CRESST REPORT 833

ESTIMATION OF CONTEXTUAL EFFECTS THROUGH NONLINEAR MULTILEVEL LATENT VARIABLE MODELING WITH A METROPOLIS-HASTINGS ROBBINS-MONRO ALGORITHM

SEPTEMBER, 2013

Ji Seung Yang

Li Cai

National Center for Research
on Evaluation, Standards, & Student Testing
UCLA | Graduate School of Education & Information Studies

Estimation of Contextual Effects through Nonlinear Multilevel Latent Variable Modeling with a Metropolis-Hastings Robbins-Monro Algorithm

CRESST Report 833

Ji Seung Yang and Li Cai
University of California, Los Angeles

September 2013

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by Institute of Education Sciences (R305D100039), the National Institute on Drug Abuse (R01DA026943 and R01DA030466), and Society of Multivariate Experimental Psychology Dissertation Support Awards.

The findings and opinions expressed here do not necessarily reflect the positions or policies of the Institute of Education Sciences, the National Institute on Drug Abuse, or the Society of Multivariate Experimental Psychology Dissertation Support Awards.

To cite from this report, please use the following as your APA reference: Yang, J.S., & Cai, L. (2013). *Estimation of Contextual Effects through Nonlinear Multilevel Latent Variable Modeling with a Metropolis-Hastings Robbins-Monro Algorithm* (CRESST Report 833). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

ESTIMATION OF CONTEXTUAL EFFECTS THROUGH NONLINEAR MULTILEVEL LATENT VARIABLE MODELING WITH A METROPOLIS-HASTINGS ROBBINS-MONRO ALGORITHM

Ji Seung Yang and Li Cai
University of California, Los Angeles

Abstract

The main purpose of this study is to improve estimation efficiency in obtaining full-information maximum likelihood (FIML) estimates of contextual effects in the framework of a nonlinear multilevel latent variable model by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). Results indicate that the MH-RM algorithm can produce FIML estimates and their standard errors efficiently, and the efficiency of MH-RM was more prominent for a cross-level interaction model, which requires five dimensional integration. Simulations, with various sampling and measurement structure conditions, were conducted to obtain information about the performance of nonlinear multilevel latent variable modeling compared to traditional hierarchical linear modeling. Results suggest that nonlinear multilevel latent variable modeling can more properly estimate and detect a contextual effect and a cross-level interaction than the traditional approach. As empirical illustrations, two subsets of data extracted from The Programme for International Student Assessment (PISA, 2000; OECD, 2000) were analyzed.

Introduction

In educational research, a contextual effect is traditionally defined as the difference between two coefficients in a hierarchical linear model (HLM) analysis framework (Raudenbush & Bryk, 1986; Willms, 1986; Lee & Bryk, 1989; Raudenbush & Willms, 1995): one from the individual-level and the other coefficient from the school-level. A representative application of this kind of contextual effect in education is discussed in Raudenbush and Bryk (2002) using a subset of High School and Beyond Data (HS&B). In this example, individual math achievement is regressed on individual-level socioeconomic status (SES) and school-level math achievement is regressed on aggregated school-level SES using multilevel modeling. The result shows that the two coefficient estimates are not the same, indicating two students who have the same SES level are expected to have different levels of math achievement depending on to which school a student belongs. Statistically significant difference between these two coefficients represents a significant compositional effect.

While hierarchical linear modeling opened the door to defining and estimating contextual effects, there have been two unresolved methodological issues. The first one is related to the attenuated coefficient estimates due to measurement error in predictors (Spearman, 1904), and

the other is biased parameter estimates due to sampling error associated with aggregating level-1 variables to form level-2 variables by simply averaging the values (Raudenbush & Bryk, 2002, chap.3). Accordingly, two regression coefficients at level-1 and level-2 tend to be attenuated when summed or averaged scores are used as predictors.

