the dissertation work of Bozhidar M. Bashkov

Christine E. DeMars   3/12/15
Christine E. DeMars   Date

Monica K. Erbacher   2/13/15
Monica K. Erbacher   Date

Daniel P. Jurich   2/13/15
Daniel P. Jurich   Date

Dena A. Pastor   2/15/15
Dena A. Pastor   Date

Received by The Graduate School

_____
Date

Examining the Performance of the Metropolis-Hastings Robbins-Monro Algorithm in the

Estimation of Multilevel Multidimensional IRT Models

Bozhidar M. Bashkov

A dissertation submitted to the Graduate Faculty of

JAMES MADISON UNIVERSITY

In

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

Assessment & Measurement

May 2015

Acknowledgements

I would like to acknowledge several people who have supported me while I was working on this project.

First, I would like to thank my family and friends for believing in me, encouraging me to persist and give my best, helping me get through the tough times, and celebrating with me the small victories I accomplished in graduate school. I feel so blessed to have each and every one of you in my life.

Second, I would like to give a big shout-out to my dissertation committee members for their invaluable feedback and insightful comments on the proposal version of this document. I greatly appreciate their time and efforts.

Last but not least, I would like to express my most sincere gratitude for my academic advisor, Dr. Christine DeMars, who tirelessly worked with me over the past few years and helped me grow both academically and professionally. It has been a great honor to work with you. Thank you so much!

Table of Contents

List of Tables

List of Figures

Abstract

The purpose of this study was to review the challenges that exist in the estimation of complex (multidimensional) models applied to complex (multilevel) data and to examine the performance of the recently developed Metropolis-Hastings Robbins-Monro (MH-RM) algorithm (Cai, 2010a, 2010b), designed to overcome these challenges and implemented in both commercial and open-source software programs. 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, which decreases estimation time tremendously. Thus, the algorithm shows great promise in the estimation of complex models applied to complex data.

To put this promise to the test, the accuracy and efficiency of MH-RM in recovering item parameters, latent variances and covariances, as well as ability estimates within and between groups (e.g., schools) was examined 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. More replications are needed to better determine the accuracy of analytical standard errors for some of the parameters. Limitations of the study, implications for educational measurement practice, and directions for future research are offered.

Chapter I

Introduction

The field of educational measurement has grown rapidly and vastly over the last few decades. A major contributor to this development is the ever-increasing power of computers to perform complex computational tasks, often in a fraction of the time needed to execute such tasks ten or twenty years ago. Sophisticated models, which are arguably a closer approximation of reality than simple models (McDonald, 2000; Reckase, 1997), are now not only possible to estimate but also viable options to employ in practice. Moreover, researchers have begun to account for the nested (hierarchical) structure of educational data by modeling the different sources of variability in test scores and their predictors using multilevel models (e.g., Adams, Wilson, & Wu, 1997). The purpose of this dissertation is to review the challenges that exist in estimating multidimensional models applied to multilevel data and examine the performance of a recently developed algorithm implemented in commercial and open-source software programs to overcome these challenges.

Chapter I serves as a brief introduction to multidimensional and multilevel models and their use in educational measurement. It also provides an overview of the algorithm under study, its applications in published research, as well as the purpose and specific research questions of this dissertation. In Chapter II, I discuss in more depth the challenges in estimating multidimensional models using popular estimation methods. In addition, I review the development, specification, and interpretation of multilevel measurement models. I conclude this chapter with a presentation of the multilevel multidimensional model under study. Chapter III lays out the method used to examine the

research questions posed at the end of the introduction and explains the choice of

conditions and specific levels for the simulation study. In Chapter IV, I present the results

with the aid of visual displays. Finally, Chapter V provides a summary of the results and

draws conclusions on the accuracy and efficiency of MH-RM in the estimation of

multilevel multidimensional measurement models and offers implications for educational

measurement practice as well as directions for future research.

Background

Assessment practitioners usually design and administer tests that measure not one

but several abilities or latent traits. For example, large-scale testing programs such as the

SAT® and the GRE® contain multiple subtests (e.g., reading/verbal reasoning,

math/quantitative reasoning, writing) to obtain a multifaceted picture of students'

readiness for college and graduate school, respectively. 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.

Once ability estimates or scaled scores are obtained, these scores are correlated with each

other and/or other measures to investigative substantive research questions. 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. That is, the

correlations among latent traits, which could help in the estimation of item and person

parameters, are essentially ignored. 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 domain is small, as it is in large testing programs such as

NAEP (Zhang, 2012).

Multidimensional Item Response Theory (MIRT)

An alternative approach to modeling multiple constructs is simultaneous

estimation via multidimensional IRT (MIRT), which extends the unidimensional model

to include multiple latent traits. Under the unidimensional three-parameter logistic (3PL)

model (Birnbaum, 1968; Lord, 1980) a typical examinee j's conditional probability of

correct response to a dichotomously scored item i ($P(U_{ij}) = 1$) is a function of a single

latent variable θ_j and the item parameters (a_i = discrimination, b_i = difficulty, and c_i =

lower asymptote[1]):

$$P(U_{ij} = 1 \mid \theta_j, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{a_i(\theta_j - b_i)}}{1 + e^{a_i(\theta_j - b_i)}}. \tag{1.1}$$

Under the 3PL MIRT model, the probability of correct response to an item is a function

of *multiple* latent variables $\boldsymbol{\theta_j} = (\theta_{j1}, \theta_{j2}, \ldots, \theta_{jm})$ related to the item via a vector of loadings

$\mathbf{a_i} = (a_{i1}, a_{i2}, \ldots, a_{im})$, where m = the number of dimensions (Reckase, 2009):

$$P(U_{ij} = 1 \mid \boldsymbol{\theta_j}, \mathbf{a_i}, c_i, d_i) = c_i + (1 - c_i)\frac{e^{\mathbf{a_i}\boldsymbol{\theta_j}' + d_i}}{1 + e^{\mathbf{a_i}\boldsymbol{\theta_j}' + d_i}}. \tag{1.2}$$

Here the exponent term $a(\theta - b)$ from Equation (1.1) is written in the slope/intercept form

$a\theta + d$ by replacing $-ab$ with the scalar d (Reckase, 2009, Chapter 4). Note that the

model presented in Equation (1.2) is a *compensatory* MIRT model, meaning that low

ability in one dimension can be compensated for by high ability in another dimension. It

is also worth noting that if the test has simple structure (i.e., the test is multidimensional

---

[1] The lower asymptote, also known as the pseudo-guessing parameter, indicates the probability of correct response for examinees of low proficiency, possibly due to guessing.

but each individual item loads onto a single dimension), only one discrimination parameter takes on a nonzero positive value at a time. In this situation, the scalar d can be converted to the familiar difficulty parameter:

$$b_i = \frac{-d_i}{\sqrt{a_i^2}} .$$

(1.3)

It is important to note, however, that even though the simple-structure MIRT model resembles a unidimensional model, the estimation of the model is still multidimensional in nature in that the dimensions with zero loadings still play a role in the estimation of parameters. This is akin to the borrowing of information in score augmentation techniques (e.g., Wainer et al., 2001). In essence, 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). 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).

Proponents of MIRT models argue that in reality items and tests are rarely strictly unidimensional; therefore multidimensional models should be used over unidimensional models to account for the multidimensionality (Ackerman, 1994). Hartig and Höhler (2009) highlighted three types of applications of MIRT models. First, MIRT models can be used to accommodate unintended multidimensionality when a unidimensional construct was originally assumed. For example, groups of items based on a common stimulus (known as "testlets") can often share variability above and beyond the main trait being measured; thus, the unidimensional model could be extended post hoc to a MIRT

model (e.g., a bifactor model) to accommodate the additional covariability within testlets after controlling for the primary trait (DeMars, 2006; Wainer, Bradlow, & Wang, 2007).

Second, when a test was intentionally designed to measure multiple dimensions (i.e., latent traits[2]), MIRT models allow the examination of the latent covariance structure among the modeled traits. In fact, the latent trait covariance matrix is an automatic byproduct of the analysis. Importantly, since these relationships are estimated at the latent level, they are stronger and more accurate than the observed correlations among subtests based on raw scores (i.e., number correct). Even the correlations among traits based on unidimensional IRT ability estimates or scaled scores tend to be underestimated, unless they are disattenuated for measurement error (see de la Torre & Patz, 2005). However, in the presence of complex structure, these relationships can be overestimated (Zhang, 2012).

Finally, MIRT models make it possible to model data with complex structure where multiple skills or solution strategies can impact the probability of correct response[3]. By far the most prominent implication of this application of MIRT models is the investigation of DIF from a multidimensional perspective (Ackerman, 1992; Jeon, Rijmen, & Rabe-Hesketh, 2013; Walker & Beretvas, 2001). Given the numerous advantages of MIRT models, it is not surprising that many methodologists recommend their use to model complex (multidimensional) constructs (e.g., Ackerman, 1994; Reckase, 2009). The numerous benefits of MIRT models do not come without a price,

---

[2] The term *latent trait* is used to refer to the substantive construct underlying the data, whereas the term *dimension* is used to refer to the statistically estimated representation of this trait in the model.
[3] Note that only simple-structure MIRT models are considered here to keep the models relatively simple given the multilevel structure of the data. See Chalmers and Flora (2014) for an extensive study of single-level complex-structure MIRT models estimated via the Metropolis-Hastings Robbins-Monro (MH-RM) algorithm.

however. In the following section, I describe the main challenge that has hindered a widespread use of these models in practice.

The "Curse of Dimensionality"

Despite the vast theoretical support for MIRT models, applications in research and especially in practice remain limited, even with 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 Gaussian quadrature points, which makes the estimation process computationally intensive and often intractable (Asparouhov & Muthén, 2012; Cai, 2010a, 2010b). For example, if each of three latent trait dimensions is integrated over 9 fixed quadrature points, the total number of quadrature points increases geometrically to $9^3 = 729$. The default number of fixed quadrature points in many statistical software packages is often much larger than nine, which makes the estimation of a model with more than four or five dimensions impossible. The default number of quadrature points can typically be lowered manually by the researcher; however, doing so may lead to convergence problems or decrease the accuracy of posterior means and standard deviations of parameter estimates (Lesaffre & Spiessens, 2001; Rabe-Hesketh, Skrondal, & Pickles, 2002). 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). I review these alternatives in Chapter II.

The Metropolis-Hastings Robbins-Monro (MH-RM) Algorithm

Overview. Recently, a new algorithm, Metropolis-Hastings Robbins-Monro (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, du Toit, & Thissen, 2011), and the "mirt" package (Chalmers, 2012) in the open-source statistical programming environment R (R Core Team, 2013).

Applications. Several published studies have employed the MH-RM algorithm. Yang and Cai (2014) used MH-RM in the estimation of contextual effects[4] via nonlinear multilevel latent variable modeling and illustrated the model using data from the Programme for International Student Assessment (PISA). Thissen (2014) applied MH-RM to estimate a correlated six-dimensional model on data from the certification exam of the American Production and Inventory Control Society (APICS). Wiley, Shavelson, and

---

[4] In multilevel modeling, contextual effects are the effects of group (i.e., Level 2) variables on the dependent variable after controlling for the effect of individual (i.e., Level 1) variables. In other words, there is a difference in the relationship between the predictor and the criterion at different levels of the analysis. An example of a contextual effect is the effect of school-level socioeconomic status after controlling for individual student socioeconomic status.

Kurpius (2014) used MH-RM to examine the factor structure of the current version of the SAT®. Finally, Wright (2013) applied MH-RM in the estimation of unidimensional and multidimensional IRT models in efforts to gather construct validity evidence for situational judgment tests.

These applications of the MH-RM algorithm to operational data 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. For example, Wiley and colleagues (2014) applied the algorithm with a 3PL model; however, no known study has examined the accuracy and efficiency of MH-RM with this model. In addition, the study by Yang and Cai (2014) is the only one to use MH-RM in the estimation of a multilevel nonlinear model. Thus, multilevel models are another important area of research with respect to MH-RM in particular and as applied to educational measurement in general.

Multilevel Models

Multilevel models are a family of statistical models developed to accommodate and appropriately model data with hierarchical (i.e., nested or clustered) structure (Raudenbush & Bryk, 2002; Hox, 2010; Snijders & Bosker, 2010). For example, students are nested within schools, nurses are nested within hospitals, employees are nested within companies, etc. Because of 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 assumed by regression. Ignoring this violation results in underestimated standard errors and inflated Type I error for inferential tests.

Fortunately, multilevel models allow for the decomposition of variance within and between clusters, so that the standard errors of parameter estimates and any inferential tests associated with them are more accurate. Furthermore, multilevel models allow for the inclusion of person- and cluster-level predictors as well as cross-level interactions to explain variability in the outcome. Because of their flexibility, multilevel models have been widely used to model educational data. For example, researchers and policy makers are often interested in students' achievement, after controlling for student background and school-level variables. Multilevel models are a natural choice for this purpose. Alternatively, one may look for contextual or compositional effects that are highly related to differences in achievement and seek ways to minimize their influence.

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. Prior efforts in this area 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 et al., 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).

An alternative approach is to extend the measurement model to multiple levels. That is, 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. This is the approach taken by Höhler, Hartig, and Goldhammer (2010), although they specified only a 2PL model and were only interested in the latent covariance structure at different levels, and how it differs from a model which ignores the nested structure of the data. However, their model could easily be extended to a 3PL model to account for the probability of correct response by examinees of low proficiency. More importantly, ability estimates could be estimated at both the examinee level and the school level. In fact, this is one of the most attractive features of the multilevel MIRT model. That is, 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 educators and policymakers.

In sum, Wiley and colleagues (2014) applied the MH-RM algorithm with a multidimensional 3PL model in a single-level analysis, and Yang and Cai (2014) used MH-RM to estimate a multilevel nonlinear 2PL unidimensional model. However, no known study has examined the use of MH-RM with multilevel *and* multidimensional data. The current study is intended to serve this purpose.

The Current Study

   Purpose. The purpose of this dissertation is to examine the performance of the

MH-RM algorithm in the estimation of a 3PL multilevel MIRT (3PL ML-MIRT) model

under different conditions. The study will conceptually represent 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 dissertation strives to answer

the following research questions:

   Research question #1: How well does the MH-RM algorithm recover the true

parameter values? Of particular interest is 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 is of interest to examine

the accuracy (i.e., lack of bias) of the standard errors 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, and sample size (i.e., number of clusters and cluster

size) affect these estimates? Additionally, how long does it take on average across

replications to estimate the model in each of these conditions?

<div style="text-align: center">

Chapter II

Review of the Literature

</div>

Popular Methods for Estimating MIRT Models

Several methods have been developed over the last few decades to allow the estimation of high-dimensional IRT models and to make the estimation process more time-efficient. In the following sections I describe each method conceptually, highlighting both its desirable features and its limitations. The first of these methods is an extension of the fixed quadrature method discussed in the previous chapter.

Adaptive quadrature MML. As discussed in the introduction, estimating the item parameters of a MIRT model via maximum marginal likelihood (MML) relies on the numerical integration of the latent trait variables by use of quadrature points. The problem of using fixed Gaussian-Hermite quadrature points arises when the number of dimensions increases to four or five (or more) because the total number of quadrature points increases by a power equal to the number of dimensions, making the evaluation of integrals extremely difficult to impossible. For example, if each dimension in a five-dimensional model is integrated by 9 quadrature points, the total number of quadrature points amounts to $9^5 = 59,049$. To circumvent this problem, methodologists proposed adaptive quadrature rules (Liu & Pierce, 1994; Naylor & Smith, 1982). Unlike fixed-point quadrature, adaptive quadrature approximates the posterior distribution by strategically placing the quadrature nodes under areas of the distribution that are more "interesting," that is, of higher density (Liu & Pierce, 1994, p. 264). As a result, fewer quadrature points are required for approximation with adequate accuracy (Schilling & Bock, 2005). Although adaptive quadrature MML is currently the most popular MIRT

estimation technique, it remains limited in the number of latent dimensions it can handle. Specifically, although the total number of quadrature points is smaller than it is for fixed quadrature, the number of quadrature points still increases geometrically as the number of dimensions increases linearly. Further, when adaptive quadrature MML is implemented with the expectation maximization (EM) algorithm, the asymptotic covariance matrix is not automatically a byproduct of the calibration; thus, standard errors must be estimated in a separate step (Cai, 2010a). This does not mean that the standard errors are any less accurate. The two-step approach simply adds to the computation time of the MML-EM method.

   MCEM. Another way of estimating a MIRT model is via the Monte Carlo expectation maximization (MCEM) algorithm (Meng & Schilling, 1996). In this algorithm, the integration in the E-step is achieved by sampling the quadrature points (i.e., Monte Carlo simulation), in place of using Gaussian-Hermite quadrature. However, as Cai (2010a, 2010b) points out, in order to achieve pointwise convergence of parameter estimates, the simulation size (i.e., number of random draws) must increase tremendously, especially in the last few iterations, as the parameter estimates get closer to the maximum of the likelihood function. Moreover, the convergence time of MCEM is increased due to the fact that for each E-step iteration, the sampler generates a new set of random draws. Given these limitations, MCEM may not be the algorithm of choice in practical applications.

   MCMC. Finally, a fully Bayesian (i.e., stochastic) approach to estimating MIRT models involves multiple imputation from the posterior distribution via Markov Chain Monte Carlo (MCMC) procedures. Specifically, a Markov chain is specified such that its

target or invariant measure (i.e., the stationary distribution to which it converges) is the

posterior distribution, from which point estimates of the parameters can be obtained (see

Keller, 2005). There are two common sampling techniques for the Markov chain—the

Gibbs sampler and the Metropolis-Hastings algorithm within Gibbs. The Gibbs sampler

operates on the Birnbaum paradigm, where one set of parameters are estimated

conditional on (i.e., holding constant) another set of parameters (see Baker & Kim, 2004,

Chapter 4). For example, let $\boldsymbol{\beta}$ represent a set of item parameters and let $\boldsymbol{\theta}$ represent a set

of latent variables. Starting with some provisional item parameters $\boldsymbol{\beta}^t$, the $(t + 1)$th

iteration of the MCMC algorithm goes through two stages. In the first stage, ability

parameters $\boldsymbol{\theta}^{(t+1)}$ are drawn from the complete conditional distribution $\boldsymbol{\theta}^{(t+1)} \sim \Pi(\boldsymbol{\theta} \mid \mathbf{Y}, \boldsymbol{\beta}^t)$,

where $\mathbf{Y}$ denotes the observed response data. In the second stage, new values for the item

parameters $\boldsymbol{\beta}^{(t+1)}$ are drawn from the complete conditional distribution

$\boldsymbol{\beta}^{(t+1)} \sim \Pi(\boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\theta}^{(t+1)})$. The simulation process continues until the Markov chain

converges to the posterior distribution; that is, after some large number of iterations and

discarding some initial draws (i.e., burn-in cycles), the draws from the complete

conditional distribution can be assumed to come from the posterior distribution.

  Although virtually any posterior distribution can be approximated using MCMC

with the Gibbs sampler, some distributions (e.g., those involving a large number of item

parameters and many dimensions) may be extremely difficult to approximate

computationally. For such cases, the Metropolis-Hastings (MH) algorithm within Gibbs

can be very useful. MH alleviates the computational burden on the Gibbs sampler in two

ways. First, it allows the specification of proposal distributions for the parameters. The

advantage here is that it is far easier to sample from a proposal distribution than from the

complete conditional distribution[5]. Second, once the proposal distributions have been specified, rather than retaining all draws from the Markov chain, each draw from the proposal distribution can be evaluated based on its probability to be also from the complete conditional distribution. Specifically, if the probability of the proposed draw is higher than that of the current state, the draw is accepted with probability 1; if the probability of the proposed draw is lower than that of the current state, the probability of accepting the draw depends on the ratio of the likelihood of the current draw to the likelihood of the previous draw (see Keller, 2005 for details).

These two features of the MH algorithm substantially speed up the estimation process compared to using only the Gibbs sampler. However, one big challenge is the specification of appropriate proposal distributions, which may need to be determined empirically by trial and error. In addition, because MCMC still approximates the entire response surface in multidimensional space, application of MCMC with MH within Gibbs to multivariate problems may still require extensive computational time, prohibiting application in practice. Finally, assessing convergence in MCMC applied to sophisticated models with many items or many latent trait dimensions can be cumbersome and requires human judgment. In response to the limitations of the popular MIRT estimation methods described above, a new method was developed.

MH-RM and Prior Research on Its Functionality

Overview. The Metropolis-Hastings Robbins-Monro (MH-RM) algorithm is a data-augmentation technique that combines the MH algorithm described above with the

---

[5] Proposal distributions are typically (multivariate) normal or *t* distributions that resemble the target parameter distributions. The Gibbs sampler is modified by the MH algorithm such that instead of repeatedly sampling from the complete conditional distribution, the proposal distribution for each parameter (i.e., discrimination, difficulty, lower asymptote) is specified at each transition based on the previous state of the Markov chain (see Patz & Junker, 1999).

Robbins-Monro (RM) stochastic approximation algorithm (Cai, 2010a, 2010b). Although the MH-RM algorithm operates differently from MCMC, it still involves Markov chain random imputation via the MH sampler. The random draws are then combined via stochastic approximation to inform how much to adjust the estimates in each iteration via the RM algorithm. As such, MH-RM can be considered an extension of the stochastic approximation EM (SAEM) algorithm (Delyon, Levielle, & Moulines, 1999). The section below describes the logic behind MH-RM and each of its steps in more detail.

How does MH-RM work? The MH-RM algorithm transforms parameter estimation into a missing data problem by use of Fisher's identity. As in the previous section, let $\mathbf{Y}$ represent the observed data, and now let $\mathbf{X}$ represent the missing data (i.e., the unknown latent variables or random effects). Thus, the complete data $\mathbf{Z} = (\mathbf{Y}, \mathbf{X})$. The complete data likelihood for a vector of parameters $\boldsymbol{\theta}$ is then $L(\boldsymbol{\theta} \mid \mathbf{Z})$, and the complete data log-likelihood is $l(\boldsymbol{\theta} \mid \mathbf{Z})$. Recall that the goal of maximum likelihood estimation is to find through an iterative process (e.g., the EM algorithm) the set of parameter estimates for which the observed data are most likely. That is, the goal is to find $\hat{\boldsymbol{\theta}}$ based on the observed data log-likelihood function $l(\boldsymbol{\theta} \mid \mathbf{Y})$. Whereas maximizing $l(\boldsymbol{\theta} \mid \mathbf{Y})$ is computationally intensive due to high-dimensional integrals, maximizing the complete data log-likelihood $l(\boldsymbol{\theta} \mid \mathbf{Z})$, which is based on products of likelihoods, is much simpler. Specifically, for a current set of parameter estimates $\boldsymbol{\theta}^*$ the expectation of the complete data log-likelihood can be expressed as

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^*) = \int_{\mathcal{E}} l(\boldsymbol{\theta} \mid \mathbf{Z}) \Pi(d\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}^*) , \qquad\qquad (2.1)$$

where $\mathcal{E}$ is some sample space to which $\mathbf{X}$ belongs, and $\Pi(d\mathbf{X}\,|\,\mathbf{Y},\boldsymbol{\theta})$ is the conditional distribution of the missing data, given the observed data. This is essentially the E-step of the familiar EM algorithm, the M-step being the maximization of $Q(\boldsymbol{\theta}\,|\,\boldsymbol{\theta}^*)$ by computing new parameter estimates.

If we denote the gradient[6] of the complete data log-likelihood as

$$\mathbf{s}(\boldsymbol{\theta}\,|\,\mathbf{Z}) \;=\nabla_{\boldsymbol{\theta}}l(\boldsymbol{\theta}\,|\,\mathbf{Z}),\tag{2.2}$$

then by Fisher's identity, the conditional expectation of $\mathbf{s}(\boldsymbol{\theta}\,|\,\mathbf{Z})$ over $\Pi(d\mathbf{X}\,|\,\mathbf{Y},\boldsymbol{\theta})$ is equal to the gradient of the observed data log-likelihood:

$$\nabla_{\boldsymbol{\theta}}l(\boldsymbol{\theta}\,|\,\mathbf{Y})=\int_{\mathcal{E}}\mathbf{s}(\boldsymbol{\theta}\,|\,\mathbf{Z})\Pi(d\mathbf{X}\,|\,\mathbf{Y},\boldsymbol{\theta}).\tag{2.3}$$

Thus, augmenting the missing data by taking draws from its posterior predictive distribution $\Pi(d\mathbf{X}\,|\,\mathbf{Y},\boldsymbol{\theta})$ and solving for Equation (2.2) amounts to evaluating the gradient of the observed data log-likelihood without actually analytically evaluating it (Cai, 2010a, 2010b).

Before delving into the specific steps comprising the MH-RM algorithm, it is helpful to review some concepts and notation following Cai (2010a). Specifically let

$$\mathbf{H}(\boldsymbol{\theta}\,|\,\mathbf{Z})=-\frac{\partial^2 l(\boldsymbol{\theta}\,|\,\mathbf{Z})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\tag{2.4}$$

denote the complete data information matrix (i.e., -1 times the second derivative matrix of the complete data log-likelihood). Also, let $\mathcal{K}(\cdot,A\,|\,\mathbf{Y},\boldsymbol{\theta})$ be a Markov transition kernel with $\Pi(\mathbf{X}\,|\,\mathbf{Y},\boldsymbol{\theta})$ as its target, such that for any subset of parameters $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and any subspace $A\in\mathcal{E}$

---

[6] The gradient is a vector based on the first-order partial derivatives. It can be thought of as the multidimensional counterpart of the derivative of a function in one dimension.

$$\int_A \Pi(d\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}) = \int_{\mathcal{E}} \Pi(d\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}) \mathcal{K}(\mathbf{X}, A \mid \mathbf{Y}, \boldsymbol{\theta}).$$ (2.5)

Keeping these expressions in mind, let us review the phases of MH-RM.

The MH-RM item calibration algorithm, as implemented in flexMIRT (Cai,

2013), involves three stages. Stage I procures starting values for the parameters via

unweighted least squares factor extraction. Stage II improves these values via "EM-like"

procedures (Houts & Cai, 2013, p. 86). Finally, assuming some initial values $\boldsymbol{\theta}^{(0)}$ and $\boldsymbol{\Gamma}_0$,

and letting $\boldsymbol{\theta}^{(t)}$ represent the parameter estimates at the tth iteration, Stage III is where

MH-RM actually occurs in the following three steps within each iteration:

1. Stochastic imputation. Draw m_k sets of imputed missing data

$\left\{\mathbf{X}_j^{(t+1)}; j = 1, ..., m_t\right\}$ from the Markov chain $\mathcal{K}(\cdot, A \mid \mathbf{Y}, \boldsymbol{\theta}^{(t)})$ to get m_t sets of complete data

$\left\{\mathbf{Z}_j^{(t+1)} = \mathbf{Y}, \mathbf{X}_j^{(t+1)}; j = 1, ..., m_t\right\}$. The MH sampler can be used for these imputations based

on the relation $\Pi(\mathbf{X} \mid \mathbf{Y}, \boldsymbol{\theta}) \propto L(\mathbf{Z} \mid \boldsymbol{\theta})$ (i.e., the posterior predictive distribution of the

missing data, given the observed data and some estimates of $\boldsymbol{\theta}$, is proportional to the

complete data likelihood).

2. Stochastic approximation. Based on Equation (2.3), approximate the observed

data gradient $\tilde{\mathbf{s}}_{t+1}$ by averaging the complete data gradients

$$\tilde{\mathbf{s}}_{t+1} = \frac{1}{m_t} \sum_{j=1}^{m_t} \mathbf{s}(\boldsymbol{\theta}^{(t)} \mid \mathbf{Z}_j^{(t+1)})$$ (2.6)

and the conditional expectation of the complete data information matrix

$$\boldsymbol{\Gamma}_{t+1} = \boldsymbol{\Gamma}_t + \gamma_t \left\{ \frac{1}{m_t} \sum_{j=1}^{m_t} \mathbf{H}(\boldsymbol{\theta}^{(t)} \mid \mathbf{Z}_j^{(t+1)}) - \boldsymbol{\Gamma}_t \right\}.$$ (2.7)

Equations (2.6) and (2.7) are conceptually similar to obtaining the first and second

derivative in MML estimation.

3. Robbins-Monro update. Set the new set of parameters to

$$\mathbf{\theta}^{(t+1)} = \mathbf{\theta}^{(t)} + \gamma_t (\mathbf{\Gamma}_{t+1}^{-1} \tilde{\mathbf{s}}_{t+1}), \qquad\qquad (2.8)$$

where γ_t $(t \leq 1)$ is a sequence of gain constants that regulates the amount of adjustment

to the parameter estimates in each iteration of the algorithm. Cai (2010a) notes that "…in

practice, [γ_t] may be taken as 1/[*t*], in which case the choice of $\mathbf{\Gamma}_0$ becomes arbitrary." (p.

40). Conceptually, the term $(\mathbf{\Gamma}_{t+1}^{-1} \tilde{\mathbf{s}}_{t+1})$ represents the ratio of the first derivative to the

second derivative. Thus, this step of the MH-RM algorithm can be thought of as a

multidimensional version of the Newton-Raphson procedure. For more details on any of

the steps above see Cai (2010a, 2010b).

Similarities and differences among MCMC, MH-RM, and MML-EM. In

terms of similarities, all three algorithms serve the general goal of producing item

parameter estimates in item factor analysis (IFA). The ways in which they do so,

however, are markedly different. In general, MCMC and MH-RM are different from

MML-EM in that MCMC and MH-RM are *stochastic,* whereas MML-EM is

deterministic in nature. That is, the EM algorithm maximizes the log-likelihood of the

item parameters, given the data, whereas in MCMC and MH-RM there is no direct

maximization taking place. Rather, these are two approximation algorithms operating on

the principle of data augmentation via repeated sampling from a target posterior

distribution. This very feature is what MCMC and MH-RM share in common (via the

MH sampler).

Though somewhat similar, MCMC and MH-RM do operate quite differently,

especially in the way the posterior distribution is approximated and parameter estimates

are obtained. Specifically, MCMC approximates the entire posterior distribution, whereas

MH-RM focuses on point estimates and standard errors (Cai, 2010b). This allows MH-RM to reach convergence[7] much faster than MCMC. More specifically, in MH-RM the "jumps" in the Markov chain serve simply to determine the direction of change throughout iterations, not to provide accurate approximation of a surface that may be off-target. Time-efficiency and other characteristics of the MH-RM algorithm are reviewed next.

Prior research on the functionality of MH-RM. The MH-RM algorithm is fairly recent, and not much research has been done to investigate its performance under different conditions. However, several studies have compared MH-RM to some of the popular estimation techniques described earlier, and the results are promising. Thus, in the following subsections I review prior research on the performance of MH-RM compared to fixed quadrature MML-EM, adaptive quadrature MML-EM, and MCMC.

MH-RM vs. fixed quadrature MML-EM. The Bock and Aitkin (1981) EM algorithm has now been used for decades in applications of IRT. As such, it is a logical choice of algorithms to which MH-RM should be compared. Cai (2010a) did so in an exploratory factor analysis (EFA) framework by examining the parameter recovery (raw bias and sampling variability) of a two-dimensional model with mixed structure. That is, five of the 10 trichotomous items loaded on a single dimension, whereas the other five items loaded on both dimensions. The sample size was 1000, and the study was based on 100 replications. The two algorithms recovered the item parameters equally well with the same root mean square deviation (0.014), though with this simple model the MML-EM algorithm converged faster.

---

[7] Convergence for MH-RM is assumed when the absolute difference of parameter estimates between iterations is $< 10^{-4}$.

Similar results were found in studies in a confirmatory factor analysis (CFA) framework. Cai (2010b) compared fixed quadrature MML-EM to MH-RM with a unidimensional and a three-dimensional correlated-factors IFA model. The unidimensional model was based on responses from 1,000 simulees to 10 items with five ordered categories. Here 49 fixed quadrature points were specified for MML-EM. The results revealed nearly identical item parameter estimates for both algorithms, with root mean squared error (RMSE) for the slopes being slightly larger (.13) for MML-EM than for MH-RM (.10). MML-EM took 0.21 seconds[8] per replication on average, whereas MH-RM took 9 seconds on average. Similar to the EFA study above, fixed quadrature MML-EM appears more time efficient in low-dimensional models.

To demonstrate the time efficiency of MH-RM over fixed quadrature MML-EM in more complex models, Cai (2010b) also compared the two algorithms on data from 500 simulees responding to 18 items mapped to three correlated dimensions. Each dimension was measured by six items (two dichotomous, two trichotomous, and two with five ordered categories). For faster convergence under MML-EM, the number of fixed quadrature points was reduced to 20. Both MML-EM and MH-RM recovered the item parameters and inter-factor correlations equally well (the overall RMSE was .17 for both algorithms). In terms of processing time, MML-EM took 49 seconds per replication on average, whereas MH-RM took 20 seconds per replication on average. Clearly the use of MH-RM becomes more advantageous as the number of dimensions increases. This point is illustrated by several studies described next[9].

---

[8] Unless stated otherwise, estimation processing is measured in CPU time.

[9] More recently, Monroe and Cai (2014) compared the performance of MH-RM and MML-EM applied to models with nonnormal (e.g., skewed, bimodal) latent distributions estimated by Ramsay-curve methods. They found that both algorithms recovered item parameters equally well in terms of average RMSE and

MH-RM vs. adaptive quadrature MML-EM. Another study by Cai (2010a)

compared MH-RM to MML-EM with adaptive quadrature using real data based on a

social quality of life scale for children, which included 24 five-category items. He fit both

a unidimensional and a five-dimensional exploratory IFA model to the data. Although the

unidimensional model did not fit well, results from both models could be used to pit the

two algorithms against each other. To obtain good approximation of the likelihood with

MML-EM, 21 quadrature points were used for the unidimensional model; however, the

number of quadrature points per factor needed to be reduced to 5 for the five-dimensional

model, which amounted to $5^5 = 3125$ quadrature points (and function evaluations) in

total. This number should foreshadow the differences in estimation time between the two

algorithms.

Specifically, for the unidimensional model adaptive quadrature MML-EM took 5

seconds, whereas MH-RM took 10 seconds. However, for the five-dimensional model

MML-EM took 1 hour and 27 minutes, whereas MH-RM took only 95 seconds. This

application with real data highlights the advantage of MH-RM in high-dimensional IFA

models over the "gold standard" estimation method in terms of time-efficiency (Cai,

2010a, p. 34). With respect to parameter estimates (intercepts and target rotated factor

loadings) both algorithms produced nearly identical results with an absolute difference of

.02 between the two methods under both models. In terms of sampling variability, the

algorithms were also comparable. Specifically, the estimated standard errors of the slopes

for the two algorithms were very similar in the unidimensional model (within |.01|

---

estimated bias. For MH-RM, the Monte Carlo standard deviations were somewhat larger than the average
standard errors estimated by the EM algorithm. No computation time differences were reported.

difference)[10], as was the root mean square deviation of rotated loadings in the five-dimensional model (.101 for adaptive quadrature MML-EM vs. .103 for MH-RM). Finally, the log-likelihoods of the two algorithms under the two models differed only in the decimals. All in all, MH-RM produced essentially the same results as the commonly accepted algorithm, but more than 50 times faster.

Cai (2010b) used the same data to compare adaptive quadrature MML-EM and MH-RM in a confirmatory IFA model, which hypothesized a general social quality of life factor, three method factors (positively worded items, negatively worded items, and items about interactions with adults), and four "doublets" (i.e., pairs of items with highly correlated residuals once controlling for the other four factors; Cai, 2010b, p. 326)[11]. For MML-EM, four adaptive quadrature points per dimension were used to approximate the log-likelihood. Both MML-EM and MH-RM produced very similar parameter estimates, standard errors, and log-likelihoods. However, the two algorithms differed widely in processing time. Adaptive quadrature MML-EM took 4.5 hours until convergence, whereas MH-RM took 145 seconds to converge. Again, this result supports the time-efficiency quality of MH-RM in high-dimensional models. Next, I review a comparison of MH-RM to another popular estimation method, MCMC.

MH-RM vs. MCMC. Cai (2010a) also compared the performance of MH-RM with that of Gibbs sampler based MCMC in a generating four-factor model consisting of 19 four-category items. An exploratory IFA model with oblique target rotation was used to evaluate item parameter recovery and inter-factor correlations. The results indicated

---

[10] Interestingly, the standard errors of the intercepts in the unidimensional model were not reported.
[11] For identification purposes, the slopes of the two items within each "doublet" were set equal; all eight factors were standardized (means of 0, variances of 1) and specified as orthogonal (i.e., all factor covariances constrained to 0).

that MH-RM and MCMC estimates were very close to one another as well as to the

generating parameter values, both for item parameters (rotated factor loadings) and inter-

factor correlations. Cai noted that the root mean square deviation from the true values

was larger for MH-RM (0.046) than it was for MCMC (0.039) and explained that this

could be due to the fact that the software running MH-RM (IRTPRO) optimizes a log-

likelihood, whereas the software used for MCMC (MultiNorm) does not. In terms of

computation time, MH-RM took seconds, whereas MCMC took 1 hour 20 minutes and

34 seconds.

MH-RM vs. other methods. Asparouhov and Muthén (2012) examined parameter

recovery (absolute bias and confidence interval coverage) and processing time for the

MH-RM algorithm and four other methods based on both ordered categorical and

dichotomous data. IRTPRO (Cai et al., 2011) was used for MH-RM. The other four

methods were estimated in M*plus* Version 7 (Muthén & Muthén, 1998-2012). These

methods were Monte Carlo with 500 integration points, Monte Carlo with 5000

integration points[12], Bayesian estimation (i.e., MCMC) with weak (noninformative)

priors, and the weighted least squares mean and variance (WLSMV) adjusted estimator.

Both the polytomous and the dichotomous data were based on 35 items; 100 samples of

500 simulees were generated following a seven-dimensional model with five items

mapped to each dimension. Asparouhov and Muthén reported very little to no bias across

the five methods. However, they found that the confidence interval coverage of the

loading estimates was significantly lower for MH-RM (54% for the polytomous data and

---

[12] It is important to note that the Monte Carlo integration method implemented in M*plus* is different from numerical integration based on (adaptive) quadrature in that Monte Carlo integration does not depend on the number of dimensions; thus, it is a viable stochastic approximation alternative to quadrature-based EM in the estimation of high-dimensional IFA models. However, the number of integration points Q does affect numerical error because numerical error is proportional to $1/\sqrt{Q}$ (see Asparouhov & Muthén, 2012).

42% in the dichotomous case) than it was for the other four methods, which maintained coverage rates close to 95%.

In terms of processing time, MH-RM was compared to the other methods in three different scenarios: 1) a real data EFA example with 17 dichotomous items and four orthogonal dimensions, 2) the simulated seven-dimensional EFA model with 35 dichotomous items presented above, and 3) a two-group CFA measurement invariance model with 25 dichotomous items and five orthogonal dimensions. Processing time was measured in seconds. Whereas the results for the full-information estimation methods were inconclusive, the limited-information WLSMV was the fastest and estimated all three scenarios in either one or two seconds[13]. The processing time for MH-RM varied considerably across scenarios relative to the other full-information techniques. Although the results appear divergent from those presented in Cai (2010a, pp. 50-51), Asparouhov and Muthén note that it is difficult to compare processing time across very similar estimation techniques such as the full-information stochastic methods examined here because convergence criteria and other user-defined options may prohibit the generalization of any comparison results to new models or data.

Summary. In summary, the MH-RM algorithm appears to be a promising tool in estimating MIRT models. Not only does MH-RM appear to have overcome the "curse of

---

[13] Despite its speed advantage, WLSMV has several limitations. First, it uses information only from the first- (i.e., the means, which here would be percent correct) and second-order moments (i.e., the standard deviations, which here would be the tetrachoric correlations based on the normally distributed latent continuous variables assumed to underlie the observed categories), whereas full-information methods incorporate entire observed response patterns. Second, WLSMV does not allow for a lower asymptote to accommodate the probability of correct response for low-ability examinees, which can lead to bias in other parameter estimates (see Jurich & DeMars, 2013; Yen, 1981). This limitation also prohibits the application of WLSMV to the 3PL data modeled in this dissertation. Further, the use of WLSMV may not be optimal when the latent distribution is not normal as is assumed by the estimator (see DeMars, 2012). Finally, it would be impossible to estimate the multilevel MIRT models discussed and examined later with a limited-information estimator such as WLSMV.

dimensionality" by making it possible to estimate high-dimensional models, but it also estimates such models remarkably fast. Although prior research on its ability to recover item parameters is not entirely unanimous, the majority of simulation studies are quite favorable. In terms of bias, MH-RM has been found to be just as accurate as the popular algorithms in use (e.g., MML-EM, MCMC). With respect to efficiency, prior research has found MH-RM to provide essentially the same standard errors as MML-EM. Importantly, with MH-RM standard errors are an automatic byproduct and do not need to be estimated in a separate step as they do with MML-EM. However, studies comparing MH-RM to other methods (Monte Carlo integration, M*plus* Bayes estimation) indicate that MH-RM standard errors tend to be noticeably underestimated. Finally, multiple studies have shown the astonishing time efficiency of MH-RM in estimating various models.

Although highly promising, the findings summarized above are based on limited research; much more research is needed to support the use of the MH-RM in practice. Despite its infancy, MH-RM has already been applied in several published studies (Thissen, 2014; Wiley et al., 2014; Wright, 2013; Yang & Cai, 2014), making the call for more research even more urgent. In addition, several extensions of the algorithm have been discussed and are likely to appear in the literature in the upcoming years (Cai, 2010a; 2010b). One of these extensions is the application of the MH-RM algorithm to multilevel models. As the current dissertation focuses on this application area, these models are discussed next.

Analyzing Data with Nested Structure

As briefly discussed in Chapter I, the structure of educational data is typically nested, meaning that individual students are not simple random samples from the population to which one wishes to generalize. Instead, students are nested or clustered within larger units, such as classrooms, schools, districts, etc. Such clustering of data occurs for practical reasons. For example, it is far less expensive to collect data from, say, 100 students in one school than it is to collect data from one student each in 100 different schools. To maximize the randomization process and the approximation of the sample to the population of interest, a two-stage complex sampling design is usually employed, whereby primary sampling units (e.g., schools) are sampled first, and lower-level sampling units (e.g., students within schools) are sampled second. Examples of nested data obtained via multi-stage sampling include large international assessment programs (e.g., the Programme for International Student Assessment [PISA], Trends in International Mathematics and Science Study [TIMSS], Progress in International Reading Literacy Study [PIRLS]), among others.

Nesting or clustering of examinees is not limited to assessment programs that involve multi-stage sampling. Natural nesting occurs for any testing program that includes virtually all lower-level units within higher-level units (i.e., census data). For example, assessment in K-12 for accountability purposes often requires that all students at certain grade levels in all schools within a state be tested. In such cases, the examinees are not sampled (i.e., the entire student population in a given grade completes the assessments), yet the examinees are nested within their respective schools and school districts.

Regardless of how the data are obtained, analyzing nested data poses some

challenges due to the shared variability among Level 1 units (e.g., students) nested within

Level 2 units (e.g., schools). This shared variability occurs because Level 1 units within

the same cluster (i.e., Level 2 unit) are typically much more similar to one another than

they are to Level 1 units in other clusters. For example, one would expect students

attending the same school to share many more background and achievement

characteristics with one another than they would with students attending other schools.

These similarities could be due to geographic, demographic, socioeconomic, curricular,

co-curricular or other factors. What is important is that Level 1 units do share some

variability, and unless modeled, this shared variability could result in confounded

parameter estimates (e.g., variance components) and underestimated standard errors. The

amount of shared variability among Level 1 units due to clustering can be summarized in

the intraclass correlation coefficient.

The Intraclass Correlation Coefficient

Overview. The intraclass correlation coefficient (ICC) is conceptually defined as

the ratio of the between-cluster variance to the total variance (between + within):

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} . \tag{2.9}$$

For a specific sample, the ICC can be computed using the mean square components of

ANOVA:

$$\hat{\rho} = \frac{MS_B - MS_W}{MS_B + (n.-1)MS_W} , \tag{2.10}$$

where *n*. is the sample size per cluster, if balanced (e.g., the same number of students in each school). When the cluster size is unbalanced, the average cluster size *n*. can be computed by

$$n. = \frac{N^2 - \sum_{k=1}^{K} n_k^2}{N(K-1)},$$ (2.11)

where N is the total sample size, K is the number of clusters, and n_k is the number of Level 1 units within cluster k (Stapleton, 2013).

Conceptually, the ICC indicates the degree of dependence among Level 1 units due to clustering. It measures the "extent to which members of the same [cluster] are more similar to one another than members of other [clusters]" (Cohen, Cohen, West, & Aiken, 2003, p. 537). The above presentation of the ICC is focused on a single dependent variable. However, the concept can be easily extended to the multivariate case. In fact, as shown in Chapter III in greater detail, in the 3PL ML-MIRT model an ICC can be specified for each latent variable based on the variances from both levels because the observed item responses are conditional on the latent variables at different levels.

Typical ICC values. Theoretically, ICC values range from 0 (complete independence) to 1 (complete dependence). Prior research has suggested that in geographically determined clusters (e.g., states, districts), ICCs tend to be relatively low for demographic variables (e.g., age, gender), somewhat higher for socioeconomic variables and attitudes, and maybe even higher for educational data involving classrooms (Stapleton, 2013). Specifically, typical ICC values for health-related variables (e.g., drinking) or attitudinal measures (e.g., career interests) are in the range of .02-.07, whereas ICC values for mathematics achievement scores from eighth-grade students have

been reported in the range .30-.40 when the Level 2 units are classrooms and .15-.20

when the Level 2 units are schools (Muthén, 1997). Hedges and Hedberg (2007)

examined existing datasets from schools across the U.S. and found that the average

unconditional ICC for mathematics and reading in K-12 was .22. Based on their findings

and prior research, Hedges and Hedberg recommended the use of ICCs in the range of

.15-.25 for cluster-randomized experiments involving diverse or low-socioeconomic-

status schools and ICC values of .05-.15 when the clusters are low-achieving schools.

Implications. Although there is substantial variability in ICCs across different

populations, it should be noted that even ICC values as small as .01 or .05 can have

serious implications for standard errors and tests of significance unless taken into

account. As mentioned earlier, performing a single-level analysis on nested data results in

underestimation of the sampling variability that would have been observed had the data

been obtained via simple random sampling. The literature suggests that although fixed

effects (e.g., regression coefficients) tend to be unbiased, the standard errors of many

parameters can be severely underestimated (Stapleton & Thomas, 2008). For example,

the effect of an ICC ρ on the standard error of the mean for a given cluster size $n.$ is

known as the *design effect* and can be computed by $[1+(n.-1)\rho]$ (Kish, 1965 as cited in

Stapleton, 2013). As a result, the standard error is underestimated by a factor equal to the

square root of the design effect. Based on this formula, when the ICC is 0, the design

effect is 1, and the standard error remains unbiased. However, when the ICC is greater

than 0, the standard error is underestimated, and more so as the cluster size increases,

resulting in overly inflated alpha levels (see Kreft & de Leeuw, 1998, p. 10). With an

inflated alpha rate, any significance tests may be biased, leading the researcher to make incorrect inferences about the model and the parameter estimates.

Requirements. Although ICCs as low as .01 can lead to substantial increase in Type I error rates, sufficiently large ICCs are needed to estimate a multilevel model. This is because when there is not enough variability between clusters, there is not enough information to estimate another set of parameters. Thus, when the ICC is near 0, it may not be possible to estimate a multilevel model (Stapleton, 2013), and doing so may not be necessary.

Another important consideration in multilevel modeling is the number of observations at different levels of the analysis. Simply because a model can be estimated does not guarantee that its parameter estimates can be trusted. Several simulation studies have investigated the question of how many Level 1 units (e.g., students) and Level 2 units (e.g., schools) are necessary to obtain stable and unbiased parameter estimates in multilevel linear models (e.g., Maas & Hox, 2005). The consensus is that a large number of clusters (e.g., ≥ 30) is much more desirable than a large number of observations within clusters (Maas & Hox, 2005; Snijders, 2005; Spybrook, 2008). Lüdtke, Marsh, Robitzsch, and Trautwein (2011) performed an extensive study investigating the effects of the number of Level 1 and Level 2 units, ICCs, and other factors in linear multilevel models for contextual effects that correct for measurement and/or sampling error in the predictor. In line with the considerations discussed here, they found that the combination of small number of clusters and low ICC resulted in unstable estimates.

Although there are no known guidelines as to the desirable number of clusters and cluster size in *nonlinear* multilevel models, several simulation studies suggest that small

numbers of clusters and small cluster sizes can be even more problematic than in linear multilevel models. Such problems have been noted and largely attributed to the imperfect estimation methods available for multilevel models with binary outcomes (Goldstein & Rasbash, 1996; Rodríguez & Goldman, 1995). A decade later, despite improvements in estimation techniques, simulation research suggests that fixed parameter estimates and their standard errors may still be biased when the cluster size is small (e.g., 10), even with a large number of clusters (Austin, 2010; Clarke, 2008; Moineddin, Matheson, & Glazier, 2007). The bias seems to disappear with 30 or more clusters of at least 30 each.

Similarly, unlike linear multilevel models, in which a large number of clusters can typically compensate for small cluster size (e.g., 5 or 10), the standard errors in nonlinear multilevel models with small cluster size tend to be substantially biased. Importantly, there has not been any research on multilevel models with multiple binary outcomes such as the multilevel measurement models examined in this dissertation. As such, this is another area to which the study is meant to contribute (e.g., how does a cluster size of 20 vs. 100 affect fixed parameter estimates such as item difficulty and discrimination and random effects such as the latent variances and covariances at different levels?). To help the reader understand the specific type of nonlinear multilevel models considered here, the next section situates this type of model in the greater family of multilevel measurement models by tracing its development and comparing it to similar models.

Multilevel Measurement Models

The notion of specifying a measurement model at multiple levels when the data have nested structure is not new. Muthén (1991) credited Cronbach for laying out the theoretical foundation of such models back in the mid-1970s and noted that their

application has been largely inhibited by the limited power of computers and software

packages. Since then, the technological aspect of educational measurement has grown

tremendously, and numerous formulations of multilevel measurement and structural

equation models have been suggested and demonstrated (e.g., Muthén & Satorra, 1995;

Mehta & Neale, 2005; Pastor, 2003; Rabe-Hesketh, Skrondal, & Pickles, 2004). In the

following section, I present a simple unidimensional CFA model following Muthén

(1991)[14]. Then I note how the model has been conceptualized as a three-level model in

the multilevel IRT literature, and how more recently Höhler and colleagues (2010)

combined the multilevel IRT and MIRT frameworks into a single ML-MIRT model.

Finally, I extend this ML-MIRT model to include a pseudo-guessing parameter and show

the model both mathematically and graphically.

Multilevel CFA. The premise of multilevel CFA lies in the decomposition of the

total variance for each observed variable into between-cluster variance and within-cluster

variance. Specifically, the observed score on item i for examinee j nested within cluster k

can be expressed as

$$y_{ijk} = \bar{y}_i + y_{ik}^{\text{B}} + y_{ijk}^{\text{W}}, \qquad\qquad (2.12)$$

where $\bar{y}_i$ is the grand mean on item i, y_{ik}^{B} is cluster k's deviation from the grand mean on

item i (which contributes to between-cluster variance), and y_{ijk}^{W} is examinee j's deviation

from cluster k's mean on item i (which contributes to the within-cluster variance).

Assuming clusters have the same number of Level 1 units, the only contribution they

have toward the total variability is via the cluster means, which can be conceptualized as

---

[14] This specific paper was chosen to illustrate multilevel CFA for its simplicity. In the analysis, Muthén combined multiple dichotomous items into "subscores" which then served as indicators (Muthén, 1991, p. 341), a procedure termed *item parceling*. Methodologists have clearly discouraged the practice of item parceling (see Bandalos & Finney, 2001).

deviations from the grand mean. Similarly, the only contribution of individual scores

toward the total variability is their deviation from the cluster means. Thus, the variance of

y is a function of between- and within-cluster variability, which are *independent* of each

other and thus additive:

$$\sigma_y^2 = \sigma_B^2 + \sigma_W^2. \tag{2.13}$$

The same principle applies to the multivariate case. There instead of decomposing the

variance of a single variable, one decomposes the entire variance-covariance matrix $\mathbf{\Sigma}_y$

into a between-cluster covariance matrix $\mathbf{\Sigma}_y^B$ and a within-cluster covariance matrix $\mathbf{\Sigma}_y^W$:

$$\mathbf{\Sigma}_y = \mathbf{\Sigma}_y^B + \mathbf{\Sigma}_y^W. \tag{2.14}$$

Traditionally, each of these matrices was estimated separately (e.g., Muthén, 1994) using

a limited information ML estimator known as the Muthén multilevel ML estimator

(MUML; Muthén & Satorra, 2005). There are some issues with this approach (see

Zyphur, Kaplan, & Christian, 2008), and several stepwise approaches to model fitting

have been proposed instead (e.g., Hox, 2010; Stapleton, 2013), using full information ML

estimation.

The decomposition of the observed variance-covariance matrix not only aids the

understanding of the multilevel model, but it also serves an important role in the

statistical identification of the model. Specifically, the different levels of analysis in

multilevel CFA are modeled explicitly as different latent factor structures. Theoretically,

each of these latent factor structures is allowed to have its own set of measurement and

structural parameters. The estimation of unique parameters and factor structures across

levels is possible due to the decomposition of the variance-covariance matrix. The

estimation of different measurement (item) parameters across levels is much more meaningful in the organizational literature, where Level 2 constructs can have a completely different meaning from Level 1 constructs (see Bliese & Jex, 2002). Given the current study focuses on educational data and applications, item parameters were assumed to be the same across levels because the latent dimensions bear the same interpretation.

Multilevel IRT. Following Adams and colleagues' (1997) conceptualization of IRT within a multilevel framework, several different multilevel IRT models have been proposed (see Kamata & Vaughn, 2011 for an overview). These include Fox and Glas' (2001) multilevel IRT model, Kamata's (2001) hierarchical generalized linear model, and Muthén and Asparouhov's (2011) multilevel CFA model with categorical indicators. Although these models may differ in estimation methods, link functions (e.g., normal ogive vs. logistic function), and scaling of the parameters, they still share many similarities. For example, the measurement part of the model is typically set up as a two-level model, where observed item responses are nested within persons. Then, to examine variation across clusters and the effects of person and cluster-level predictors, the model is usually extended to a three-level model (e.g., items nested within students nested within schools). Overall, it appears that the focus of multilevel IRT developments has been on specifying a single latent dimension measured by a set of items and modeling its variance as a function of predictors at different levels.

It is important to distinguish between two different types of multilevel item response models: *measurement* models, which focus on the measurement of individual examinees, and *explanatory* models, which are not concerned with the measurement of

individual examinees and instead focus on the explanation of item responses in terms of examinee- and item-level predictors (see De Boeck & Wilson, 2004, Chapter 1 for an in-depth treatment of this topic). The multilevel IRT models described above fall within the explanatory type, whereas the models examined in this dissertation fall within the measurement type. That is, here one is interested in the descriptive measurement of ability at the student and the school level. As such, the model is descriptive in nature (i.e., not explanatory); the clustering of students within schools is simply seen as a nuance of the data which the model can accommodate.