To handle measurement error and sampling error more properly, multilevel latent variable modeling has been suggested as an alternative to traditional methods (e.g. Lüdtke et al., 2008; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Marsh et al., 2009). For example, Lüdtke et al. (2008) proposed a multilevel latent variable modeling framework for contextual analysis. Lüdtke et al. (2008) examined the relative bias in contextual effect estimates when the traditional HLM is used under different data conditions. The results showed that the relative percentage bias of contextual effect was less than 10% across varying data conditions when a multilevel latent variable model was used. On the other hand, the relative percentage bias of contextual effect was up to 80% when the traditional HLM model was used. However, the traditional HLM can yield less than 10% relative bias under favorable data conditions—that is, when level-1 and level-2 units exceed 30 and 500, respectively, and when there is substantial intra-class correlation (ICC) in the predictor (e.g., 0.3). While the manifest variables are limited to only continuous variables in Lüdtke et al. (2008), multiple categorical variables are used as manifest variables for both latent predictor and outcome variables in the current study.

Another study using multilevel latent variable modeling for contextual effect analysis was conducted by Marsh et al. (2009). Marsh and colleagues examined and compared several contextual modeling options related to ”big fish-little-pond effect (BFLPE)” estimates using an empirical data set in which academic achievement and self-concept were measured by three and four continuous manifest variables, respectively. Among the tested models, a multilevel latent variable model that takes both measurement and sampling error into account yielded the largest BFLPE estimate. The authors described this model as a doubly latent variable contextual model. Such a model is theoretically the most desirable choice for researchers, since the model tries to take both measurement and sampling error into account by utilizing information from the manifest variables, rather than using summed or averaged scores of those manifest variables. Again, Marsh et al.’s (2009) study was limited, using three continuous manifest variables.

While nonlinear multilevel latent variable modeling can deal with measurement and sampling error properly, this approach presents significant computational difficulties with categorical manifest variables. Standard approaches such as numerical integration (e.g., adaptive quadrature) or Markov chain Monte Carlo (MCMC; e.g., Gibbs Sampling) based estimation methods have important limitations that make them less practical for routine use, because their computational efficiency drops dramatically when the dimensionality is high. Lüdtke et al.

(2011) also reported the occurrence of unstable estimates. The model has difficulty converging when sample size is small and the intraclass correlation coefficient (ICC) in a predictor is small. Therefore, further research efforts are needed to improve estimation of contextual effect in the nonlinear multilevel latent variable modeling framework.

The main objective of this study was to develop a more efficient estimation method for contextual effects in the nonlinear multilevel latent variable modeling framework, by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). Computational efficiency and parameter recovery were assessed in a comparison with an existing EM algorithm using adaptive Gauss-Hermite quadrature for numerical integration (e.g., Mplus; Muthén & Muthén, 2008). Another objective was to find, through a simulation study, how much measurement error and sampling error can influence contextual effect estimates under different conditions. The results provide the rationale for using computationally demanding nonlinear multilevel latent variable models. The last objective of the proposed study was to provide an empirical illustration of estimating contextual effects by applying nonlinear multilevel latent variable models to real data that contain more complex measurement structures and unbalanced data. Subsets from The Programme for International Student Assessment (PISA; Adams & Wu, 2002) were analyzed to illustrate a contextual effect model and a cross-level interaction model.

The particular contextual effect of interest in this study is one that occurs when a group-level characteristic of interest is measured by individual-level characteristics, and the individual-level characteristics are measured by categorical manifest variables. This study considers a contextual effect not only as a compositional effect that captures the influence of contextual variables on individual level outcomes, but also cross-level interactions that capture the influence of contextual variables on within-group slopes.

Contextual Effects in a Nonlinear Multilevel Latent Variable Model

Structural Models

The traditional HLM defines a compositional effect β_c as follows:

$$\begin{aligned} Y_{ij} &= \beta_{0j} + \beta_{1j}(X_{ij} - X_{.j}) + r_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}(X_{.j} - X_{..}) + u_{0j}, \\ \beta_{1j} &= \gamma_{10}, \\ \gamma_{10} &= \beta_w, \\ \gamma_{01} &= \beta_b, \\ \beta_c &= \gamma_{01} - \gamma_{10} \end{aligned} \tag{1}$$

In Equation (1), Y_{ij} and X_{ij} denote outcome and predictor values of student i in school j, respectively. Y_{ij} and X_{ij} are typically constructed by summing item scores on self-report responses. The random effects r_{ij} and u_{0j} are assumed to be normally distributed with zero means and variances (σ^2 and τ). In this particular definition of a contextual effect as a *compositional effect*, the within-slope, γ_{10}, is the same across groups as a fixed effect, which may or may not be appropriate, depending on the context.