Höhler and colleagues (2010) took the latter approach by focusing on the estimation of a MIRT model in a multilevel framework and interpreting the correlations among latent traits at different levels compared to a single-level MIRT model in which the nested nature of the data was ignored. This conceptualization of a ML-MIRT model is much more similar to the multilevel CFA model discussed earlier than to the typical multilevel IRT formulations mentioned above. Specifically, Höhler and colleagues (2010) applied a two-level, three-dimensional model to the language test scores of 9th-grade students nested within classrooms. They indicated that all analyses were carried out in M*plus* Version 5.1, using ML estimator with robust standard errors for the MIRT models and Monte Carlo integration with 1000 points per dimension for the ML-MIRT models. It appears that Höhler and colleagues specified a 1PL and/or a 2PL model; however, this was not explicitly stated. Importantly, failure to model the pseudo-guessing parameter to accommodate the probability of correct guessing for examinees of low proficiency can lead to underestimation of the loadings for difficult items (Jurich & DeMars, 2013; Yen, 1981). The model considered in this dissertation builds on Höhler

and colleagues' (2010) model to include a pseudo-guessing parameter and more than three dimensions. This is made possible by using the MH-RM algorithm for estimation. The model is described in more detail next.

The 3PL ML-MIRT model. Similar to the decomposition of an observed score presented in Equation (2.12), the latent ability level of examinee j from cluster k on dimension g can be decomposed as

$$\theta_{gjk} = \bar{\theta}_g + \theta_{gk}^{\mathrm{B}} + \theta_{gjk}^{\mathrm{W}}, \qquad (2.15)$$

where $\bar{\theta}_g$ is the grand mean on dimension g (which is typically constrained to 0 for identification purposes), $\theta_{gk}^{\mathrm{B}} = \bar{\theta}_{gk} - \bar{\theta}_g$ is the deviation of cluster k's mean from the grand mean, and $\theta_{gjk}^{\mathrm{W}} = \theta_{gjk} - \bar{\theta}_{gk}$ is examinee j's deviation from cluster k's mean on dimension g. Then the 3PL ML-MIRT model as an extension of Equation (1.2) becomes

$$P(U_{ijk} = 1 \mid \boldsymbol{\theta}_{\mathbf{k}}^{\mathbf{B}}, \boldsymbol{\theta}_{\mathbf{jk}}^{\mathbf{W}} \mathbf{a}_{\mathbf{i}}^{\mathbf{B}}, \mathbf{a}_{\mathbf{i}}^{\mathbf{W}} c_i, d_i) = c_i + (1 - c_i) \frac{e^{\mathbf{a}_{\mathbf{i}}^{\mathbf{B}} \boldsymbol{\theta}_{\mathbf{k}}^{\mathbf{B'}} + \mathbf{a}_{\mathbf{i}}^{\mathbf{W}} \boldsymbol{\theta}_{\mathbf{jk}}^{\mathbf{W'}} + d_i}}{1 + e^{\mathbf{a}_{\mathbf{i}}^{\mathbf{B}} \boldsymbol{\theta}_{\mathbf{k}}^{\mathbf{B'}} + \mathbf{a}_{\mathbf{i}}^{\mathbf{W}} \boldsymbol{\theta}_{\mathbf{jk}}^{\mathbf{W'}} + d_i}}. \qquad (2.16)$$

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

$$P(U_{ijk} = 1 \mid \theta_k^{\mathrm{B}}, \theta_{jk}^{\mathrm{W}}, a_i, c_i, d_i) = c_i + (1 - c_i) \frac{e^{a_i(\theta_k^{\mathrm{B'}} + \theta_{jk}^{\mathrm{W'}}) + d_i}}{1 + e^{a_i(\theta_k^{\mathrm{B'}} + \theta_{jk}^{\mathrm{W'}}) + d_i}}. \qquad (2.17)$$

It is important to reiterate that 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 $\hat{\theta}_k^{\mathrm{B}}$ associated

with Level 2 and a within-cluster ability estimate $\hat{\theta}_{jk}^{\mathrm{w}}$ associated with Level 1.

Conceptually, these represent the school-level and the student-level ability estimates. That is, the model not only estimates each student's ability estimate (as in single-level models) but also school-average ability estimates, which are direct estimates of the model (in the sense that one "borrows" information from the other schools assumed to come from the same population of schools in a two-level analysis) and take on the same value for each student within a given school. A graphical depiction of the 3PL ML-MIRT model is presented in Figure 1.

The 3PL ML-MIRT model is fairly sophisticated and may be even impossible to estimate via the popular estimation techniques and algorithms discussed in this chapter. Fortunately, the MH-RM algorithm was designed to overcome estimation challenges posed by complex models, and its capabilities were put to the test in the current study. As discussed throughout this chapter, MH-RM is very flexible and has shown promising results in handling both MIRT and multilevel models. Importantly, only Wiley and colleagues (2014) have applied the MH-RM algorithm to 3PL MIRT data; however they performed a single-level analysis. Thus, no one has examined the performance of MH-RM with the 3PL ML-MIRT model, hence the need for the current study.
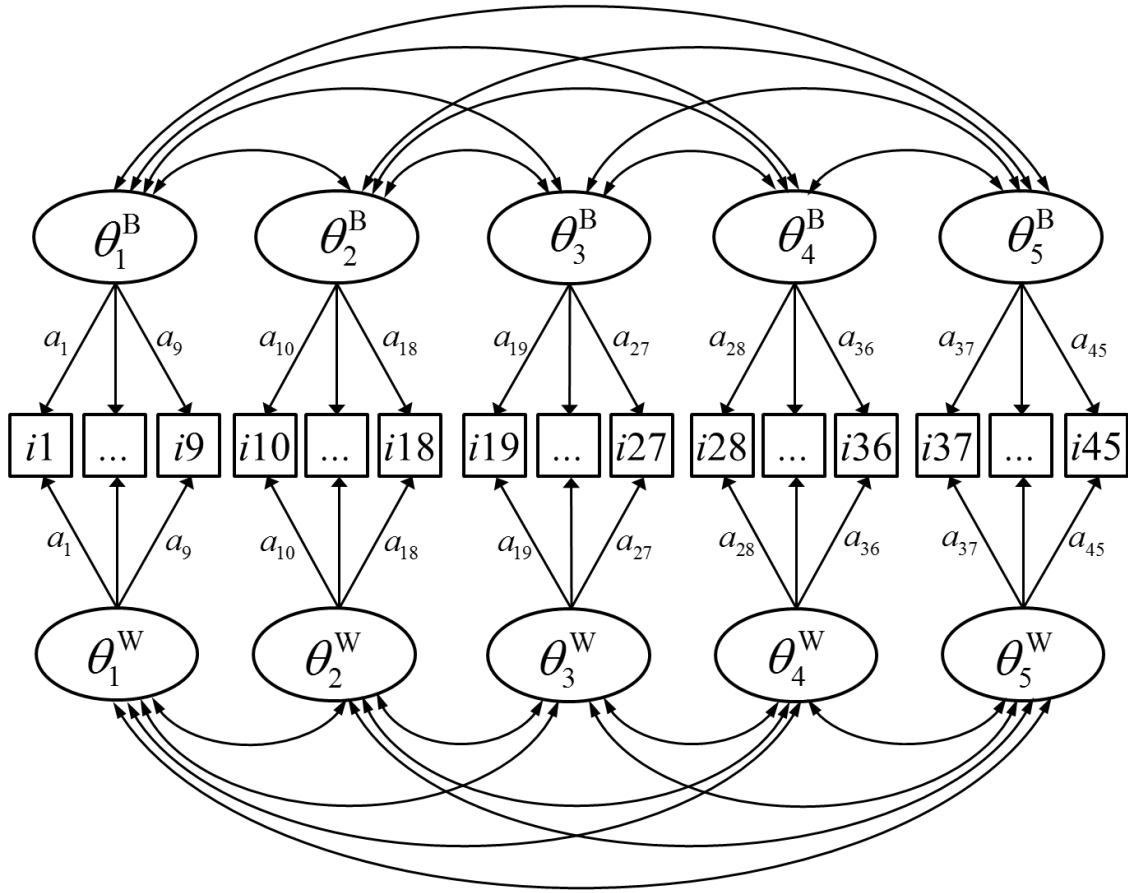
Figure 1. A graphical representation of the 3PL ML-MIRT model with five dimensions and 45 dichotomous items.

The top half shows the measurement model between clusters, whereas the bottom half shows the model within clusters. Within-cluster variances are fixed to 1.0 to identify the model; between-cluster variances are freely estimated. All covariances within the same level (shown as double-headed arrows for simplicity) are freely estimated as well. No level-specific superscript is used for the discrimination parameters to highlight the fact that they are the same for each item across levels (i.e., one loading is estimated per item and is fixed to be the same across levels). Squares represent dichotomous item responses, and ellipses (…) indicate items not shown in the graph for simplicity.

Chapter III

Method

Given no prior research on the performance of the MH-RM algorithm applied to 3PL ML-MIRT models, the focus of this study is 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 if and how well MH-RM can estimate 3PL MIRT models is 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 ICC level). 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. These conditions are presented next.

Conditions

The current study varied four different factors: one pertaining to multidimensional models and three factors pertaining to multilevel models. Specifically, the design 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 results in 2 dimension levels × 3 ICC levels × 2 numbers of clusters × 2 cluster sizes = 24 conditions total (see Table 1).

Table 1
Breakdown of the 24 Simulation Conditions

Condition #	Dimensions	ICC	K	n.	N
1	3	.15	40	20	800
2	3	.15	40	100	4,000
3	3	.15	200	20	4,000
4	3	.15	200	100	20,000
5	3	.25	40	20	800
6	3	.25	40	100	4,000
7	3	.25	200	20	4,000
8	3	.25	200	100	20,000
9	3	.35	40	20	800
10	3	.35	40	100	4,000
11	3	.35	200	20	4,000
12	3	.35	200	100	20,000
13	5	.15	40	20	800
14	5	.15	40	100	4,000
15	5	.15	200	20	4,000
16	5	.15	200	100	20,000
17	5	.25	40	20	800
18	5	.25	40	100	4,000
19	5	.25	200	20	4,000
20	5	.25	200	100	20,000
21	5	.35	40	20	800
22	5	.35	40	100	4,000
23	5	.35	200	20	4,000
24	5	.35	200	100	20,000

Note. K = number of clusters; $n.$ = cluster size; N = total sample size.

The three versus five dimensions were chosen for two reasons. First, to truly examine the performance of MH-RM in estimating MIRT models, three or more dimensions would be desired. As discussed in Chapter II, other methods (e.g., adaptive quadrature MML-EM) are equally or more time efficient than MH-RM when estimating one- or two-dimensional models. Thus, the benefits of MH-RM become more evident in the estimation of models with more dimensions. On the other hand, given the models considered here are also multilevel models, examining models with more than five dimensions may require too much time for the timely completion of the study. Second, and related to the first reason, examining models with three to five dimensions is what one might encounter in practice. For example, one can conceive of the three-dimensional

model as representing three subtests (e.g., English Language Arts, math, and science). Similarly, one can apply the five-dimensional model when one wishes to calibrate data collected on the different domains of a subject area. An example is the five different domains of mathematics as defined by the Common Core State Standards for grades 3-5: Operations and Algebraic Thinking, Number and Operations in Base Ten, Number and Operations—Fractions, Measurement and Data, and Geometry (National Governors Association, 2010).

With respect to the number of clusters and cluster size, the values chosen for the study could represent different combinations of sample or population compositions encountered by assessment practitioners. For example, the larger number of clusters (200) could represent schools, whereas the smaller number of clusters (40) could represent classrooms within the same school or school district. Similarly, the larger cluster size (100) could represent students nested within the same school, whereas the smaller cluster size (20) could represent students nested within a smaller Level 2 unit (e.g., a classroom). Multiplying the number of clusters by cluster size results in three possible overall sample sizes ranging from 800 to 20,000. This range should cover a good number of the typical sample sizes found in large international assessment and state K-12 testing programs. It should be noted that the design considered here is balanced, meaning that for a given condition all clusters consist of the same number of Level 1 units. Maas and Hox (2005) reported that having a balanced versus unbalanced design had little to no effect on parameter estimates and standard errors. In addition, for simplicity the design does not incorporate sampling weights as might be done in practice. Finally, the ICC values, which are specified at the latent level and for simplicity were assumed to be the

same for all dimensions at both levels, were chosen based on prior research to accommodate typical classroom- as well as school-level ICCs (Hedges & Hedberg, 2007; Muthén, 1997). More detail on the specification of ICC values is provided in the next section.

Data Generation

The data for all conditions were generated via the "Simulation" mode in flexMIRT (Houts & Cai, 2013). Specifically, batch-mode input files were generated in R (R Core Team, 2013) to generate and calibrate the data in flexMIRT. The generating model is based on user-supplied item parameters and a latent variance-covariance matrix. One of the advantages of flexMIRT is that it can easily generate multilevel data with a user-specified cluster size. The ICC values can be specified via the generating latent between- and within-cluster variance components. Each of these features of the generating models is described in detail below.

Item parameters. Given the number of dimensions (three or five), the hypothetical test length was set at 45 items, which allows 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. The odd number of items (45) and the number of items per dimension (15 or 9) allow distributing the three discrimination parameter values equally across the items under the two different models, which was useful in summarizing the results.

The item difficulties were set at nine different values (0, ±0.380, ±0.787, ±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

$$d_i = -b_i a_i . \tag{3.1}$$

It is important to note that the nine difficulty values were spread over the three and five dimensions strategically, so that each dimension had about the same number of easy, medium, and difficult items; however, not all difficulty levels were fully crossed with the three item discrimination values.

The pseudo-guessing parameters for all items were fixed to .20, which is typical for multiple-choice items with five response options. In flexMIRT this is done by specifying the logit of the lower asymptote to be equal to -1.4 (see Houts & Cai, 2013).

Latent variance-covariance matrix and ICC values. Since the within-cluster variances were set to 1.0 for model identification purposes, the within-cluster covariances are on the correlation metric. However, this assumes that the variability within cluster is the same for all Level 2 units. For example, this implies that the within-school variability is the same across schools, which is a serious assumption that may or may not be true in reality. 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. In the three-dimensional model, there are only three correlations (Level 1) or covariances (Level 2). Thus, these three values were used (see Table 2). In the five-dimensional model, each of these values was repeated three times, with the exception of

.70 which was repeated four times (for a total of $5*(5-1)/2 = 10$ correlations per level; see Table 3).

Importantly, to set the ICC at a specific value, the generating between-cluster variances were specified such that the desired ICC was obtained via Equation (2.9). For example, to obtain an ICC of .25, one would plug this value and the within-cluster variance (1.0) into Equation (2.9) and solve for the between-cluster variance:

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2}$$

$$\frac{\sigma_B^2}{\sigma_B^2 + 1} = .25$$

$$\sigma_B^2 = .25*(\sigma_B^2 + 1)$$

$$\sigma_B^2 - .25\sigma_B^2 = .25$$

$$.75\sigma_B^2 = .25$$

$$\sigma_B^2 \approx .333.$$

It should be noted that because the ICCs are based on the ratio of the latent between- and within-cluster variance components, these ICCs were somewhat higher than the ICCs based on observed scores (see Höhler et al, 2010). Thus, although an ICC of .35 may appear large, its observed counterpart would be lower and thus closer to the typical ICC values found in educational data (e.g., Hedges & Hedberg, 2007).

Latent means and distributions. The abilities for each dimension in both levels were generated to be multivariate normal, with a mean of 0 and a standard deviation or variance of 1.0 at Level 1 and variance as described above at Level 2.

Table 2
Generating Variances and Covariances for the Three-Dimensional Models

ICC = .15	θ_1^B	θ_2^B	θ_3^B	θ_1^W	θ_2^W	θ_3^W
θ_1^B	0.176					
θ_2^B	0.123	0.176				
θ_3^B	0.141	0.158	0.176			
θ_1^W	0	0	0	1		
θ_2^W	0	0	0	0.7	1	
θ_3^W	0	0	0	0.8	0.9	1
ICC = .25	θ_1^B	θ_2^B	θ_3^B	θ_1^W	θ_2^W	θ_3^W
θ_1^B	0.333					
θ_2^B	0.233	0.333				
θ_3^B	0.266	0.300	0.333			
θ_1^W	0	0	0	1		
θ_2^W	0	0	0	0.7	1	
θ_3^W	0	0	0	0.8	0.9	1
ICC = .35	θ_1^B	θ_2^B	θ_3^B	θ_1^W	θ_2^W	θ_3^W
θ_1^B	0.538					
θ_2^B	0.377	0.538				
θ_3^B	0.430	0.484	0.538			
θ_1^W	0	0	0	1		
θ_2^W	0	0	0	0.7	1	
θ_3^W	0	0	0	0.8	0.9	1

Note. Variances are on the main diagonal. Covariances are on the lower off-diagonal. Numbers in the subscripts differentiate the latent dimensions within each level. Letters in the superscripts indicate the level (B = between or Level 2; W = within or Level 1). By definition, covariances across levels are fixed at zero. When converted to correlations, the covariances at Level 2 match the correlations at Level 1 (.70, .80, and .90).

Table 3

Generating Variances and Covariances for the Five-Dimensional Models

ICC = .15	θ_1^B	θ_2^B	θ_3^B	θ_4^B	θ_5^B	θ_1^W	θ_2^W	θ_3^W	θ_4^W	θ_5^W
θ_1^B	0.176									
θ_2^B	0.123	0.176								
θ_3^B	0.141	0.158	0.176							
θ_4^B	0.123	0.141	0.158	0.176						
θ_5^B	0.123	0.141	0.158	0.123	0.176					
θ_1^W	0	0	0	0	0	1				
θ_2^W	0	0	0	0	0	0.7	1			
θ_3^W	0	0	0	0	0	0.8	0.9	1		
θ_4^W	0	0	0	0	0	0.7	0.8	0.9	1	
θ_5^W	0	0	0	0	0	0.7	0.8	0.9	0.7	1
ICC = .25	θ_1^B	θ_2^B	θ_3^B	θ_4^B	θ_5^B	θ_1^W	θ_2^W	θ_3^W	θ_4^W	θ_5^W
θ_1^B	0.333									
θ_2^B	0.233	0.333								
θ_3^B	0.266	0.300	0.333							
θ_4^B	0.233	0.266	0.300	0.333						
θ_5^B	0.233	0.266	0.300	0.233	0.333					
θ_1^W	0	0	0	0	0	1				
θ_2^W	0	0	0	0	0	0.7	1			
θ_3^W	0	0	0	0	0	0.8	0.9	1		
θ_4^W	0	0	0	0	0	0.7	0.8	0.9	1	
θ_5^W	0	0	0	0	0	0.7	0.8	0.9	0.7	1
ICC = .35	θ_1^B	θ_2^B	θ_3^B	θ_4^B	θ_5^B	θ_1^W	θ_2^W	θ_3^W	θ_4^W	θ_5^W
θ_1^B	0.538									
θ_2^B	0.377	0.538								
θ_3^B	0.430	0.484	0.538							
θ_4^B	0.377	0.430	0.484	0.538						
θ_5^B	0.377	0.430	0.484	0.377	0.538					
θ_1^W	0	0	0	0	0	1				
θ_2^W	0	0	0	0	0	0.7	1			
θ_3^W	0	0	0	0	0	0.8	0.9	1		
θ_4^W	0	0	0	0	0	0.7	0.8	0.9	1	
θ_5^W	0	0	0	0	0	0.7	0.8	0.9	0.7	1

Note. Variances are on the main diagonal. Covariances are on the lower off-diagonal. Numbers in the subscripts differentiate the latent dimensions within each level. Letters in the superscripts indicate the level (B = between or Level 2; W = within or Level 1). By definition, covariances across levels are fixed at zero. When converted to correlations, the covariances at Level 2 match the correlations at Level 1 (.70, .80, and .90).

Dependent Variables of Interest

Parameter recovery under the MH-RM algorithm was examined over 100 replications.[15] Specifically, the accuracy and efficiency of parameters (item discrimination, item difficulty, between-cluster variances and covariances, and within-cluster covariances) 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 standard errors described in Chapter II, standard error accuracy is of particular interest. In addition, the processing time was reported (in real time) for each condition across replications. 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

$$Bias_\beta = \frac{\sum_{r=1}^{R} (\hat{\beta}_r - \beta)}{R},\qquad(3.2)$$

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

---

[15] Having more replications (e.g., 500 or 1000) would be desirable, especially for the standard errors. However, one preliminary run of data generation and model calibration across all conditions took about 27 hours on a computer with dual-core i7-4500U CPU processor at 16GB with up to 2.40GHz RAM. Thus, for the timely completion of this dissertation a second computer with quad-core i5-2400U CPU processor at 4GB with up to 3.10GHz RAM was used to run some of the replications.

RMSE. The root mean squared error (RMSE) combines both bias and sampling

variability of parameter estimates across replications. Specifically, RMSE was computed

as

$$RMSE_\beta = \sqrt{\frac{\sum_{r=1}^{R}(\hat{\beta}_r - \beta)^2}{R}} = \sqrt{Bias_\beta^2 + SE_\beta^2} \; , \tag{3.3}$$

where SE_β is the empirical standard error of the parameter (i.e., the standard deviation of

the parameter estimates across replications), and all elements are as defined above.

Standard error accuracy. The accuracy of standard errors was examined in

terms of confidence interval coverage probability, which is the proportion of replications

in which the 95% confidence interval contains the generating (true) parameter.

Specifically, based on the analytical standard error from each replication, a 95%

confidence interval around the parameter estimate from that replication was constructed.

Then an indicator variable was created such that it took on a value of 1 when the

confidence interval contained the true parameter and 0 otherwise. Averaging the values

of this variable across all replications returned the confidence interval coverage for the

parameter in question. Confidence interval coverage rates near 95% are desirable because

they would indicate the analytical standard errors and Type I error rates are accurate. It is

important to note, however, that if the parameter estimates were biased, the 95%

confidence interval would not cover 95% of the estimates, even if the standard errors

were accurate.

Chapter IV

Results

All data management work and statistical analyses were performed in SAS

software, version 9.4. All figures displayed in the results were created in the R

programming environment (R Core Team, 2013). To determine which combinations of

condition factors had the greatest impact on the dependent variables of interest, I

estimated the proportion of variance accounted for by each factor through a series of

regression models using the *proc glm* procedure in SAS software, which allows the

inclusion of both categorical and continuous predictors. More specifically, the condition

factors (e.g., cluster size), generating value of the parameter where applicable (e.g.,

generating item discrimination a), as well as the two-way and three-way interactions

among these factors served as predictors of the dependent variable (e.g., item difficulty

bias). Four-way (or higher-order) interactions were not examined because they can result

in estimation difficulties and can be nearly impossible to display and interpret. See

Appendix A for more detail on the procedures used to examine the regression models as

well as the output from the full models containing all main effects, two-way interactions,

and three-way interactions.

Bias

Overall, item parameters (which were fixed to be equivalent across levels), latent

variances and covariances, and the abilities were reproduced well. In the following, I

break down the results regarding bias by item parameters (item difficulty and item

discrimination), latent variances (only for Level 2 since the Level 1 variances were fixed

at 1.0 for identification), latent covariances/correlations at both levels, as well as ability

estimates at both levels. The 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. As a reminder, bias was defined as the average difference between the estimated value and the generating value of a parameter. Thus, positive bias indicates the parameter was overestimated, whereas negative bias indicates the parameter was underestimated.

Item difficulty. The item difficulty values (which as explained in Appendix A are somewhat confounded by item discrimination) were slightly positively biased on average (mean bias across conditions was 0.128, $SD = 0.059$), indicating that, on average, the items were estimated to be easier than they actually were (see Equation 1.3 for the relationship between the item "easiness" parameter [d] considered here and the traditional item difficulty parameter [b]). The full regression model with all main effects, two-way interactions, and three-way interactions explained the majority of the variability in this bias (see Table A1 in Appendix A), with generating item discrimination value (labeled "aval") being the most significant predictor of this variability, followed by the number of dimensions ("dim"), generating item difficulty value ("dval"), number of clusters, cluster size, the ICC level, and some interactions, each explaining at least 1% of the variability in item difficulty bias. I interpret these effects next with the aid of visual displays.

As shown in Figure 2, there was a positive relationship between generating *a* value and bias in item difficulty. Figure 2 also shows the main effect of the number of dimensions, with bias being consistently higher in the three-dimensional models than in

the five-dimensional models. Finally, one should note the effect of sample size. Specifically, 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 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.
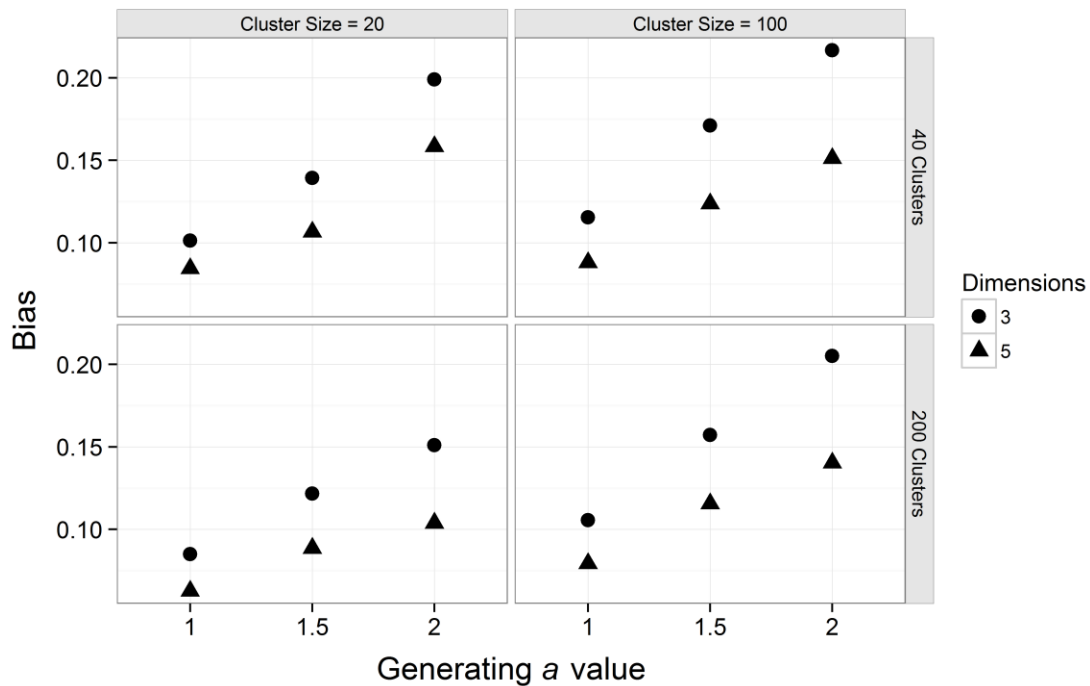
Figure 2. Item difficulty bias (*y* axis) as a function of generating *a* value (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

There was a significant quadratic effect of generating *d* value on item difficulty bias. However, as shown in Figure 3, this effect was likely due to the generating *a* values,

with more discriminating items showing greater bias, especially for middle-difficulty items. Nevertheless, the higher the generating *d* value (i.e., the easier the item), the greater the bias, above and beyond the effect of item discrimination. In addition, the effect of the ICC level was present only when cluster size was small (top panels): the lower the ICC level, the smaller the bias in item difficulty; ICC level did not appear to affect item difficulty bias when clusters were large (bottom panels).
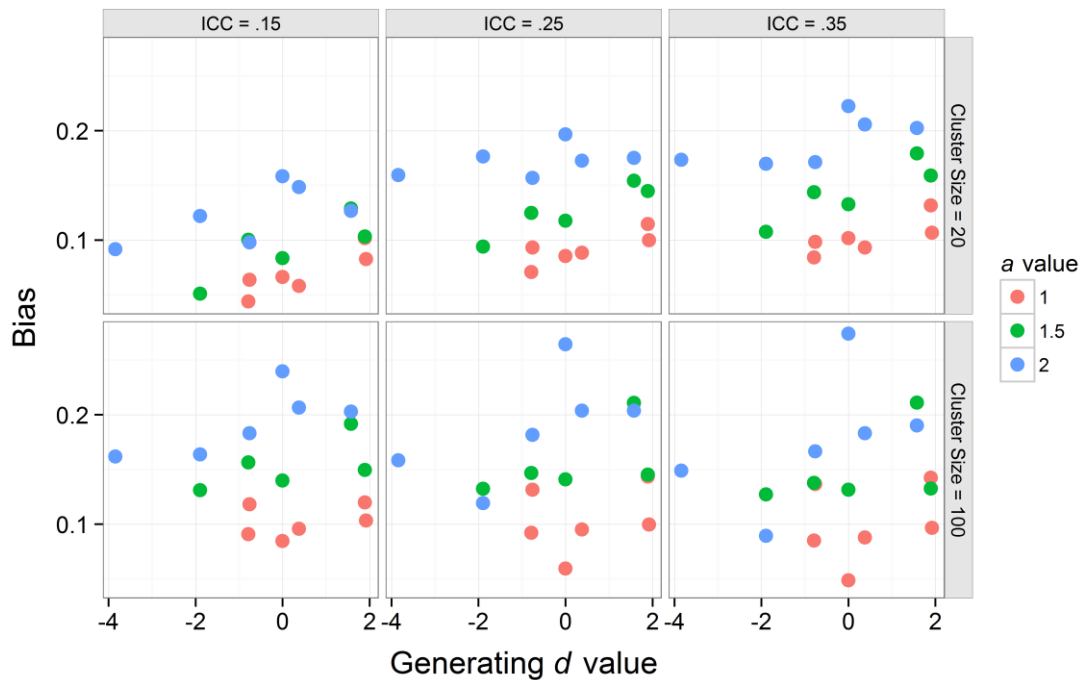
Figure 3. Item difficulty bias (*y* axis) as a function of generating *d* value (*x* axis), cluster size (top vs. bottom panels), ICC level (columns of panels), and generating *a* value (colors).
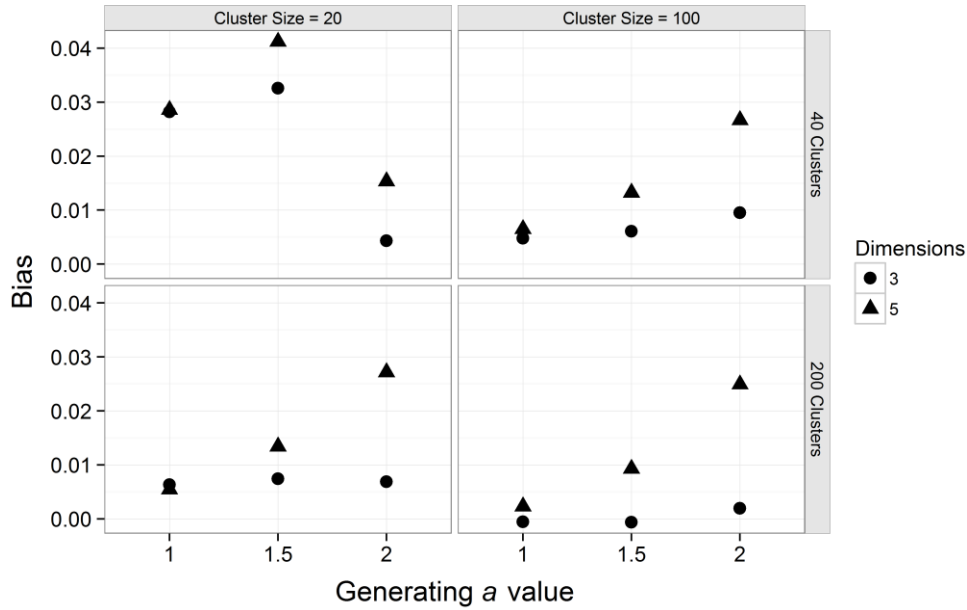
Figure 4. Item discrimination bias (*y* axis) as a function of generating *a* value (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

Item discrimination. The bias in item discrimination across replications and conditions was small (mean bias = 0.013, *SD* = 0.027). Linear regression (Table A2) revealed that the most important factors affecting item discrimination bias were sample size, the number of dimensions, and generating *a* value.

Similar to the bias in item difficulty, the bias in item discrimination was larger for higher generating *a* values (see Figure 4), with the exception of the smallest sample size combination (top left panel, where *a* = 2). However, this could be due to chance. Unlike item difficulty bias, item discrimination bias tended to be smaller in the three-dimensional models than in the five-dimensional models. In addition, item discrimination bias was the smallest when the overall sample size was the largest (bottom right panel),

and the make-up of number of clusters versus cluster size appeared to be of little

significance.
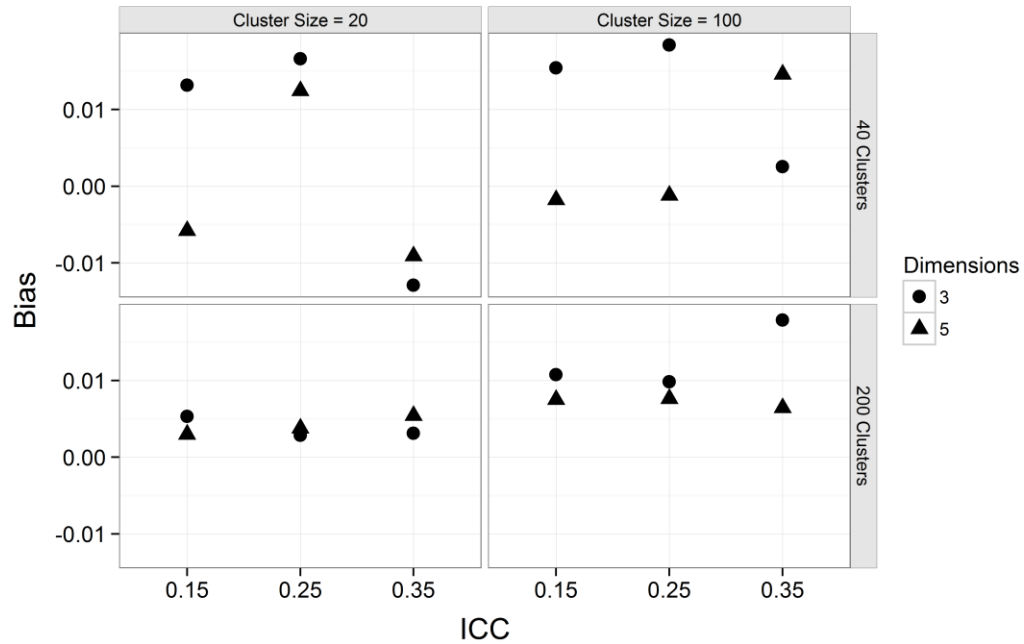
Figure 5. Level 2 (between) variance bias (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

Variances and covariances. Recall that only Level 2 (between) variances were estimated; Level 1 (within) variances were fixed to 1.0 for identification. Thus, the three parameters of interest here were Level 2 (between) variances, Level 2 (between) covariances, and Level 1 (within) covariances.

Level 2 (between) variances. Overall, Level 2 (between) variances were very slightly positively biased across all conditions (mean bias = 0.005, *SD* = 0.010). More than half of the variability in this bias was explained by the number of dimensions, the ICC level, the number of clusters, cluster size, and the interactions of these factors (see

Table A3). As shown in Figure 5, when the number of clusters was small (top two panels), bias in the Level 2 (between) variances tended to be smaller in the five-dimensional models than in the three-dimensional models, whereas the number of dimensions did not appear important when there was a large number of clusters (bottom two panels). Finally, there was no clear pattern in terms of the ICC level. The effect for ICC displayed in the top left panel could simply be due to chance because of the small sample size there (800).

Level 2 (between) correlations. Correlations rather than covariances among the Level 2 dimensions were examined for two reasons. First, one could argue that correlations allow for a more accurate examination of bias in the relationships among the latent dimensions than covariances because covariances contain bias due to the bias in the Level 2 variances. By converting the estimated covariances to correlations and subtracting the generating correlation values (.7, .8, and .9) in the calculation of bias, the bias due to the Level 2 variances cancels out in the conversion formula and does not carry over into the result. As such, one obtains a pure estimate of the discrepancy between the estimated and generating values. The second reason is that correlations are far easier to interpret than covariances. In addition, the Level 1 (within) covariances are already on the correlation metric because the variances there were set to 1.0. Thus, examining correlations rather than covariances allows for direct comparison across Level 1 and Level 2 correlation bias.

Similar to the Level 2 (between) variances, Level 2 correlations were very slightly positively biased across all conditions (mean bias = .003, *SD* = .008). Again, the factors accounting for the majority of the variability in this bias were the number of dimensions,

the ICC level, the number of clusters, cluster size, and the interactions among them (see

Table A4). Although Figure 6 does not reveal a clear pattern of these effects, it appears as

though bias was lower when there were more clusters (bottom two panels), and especially

when the clusters were small (bottom left panel). With respect to the ICC level and the

number of dimensions, it is difficult to draw any conclusions with certainty. It is

important to note, however, that the bias values across different factors displayed in

Figure 6 were very small and relatively close to one another. One could argue that such

small differences are not important.

Figure 6. Level 2 (between) correlation bias (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

Level 1 (within) correlations. Overall, the Level 1 (within) correlation bias was

small and negative (mean bias = -.020 *SD* = .018), indicating that the relationships

between the dimensions at Level 1 tended to be slightly underestimated. Linear

regression revealed that almost all of the variability in this bias was due to the main effects of generating value, the number of dimensions, and their interaction (see Table A5). As shown in Figure 7, bias was smaller in the three-dimensional models than in the five-dimensional models. In addition, the larger the generating values for the Level 1 (within) correlations, the more those correlations were underestimated (i.e., greater bias).
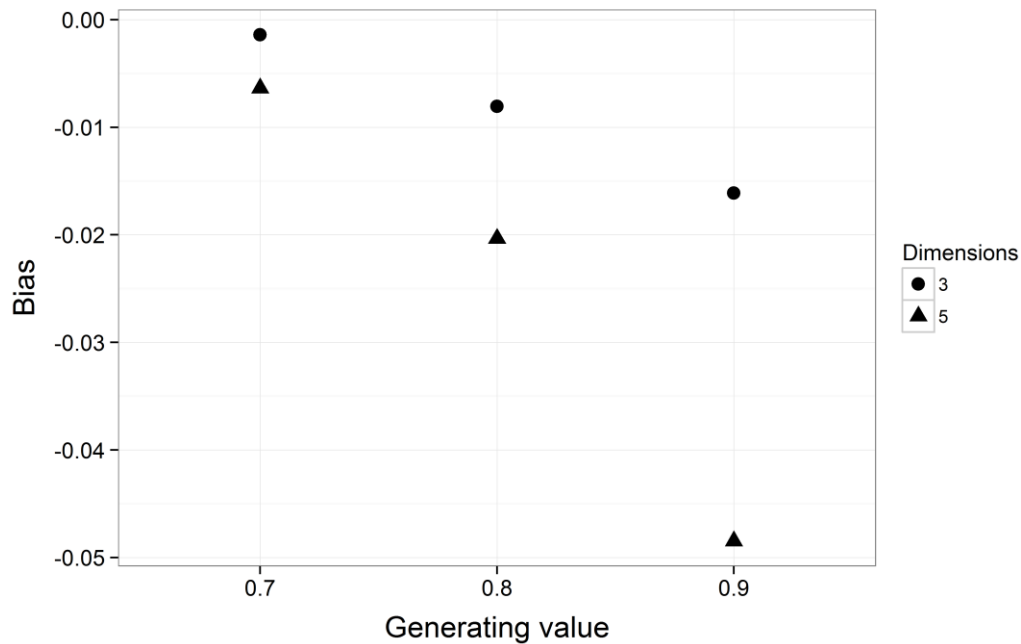
Figure 7. Level 1 (within) correlation bias (*y* axis) as a function of generating value (*x* axis) and number of dimensions (shapes).

Ability estimates. Level 2 (between) and Level 1 (within) ability estimates were biased very slightly across conditions. On average, Level 2 (between) abilities were slightly underestimated (mean = -0.074, *SD* = 0.026), whereas Level 1 (within) abilities were unbiased (mean = -0.008, *SD* = 0.005). Again, I used linear regression to identify the most important factors that impacted bias in the ability estimates. The results from the full models with all predictors are presented in Table A6 and A7 in Appendix A.
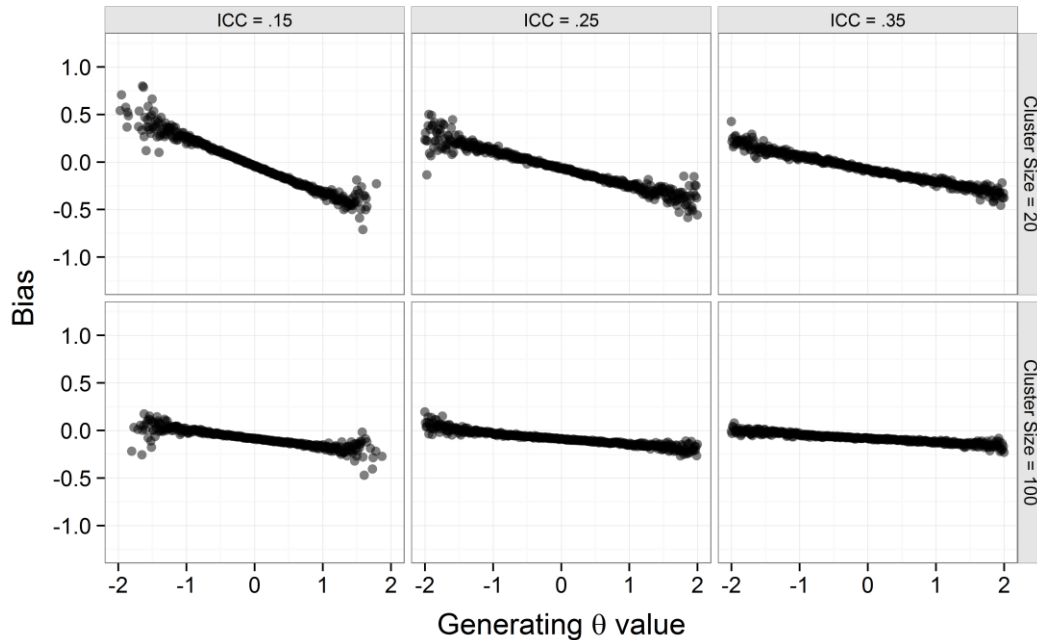
Figure 8. Level 2 (between) ability estimate bias (*y* axis) as a function of rounded generating θ value (*x* axis), cluster size (top vs. bottom panels), and ICC level (columns of panels from left to right).

As shown in Figure 8, 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 was 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 increased as the ICC level increased. However, the effect of the ICC level was not so profound when cluster size was large.

Bias in the ability estimates at Level 1 (within) was essentially a function of generating θ value (see Figure 9). Again, ability at the low end of the proficiency

continuum were overestimated, whereas abilities at the high end of the proficiency continuum were underestimated. The majority of mid-level abilities were unbiased.
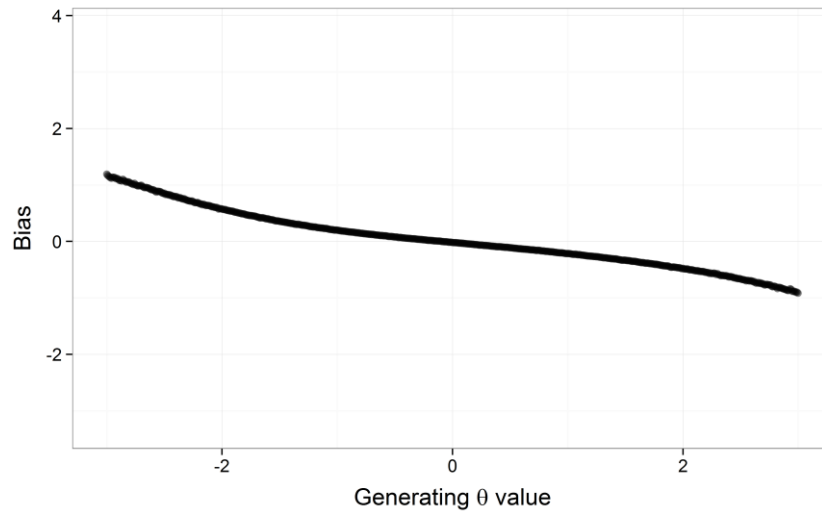
Figure 9. Level 1 (within) ability estimate bias (*y* axis) as a function of rounded generating θ value (*x* axis).

RMSE

Root mean squared error (RMSE) as defined in Equation 3.3 is a combination of both bias and sampling variability, thus providing insight not only into the average accuracy of MH-RM in recovering the parameters of the model, but also into its efficiency (i.e., extent to which estimates were stable across replications). Again, results are presented for item difficulty, item discrimination, Level 2 (between) variances, Level 2 (between) correlations, Level 1 (within) correlations, and ability estimates.

Item difficulty. On average, item difficulty RMSE was not overly large (mean RMSE = 0.193, *SD* = 0.110). Linear regression revealed that several simulation condition factors, generating *d* and *a* values, as well as interactions accounted for the majority of variance in item difficulty RMSE (see Table A8). As was the case with bias, a sizeable part of the variance in item difficulty RMSE was accounted for by the quadratic effect of
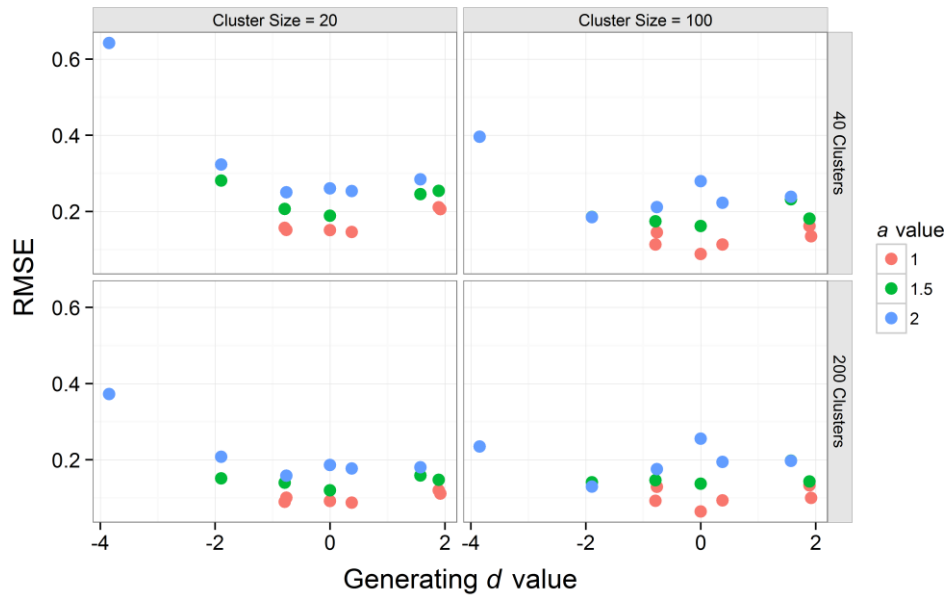
Figure 11. Item difficulty RMSE (*y* axis) as a function of generating *a* value (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

As expected, RMSE tended to be the smallest when the overall sample size was the largest (bottom right panel), with a slight advantage of a large number of small clusters (bottom left panel) over a smaller number of large clusters (top right panel). Figure 11 also shows this pattern. Again, one can see the positive relationship between generating *a* value and RMSE. Finally, the number of dimensions affected item difficulty RMSE such that the three-dimensional models had slightly higher item difficulty RMSE than the five-dimensional models. The same pattern was observed in item difficulty bias. However, there the effect was consistent across all four combinations of number of clusters and cluster size. Here, the effect was more visible when cluster size was large (right-hand-side panels in Figure 11).

Item discrimination. Overall, item discrimination RMSE was small across conditions and replications (mean RMSE = 0.123, *SD* = 0.085). Almost all of the variance in item discrimination RMSE was accounted for by sample size (number of clusters and cluster size and their interaction), the main effect of generating *a* value, and the main and quadratic effects of generating *d* value (see Table A9). The quadratic effect of generating *d* value is somewhat visible in Figure 12. But again, generating item discrimination is already part of item difficulty, which may be why the quadratic effect of *d* was significant (e.g., examine one color at a time to see how the RMSE dips for *d* = 0 and increases slightly for easier and more difficult items).
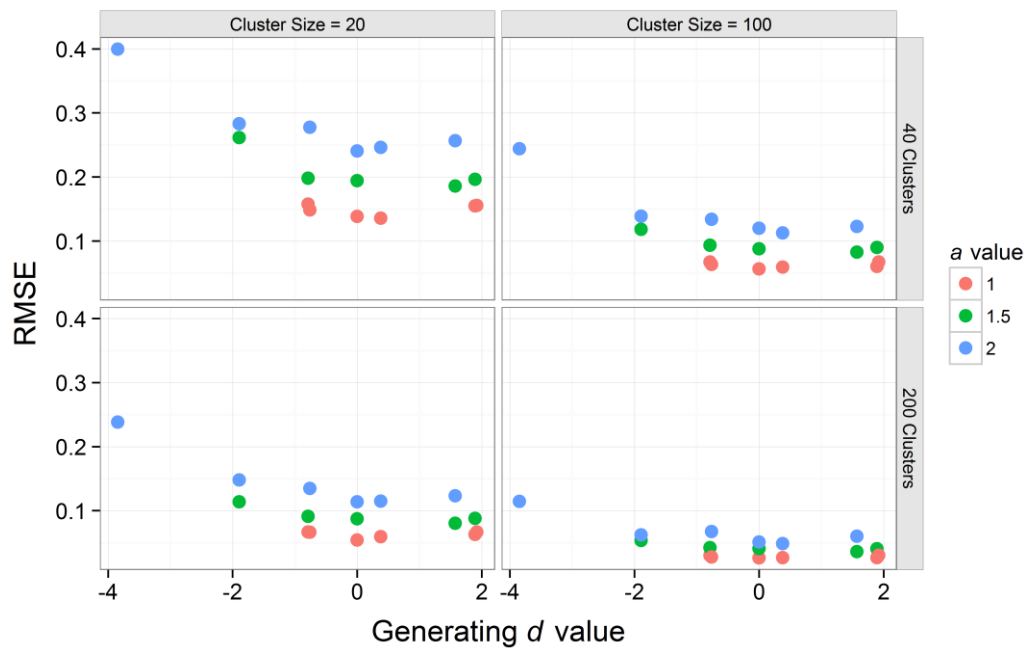
Figure 12. Item discrimination RMSE (*y* axis) as a function of generating *d* value (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and generating *a* value (colors).

In addition, there were no extremely easy items; otherwise, one might expect to see high RMSE on the right-hand-side of each panel as well, making the plot symmetric and revealing the quadratic effect of *d.* Still, highly discriminating *and* difficult items resulted in larger RMSE. By contrast, the effect of generating *a* value alone was much more prominent; the higher the generating *a*, the greater the item discrimination RMSE. In terms of sample size, a large total sample size (bottom right panel) appeared to trump the effect of either number of clusters or cluster size.

Variances and covariances. Following the structure of the results for bias, I present RMSE for the Level 2 (between) variances, Level 2 (between) correlations, and Level 1 (within) correlations. Note that because Level 2 variance and correlation bias was so small, RMSE is predominantly a function of sampling variability.
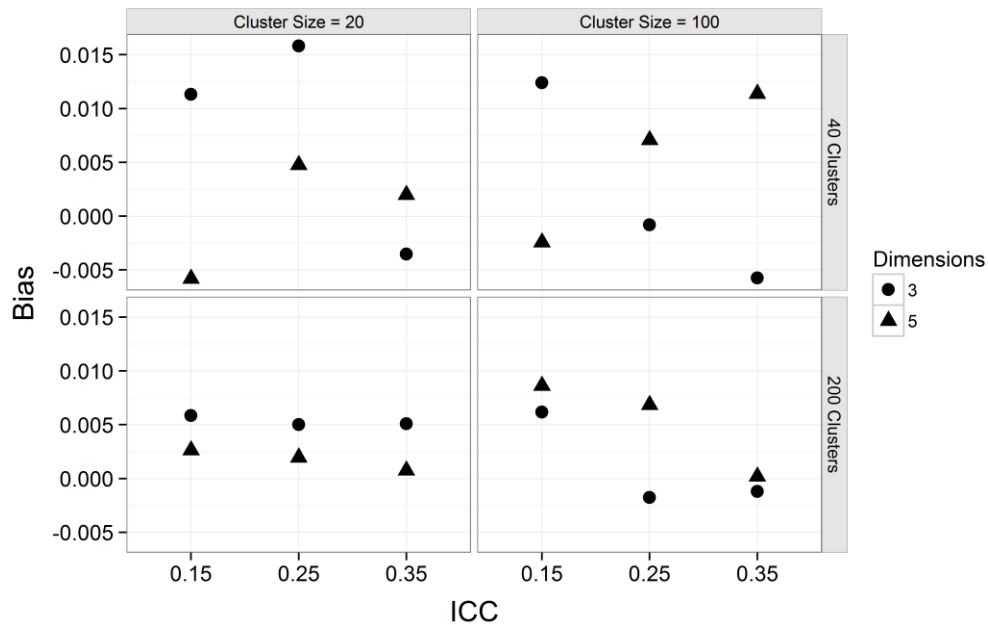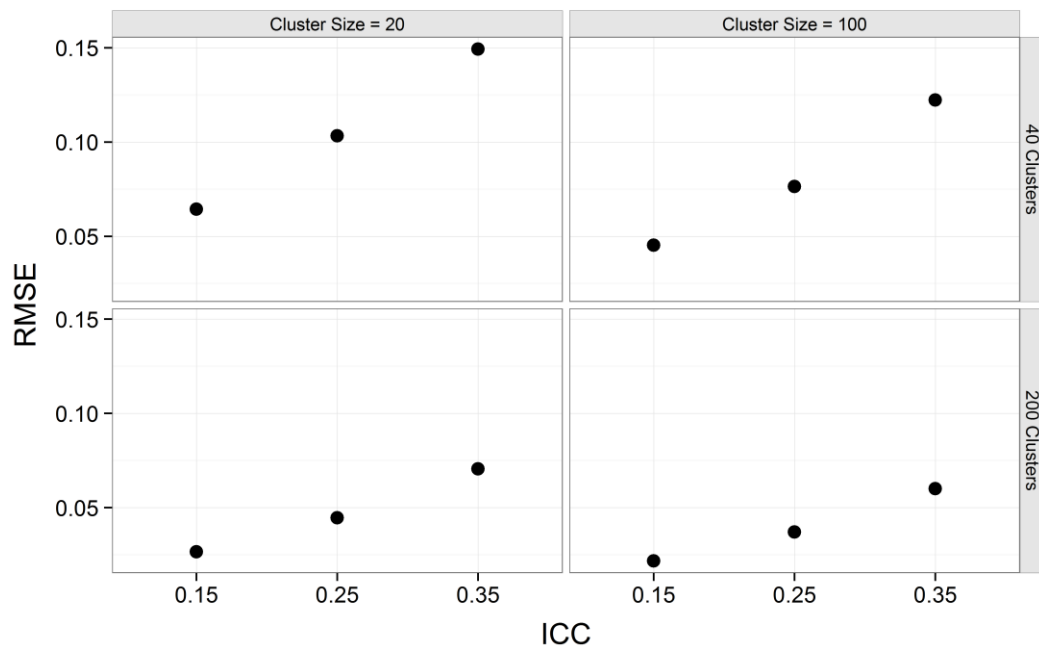

Figure 13. Level 2 (between) variance RMSE (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), and cluster size (left-hand-side vs. right-hand-side panels).