In a nonlinear multilevel latent variable model, instead of using Y_{ij} and X_{ij} that are observed variables, we substitute them with latent variables η_{ij} and ξ_{ij} for individual i in group j. Those latent variables are connected to manifest variables through measurement models. For notational simplicity, latent individual deviations from latent group means ($\xi_{ij} - \xi_{.j}$) can be defined as δ_{ij}, and group mean deviations from the latent grand mean ($\xi_{.j} - \xi_{..}$) can be defined as $\delta_{.j}$. Then Equation (1) translates into the following compositional effect model:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}\delta_{ij} + r_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\delta_{.j} + u_{0j}, \\ \beta_{1j} &= \gamma_{10}, \\ \gamma_{10} &= \beta_w, \\ \gamma_{01} &= \beta_b, \\ \beta_c &= \gamma_{01} - \gamma_{10} \end{aligned} \tag{2}$$

β_c is the compositional effect of this research interest. Similar to Equation (1), the random effects r_{ij} and u_{0j} are assumed to be normally distributed with zero means and variances σ^2 and τ_{00}, respectively.

Now consider a contextual effect as a *cross-level interaction*. The grand-mean-centered contextual variable ($\xi_{.j}$) is included in Equation (2) as a predictor for β_{1j}. Therefore, β_{1j} is re-defined as follows:

$$\begin{aligned} \eta_{ij} &= \beta_{0j} + \beta_{1j}\delta_{ij} + r_{ij}, \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\delta_{.j} + u_{0j}, \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}\delta_{.j} + u_{1j}, \end{aligned} \tag{3}$$

In Equation (3), γ_{11} is the parameter of research interest, which is the regression coefficient for the cross-level interaction term between level-1 and level-2 predictors.

Measurement Models

The measurement models define the relationship between observed (manifest) variables and latent variables. For simplicity, only the measurement models of level-1 latent predictor variable ξ_{ij} will be described in this section, since the measurement models for other variables such as the latent outcome η_{ij} follow the same principles.

When manifest variables are graded response variables with multiple categories, Samejima's (1969) model can be utilized. Let $x_{ijl} \in \{0, 1, 2, ..., K_l - 1\}$ be an element of ith individual's response in jth group to lth item that has K_l ordered categories. Then the logistic conditional cumulative response probability for each category is listed as follows:

$$\begin{aligned} P_\theta(x_{ijl} \geq 0|\xi_{ij}) &= 1, \\ P_\theta(x_{ijl} \geq 1|\xi_{ij}) &= \frac{1}{1+\exp[-(b_{1,l} + a_l\xi_{ij})]}, \\ P_\theta(x_{ijl} \geq 2|\xi_{ij}) &= \frac{1}{1+\exp[-(b_{2,l} + a_l\xi_{ij})]}, \\ &\vdots \\ P_\theta(x_{ijl} \geq K_l - 1|\xi_{ij}) &= \frac{1}{1+\exp[-(b_{K_l-1,l} + a_l\xi_{ij})]}, \end{aligned} \tag{4}$$

The category response probability is defined as the difference between two adjacent cumulative probabilities:

$$P_\theta\,(x_{ijl} = k|\xi_{ij}) = P_\theta\,(x_{ijl} \geq k|\xi_{ij}) - P_\theta\,(x_{ijl} \geq k+1|\xi_{ij}), \tag{5}$$

where $P_\theta\,(x_{ijl} \geq K_l\,|\xi_{ij})$ is zero. χ_k is an indicator function in which χ_k is 1 if $x_{ijl} = k$, or 0 otherwise. The conditional density for x_{ijl} follows a multinomial with trial size 1 in K_l categories:

$$f_\theta(x_{ijl}\,|\xi_{ij}) = \prod_{k=0}^{K_l-1} P_\theta(x_{ijl} = k|\xi_{ij})^{\chi_k(x_{ijl})}. \tag{6}$$

The observed and complete data likelihoods of are suggested in Appendix A.