Level 2 (between) variances. On average, Level 2 (between) variance RMSE was very small across conditions and replications (mean RMSE = 0.068, *SD* = 0.038). Its variability was predominantly a function of ICC level, the number of clusters, cluster size, and the interactions of these factors (see Table A10). As shown in Figure 13, ICC level and the number of clusters appeared to affect Level 2 (between) variance RMSE the most, such that as the ICC level increased, so did RMSE. Additionally, a larger number of clusters (regardless of size) was accompanied by lower RMSE.

Level 2 (between) correlations. Similar to the variances, the Level 2 (between) correlations had small RMSE overall (mean RMSE = .056, *SD* = .036). Linear regression revealed that the variability in RMSE was almost completely accounted for by the number of clusters, generating value, cluster size, ICC level, and their interactions (see Table A11). Figure 14 clearly shows the effect of number of clusters: the more clusters (regardless of size and ICC level), the lower the RMSE (bottom two panels). When the number of clusters was small, however, ICC did play a role (the higher, the better), especially when the clusters were small (top left panel). Finally, the higher the generating value of the Level 2 (between) correlation, the smaller the RMSE. Interestingly, the last pattern was also observed with the Level 1 (within) correlation bias, but not with the Level 2 (between) correlation bias.
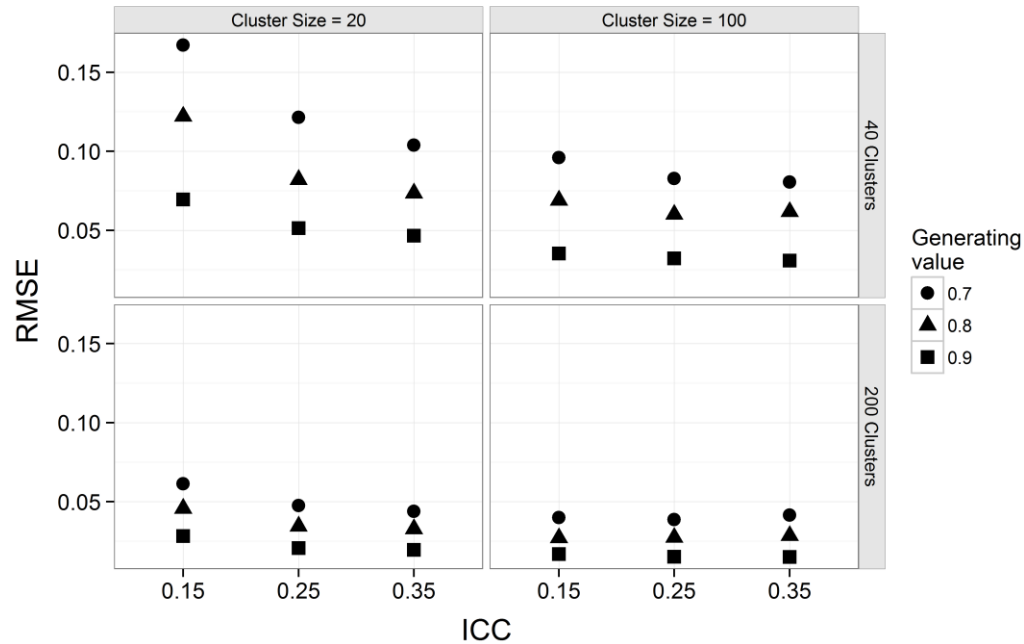
Figure 14. Level 2 (between) correlation RMSE (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and generating value (shapes).

Level 1 (within) correlations. In line with Level 2 (between) variances and correlations, Level 1 (within) correlations had a small RMSE across conditions and replications (mean RMSE = .030, *SD* = .015). Linear regression revealed that nearly all of the variability in Level 1 (within) correlation RMSE could be accounted for by four factors and their interactions: the number of dimensions, the generating value, the number of clusters, and cluster size (see Table A12). As shown in Figure 15, unlike the effect of the number of dimensions on Level 1 (within) correlation bias, RMSE was consistently smaller in the three-dimensional models than in the five-dimensional models, and especially at higher generating values. Overall, the higher the generating value, the larger

the Level 1 (within) correlation RMSE. Finally, the greater the total sample size, the

smaller the RMSE (regardless of number of clusters vs. cluster size).
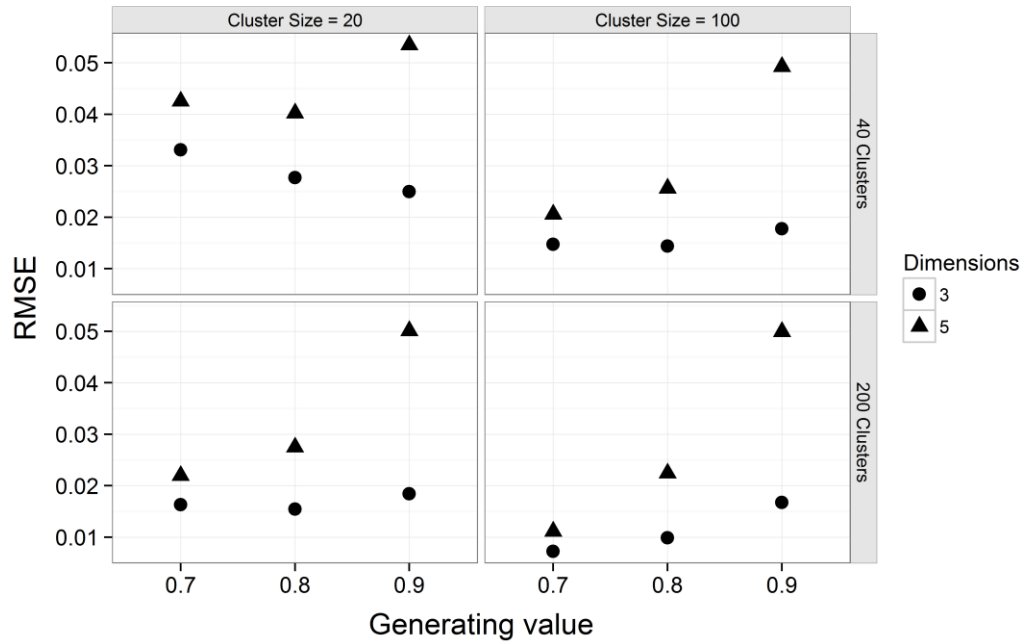
Figure 15. Level 1 (within) correlation RMSE (*y* axis) as a function of generating value (*x* axis), number of

clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of

dimensions (shapes).

Ability estimates. Overall, RMSE for the Level 2 (between) ability estimates was

small (mean = 0.207, *SD* = 0.056). However, RMSE was noticeably larger for the Level 1

(within) ability estimates (mean = 0.495, *SD* = 0.041) 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

(see Table A13), whereas for Level 1 (within) abilities RMSE was essentially a function

of only the generating θ value (see Table A14). Figure 16 clearly shows the strong effect

of cluster size on RMSE for the Level 2 (between) ability estimates. RMSE was much smaller for cluster sizes of 100 than it was for cluster sizes of 20. In addition, RMSE was smaller for generating θ values near the middle range of proficiency and tended to get larger for very low or very high proficiency levels, especially for small clusters.
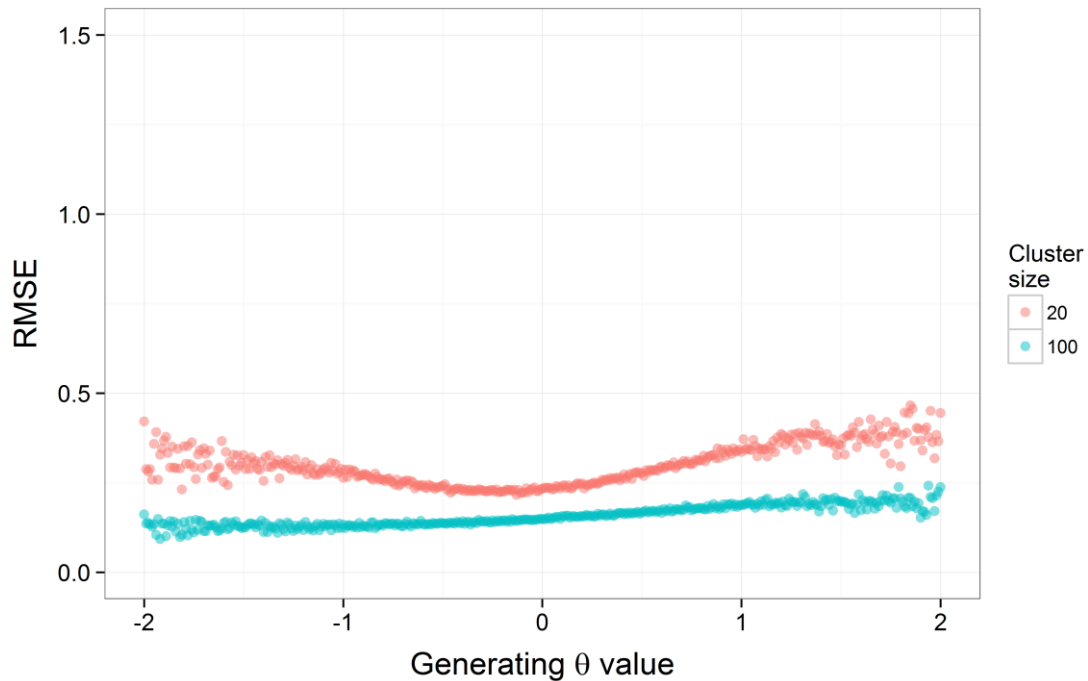
Figure 16. Level 2 (between) ability estimate RMSE (*y* axis) as a function of generating θ value (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

By contrast, RMSE for the Level 1 (within) ability estimates was not affected by sample size and was purely a function of generating θ value (see Figure 17). Here, even small deviations from 0 (or the average proficiency in each cluster) were associated with large RMSE, and the effect was even stronger at the extremes of the proficiency continuum.
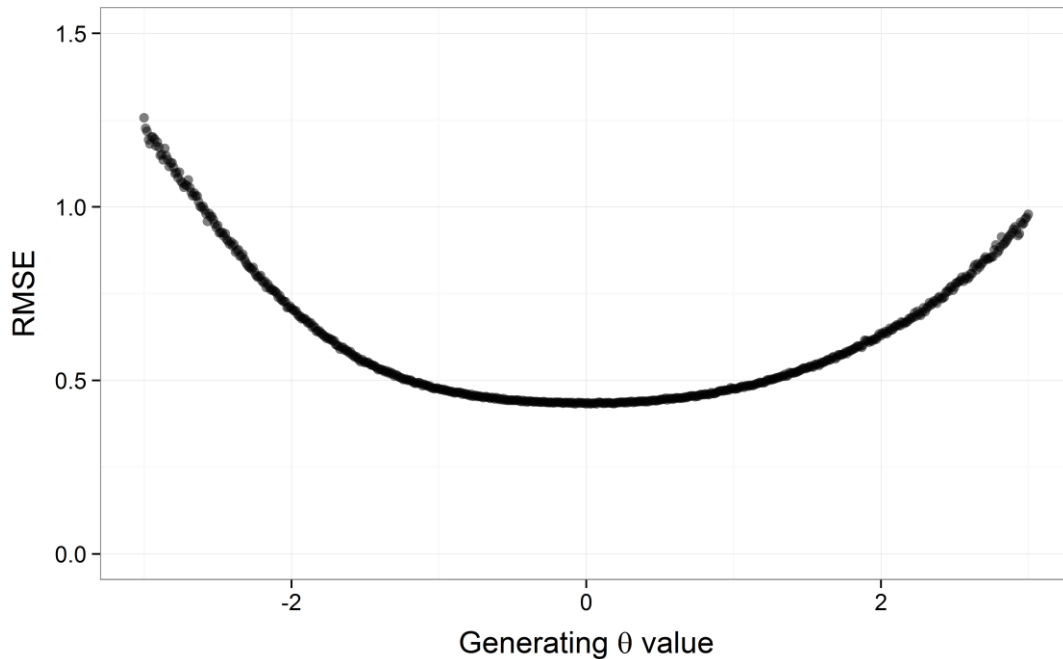
Figure 17. Level 1 (within) ability estimate RMSE (*y* axis) as a function of generating *θ* value (*x* axis).

Standard Error Accuracy

As described in Chapter III, the accuracy of the standard errors for the item

parameters and latent variances and covariances was examined by constructing 95%

confidence intervals based on the parameter estimate from each replication and the

analytical standard error from the flexMIRT output. A coverage rate was then computed

as the number of replications in which the 95% confidence interval contained the

generating parameter value. Coverage rates close to 95% indicate that the analytical

standard errors across replications tended to be accurate. Values greater than 95%

indicate that the standard errors were too large, whereas values smaller than 95% indicate

that the standard errors were too small.

It is important to note, however, that in the presence of bias, the coverage rates

computed as described above may not reflect the accuracy of analytical standard errors

due to bias in the parameter estimates. Specifically, the greater the bias, the smaller the

confidence interval coverage rates would be. This is because the generating value is more

likely to be outside the conference interval constructed around a biased estimate. The

results presented so far indicated that bias was small overall (see Appendix B). Thus, if

the standard errors were accurate, the coverage rates should not be much lower than 95%.

       Another factor that could impact confidence interval coverage rates is the number

of observations on which they are based. In this case, the coverage rates were computed

based on the estimates and standard errors for each parameter in the 100 replications. As

discussed in Chapter III, more observations are desirable, in order to assess the

trustworthiness of standard errors 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 standard errors) of certain

parameters and conditions were negative in some replications. Because of this, the

standard errors for these parameters and conditions were treated as missing. Thus, the

confidence interval coverage rate for a given parameter was based only on the

replications with a nonnegative error variance for that parameter. In the case of

parameters with the same generating value, only those parameters with a negative error

variance were excluded from the computation of coverage.

       The analytical error variances for item discrimination were negative under only a

few conditions, and in no more than 1% of the replications. For item difficulties, the

analytical error variances were also negative only under certain conditions and in no

more than 3% of the replications. Negative analytical error variances were much more

prevalent for the latent variances and covariances. In the three-dimensional models,

negative analytical error variances occurred in up to 1% of the Level 2 (between) variances and covariances, and in up to 24% of the Level 1 (within) covariances/correlations, usually in the conditions with the largest sample size (20,000). In the five-dimensional models, negative analytical error variances occurred much more frequently: in up to 23% of the Level 2 (between) variances and covariances, and in up to 56% of the Level 1 (within) covariances/correlations.

These results indicated that reasonable analytical standard errors were produced most of the time for item difficulty and item discrimination regardless of sample size and the number of dimensions. However, for Level 2 (between) variances and covariances, and especially for the Level 1 (within) covariances/correlations, analytical error variances were sometimes negative in three-dimensional models with large sample sizes, and frequently negative in the five-dimensional models, indicating the confidence interval coverage rates for these parameters may not be trustworthy.

Nevertheless, below I examine the confidence interval coverage rates for item parameters, variances and covariances, and ability estimates (following the same structure as bias and RMSE) for several reasons. First, standard errors are important because they speak to the amount of variability in an estimate we could expect upon replications with a similar sample size. Second, prior studies have provided conflicting evidence regarding the accuracy of standard errors obtained via MH-RM; thus, any additional information about standard errors with MH-RM is welcome. And finally, no prior research has examined the accuracy of standard errors under the specific conditions considered in this study.

Item difficulty. Across conditions, confidence interval coverage for item difficulty was much lower than the nominal rate of 95% (mean = .560, *SD* = .117), indicating the analytical standard errors were too small. Regression analysis revealed that almost all of the variability in confidence interval coverage could be accounted for by cluster size, the number of clusters, generating *a* value, the number of dimensions, generating *d* value, the ICC level, as well as some interactions among these factors (see Table A14).
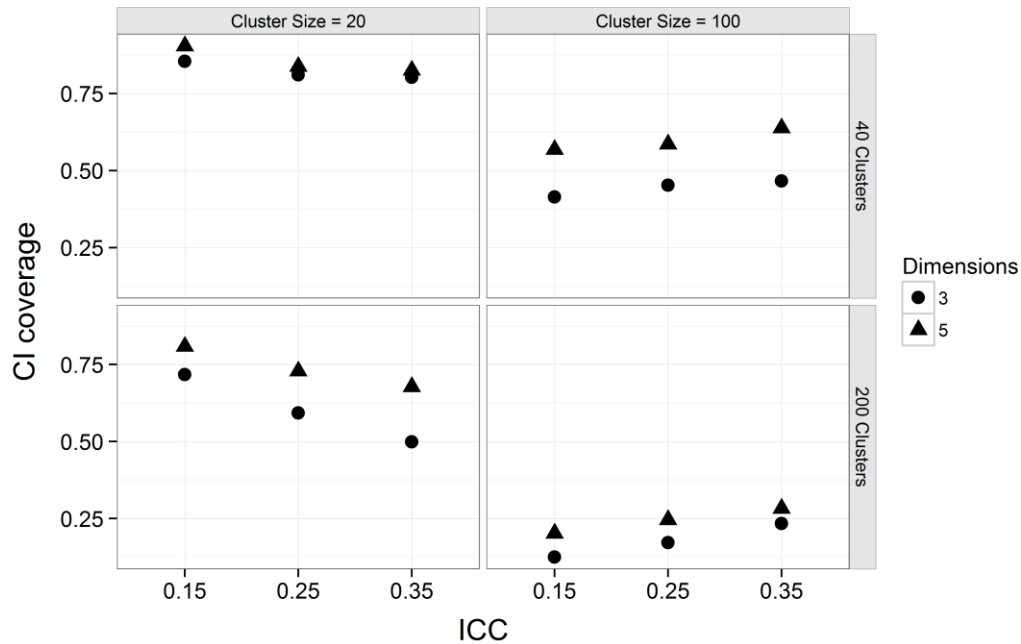
Figure 18. Item difficulty confidence interval coverage (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and number of dimensions (shapes).

Figure 18 shows an interaction between cluster size and the ICC: for small clusters, as the ICC increased, the confidence interval coverage decreased, whereas for large clusters, as the ICC increased, the confidence interval coverage increased. Most

importantly, Figure 18 clearly shows that the most important factor affecting confidence interval coverage was sample size. Specifically, item difficulty confidence interval coverage rates were closest to their desired value for the smallest sample size of 800 (top left panel); they were extremely small for a total sample size of 20,000 (bottom right panel), and somewhere in between for a sample size of 4,000 (the remaining two panels). This pattern did not appear to be related to bias; therefore, the results speaks directly to the accuracy of the standard errors for item difficulty, which were too small for large sample sizes.
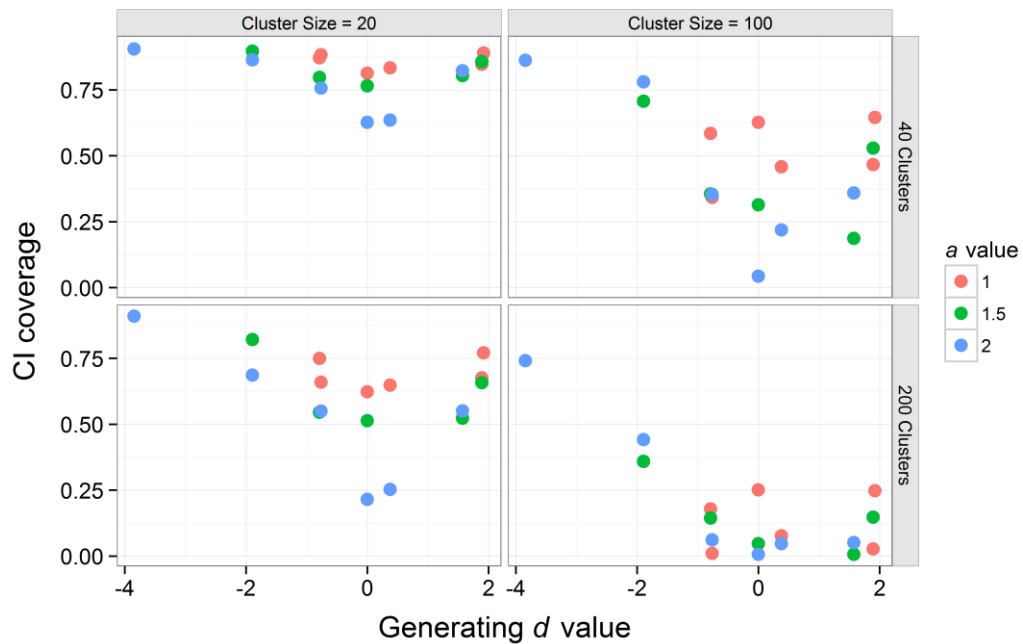
Figure 19. Item difficulty confidence interval coverage (*y* axis) as a function of generating *d* value (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and generating *a* value (colors).

Figure 19 reiterates the effect of sample size and also shows the interaction between generating *a* and *d* values. Similar to bias and RMSE, there was a significant

quadratic effect of generating *d*, and this effect appears to be confounded by the effect of generating *a* value. For example, the quadratic effect of generating *d* value on item difficulty confidence interval coverage is more visible for highly discriminating items (blue color) than less discriminating items (green and pink) and indicates that confidence interval coverage was slightly higher (i.e., standard errors are more accurate) for extremely easy (right within each panel) or extremely difficult (left within each panel) items. By itself, higher discrimination was associated with smaller confidence interval coverage rates for item difficulty.
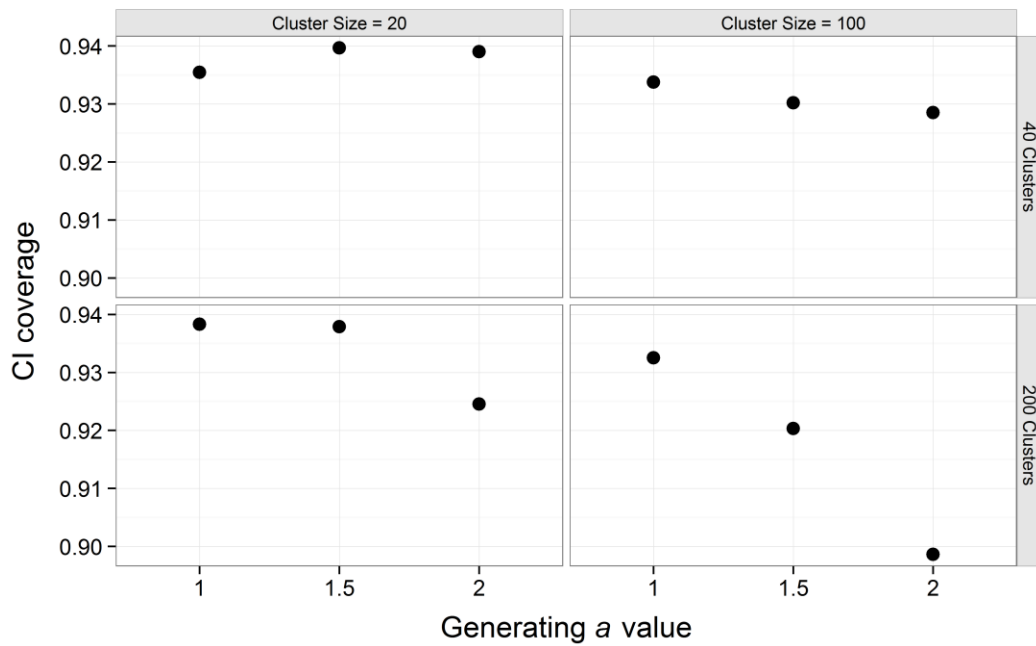
Figure 20. Item discrimination confidence interval coverage (*y* axis) as a function of generating *a* value (*x* axis), number of clusters (top vs. bottom panels), and cluster size (left-hand-side vs. right-hand-side panels).

Item discrimination. Unlike item difficulty, confidence interval coverage for item discrimination across conditions was very close to the nominal rate of 95% (mean =

0.930, *SD* = 0.016), indicating that the analytical standard errors were accurate. Several predictors and their interactions were able to explain a small proportion of the variability in confidence interval coverage (see Table A16).

As shown in Figure 20, similar to item difficulty, confidence interval coverage rates for item discrimination were closest to 95% for the smallest total sample size (800). Coverage rates were lower for sample sizes of 4,000, and even lower for the sample size of 20,000. Still, item discrimination confidence interval coverage rates were not nearly as low for the larger sample sizes as they were for item difficulty. In terms of generating *a* values, which again appeared to affect the confidence interval coverage rates the most, higher generating *a* values were associated with lower confidence interval coverage rates (i.e., smaller analytical standard errors relative to the empirical standard errors).

Variances and covariances. Following the layout of the results so far, the regression analyses of confidence interval coverage rates associated with the latent variances and covariances/correlations are presented next.

Level 2 (between) variances. On average, across conditions and replications, the confidence interval coverage rates for the Level 2 (between) variances were very close to 95% (mean = .925, *SD* = .004), again indicating that the standard errors were accurate. Linear regression revealed that half of the variability in confidence interval coverage could be accounted for by cluster size, the number of clusters, the number of dimensions, the ICC level, and interactions among these factors (see Table A17). Figure 21 shows no clear pattern of how the factors interact to affect confidence interval coverage. However, it appears that larger sample size (especially larger cluster size) was associated with confidence interval coverage rates closer to 95%, and more so in the five-dimensional

models. The effect of ICC level is not clear. However, it appears to interact with the other factors via three-way interactions. The interpretation of those is moot, especially considering the small deviations of the confidence interval coverage rates from 95%.
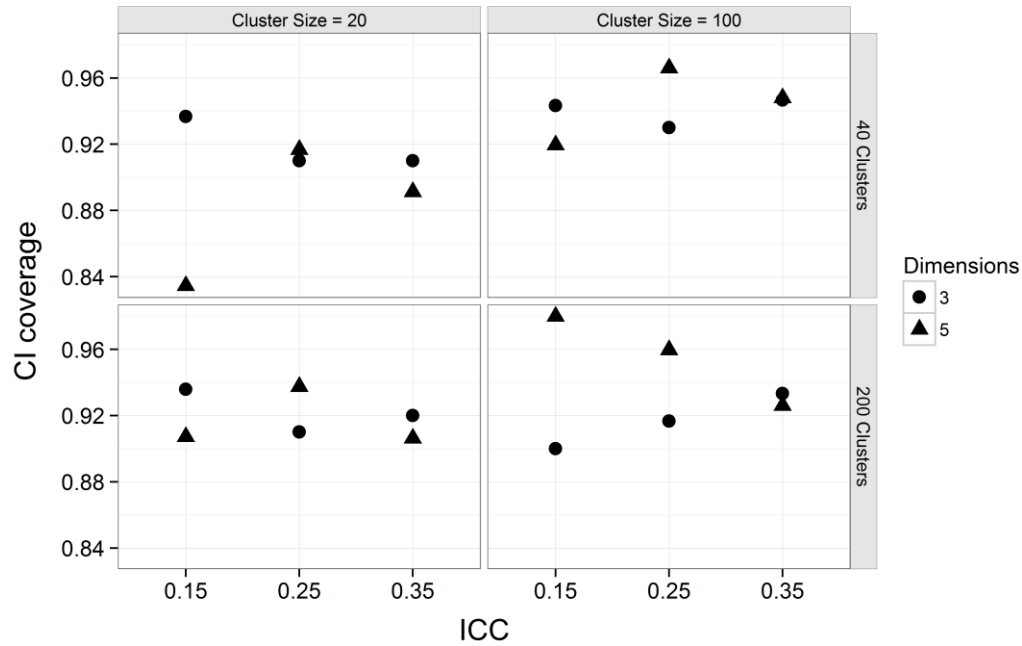
Figure 21. Level 2 (between) variance confidence interval coverage (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and the number of dimensions (shapes).

Level 2 (between) covariances. Similar to the variances, the confidence interval coverage rates for the Level 2 (between) covariances were close to 95% across conditions and replications (mean = .927, *SD* = .005), indicating that the standard errors tended to be accurate. The majority of the variability in confidence interval coverage was explained by cluster size, the number of clusters, the ICC level, the number of dimensions, and the interactions of these factors (see Table A18). Figure 22 shows that for the most part, confidence interval coverage rates were closer to 95% when the sample size was large.

This result is consistent with the effect of sample size on confidence interval coverage for the Level 2 variances. The effects of the number of dimensions and ICC level were not clear, although interestingly enough the pattern was almost identical to the one earlier for the Level 2 variances.
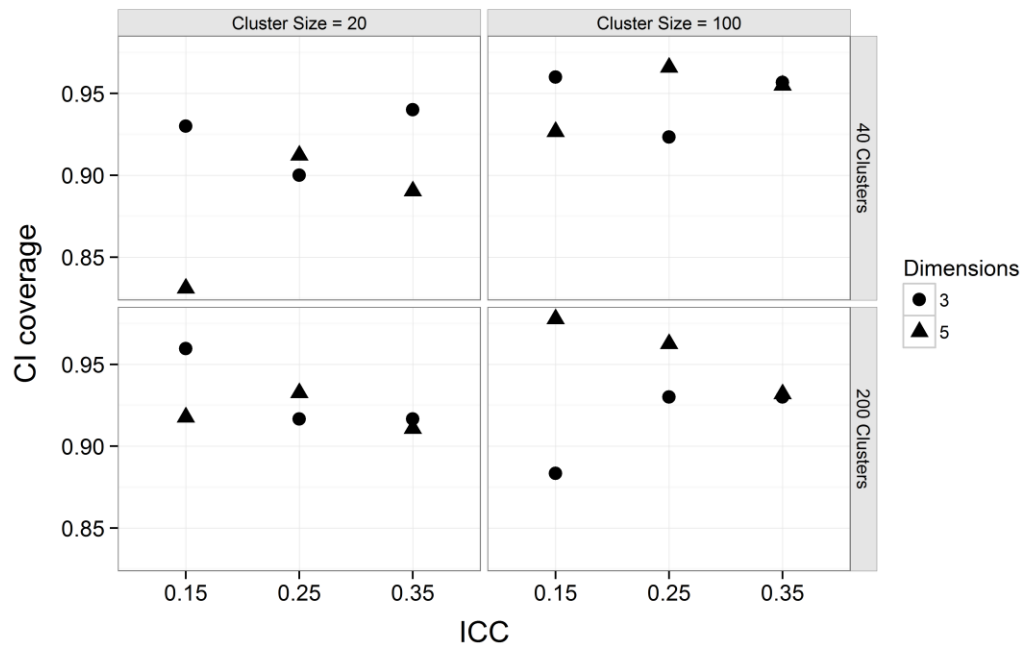
Figure 22. Level 2 (between) covariance confidence interval coverage (*y* axis) as a function of ICC level (*x* axis), number of clusters (top vs. bottom panels), cluster size (left-hand-side vs. right-hand-side panels), and the number of dimensions (shapes).

Level 1 (within) covariances. Unlike the Level 2 (between) variances and covariances, confidence interval coverage for the Level 1 (within) covariances was much lower than 95% (mean = .563, *SD* = .285), indicating that, on average, the analytical standard errors were much lower than their empirical counterparts. The majority of the variability in confidence interval coverage was explained primarily by the generating value and its interaction with cluster size and the number of clusters (see Table A19).
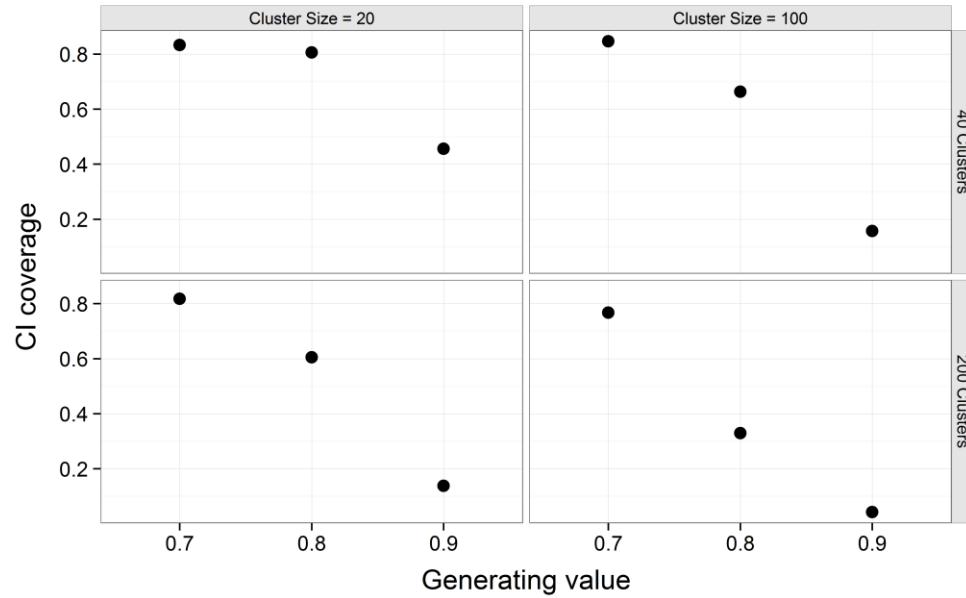
Figure 23. Level 1 (within) covariance confidence interval coverage (*y* axis) as a function of generating value (*x* axis), number of clusters (top vs. bottom panels), and cluster size (left-hand-side vs. right-hand-side panels).

As shown in Figure 23, the larger the generating value for the Level 1 (within) covariances, the smaller the confidence interval coverage. In other words, the standard errors for Level 1 (within) covariances became too small as the level of correlation among the latent dimensions increased. There was also a relationship between total sample size and confidence interval coverage. Surprisingly, the smaller the sample size, the more accurate the standard errors. It is important to note, however, that the larger the sample size, the more negative error variances were found; thus the smaller number of observations on which Figure 23 was based. As such these results should be interpreted with caution.

Ability estimates. Confidence interval coverage rates for the ability estimates were computed similarly to the item parameters and variances and covariances.

Specifically, for each Level 2 unit (e.g., school), the 95% confidence interval was constructed from the standard error. A coverage rate was then computer as the number of schools in which the 95% confidence interval contained the generating parameter value. The same procedure was followed for Level 1 units. In addition, unlike the item parameters and latent variances and covariances, there were no negative standard errors for the ability estimates at either level, so the coverage rates here were based on all Level 2 and Level 1 estimates from all replications.

Across conditions, the mean confidence interval coverage was .887 (SD = .043) for Level 2 (between) ability estimates and .950 (SD = .005) for Level 1 (within) ability estimates. Tables A20 and A21 show the output from the full regression models with all predictors and interactions.
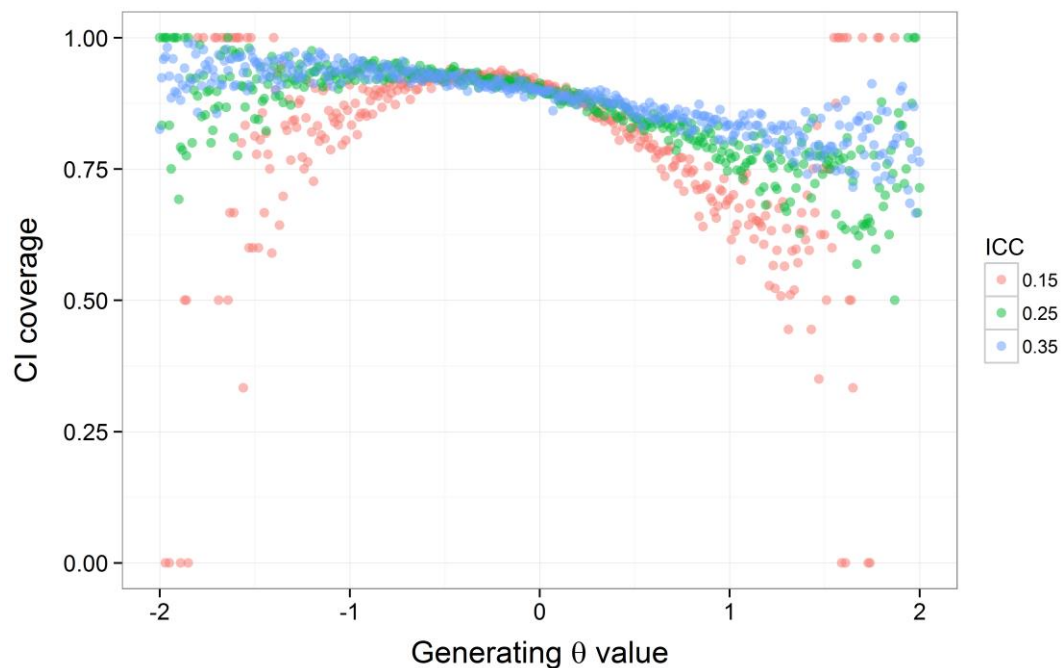
Figure 24. Level 2 (between) ability estimate confidence interval coverage (*y* axis) as a function of generating θ value (*x* axis) and ICC level (colors).

As shown in Figure 24[16], Level 2 (between) ability estimate confidence interval coverage was primarily a function of generating θ value. Specifically, at extremely low and extremely high levels of proficiency, confidence interval coverage rates were too high (and in a few occasions too low), whereas for ability levels closer to the mid-range, confidence interval coverage was still lower than 95%, but closer to this value (i.e., analytical standard errors were more accurate). Moreover, Figure 24 shows that coverage was closer to 95% for low proficiency (e.g., -2 to 0) than it was for high proficiency. Finally, this effect of generating θ value was moderated by the ICC level: the higher the ICC level, the better the coverage rates (i.e., more accurate standard errors).

Unlike Level 2, confidence interval coverage rates for Level 1 (within) ability estimates were at the desired level of 95%, on average. In other words, the relative size of the analytical standard errors for the Level 1 (within) ability estimates was accurate. Figure 25 shows the effect of generating θ value, which was the only significant predictor of the variability in confidence interval coverage for the Level 1 (within) ability estimates. Consistent with the results for bias and RMSE, confidence interval coverage rates were best for Level 1 (within) ability levels near the mid-range of the proficiency continuum. Standard errors were too small at the extremes. The average bias, RMSE, and confidence interval coverage across conditions for item parameters, Level 2 variances, Level 2 and Level 1 covariances/correlations, and ability estimates at both levels are presented in Table B1 in Appendix B. A summary of these results is provided in the next chapter.

---

[16] The points in Figure 24 (and all figures displaying ability estimates) are based on groups created by rounding the generating θ values at the second decimal (.01). Thus, points with perfect and zero coverage are most likely based on one school (Level 2) or one student (Level 1) within that generating ability interval.
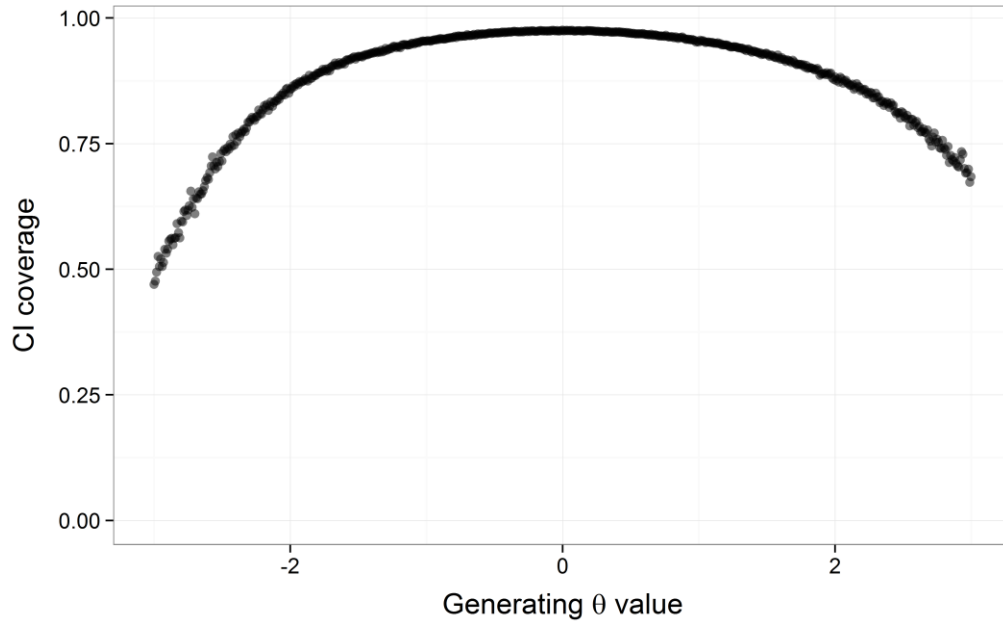
Figure 25. Level 1 (within) ability estimate confidence interval coverage (*y* axis) as a function of generating θ value (*x* axis).

Processing Time

Although MH-RM was not compared to any other algorithm in this study, it was still of interest to examine the average processing time across replications for each condition. The simulation study was conducted on two different computers: replications 1 through 25 of all 24 conditions were conducted on PC 1; replications 26 through 100 were conducted on PC 2. It is important to note that PC 2 was much more powerful (i.e., computationally faster) than PC 1. Thus, there was a lot of variability in estimation time[17] not only across the simulation conditions, but also across machines.

---

[17] Simulation of the data was also performed in flexMIRT, but because each simulation run took only 1-3 seconds, even for the five dimensional models with 20,000 simulees, only estimation times were examined.
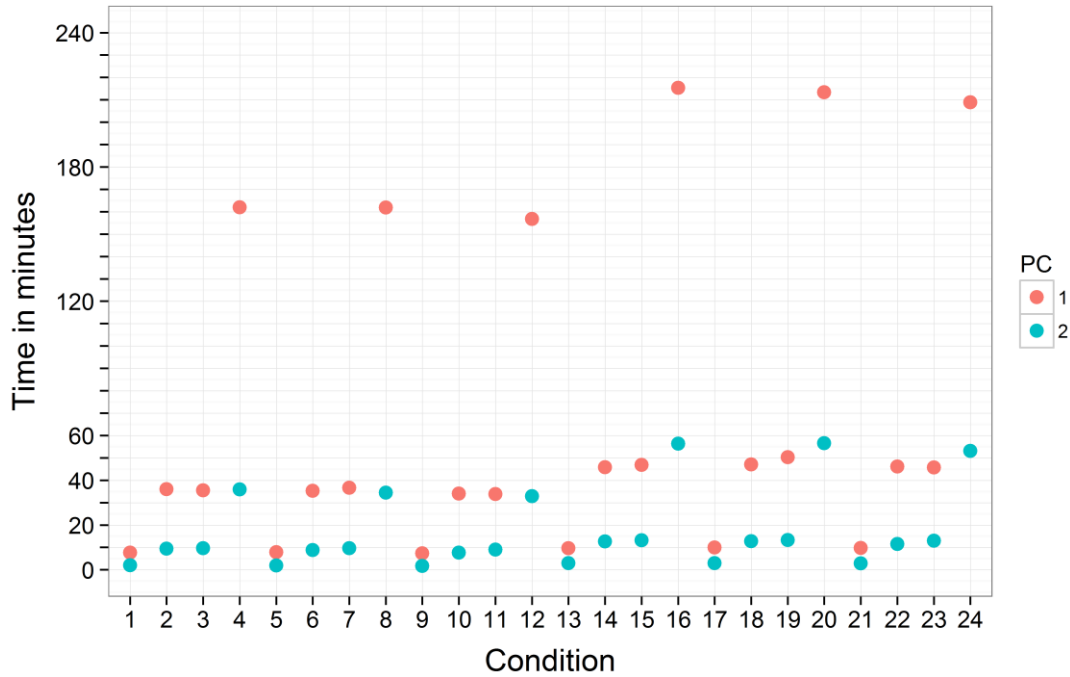
Figure 26. Average processing time in minutes (*y* axis) across conditions (*x* axis) by personal computer (1 = two cores, four logical processors; 2 = four cores, eight logical processors).

The average processing time by condition is provided in real time in Figure 26 for each of the two computers across replications. For PC 1 each point is based on 25 replications, whereas for PC 2 each point is based on 75 replications. Processing time for the estimation of the models was mostly a function of sample size. The models with 20,000 simulees took the most time: over 3.5 hours for PC 1 and almost an hour for PC 2. By contrast, the models with the smallest sample size (800), took 7-10 minutes for PC 1 and 2-3 minutes for PC 2, which is remarkably fast. It appears that when the sample size is small, the number of dimensions is not an issue. However, for larger sample sizes, a higher number of dimensions does add to the processing time. The make-up of the sample (i.e., number of clusters vs. cluster size) does not appear important, nor does the

ICC level. Overall, MH-RM shows great processing time efficiency in estimating these

complex models.

Chapter V

Discussion

The overarching goal of this chapter is to bring together the main ideas from the previous four chapters in a meaningful way. The chapter consists of five sections, each with a specific objective. I begin Chapter V with a summary of the results presented in the previous chapter. In addition to highlighting the main findings, this section elaborates on the interpretation of the results. Next, I discuss the limitations of the current study, and how they can impact the inferences one could draw from the results. Then, I place MH-RM in the spotlight and compare the findings from the current study to those from prior research, drawing conclusions specific to the accuracy and efficiency of MH-RM as an estimation algorithm. In the final two sections, I discuss the implications of this work for practice and point to possible directions for future research.

Summary

The purpose of this dissertation was to examine the performance of the MH-RM algorithm in the estimation of a 3PL ML-MIRT model under different conditions. Specifically, of particular interest was the bias, efficiency, and confidence interval coverage associated with item parameters (i.e., item difficulty and discrimination), latent variances and covariances/correlations at Level 2 (e.g., schools) and Level 1 (e.g., students), as well as the ability estimates at both levels. Each of these dependent variables was regressed on the simulation condition factors (i.e., the number of dimensions, the ICC level, the number of clusters, and cluster size), relevant generating parameter values, and their interactions, to identify the features of the model and the sample that had the greatest impact on the dependent variables (i.e., bias, RMSE, and standard error

accuracy). In addition, I examined the average estimation time under each condition for two different computers used in the study to inform researchers and practitioners using the MH-RM algorithm as implemented in flexMIRT as to the time demands to estimate similar models.

First, it is important to note that MH-RM was able to estimate all 100 replications for each of the 24 conditions. That is, the algorithm exhibited 100% convergence rate[18] and generally produced meaningful analytical standard errors. As described in Chapter IV, the error variances for some item parameters and latent variances/covariances were negative, usually when the overall sample size was large, and especially in the five-dimensional models. One should note that when this occurred, the error variances were extremely small in the replications in which they were positive. When only the item parameters are of interest, negative error variances are not likely to be a problem.

In terms of bias, or how far off estimates are from the generating (true) value for a given parameter on average, there was a small positive bias for item difficulty and Level 2 (between) ability estimates, and very little bias for item discrimination. The bias for latent variances and covariances and Level 1 (within) ability estimates was extremely small. Given the latent variances and covariances and Level 1 (e.g., student-level) ability estimates were essentially unbiased, they will not be considered further.

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. Item difficulties were the least biased when the ICC was small (.15), and cluster size was small (20).

---

[18] Out of 24,000 model calibrations, only two iterations failed to converge. Once the random seed number was changed, both of these iterations finished successfully.

Given highly discriminating items are beneficial for reliability and the overall integrity of the model, the only clear advantage here is that of having more dimensions.

By contrast, bias for item discrimination was very small, and actually smaller in the three-dimensional models. However, similar to bias for the difficulties, highly discriminating items tended to be more biased. It is important to note that logically, one might expect greater bias for higher generating values. That is, when the true value of a parameter is a larger number, there is more room for error, which manifests as greater bias. Moreover, as discussed in Chapter IV, item discrimination is also part of item difficulty. Thus, bias in one parameter also carries over into the other. On the whole, however, bias in item discrimination was very small, and there is no reason for concern.

What is somewhat concerning is the bias in the ability estimates (specifically at Level 2). On average, Level 2 (e.g., school) ability estimates were underestimated, (again, on average Level 1 [e.g., student] ability estimates were unbiased). The reason for the direction of bias was not clear. However within each level, bias behaved as expected. Specifically, for both levels, low generating abilities were overestimated, whereas high generating abilities were underestimated. That is, ability estimates at both levels were pulled toward their respective means, which was to be expected of Bayes estimates. 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 8 (Chapter IV), bias appeared greater when the ICC was lower, but this was largely because the -2 to 2 range included values that were relatively rarer for the lower ICC levels. Thus, overall MH-RM recovered the vast majority of ability estimates quite well at both levels.

Root mean squared error (RMSE) showed a pattern similar to bias for each of the parameters. This was not surprising, since RMSE combines bias and sampling variability (i.e., the empirical standard errors or average standard deviations of the estimates across replications). 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).

As 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).

Confidence interval coverage was assessed to determine the accuracy of the analytical standard errors produced by MH-RM in flexMIRT. It is important to note that the interpretation of confidence interval coverage 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 negative, especially in the five-dimensional models. Because of this, the error variances (and standard errors) from those replications were treated as missing in the construction of confidence intervals by which coverage rates were evaluated. Therefore, one should interpret the coverage rates for the latent variances and covariances with great caution, since they were based on fewer replications, and even 100 replications might be considered too few in the evaluation of standard errors.

That said, the results revealed that, on average, standard errors were fairly accurate for item discrimination, Level 2 variances and covariances, and Level 1 ability estimates; a little too small for Level 2 ability estimates; and extremely small for item difficulty and Level 1 covariances/correlations. Since the standard errors for item discrimination, Level 2 variances and covariances, and Level 1 ability estimates were essentially accurate, I do not consider them further. What is more interesting are the standard errors for item difficulty, Level 1 covariances, and Level 2 ability estimates.

The standard errors for ability estimates behaved as one might expect. Specifically, standard errors for Level 2 abilities near 0 or a little below 0 were the most accurate, whereas for abilities near the extremes, standard errors tended to be too small. The same pattern was observed for the standard errors of Level 1 ability estimates, except that here standard errors for abilities near 0 or slightly above zero were the most accurate, whereas standard errors of very low or very high abilities were too small. Recall that 95% coverage rates indicate the standard errors are accurate in the absence of bias. When bias is present, coverage rates may be too small, even when the analytical standard errors are very close to the empirical standard errors. So the asymmetry described above may well be due to bias.

The notion of bias also helps explain in part the extremely low confidence interval coverage for item difficulty. Recall that, on average, there was a sizeable bias in item difficulties. When bias is present, the 95% confidence intervals on which the coverage rates are based are sometimes constructed around biased estimates, which makes it less likely for the generating value to fall within the confidence intervals. As a result, confidence interval coverage rates are too small.

However, the low confidence interval coverage was not merely a function of bias. Sample size also influenced confidence interval coverage for item difficulty, such that when sample size was large, the estimated standard errors were too small. As an aside, recall that sample size did not affect bias in the item difficulties. Clearly, the analytical standard errors of item difficulties are too small in large samples, and this is not simply a function of bias. Given no prior research on the standard errors specific to multilevel measurement models, this finding is extremely important. Confidence interval coverage

rates were also extremely small for the Level 1 covariances, and there was no large bias associated with this parameter. A far more likely explanation here is that there were a lot of negative error variances. As such, the confidence interval coverage rate computed here was based on a small number of observations. More replications may be needed to obtain a more accurate estimate of standard error accuracy for the Level 1 covariances/correlations. This leads us to the next section, which covers the limitations of this dissertation.

Limitations

Despite the extensive scope of this dissertation, 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 of the models item parameters were constrained to be of the same magnitude across levels. That is, a single item difficulty and item discrimination was estimated for each item for both levels of the measurement model. By applying this constraint on the model, 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. One can easily imagine Level 2 factors, such as school type (e.g., public vs. private), or school-level socioeconomic status (SES) among others, having an impact on the variability in student achievement and consequently ability estimates. For example, imagine that School X is a public, urban school, with half of its students qualifying for free or reduced lunch. One would expect a good amount of variability in achievement and ability estimates across students. Now imagine that School Y is a private school in the suburbs, where no students qualify for free or reduced lunch (i.e., high SES). Here, all students are high achievers, and as a result, there is little variability in ability estimates. Clearly, constraining the variability within schools to be the same would not be reflective of reality. A less likely but noteworthy argument is that regardless of school type and SES, there will always be variability in achievement across students within schools, and that the model constraint of setting the within-school variance to 1 is not farfetched. Although setting the Level 1 (within) variances to 1 was required for model identification, it is important to acknowledge what this constraint implies in practice.

Another limitation, though not particularly significant, was that item difficulties were not fully crossed with item discrimination values. Specifically, there were no combinations of very easy, yet highly discriminating items. This did not appear to have

any effect on the results. However, it made the interpretation of the quadratic effect of generating difficulty less straightforward because $d = -ab$ was not symmetrical around 0.

Finally, the results of the study were based on 100 replications for each condition. Although this number of replications may be sufficient for the examination of bias and RMSE, a much larger number of replications (e.g., 1000 or more) is desirable for the proper assessment of standard error accuracy. This is especially true for the latent variances and covariances whose error variances were often negative in the five-dimensional models. Many more replications are needed there.

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 dissertation. As mentioned above, MH-RM was able to estimate all 100 replications of all 24 conditions in a reasonable amount of time, especially with a more powerful computer. Given the specific conditions examined here and the fact that MH-RM has not been studied before with multilevel measurement models, the results I obtained cannot be compared directly to how MH-RM performed with different models (e.g., single-level exploratory or confirmatory IFA). However, a crude comparison of the results from this study and the research on the functionality of MH-RM compared to other estimation methods (see Chapter II) 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 standard errors can be trusted. No known study has examined the accuracy of ability estimates produced by MH-RM. Therefore, the current study is the first to shed some light in this area. In terms of bias, I 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 other hand, is that the standard errors for some parameters may be inaccurate. For example, the standard errors for item discrimination were found to be essentially accurate, whereas those for item difficulty were too small. Interestingly, Asparouhov and Muthén (2012) found the opposite: standard errors of item thresholds (i.e., difficulties) were more or less accurate, whereas the standard errors of item loadings (i.e., discriminations) were too small. Again, the examination of standard error accuracy in the current study was somewhat limited therefore this disagreement 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 multilevel

multidimensional IRT models. Next I discuss the potential benefits of such applications

in practice and provide possible paths for future research.

Implications for Practice

Being the first of its kind, the 3PL ML-MIRT model presented in this dissertation

has enormous potential for educational measurement practice. K-12 education

practitioners continuously demand more diagnostic feedback from assessments for

accountability to 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. A relevant example today is

the Common Core State Standards. Now that the "curse of dimensionality" has been

lifted by the MH-RM algorithm, multidimensional models can be easily applied in

practice. 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 to unidimensional models, which typically require

more items to achieve the same measurement precision. This is due to the borrowing of

information across dimensions, which is only possible with multidimensional models.