Metropolis-Hastings Robbins-Monro Algorithm for Contextual Models

An MH-RM algorithm was initially proposed by Cai (2008) for nonlinear latent structure analysis with a comprehensive measurement model, and the application of algorithm has been expanded to further measurement and statistical models (e.g., Cai, 2010a, 2010b). The MH-RM algorithm was motivated by Fisher's Identity (Fisher, 1925), which proved that the gradient of the observed likelihood is the expectation of the gradient of the complete likelihood. While maximizing the observed likelihood, denoted as $L(\theta|\mathbf{Y}_o)$, involves high-dimensional integrals, the complete data likelihood, denoted as $L(\theta|\mathbf{Y})$, involves a series of products of likelihoods that are fairly simple to maximize. Therefore, having plausible values of random effects and latent variables makes the estimation problem simpler. This also allows straightforward optimization of the complete data likelihood with respect to θ. However, proper imputation requires the distribution of the missing data to be conditional on the observed data. As the model is nonlinear, analytical derivation of the distribution of missing data conditional on the observed data is difficult. Nevertheless, a property of the posterior of the missing data enables us to have appropriate imputation. That is, the posterior of missing data, given observed data and a provisional θ, is proportional to the complete data likelihood. To utilize this property, Metropolis-Hastings sampler (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) is adopted to produce the imputations from a Markov chain with the missing data posterior as the target. Then, the random imputations are combined into Stochastic Approximation using the Robbins-Monro algorithm (RM; Robbins & Monro, 1951).

The $(k + 1)$th iteration of the MH-RM algorithm consists of 3 steps: Stochastic Imputation, Stochastic Approximation, and Robbins-Monro Update.

Step 1. Stochastic Imputation

Draw m_k sets of missing data, which are the random effects and latent variables, from a Markov chain that has the distribution of missing data conditional on observed data as the target. Then, m_k sets of complete data are as follows:

$$\left\{ \mathbf{Y}_j^{k+1}; j = 1, ..., m_k \right\} \tag{7}$$

Step 2. Stochastic Approximation

Using Fishier's Identity, a Monte Carlo approximation to $\nabla_\theta l(\boldsymbol{\theta}^k | \mathbf{Y}_o)$ can be computed as the sample average of complete data gradients. We also compute a recursive approximation of the conditional expectation of the information matrix of the complete data log-likelihood. For simplicity, let $\mathbf{s}(\boldsymbol{\theta}|\mathbf{Y})$ stand for $\nabla_\theta l(\boldsymbol{\theta}|\mathbf{Y})$, and the sample average of complete data gradients can be written as:

$$\tilde{\mathbf{s}}_{k+1} = \frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{s}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}), \tag{8}$$

and $\boldsymbol{\Gamma}_{k+1}$ is

$$\boldsymbol{\Gamma}_{k+1} = \boldsymbol{\Gamma}_k + \gamma_k \left[\frac{1}{m_k} \sum_{j=1}^{m_k} \mathbf{H}(\boldsymbol{\theta}^k | \mathbf{Y}_j^{k+1}) - \boldsymbol{\Gamma}_k \right], \tag{9}$$

where $\mathbf{H}(\boldsymbol{\theta}|\mathbf{Y})$ is the complete data information matrix, which is $^{-}1$ times the second derivative matrix of the complete data log-likelihood. The first and second order derivatives of the complete data models are suggested in Appendix B.

Step 3. Robbins-Monro Update

Now new parameters are estimated through the following update:

$$\boldsymbol{\theta}^{k+1} = \boldsymbol{\theta}^k + \gamma_k (\boldsymbol{\Gamma}_{k+1}^{-1} \tilde{\mathbf{s}}_{k+1}) \tag{10}$$

The whole iteration process is composed of three stages: initial stage in which parameters are not updated (M1), constant gain stage in which parameters are updated with a constant gain (M2), and the decreasing gain stage in which parameters are updated with a decreasing constant gain so that they stop oscillating around MLE (M3). The iterations can be stopped upon convergence when the changes in parameter estimates are sufficiently small. Cai (2008) verified the asymptotic behaviors of MH-RM in time and that it converges to MLE. For further details about the algorithm itself, readers can refer to Cai (2008, 2010a, 2010b).

Approximation to the Observed Information Matrix

One of the benefits of using the MH-RM algorithm is that the observed data information matrix can be recursively approximated as a byproduct of the iterations. The inverse of the observed data information matrix becomes the large-sample covariance matrix of parameter estimates. The square root of the diagonal elements are the standard errors. Another practical option for approximating the observed information matrix is a direct application of Louis's (1982) approach, in which the score vector and the conditional expectation are approximated directly after they converge. In this study, the first method is called *recursively approximated standard errors* and the latter is called *post-convergence approximated standard errors*.

Simulation Studies

Simulation Study 1: Comparison of Estimation Algorithms

The first simulation study was to examine the parameter recovery and standard errors when an MH-RM algorithm is implemented in comparison to those from an existing EM algorithm.

Methods. The data-generating and fitted models followed Equation (2) for a compositional effect model and Equation (3) for a cross-level interaction model. The simulated data are balanced in that the number of level-2 units (*ng*) is 100 and the number of level-1 units per group (*np*) is 20. The generating ICC value for the latent predictor was 0.3.

For the measurement model, five dichotomously scored manifest variables were generated for each latent trait (i.e., η, and ξ) using a 2-PL model. The item parameters were the same across levels, representing cross-level measurement invariance.

100 data sets were generated with the same parameters but with 100 different random seeds for each model. The first 10 data sets were analyzed using two methods: an MH-RM algorithm implemented in R (R Core Team, 2012) and an adaptive quadrature EM approach implemented in Mplus (Muthén & Muthén, 2010). Then the other 90 data sets are all analyzed using the MH-RM algorithm.

For compositional effect estimation, the MH-RM algorithm's convergence criterion was 5.0×10^{-5}, and the maximum numbers of iterations for each stage were $M1 = 100$, $M2 = 500$, and $M3 = 600$. For the cross-level interaction model, the MH-RM algorithm convergence criterion was 5.0×10^{-5} and the maximum numbers of iterations for each stage were M1 = 100, $M2 = 800$, and $M3 = 800$. To calculated post-convergence approximated standard errors, 100 to 500 samples were used for the compositional effect model, and 100 to 800 samples were used for the cross-level interaction model. The convergence rates at the given number of iterations were

100% and 52% for the compositional effect model and the cross-level interaction model, respectively.

Results: Compositional effect model. The generating values and the corresponding estimates for the compositional effect model from different algorithms are summarized in Table 1. The first column contains the true parameters for the measurement and structural parameters. The second set of columns and the third set of columns include the estimates and SEs from EM with different numbers of adaptive quadrature points (5 and 14). The means of point estimates and standard errors from different algorithms are generally very close to one another. For structural parameter estimates in the first panel, the number of quadrature points does not appear to make a large difference, though 14-quadrature-point estimates are slightly closer to the MH-RM estimates and the generating values in terms of τ_{00} and $var(\xi_{.j})$. For measurement parameter estimates, both the means of point estimates and the standard errors were the same up to the second decimal point across different numbers of quadrature points.

Table 1

Generating values and estimates for a compositional effect model (N=2,000, ng=100, np=20, 10/10 converged)

		EM (5qp)		EM (14qp)		MH-RM	
	θ	E(θˆ)	E{se(θˆ)}	E(θˆ)	E{se(θˆ)}	E(θˆ)	E{se(θˆ)}
Structural parameters							
$\gamma 01$	1.00	1.02	0.19	1.01	0.19	1.00	0.18
$\gamma 10$	0.50	0.52	0.05	0.51	0.05	0.52	0.09
$\tau 00$	1.00	0.90	0.16	0.91	0.17	0.93	0.16
$var(\xi_{.j})$	0.43	0.40	0.07	0.42	0.07	0.42	0.07
Measurement parameters							
a_{x1}	0.80	0.79	0.07	0.79	0.07	0.79	0.08
a_{x2}	1.00	1.01	0.08	1.01	0.08	1.00	0.09
a_{x3}	1.20	1.24	0.09	1.24	0.09	1.24	0.11
a_{x4}	1.40	1.39	0.10	1.39	0.10	1.39	0.12
a_{x5}	1.60	1.67	0.14	1.67	0.14	1.69	0.15
a_{y1}	0.80	0.78	0.06	0.78	0.06	0.78	0.06
a_{y2}	1.00	1.00	0.07	1.00	0.07	1.00	0.07
a_{y3}	1.20	1.23	0.09	1.23	0.09	1.23	0.08
a_{y4}	1.40	1.40	0.11	1.40	0.11	1.40	0.10
a_{y5}	1.60	1.61	0.13	1.61	0.13	1.60	0.12
c_{x1}	-0.80	-0.75	0.08	-0.75	0.08	-0.75	0.06
c_{x2}	0.00	0.02	0.08	0.02	0.08	0.02	0.05
c_{x3}	1.20	1.30	0.11	1.30	0.11	1.29	0.08
c_{x4}	-0.70	-0.61	0.11	-0.61	0.11	-0.62	0.07
c_{x5}	0.80	0.92	0.14	0.92	0.14	0.92	0.08