Given that achievement in one subject area is usually highly correlated with achievement

in other subject areas, the dimensions in a MIRT model need not be limited to

subdomains within the same subject area—one can model the response data from

multiple subtests (e.g., English Language Arts, mathematics, and science), not just the

subdomains within a single content area (e.g., mathematics).

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 date 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) level. 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. The sample sizes and combinations of number of clusters and cluster sizes examined in this study revealed that the model can be applied to a variety of sampling designs. For instance, one could model a large number of schools (e.g., 200) or a smaller number of classrooms (e.g., 40) with students nested within them (e.g., 100 within each school or 20 within each classroom, respectively). Furthermore, the models examined in this dissertation covered a fairly large range of ICC levels that one is likely to observe in educational data. When applying the 3PL ML-MIRT model in practice, one should always consider the level of dependency of observations within clusters. This is

important because if the ICC is too small (which is unlikely in educational data), the model may not converge due to lack of sufficient information to estimate the parameters of the model at both levels. Another reason why the ICC level is so important is that it can affect the parameters differently. For example, if Level 2 (e.g., school level) ability estimates are of primary interest, the results of this study showed that higher ICC level was associated with smaller bias, especially in small clusters (e.g., schools). However, recall that this effect was in part an artifact of the different school mean ranges at different ICC levels (e.g., a school mean of 1 is much more extreme when ICC = .15 than it is when ICC = .35). On the other hand, if the Level 2 (between) variances were of primary interest, the RMSE was higher for higher ICCs. Therefore, one should take into consideration the ICC level associated with each dimension and how it might impact the parameter estimates associated with that dimension.

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 dissertation 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 the education practitioners and policymakers 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

Although the MH-RM algorithm was developed fairly recently, researchers have already used it in several studies with real data. Importantly, very few studies have examined the performance of MH-RM with various models. Thus, there are many opportunities for further research on the functionality of MH-RM in general. In this dissertation, I specifically examined the accuracy and efficiency of MH-RM 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 I point to several directions for future research.

The research design of the current study could be enhanced in several ways. For example, the examination of the accuracy of standard errors for various parameters of the model presented here was limited. Specifically, many more replications are needed in order to obtain stable confidence interval coverage rates. This would be a great way to supplement the findings of the current study and expand the body of empirical support for the application of MH-RM with multilevel measurement models in practice. A recommendation for researchers who wish to replicate some form of the design employed in this dissertation is to fully cross the generating item difficulty and item discrimination values, so that all item types that one may encounter in practice are covered. Another 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 impacts the results in terms of bias, RMSE, and possibly standard errors 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 dissertation) 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 applied in practice.

Yet another possibility for future research is to examine the performance of MH-RM in the estimation of 3PL ML-MIRT models with complex-structure items. Recall that all items in this study had simple structure, meaning that the probability of correct

response to an item was a function of a single latent dimension (modeled at two levels),

as well as item parameters constrained to be the same across levels. A logical extension

of this framework is to include items that require combinations of skills or latent traits

(i.e., a compensatory model) or an exact set of multiple skills (i.e., noncompensatory

model).

 Finally, since MH-RM is now implemented in the "mirt" package in R (Chalmers,

2012), a future study could compare the estimation of a multilevel IRT, MIRT, or ML-

MIRT model (if possible) in the "mirt" package using MH-RM, and then compare the

results of the same model estimated in flexMIRT, again using MH-RM as the calibration

algorithm. Possible dependent variables of interest include convergence rate, bias,

RMSE, and standard error accuracy for item parameters, latent variances (and

covariances, where applicable), ability estimates (at different levels, where applicable), as

well as processing time.

Appendix A

**Regression analysis procedures and output from full models including all
main effects, two-way interactions, and three-way interactions**

Procedures

First, I built a full linear regression model including all main effects, two-way interactions, and three-way interactions. Then, I examined the statistical significance and effect size (semi-partial η^2) for each main effect or interaction. As a general rule, an effect had to be statistically significant, and, more importantly, explain at least 1% of the variance in the criterion, in order to be retained in the model.

To make interpretation of the significant predictors easier, nonsignificant predictors were removed in groups, starting with the nonsignificant three-way interactions, then the nonsignificant two-way interactions, and finally any nonsignificant main effects. Nonsignificant main effects and interactions were retained in the model in the presence of a significant higher-level interaction that explained at least 1% of the variance in the criterion. An exception to this cutoff (1% variance explained) was made for the main effect for the generating value of the "difficulty" parameter d, which was confounded by a (see Equation 1.3). More specifically, when the main effect of the generating value of a on the criterion was much larger than that of the generating value of d, then only the effect of a was considered in the interpretation of the results, since a is already in d (except when the generating $a = 1$). Thus, a higher minimum percentage of variance explained was used, in order to consider the main effect of generating d value meaningful.

The output from the full regression models including all main effects, two-way interactions, and three-way interactions is provided further below in Tables A1 through A21. The reduced models, in which nonsignificant predictors were removed using the procedure described above, are not presented, since they were used to simply identify the practically significant predictors to plot and interpret.

Multicollinearity was assessed by examining the difference between R^2 (i.e., the total percentage of variance explained in the criterion by the predictors) and the sum of the semi-partial eta squares (i.e., the sum of the unique contributions of the predictors, controlling for one another), where the semi-partial correlation was computed as the ratio of the Type III sum of squares to the corrected total sum of squares:

$$\eta^2 = \frac{SS}{SS_{Total}}.$$

Specifically, a relatively large positive difference between R^2 and the sum of the semi-partial correlations ($\Sigma\eta^2$) indicated multicollinearity (i.e., redundancy among the predictors). In other words, at least some predictors were highly correlated with one another; thus the sum of their unique contributions was noticeably smaller than their combined predictive power (R^2). However, sometimes $\Sigma\eta^2$ exceeded R^2. This phenomenon is known as "cooperative suppression" (Cohen & Cohen, 1975, pp. 90-91) and occurs when all predictors are positively correlated with the criterion, but some are

negatively correlated with one another. As a result, the semi-partial correlations of some predictors with the criterion exceed their zero-order counterparts.

Recall that the Level 2 variances (between) were generated such that the ICCs were set at desired values (see Tables 2 and 3 in Chapter III). Because of this, the ICC magnitude as a condition factor in the simulation study and the generating Level 2 (between) variance values were nearly perfectly correlated (r was not 1.0 due to rounding error). In other words, the two predictors were completely redundant with one another, which was reflected in the output as degrees of freedom = 0, effects of 0.00 and missing significance statistics for these effects. Thus, in order to obtain more meaningful results, out of the two predictors only the ICC was retained in the model as a predictor of bias, RMSE, and confidence interval coverage for the Level 2 variances. A similar problem occurred for the Level 2 (between) covariances in the regression model for confidence interval coverage. Here, the correlation of generating Level 2 (between) covariance values and the ICC was not too high ($r = .963$), but high enough to cause estimation issues. Once removed, the effects of all other predictors in the model could be estimated.

Descriptions of the predictors in Tables A1-A21

Predictor	Description
dim	number of dimensions
icc	intraclass correlation coefficient value
numclust	number of clusters
clustsize	cluster size
aval	generating item discrimination (a) value
dval	generating item difficulty (d) value
dvalsq	generating item difficulty (d) value squared (quadratic effect)
genval	generating value (variance, covariance/correlation)
roundt	generating θ value rounded to the second decimal

Table A1

Linear Regression of Item Difficulty Bias on Condition Factors, Generating d *Value, Generating* a *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.379	0.379	286.450	<.0001	0.102
icc	2	0.100	0.050	37.830	<.0001	0.027
numclust	1	0.108	0.108	81.420	<.0001	0.029
clustsize	1	0.148	0.148	111.650	<.0001	0.040
aval	2	0.918	0.459	346.690	<.0001	0.246
dval	1	0.139	0.139	105.380	<.0001	0.037
dvalsq	1	0.004	0.004	3.150	0.0764	0.001
dim*icc	2	0.003	0.002	1.270	0.2822	0.001
dim*numclust	1	0.000	0.000	0.030	0.8633	0.000
dim*clustsize	1	0.013	0.013	10.120	0.0015	0.004
dim*aval	2	0.020	0.010	7.370	0.0007	0.005
dval*dim	1	0.035	0.035	26.390	<.0001	0.009
icc*numclust	2	0.004	0.002	1.370	0.2553	0.001
icc*clustsize	2	0.183	0.091	69.040	<.0001	0.049
icc*aval	4	0.006	0.002	1.160	0.3274	0.002
dval*icc	2	0.001	0.001	0.410	0.6623	0.000
numclust*clustsize	1	0.022	0.022	16.490	<.0001	0.006
numclust*aval	2	0.017	0.008	6.410	0.0017	0.005
dval*numclust	1	0.000	0.000	0.370	0.5426	0.000
clustsize*aval	2	0.006	0.003	2.280	0.1026	0.002
dval*clustsize	1	0.010	0.010	7.370	0.0068	0.003
dval*aval	2	0.005	0.002	1.720	0.1789	0.001
dim*icc*numclust	2	0.004	0.002	1.650	0.1929	0.001
dim*icc*clustsize	2	0.003	0.001	0.980	0.3775	0.001
dim*icc*aval	4	0.001	0.000	0.160	0.9603	0.000
dval*dim*icc	2	0.003	0.002	1.220	0.2966	0.001
icc*numclust*clustsize	2	0.002	0.001	0.810	0.4468	0.001
icc*numclust*aval	4	0.000	0.000	0.020	0.9993	0.000
dval*icc*numclust	2	0.000	0.000	0.000	0.9971	0.000
numclust*clustsize*aval	2	0.008	0.004	3.190	0.0414	0.002
dval*numclust*clustsize	1	0.005	0.005	3.440	0.0640	0.001
dval*clustsize*aval	2	0.027	0.013	10.140	<.0001	0.007

Note. Corrected Total SS = 3.728; R^2 = .638; $\Sigma\eta^2$ = .583.

Table A2

Linear Regression of Item Discrimination Bias on Condition Factors, Generating d *Value, Generating* a *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.022	0.022	35.630	<.0001	0.027
icc	2	0.001	0.000	0.470	0.6278	0.001
numclust	1	0.025	0.025	41.340	<.0001	0.031
clustsize	1	0.031	0.031	51.150	<.0001	0.039
aval	2	0.008	0.004	6.560	0.0015	0.010
dval	1	0.000	0.000	0.000	0.9707	0.000
dvalsq	1	0.001	0.001	1.450	0.2283	0.001
dim*icc	2	0.000	0.000	0.150	0.8600	0.000
dim*numclust	1	0.000	0.000	0.710	0.4008	0.001
dim*clustsize	1	0.001	0.001	1.090	0.2974	0.001
dim*aval	2	0.013	0.006	10.580	<.0001	0.016
dval*dim	1	0.001	0.001	0.960	0.3281	0.001
icc*numclust	2	0.003	0.001	2.140	0.1187	0.003
icc*clustsize	2	0.001	0.001	1.150	0.3171	0.002
icc*aval	4	0.002	0.000	0.640	0.6307	0.002
dval*icc	2	0.001	0.001	1.010	0.3647	0.002
numclust*clustsize	1	0.007	0.007	10.850	0.0010	0.008
numclust*aval	2	0.010	0.005	8.370	0.0002	0.013
dval*numclust	1	0.002	0.002	3.710	0.0543	0.003
clustsize*aval	2	0.006	0.003	5.200	0.0056	0.008
dval*clustsize	1	0.002	0.002	2.590	0.1082	0.002
dval*aval	2	0.003	0.001	2.450	0.0866	0.004
dim*icc*numclust	2	0.001	0.000	0.440	0.6460	0.001
dim*icc*clustsize	2	0.000	0.000	0.370	0.6938	0.001
dim*icc*aval	4	0.001	0.000	0.240	0.9161	0.001
dval*dim*icc	2	0.000	0.000	0.250	0.7778	0.000
icc*numclust*clustsize	2	0.002	0.001	1.550	0.2132	0.002
icc*numclust*aval	4	0.000	0.000	0.190	0.9457	0.001
dval*icc*numclust	2	0.000	0.000	0.290	0.7502	0.000
numclust*clustsize*aval	2	0.008	0.004	6.360	0.0018	0.010
dval*numclust*clustsize	1	0.006	0.006	9.470	0.0021	0.007
dval*clustsize*aval	2	0.009	0.005	7.570	0.0005	0.011

Note. Corrected Total *SS* = 0.810; R^2 = .231; $\Sigma\eta^2$ = .206.

Table A3

Linear Regression of Level 2 (Between) Variance Bias on Condition Factors and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.001	0.001	10.830	0.0015	0.059
icc	2	0.000	0.000	4.060	0.0213	0.044
numclust	1	0.000	0.000	1.320	0.2534	0.007
clustsize	1	0.001	0.001	14.780	0.0003	0.080
dim*icc	2	0.001	0.000	5.460	0.0061	0.059
dim*numclust	1	0.000	0.000	2.360	0.1286	0.013
dim*clustsize	1	0.000	0.000	1.620	0.2074	0.009
icc*numclust	2	0.001	0.000	8.080	0.0007	0.088
icc*clustsize	2	0.001	0.000	7.390	0.0012	0.080
numclust*clustsize	1	0.000	0.000	0.000	0.9954	0.000
dim*icc*numclust	2	0.001	0.000	8.120	0.0006	0.088
dim*icc*clustsize	2	0.000	0.000	0.880	0.4187	0.010
icc*numclust*clustsize	2	0.001	0.000	7.230	0.0013	0.079

Note. Corrected Total SS = 0.010; R^2 = .592; $\Sigma\eta^2$ = .617.

Table A4

Linear Regression of Level 2 (Between) Correlation Bias on Condition Factors, Generating Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.000	0.000	0.000	0.9890	0.000
icc	2	0.000	0.000	4.360	0.0176	0.055
numclust	1	0.000	0.000	3.040	0.0870	0.019
clustsize	1	0.000	0.000	0.060	0.8098	0.000
genval	2	0.000	0.000	0.960	0.3885	0.012
dim*icc	2	0.001	0.000	9.710	0.0003	0.123
dim*numclust	1	0.000	0.000	0.130	0.7182	0.001
dim*clustsize	1	0.001	0.001	14.940	0.0003	0.095
dim*genval	2	0.000	0.000	3.080	0.0544	0.039
icc*numclust	2	0.000	0.000	2.360	0.1047	0.030
icc*clustsize	2	0.000	0.000	1.670	0.1989	0.021
icc*genval	4	0.000	0.000	0.710	0.5896	0.018
numclust*clustsize	1	0.000	0.000	0.120	0.7341	0.001
numclust*genval	2	0.000	0.000	0.160	0.8505	0.002
clustsize*genval	2	0.000	0.000	0.910	0.4074	0.012
dim*icc*numclust	2	0.001	0.000	9.440	0.0003	0.120
dim*icc*clustsize	2	0.000	0.000	2.740	0.0739	0.035
dim*icc*genval	4	0.000	0.000	0.250	0.9083	0.006
icc*numclust*clustsize	2	0.000	0.000	3.630	0.0333	0.046
icc*numclust*genval	4	0.000	0.000	1.190	0.3240	0.030
numclust*clustsize*genval	2	0.000	0.000	0.280	0.7548	0.004

Note. Corrected Total $SS = 0.005$; $R^2 = .664$; $\Sigma\eta^2 = .668$.

106

Table A5
Linear Regression of Level 1 (Within) Correlation Bias on Condition Factors, Generating Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.005	0.005	182.010	<.0001	0.326
icc	2	0.000	0.000	0.270	0.7660	0.001
numclust	1	0.000	0.000	2.880	0.0960	0.005
clustsize	1	0.000	0.000	0.050	0.8190	0.000
genval	2	0.008	0.004	142.140	<.0001	0.509
dim*icc	2	0.000	0.000	0.060	0.9410	0.000
dim*numclust	1	0.000	0.000	0.730	0.3980	0.001
dim*clustsize	1	0.000	0.000	0.190	0.6660	0.000
dim*genval	2	0.001	0.001	23.540	<.0001	0.084
icc*numclust	2	0.000	0.000	0.630	0.5370	0.002
icc*clustsize	2	0.000	0.000	0.430	0.6520	0.002
icc*genval	4	0.000	0.000	0.110	0.9770	0.001
numclust*clustsize	1	0.000	0.000	1.210	0.2770	0.002
numclust*genval	2	0.000	0.000	0.030	0.9700	0.000
clustsize*genval	2	0.000	0.000	0.010	0.9920	0.000
dim*icc*numclust	2	0.000	0.000	0.020	0.9830	0.000
dim*icc*clustsize	2	0.000	0.000	0.280	0.7570	0.001
dim*icc*genval	4	0.000	0.000	0.030	0.9990	0.000
icc*numclust*clustsize	2	0.000	0.000	0.560	0.5730	0.002
icc*numclust*genval	4	0.000	0.000	0.010	1.0000	0.000
numclust*clustsize*genval	2	0.000	0.000	0.090	0.9180	0.000

Note. Corrected Total $SS = 0.017$; $R^2 = .905$; $\Sigma\eta^2 = .938$.

Table A6
Linear Regression of Level 2 (Between) Ability Estimate Bias on Condition Factors, Rounded Generating θ Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	4.003	4.003	436.060	<.0001	0.004
icc	2	1.415	0.707	77.060	<.0001	0.001
numclust	1	0.937	0.937	102.030	<.0001	0.001
clustsize	1	6.150	6.150	669.970	<.0001	0.006
roundt	1	437.718	437.718	47680.400	<.0001	0.426
dim*icc	2	0.172	0.086	9.380	<.0001	0.000
dim*numclust	1	0.059	0.059	6.410	0.0113	0.000
dim*clustsize	1	0.339	0.339	36.970	<.0001	0.000
roundt*dim	1	2.987	2.987	325.340	<.0001	0.003
icc*numclust	2	0.149	0.075	8.120	0.0003	0.000
icc*clustsize	2	1.598	0.799	87.020	<.0001	0.002
roundt*icc	2	34.418	17.209	1874.580	<.0001	0.034
numclust*clustsize	1	0.095	0.095	10.390	0.0013	0.000
roundt*numclust	1	0.016	0.016	1.750	0.1855	0.000
roundt*clustsize	1	131.636	131.636	14339.100	<.0001	0.128
dim*icc*numclust	2	0.078	0.039	4.250	0.0143	0.000
dim*icc*clustsize	2	0.006	0.003	0.320	0.7287	0.000
roundt*dim*icc	2	0.564	0.282	30.740	<.0001	0.001
icc*numclust*clustsize	2	0.002	0.001	0.090	0.9151	0.000
roundt*icc*numclust	2	0.073	0.036	3.960	0.0190	0.000
roundt*numclust*clustsize	1	0.011	0.011	1.190	0.2753	0.000

Note. Corrected Total SS = 1026.960; R^2 = .701; $\Sigma\eta^2$ = .606.

Table A7
Linear Regression of Level 1 (Within) Ability Estimate Bias on Condition Factors, Rounded Generating θ
Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.324	0.324	5.850	0.0156	0.000
icc	2	0.079	0.040	0.720	0.4884	0.000
numclust	1	3.110	3.110	56.170	<.0001	0.000
clustsize	1	10.672	10.672	192.760	<.0001	0.000
roundt	1	46288.421	46288.421	836071.000	<.0001	0.806
dim*icc	2	0.003	0.002	0.030	0.9724	0.000
dim*numclust	1	0.183	0.183	3.310	0.0689	0.000
dim*clustsize	1	0.025	0.025	0.450	0.5029	0.000
roundt*dim	1	60.749	60.749	1097.270	<.0001	0.001
icc*numclust	2	0.052	0.026	0.470	0.6231	0.000
icc*clustsize	2	0.014	0.007	0.120	0.8846	0.000
roundt*icc	2	5.402	2.701	48.790	<.0001	0.000
numclust*clustsize	1	0.203	0.203	3.660	0.0558	0.000
roundt*numclust	1	56.675	56.675	1023.670	<.0001	0.001
roundt*clustsize	1	2.388	2.388	43.130	<.0001	0.000
dim*icc*numclust	2	0.101	0.051	0.910	0.4008	0.000
dim*icc*clustsize	2	0.274	0.137	2.480	0.0841	0.000
roundt*dim*icc	2	0.474	0.237	4.280	0.0139	0.000
icc*numclust*clustsize	2	0.187	0.093	1.680	0.1856	0.000
roundt*icc*numclust	2	0.010	0.005	0.090	0.9170	0.000
roundt*numclust*clustsize	1	0.000	0.000	0.000	0.9479	0.000

Note. Corrected Total SS = 57403.267; R^2 = .926; $\Sigma\eta^2$ = .809.

Table A8
Linear Regression of Item Difficulty RMSE on Condition Factors, Generating d *Value, Generating* a *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.128	0.128	97.960	<.0001	0.010
icc	2	0.056	0.028	21.390	<.0001	0.004
numclust	1	1.506	1.506	1153.780	<.0001	0.116
clustsize	1	0.170	0.170	129.870	<.0001	0.013
aval	2	1.092	0.546	418.350	<.0001	0.084
dval	1	0.029	0.029	22.330	<.0001	0.002
dvalsq	1	0.913	0.913	699.520	<.0001	0.070
dim*icc	2	0.003	0.001	1.140	0.3197	0.000
dim*numclust	1	0.006	0.006	4.240	0.0398	0.000
dim*clustsize	1	0.042	0.042	32.340	<.0001	0.003
dim*aval	2	0.006	0.003	2.200	0.1110	0.000
dval*dim	1	0.000	0.000	0.340	0.5613	0.000
icc*numclust	2	0.002	0.001	0.820	0.4412	0.000
icc*clustsize	2	0.092	0.046	35.200	<.0001	0.007
icc*aval	4	0.004	0.001	0.840	0.4977	0.000
dval*icc	2	0.006	0.003	2.190	0.1126	0.000
numclust*clustsize	1	0.253	0.253	193.590	<.0001	0.019
numclust*aval	2	0.069	0.035	26.470	<.0001	0.005
dval*numclust	1	0.187	0.187	143.100	<.0001	0.014
clustsize*aval	2	0.008	0.004	2.970	0.0520	0.001
dval*clustsize	1	0.063	0.063	48.220	<.0001	0.005
dval*aval	2	0.010	0.005	3.990	0.0189	0.001
dim*icc*numclust	2	0.002	0.001	0.760	0.4688	0.000
dim*icc*clustsize	2	0.000	0.000	0.130	0.8821	0.000
dim*icc*aval	4	0.001	0.000	0.140	0.9657	0.000
dval*dim*icc	2	0.007	0.003	2.670	0.0696	0.001
icc*numclust*clustsize	2	0.001	0.000	0.260	0.7741	0.000
icc*numclust*aval	4	0.001	0.000	0.130	0.9718	0.000
dval*icc*numclust	2	0.000	0.000	0.130	0.8758	0.000
numclust*clustsize*aval	2	0.003	0.002	1.250	0.2869	0.000
dval*numclust*clustsize	1	0.005	0.005	4.120	0.0425	0.000
dval*clustsize*aval	2	0.263	0.132	100.780	<.0001	0.020

Note. Corrected Total SS = 12.966; R^2 = .897; $\Sigma\eta^2$ = .380.

Table A9

Linear Regression of Item Discrimination RMSE on Condition Factors, Generating d *Value, Generating* a *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.033	0.033	160.920	<.0001	0.004
icc	2	0.007	0.004	17.670	<.0001	0.001
numclust	1	1.992	1.992	9826.560	<.0001	0.253
clustsize	1	1.613	1.613	7954.200	<.0001	0.205
aval	2	0.610	0.305	1504.050	<.0001	0.078
dval	1	0.127	0.127	625.870	<.0001	0.016
dvalsq	1	0.125	0.125	614.960	<.0001	0.016
dim*icc	2	0.000	0.000	0.370	0.6922	0.000
dim*numclust	1	0.000	0.000	0.860	0.3526	0.000
dim*clustsize	1	0.001	0.001	6.010	0.0144	0.000
dim*aval	2	0.010	0.005	25.140	<.0001	0.001
dval*dim	1	0.000	0.000	0.100	0.7503	0.000
icc*numclust	2	0.001	0.000	1.250	0.2878	0.000
icc*clustsize	2	0.000	0.000	0.740	0.4788	0.000
icc*aval	4	0.001	0.000	1.820	0.1233	0.000
dval*icc	2	0.002	0.001	4.430	0.0121	0.000
numclust*clustsize	1	0.230	0.230	1133.410	<.0001	0.029
numclust*aval	2	0.074	0.037	182.440	<.0001	0.009
dval*numclust	1	0.037	0.037	184.640	<.0001	0.005
clustsize*aval	2	0.074	0.037	182.930	<.0001	0.009
dval*clustsize	1	0.013	0.013	64.420	<.0001	0.002
dval*aval	2	0.003	0.001	7.170	0.0008	0.000
dim*icc*numclust	2	0.000	0.000	0.180	0.8368	0.000
dim*icc*clustsize	2	0.000	0.000	0.140	0.8693	0.000
dim*icc*aval	4	0.000	0.000	0.160	0.9607	0.000
dval*dim*icc	2	0.002	0.001	4.490	0.0115	0.000
icc*numclust*clustsize	2	0.000	0.000	0.560	0.5705	0.000
icc*numclust*aval	4	0.001	0.000	1.420	0.2241	0.000
dval*icc*numclust	2	0.000	0.000	0.300	0.7408	0.000
numclust*clustsize*aval	2	0.003	0.002	8.480	0.0002	0.000
dval*numclust*clustsize	1	0.001	0.001	6.600	0.0104	0.000
dval*clustsize*aval	2	0.006	0.003	15.550	<.0001	0.001

Note. Corrected Total *SS* = 7.864; R^2 = .974; $\Sigma\eta^2$ = .632.

Table A10

Linear Regression of Level 2 (Between) Variance Root Mean Squared Error (RMSE) on Condition Factors, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.000	0.000	0.010	0.9210	0.000
icc	2	0.055	0.028	832.320	<.0001	0.399
numclust	1	0.057	0.057	1712.890	<.0001	0.410
clustsize	1	0.005	0.005	148.660	<.0001	0.036
dim*icc	2	0.000	0.000	3.670	0.0301	0.002
dim*numclust	1	0.000	0.000	0.240	0.6284	0.000
dim*clustsize	1	0.001	0.001	16.800	0.0001	0.004
icc*numclust	2	0.006	0.003	88.570	<.0001	0.042
icc*clustsize	2	0.000	0.000	2.950	0.0583	0.001
numclust*clustsize	1	0.002	0.002	50.530	<.0001	0.012
dim*icc*numclust	2	0.000	0.000	1.010	0.3706	0.000
dim*icc*clustsize	2	0.000	0.000	0.010	0.9890	0.000
icc*numclust*clustsize	2	0.000	0.000	0.390	0.6787	0.000

Note. Corrected Total $SS = 0.138$; $R^2 = .982$; $\Sigma\eta^2 = .907$.

Table A11

Linear Regression of Level 2 (Between) Correlation Root Mean Squared Error (RMSE) on Condition Factors, Generating Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.001	0.001	15.030	0.0003	0.005
icc	2	0.004	0.002	55.780	<.0001	0.036
numclust	1	0.039	0.039	1094.670	<.0001	0.350
clustsize	1	0.007	0.007	185.030	<.0001	0.059
genval	2	0.027	0.014	378.680	<.0001	0.242
dim*icc	2	0.000	0.000	1.480	0.2369	0.001
dim*numclust	1	0.000	0.000	5.400	0.0240	0.002
dim*clustsize	1	0.000	0.000	13.540	0.0005	0.004
dim*genval	2	0.000	0.000	0.300	0.7418	0.000
icc*numclust	2	0.002	0.001	23.570	<.0001	0.015
icc*clustsize	2	0.002	0.001	31.990	<.0001	0.020
icc*genval	4	0.000	0.000	2.570	0.0483	0.003
numclust*clustsize	1	0.002	0.002	64.740	<.0001	0.021
numclust*genval	2	0.005	0.003	73.970	<.0001	0.047
clustsize*genval	2	0.000	0.000	5.680	0.0058	0.004
dim*icc*numclust	2	0.000	0.000	0.230	0.7991	0.000
dim*icc*clustsize	2	0.000	0.000	1.330	0.2740	0.001
dim*icc*genval	4	0.000	0.000	0.280	0.8899	0.000
icc*numclust*clustsize	2	0.000	0.000	5.240	0.0084	0.003
icc*numclust*genval	4	0.000	0.000	3.110	0.0226	0.004
numclust*clustsize*genval	2	0.000	0.000	4.440	0.0164	0.003

Note. Corrected Total $SS = 0.112$; $R^2 = .983$; $\Sigma\eta^2 = .759$.

Table A12
Linear Regression of Level 1 (Within) Correlation Root Mean Squared Error (RMSE) on Condition Factors, Generating Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.005	0.005	273.070	<.0001	0.366
icc	2	0.000	0.000	0.500	0.6118	0.001
numclust	1	0.001	0.001	72.080	<.0001	0.097
clustsize	1	0.002	0.002	102.560	<.0001	0.138
genval	2	0.002	0.001	45.570	<.0001	0.122
dim*icc	2	0.000	0.000	0.020	0.9821	0.000
dim*numclust	1	0.000	0.000	0.050	0.8170	0.000
dim*clustsize	1	0.000	0.000	0.050	0.8269	0.000
dim*genval	2	0.001	0.001	31.680	<.0001	0.085
icc*numclust	2	0.000	0.000	0.350	0.7032	0.001
icc*clustsize	2	0.000	0.000	0.150	0.8571	0.000
icc*genval	4	0.000	0.000	0.140	0.9677	0.001
numclust*clustsize	1	0.000	0.000	17.060	0.000	0.023
numclust*genval	2	0.000	0.000	13.080	<.0001	0.035
clustsize*genval	2	0.001	0.000	15.050	<.0001	0.040
dim*icc*numclust	2	0.000	0.000	0.020	0.9849	0.000
dim*icc*clustsize	2	0.000	0.000	0.140	0.8689	0.000
dim*icc*genval	4	0.000	0.000	0.010	0.9996	0.000
icc*numclust*clustsize	2	0.000	0.000	0.720	0.4894	0.002
icc*numclust*genval	4	0.000	0.000	0.050	0.9948	0.000
numclust*clustsize*genval	2	0.000	0.000	1.220	0.3028	0.003

Note. Corrected Total $SS = 0.014$; $R^2 = .929$; $\Sigma\eta^2 = .915$.

Table A13
Linear Regression of Level 2 (Between) Ability Estimate Root Mean Squared Error (RMSE) on Condition Factors, Rounded Generating θ *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.409	0.409	56.380	<.0001	0.001
icc	2	0.471	0.236	32.520	<.0001	0.001
numclust	1	0.147	0.147	20.300	<.0001	0.000
clustsize	1	134.454	134.454	18551.100	<.0001	0.314
roundt	1	19.013	19.013	2623.280	<.0001	0.044
dim*icc	2	0.157	0.078	10.810	<.0001	0.000
dim*numclust	1	0.037	0.037	5.130	0.0235	0.000
dim*clustsize	1	1.168	1.168	161.220	<.0001	0.003
roundt*dim	1	0.107	0.107	14.740	0.0001	0.000
icc*numclust	2	0.044	0.022	3.040	0.0479	0.000
icc*clustsize	2	0.124	0.062	8.520	0.0002	0.000
roundt*icc	2	0.933	0.466	64.360	<.0001	0.002
numclust*clustsize	1	0.235	0.235	32.490	<.0001	0.001
roundt*numclust	1	0.233	0.233	32.080	<.0001	0.001
roundt*clustsize	1	0.074	0.074	10.280	0.0013	0.000
dim*icc*numclust	2	0.002	0.001	0.140	0.8711	0.000
dim*icc*clustsize	2	0.021	0.010	1.420	0.2414	0.000
roundt*dim*icc	2	0.033	0.017	2.280	0.1022	0.000
icc*numclust*clustsize	2	0.017	0.008	1.140	0.3183	0.000
roundt*icc*numclust	2	0.024	0.012	1.690	0.1846	0.000
roundt*numclust*clustsize	1	0.066	0.066	9.040	0.0026	0.000

Note. Corrected Total SS = 428.561; R^2 = .434; $\Sigma\eta^2$ = .368.

Table A14
Linear Regression of Level 1 (Within) Ability Estimate Root Mean Squared Error (RMSE) on Condition Factors, Rounded Generating θ Value, and Interactions

Predictor	df	Type III SS	Mean Sq	F	p	η^2
dim	1	53.788	53.788	263.030	<.0001	0.003
icc	2	10.388	5.194	25.400	<.0001	0.001
numclust	1	158.034	158.034	772.800	<.0001	0.010
clustsize	1	56.915	56.915	278.320	<.0001	0.003
roundt	1	558.164	558.164	2729.470	<.0001	0.034
dim*icc	2	0.448	0.224	1.100	0.3345	0.000
dim*numclust	1	0.184	0.184	0.900	0.3426	0.000
dim*clustsize	1	0.051	0.051	0.250	0.6184	0.000
roundt*dim	1	0.331	0.331	1.620	0.2031	0.000
icc*numclust	2	0.313	0.156	0.760	0.4655	0.000
icc*clustsize	2	0.082	0.041	0.200	0.8191	0.000
roundt*icc	2	0.047	0.023	0.110	0.8915	0.000
numclust*clustsize	1	0.080	0.080	0.390	0.5326	0.000
roundt*numclust	1	1.268	1.268	6.200	0.0128	0.000
roundt*clustsize	1	8.264	8.264	40.410	<.0001	0.000
dim*icc*numclust	2	0.004	0.002	0.010	0.9903	0.000
dim*icc*clustsize	2	0.236	0.118	0.580	0.5616	0.000
roundt*dim*icc	2	0.465	0.232	1.140	0.3209	0.000
icc*numclust*clustsize	2	0.206	0.103	0.500	0.6048	0.000
roundt*icc*numclust	2	0.943	0.471	2.300	0.0998	0.000
roundt*numclust*clustsize	1	0.353	0.353	1.730	0.1889	0.000

Note. Corrected Total SS = 16541.478; R^2 = .058; $\Sigma\eta^2$ = .051.

Table A15
Linear Regression of Item Difficulty Confidence interval coverage on Condition Factors, Generating d
Value, Generating a *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	2.430	2.430	151.670	<.0001	0.023
icc	2	0.115	0.058	3.590	0.0279	0.001
numclust	1	15.988	15.988	997.880	<.0001	0.154
clustsize	1	40.774	40.774	2544.940	<.0001	0.392
aval	2	3.368	1.684	105.100	<.0001	0.032
dval	1	0.870	0.870	54.290	<.0001	0.008
dvalsq	1	5.592	5.592	349.010	<.0001	0.054
dim*icc	2	0.008	0.004	0.250	0.7762	0.000
dim*numclust	1	0.004	0.004	0.270	0.6017	0.000
dim*clustsize	1	0.059	0.059	3.670	0.0556	0.001
dim*aval	2	0.009	0.005	0.290	0.7455	0.000
dval*dim	1	0.001	0.001	0.050	0.8260	0.000
icc*numclust	2	0.071	0.035	2.210	0.1103	0.001
icc*clustsize	2	1.780	0.890	55.560	<.0001	0.017
icc*aval	4	0.007	0.002	0.110	0.9807	0.000
dval*icc	2	0.021	0.011	0.670	0.5144	0.000
numclust*clustsize	1	1.285	1.285	80.200	<.0001	0.012
numclust*aval	2	0.024	0.012	0.750	0.4739	0.000
dval*numclust	1	0.652	0.652	40.720	<.0001	0.006
clustsize*aval	2	0.092	0.046	2.870	0.0571	0.001
dval*clustsize	1	0.365	0.365	22.800	<.0001	0.004
dval*aval	2	0.560	0.280	17.490	<.0001	0.005
dim*icc*numclust	2	0.022	0.011	0.700	0.4963	0.000
dim*icc*clustsize	2	0.014	0.007	0.450	0.6400	0.000
dim*icc*aval	4	0.015	0.004	0.230	0.9210	0.000
dval*dim*icc	2	0.009	0.005	0.290	0.7481	0.000
icc*numclust*clustsize	2	0.231	0.115	7.200	0.0008	0.002
icc*numclust*aval	4	0.023	0.006	0.360	0.8402	0.000
dval*icc*numclust	2	0.017	0.008	0.520	0.5922	0.000
numclust*clustsize*aval	2	0.490	0.245	15.280	<.0001	0.005
dval*numclust*clustsize	1	0.078	0.078	4.850	0.0278	0.001
dval*clustsize*aval	2	0.892	0.446	27.830	<.0001	0.009

Note. Corrected Total SS = 103.962; R^2 = .843; $\Sigma\eta^2$ = .730.

Table A16
Linear Regression of Item Discrimination Confidence interval coverage on Condition Factors, Generating d *Value, Generating* a *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.000	0.000	0.050	0.8237	0.000
icc	2	0.003	0.001	0.940	0.3913	0.002
numclust	1	0.023	0.023	14.740	0.0001	0.013
clustsize	1	0.037	0.037	23.010	<.0001	0.020
aval	2	0.036	0.018	11.280	<.0001	0.019
dval	1	0.001	0.001	0.390	0.5300	0.000
dvalsq	1	0.006	0.006	3.720	0.0540	0.003
dim*icc	2	0.000	0.000	0.080	0.9269	0.000
dim*numclust	1	0.027	0.027	16.910	<.0001	0.014
dim*clustsize	1	0.017	0.017	10.510	0.0012	0.009
dim*aval	2	0.016	0.008	5.100	0.0063	0.009
dval*dim	1	0.009	0.009	5.380	0.0205	0.005
icc*numclust	2	0.003	0.001	0.910	0.4033	0.002
icc*clustsize	2	0.001	0.001	0.410	0.6641	0.001
icc*aval	4	0.001	0.000	0.150	0.9615	0.001
dval*icc	2	0.001	0.000	0.280	0.7559	0.000
numclust*clustsize	1	0.006	0.006	3.650	0.0565	0.003
numclust*aval	2	0.027	0.014	8.560	0.0002	0.015
dval*numclust	1	0.002	0.002	1.450	0.2286	0.001
clustsize*aval	2	0.012	0.006	3.780	0.0232	0.006
dval*clustsize	1	0.001	0.001	0.740	0.3911	0.001
dval*aval	2	0.003	0.002	0.970	0.3785	0.002
dim*icc*numclust	2	0.000	0.000	0.090	0.9156	0.000
dim*icc*clustsize	2	0.002	0.001	0.670	0.5108	0.001
dim*icc*aval	4	0.009	0.002	1.410	0.2282	0.005
dval*dim*icc	2	0.003	0.002	0.980	0.3760	0.002
icc*numclust*clustsize	2	0.001	0.000	0.310	0.7302	0.001
icc*numclust*aval	4	0.003	0.001	0.420	0.7934	0.001
dval*icc*numclust	2	0.001	0.001	0.390	0.6756	0.001
numclust*clustsize*aval	2	0.001	0.001	0.460	0.6311	0.001
dval*numclust*clustsize	1	0.000	0.000	0.030	0.8592	0.000
dval*clustsize*aval	2	0.001	0.000	0.250	0.7776	0.000

Note. Corrected Total $SS = 1.857$; $R^2 = .127$; $\Sigma\eta^2 = .136$.

Table A17
Linear Regression of Level 2 (Between) Variance Confidence interval coverage on Condition Factors and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.000	0.000	0.000	0.9988	0.000
icc	2	0.002	0.001	1.730	0.1847	0.015
numclust	1	0.001	0.001	1.700	0.1957	0.007
clustsize	1	0.019	0.019	33.460	<.0001	0.145
dim*icc	2	0.009	0.005	8.010	0.0007	0.070
dim*numclust	1	0.006	0.006	10.890	0.0015	0.047
dim*clustsize	1	0.010	0.010	17.870	<.0001	0.078
icc*numclust	2	0.003	0.001	2.370	0.1001	0.021
icc*clustsize	2	0.000	0.000	0.230	0.7924	0.002
numclust*clustsize	1	0.004	0.004	6.880	0.0106	0.030
dim*icc*numclust	2	0.009	0.004	7.520	0.0011	0.065
dim*icc*clustsize	2	0.007	0.004	6.200	0.0032	0.054
icc*numclust*clustsize	2	0.000	0.000	0.100	0.9074	0.001

Note. Corrected Total SS = 0.134; R^2 = .674; $\Sigma\eta^2$ = .535.

Table A18

Linear Regression of Level 2 (Between) Covariance Confidence Interval Coverage on Condition Factors and Interactions

Predictor	*df*	Type III SS	Mean Square	*F*	*p*	η^2
dim	1	0.000	0.000	0.450	0.5042	0.001
icc	2	0.001	0.001	1.220	0.2985	0.004
numclust	1	0.001	0.001	2.750	0.0996	0.005
clustsize	1	0.023	0.023	53.220	<.0001	0.093
dim*icc	2	0.011	0.006	13.200	<.0001	0.046
dim*numclust	1	0.010	0.010	22.760	<.0001	0.040
dim*clustsize	1	0.018	0.018	41.400	<.0001	0.073
icc*numclust	2	0.006	0.003	7.080	0.0012	0.025
icc*clustsize	2	0.000	0.000	0.040	0.9598	0.000
numclust*clustsize	1	0.012	0.012	28.560	<.0001	0.050
dim*icc*numclust	2	0.011	0.006	13.010	<.0001	0.046
dim*icc*clustsize	2	0.009	0.004	10.110	<.0001	0.035
icc*numclust*clustsize	2	0.002	0.001	1.860	0.1593	0.007

Note. Corrected Total $SS = 0.246$; $R^2 = .763$; $\Sigma\eta^2 = .419$.

Table A19

Linear Regression of Level 1 (Within) Covariance Confidence Interval Coverage on Condition Factors, Generating Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.506	0.506	65.330	<.0001	0.035
icc	2	0.011	0.005	0.700	0.4964	0.001
numclust	1	0.789	0.789	101.920	<.0001	0.054
clustsize	1	0.627	0.627	81.000	<.0001	0.043
genval	2	6.392	3.196	412.700	<.0001	0.439
dim*icc	2	0.018	0.009	1.170	0.3144	0.001
dim*numclust	1	0.006	0.006	0.820	0.3673	0.000
dim*clustsize	1	0.008	0.008	1.040	0.3091	0.001
dim*genval	2	0.139	0.069	8.950	0.0002	0.010
icc*numclust	2	0.007	0.004	0.480	0.6211	0.001
icc*clustsize	2	0.011	0.006	0.720	0.4888	0.001
icc*genval	4	0.001	0.000	0.040	0.9965	0.000
numclust*clustsize	1	0.000	0.000	0.010	0.9401	0.000
numclust*genval	2	0.369	0.184	23.820	<.0001	0.025
clustsize*genval	2	0.309	0.155	19.960	<.0001	0.021
dim*icc*numclust	2	0.007	0.004	0.470	0.6266	0.000
dim*icc*clustsize	2	0.013	0.006	0.810	0.4478	0.001
dim*icc*genval	4	0.002	0.001	0.070	0.9914	0.000
icc*numclust*clustsize	2	0.003	0.002	0.200	0.8159	0.000
icc*numclust*genval	4	0.005	0.001	0.170	0.9541	0.000
numclust*clustsize*genval	2	0.193	0.097	12.480	<.0001	0.013

Note. Corrected Total SS = 14.549; R^2 = .940; $\Sigma\eta^2$ = .647.

Table A20
Linear Regression of Level 2 (Between) Ability Estimate Confidence Interval Coverage on Condition Factors, Rounded Generating θ Value, and Interactions

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.003	0.003	0.080	0.7748	0.000
icc	2	11.259	5.630	170.660	<.0001	0.009
numclust	1	1.945	1.945	58.950	<.0001	0.002
clustsize	1	0.146	0.146	4.420	0.0355	0.000
roundt	1	107.958	107.958	3272.750	<.0001	0.085
dim*icc	2	2.896	1.448	43.890	<.0001	0.002
dim*numclust	1	0.454	0.454	13.760	0.0002	0.000
dim*clustsize	1	7.895	7.895	239.350	<.0001	0.006
roundt*dim	1	0.796	0.796	24.120	<.0001	0.001
icc*numclust	2	0.412	0.206	6.240	0.0020	0.000
icc*clustsize	2	0.252	0.126	3.820	0.0219	0.000
roundt*icc	2	10.489	5.245	158.990	<.0001	0.008
numclust*clustsize	1	0.591	0.591	17.900	<.0001	0.000
roundt*numclust	1	1.763	1.763	53.460	<.0001	0.001
roundt*clustsize	1	3.352	3.352	101.610	<.0001	0.003
dim*icc*numclust	2	0.243	0.121	3.680	0.0253	0.000
dim*icc*clustsize	2	0.296	0.148	4.480	0.0113	0.000
roundt*dim*icc	2	0.068	0.034	1.030	0.3583	0.000
icc*numclust*clustsize	2	0.118	0.059	1.780	0.1682	0.000
roundt*icc*numclust	2	0.357	0.178	5.400	0.0045	0.000
roundt*numclust*clustsize	1	0.033	0.033	0.990	0.3202	0.000

Note. Corrected Total SS = 1263.688; R^2 = .126; $\Sigma\eta^2$ = .120.

Table A21

Linear Regression of Level 1 (Within) Ability Estimate Confidence Interval Coverage on Condition Factors, Rounded Generating θ *Value, and Interactions*

Predictor	df	Type III SS	Mean Square	F	p	η^2
dim	1	0.542	0.542	5.670	0.0173	0.000
icc	2	0.597	0.298	3.120	0.0442	0.000
numclust	1	57.331	57.331	599.080	<.0001	0.008
clustsize	1	45.433	45.433	474.760	<.0001	0.006
roundt	1	156.012	156.012	1630.250	<.0001	0.021
dim*icc	2	0.334	0.167	1.740	0.1749	0.000
dim*numclust	1	0.082	0.082	0.860	0.3545	0.000
dim*clustsize	1	0.075	0.075	0.790	0.3753	0.000
roundt*dim	1	0.038	0.038	0.400	0.5269	0.000
icc*numclust	2	0.086	0.043	0.450	0.6366	0.000
icc*clustsize	2	0.049	0.025	0.260	0.7727	0.000
roundt*icc	2	0.046	0.023	0.240	0.7874	0.000
numclust*clustsize	1	0.003	0.003	0.030	0.8651	0.000
roundt*numclust	1	0.221	0.221	2.310	0.1288	0.000
roundt*clustsize	1	0.732	0.732	7.650	0.0057	0.000
dim*icc*numclust	2	0.029	0.015	0.150	0.8572	0.000
dim*icc*clustsize	2	0.105	0.052	0.550	0.5787	0.000
roundt*dim*icc	2	0.269	0.135	1.410	0.2446	0.000
icc*numclust*clustsize	2	0.135	0.067	0.700	0.4946	0.000
roundt*icc*numclust	2	0.490	0.245	2.560	0.0774	0.000
roundt*numclust*clustsize	1	0.537	0.537	5.610	0.0178	0.000

Note. Corrected Total SS = 7584.077; R^2 = .038; $\Sigma\eta^2$ = .035.

Appendix B

Table B1
Mean Bias, Root Mean Squared Error (RMSE), and Confidence Interval Coverage for Item Parameters, Latent Variances and Covariances, and Ability Estimates

Parameter	Bias (*SD*)	RMSE (*SD*)	Confidence interval coverage (*SD*)
Item difficulty	0.128 (0.059)	0.193 (0.110)	0.560 (0.117)
Item discrimination	0.013 (0.027)	0.123 (0.085)	0.930 (0.016)
Level 2 (between) variance	0.005 (0.010)	0.068 (0.038)	0.925 (0.004)
Level 2 (between) covariance/correlation	0.003 (0.008)	0.056 (0.036)	0.927 (0.005)
Level 1 (within) covariance/correlation	-0.020 (0.018)	0.030 (0.015)	0.563 (0.285)
Level 2 (between) ability estimate	-0.074 (0.026)	0.207 (0.056)	0.887 (0.043)
Level 1 (within) ability estimate	-0.008 (0.005)	0.495 (0.041)	0.950 (0.005)

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67-91.

Ackerman, T. A. (1994). Using multidimensional item response theory to understand what items and tests are measuring. *Applied Measurement in Education*, *7*, 255-278.

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.

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of clusters is low: A comparison of different statistical software procedures. *The International Journal of Biostatistics*, *6*, *Article 16*.

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

Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker, *New developments and techniques in structural equation modeling* (pp. 269-296). Mahwah, NJ: Erlbaum.

Bellman, R. (2003). *Dynamic programming*. Mineola, NY: Dover Publications.

Birnbaum, A. (1968). Some latent trait models. Chapter 17 in F. Lord & M. Novick's (Eds.) *Statistical Theories of Mental Test Scores* (pp. 397-424). Reading, MA: Addison, Wesley.

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.

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

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 – 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., & du Toit, S. (2011). IRTPRO for Windows [Computer software]. Chicago, IL: Scientific Software International.

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., & Flora, D. B. (2014). Maximum-likelihood estimation of noncompensatory IRT models with the MH-RM algorithm. *Applied Psychological Measurement*, *38*, 339-358.

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, *62*, 752-758.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression/correlation analyses for the behavioral sciences.* New York, NY: Wiley.

Cohen, J., Cohen, P, West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (3rd Ed.).* Mahwah, NJ: Lawrence Erlbaum Associates.

De Boeck, P., & Wilson, M. (2004). A framework for item response models. Chapter 1 in P. De Boeck & M. Wilson's (Eds.) *Explanatory Item Response Models* (pp. 3-41). New York, NY: Springer.

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.

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *The Annals of Statistics*, *27*, 94-128.

DeMars, C. E. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement*, *43*, 145-168.

DeMars, C. E. (2012). A comparison of limited-information and full-information methods in M*plus* for estimating item response theory parameters for nonnormal

populations. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 610-632.

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *159*, 505-513.

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation*, *35*(2), 57-63.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60-87.

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.

Jeon, M., Rijmen, F., & Rabe-Hesketh, S. (2013). Modeling differential item functioning using a generalization of the multiple-group bifactor model. *Journal of Educational and Behavioral Statistics*, *38*, 32-60.

Jurich, J. P., & DeMars, C. E. (2013, April). *Confirmatory factor analysis with dichotomous data: Does unmodeled guessing affect fit and parameter recovery?*

Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

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.

Kamata, A., & Vaughn, B. K. (2011). Multilevel IRT modeling. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 41-58). New York, NY: Routledge.

Keller, L. A. (2005). Markov chain Monte Carlo item response theory estimation. In B. S. Everitt & D. C. Howell (Eds.) *Encyclopedia of Statistics in Behavioral Science. Volume 3* (pp. 1143-1148). Chichester, UK: John Wiley & Sons Ltd.

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Newbury Park, CA: Sage Publications.

Lesaffre, E., & Spiessens, B. (2001). On the effect of the number of quadrature points in a logistic random effects model: an example. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *50*, 325-335.

Liu, Q., & Pierce, D. A. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, *81*, 624-629.

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

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2×2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological methods*, *16*, 444-467.

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*(3), 86-92.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*, 99-114.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equation modeling. *Psychological Methods*, *10*, 259-284.

Meng, X.-L., & Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *Journal of the American Statistical Association*, *91*, 1254-1267.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7*, *Article 34*.

Monroe, S., & Cai, L. (2014). Estimation of a Ramsay-curve item response theory model by the Metropolis-Hastings Robbins-Monro algorithm. *Educational and Psychological Measurement*, *74*, 343-369.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, *28*, 338-354.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods and Research*, *22*, 376-398.

Muthén, B. O. (1997). Latent variable modeling of longitudinal and multilevel data. In A. E. Raftery (Ed.), *Sociological Methodology 1997* (pp. 453-480). Cambridge, MA: Blackwell.

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

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. In P. V. Marsden (Ed.), *Sociological methodology 1995* (pp. 216-316). Boston, MA: Blackwell Publishers.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 41-58). New York, NY: Routledge.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for Mathematics.* Retrieved from http://www.corestandards.org/wp-content/uploads/Math_Standards.pdf

Naylor, J. C., & Smith, A. F. (1982). Applications of a method for the efficient computation of posterior distributions. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *31*, 214-225.

Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: An illustration. *Applied Measurement in Education*, *16*, 223-243.

Patz, R. J., & Junker, B. W. (1999). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics*, *24*, 146-178.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, *2*(1), 1-21.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modelling. *Psychometrika, 69,* 167–190.

Raudenbush, S., & Bryk, A. (2002). *Hierarchical linear models* (2nd ed.). Newbury Park, CA: Sage.

Reckase, M. D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York, NY: Springer.

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

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *158*, 73-89.

Schilling, S., & Bock, R. D. (2005). High-dimensional maximum marginal likelihood item factor analysis by adaptive quadrature. *Psychometrika*, *70*, 533-555.

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*.

    Volume 3, (pp. 1570-1573). Chichester, UK: John Wiley & Sons Ltd.

Snijders, T. A. B., & Bosker, R. J. (2010). *Multilevel analysis: An introduction to basic*

    and advanced multilevel modeling (2nd ed.). London, UK: Sage Publications.

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.

Stapleton, L. M. (2013). Multilevel structural equation modeling with complex sample

    data. In G. R. Hancock & R. O. Mueller (Eds.), *A second course in structural*

    equation modeling (pp. 521-562). Greenwich, CT: Information Age.

Stapleton, L. M., & Thomas, S. L. (2008). The use of national datasets for teaching and

    research. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of*

    educational data (pp. 11-57). Charlotte, NC: Information Age Publishing.

Thissen, D. (2013). Using the testlet response model as a shortcut to multidimensional

    item response theory subscore computation. In R. E. Millsap, L. A. van der Ark,

    D. M. Bolt, & C. M. Woods (Eds.), *New Developments in Quantitative*

    Psychology (pp. 29-40). New York, NY: Springer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its*

    applications. Cambridge University Press.

Wainer, H., Vevea, J. L., Camacho, F., Reeve III, B. B., Rosa, K., Nelson, L., Swygert,

    K. A., & Thissen, D. (2001). Augmented scores—"Borrowing strength" to

compute scores based on small numbers of items. In D. Thissen & H. Wainer

(Eds.), *Test scoring* (pp. 343-388). Mahwah, NJ: Erlbaum.

Walker, C. M., & Beretvas, S. N. (2001). An empirical investigation demonstrating the

multidimensional DIF paradigm: A cognitive explanation for DIF. *Journal of

Educational Measurement*, *38*, 147-163.

Wiley, E. W., Shavelson, R. J., & Kurpius, A. A. (2014). On the factor structure of the

SAT and implications for next-generation college readiness assessments.

Educational and Psychological Measurement, *74*, 859-874.

Wright, N. A. (2013). *New study, old question: Using multidimensional item response

theory to examine the construct validity of situational judgment tests.*

Unpublished doctoral dissertation, Department of Psychology, University of

North Carolina – Raleigh.

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.

Yen, W. M. (1981). Using simulation results to choose a latent trait model. *Applied

Psychological Measurement*, *5*, 245-262.

Zhang, J. (2012). Calibration of response data using MIRT models with simple and

mixed structures. *Applied Psychological Measurement*, *36*, 375-398.

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level

measurement and structural invariance in the analysis of multilevel data:

Problems and solutions. *Group Dynamics: Theory, Research, Practice*, *12*, 127-

140.