$c_{y}1$	-0.80	-0.80	0.11	-0.80	0.11	-0.81	0.06
$c_{y}2$	0.00	0.01	0.13	0.01	0.13	0.00	0.05
$c_{y}3$	1.20	1.19	0.16	1.19	0.16	1.18	0.08
$c_{y}4$	-0.70	-0.74	0.18	-0.74	0.18	-0.75	0.07
$c_{y}5$	0.80	0.79	0.21	0.79	0.21	0.78	0.08

Efficiency

When 1 processor is used	5~7 min	60~100 min	35~40 min

Note. θ = Generating values; E(θˆ) = mean of point estimates; E{se(θˆ)} = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameters; c = item threshold parameters; qp = number of quadrature points used in estimation.

In contrast, mean standard error estimates are slightly different between MH-RM and EM results in that the standard error estimates from MH-RM algorithm for intercepts are smaller than those from the EM algorithm. The log of standard error estimates from the EM algorithm and log of post-convergence approximated standard errors from the MH-RM algorithm are plotted against log standard deviations of point estimates in Figure 1. The estimates are clustered on the diagonal line, indicating that estimated standard errors are generally close to the Monte Carlo standard deviations of the point estimates, except for the intercept parameter standard errors, which appear to be underestimated when the post-convergence approximation is used for the MH-RM algorithm.

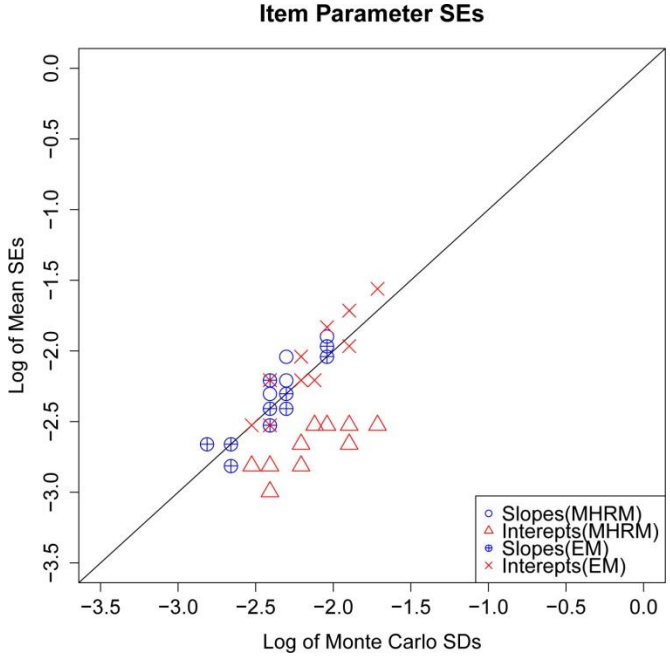

Figure 1. Comparisons of standard errors for item parameters.

When one processor was used for estimation, the 5 quadrature point EM required a very short time, while the 14 quadrature point EM required over an hour. The MH-RM algorithm

required about 40 minutes. Given that the MH-RM is implemented in R (an interpreted language) and Mplus is written in FORTRAN (a compiled language), the estimation time can be even more substantially shortened if the MH-RM is implemented with a compiled language.

To examine the performance of the MH-RM algorithm further, 100 generated data sets were analyzed, and the results are summarized in Table 2. The means of point estimates are reasonably close to generating values in general, with slight underestimation of variance estimates in the structural parameters. For structural parameters, the Monte Carlo standard deviations of parameter estimates (column 5) are also similar to both standard error estimates (column 4 and 6) as expected; the largest difference is 0.02.

Table 2

Generating values and estimates for a compositional effect model (N=2,000, ng=100, np=20)

	θ	$E(\hat{\theta})$	$E\{se1(\hat{\theta})\}$	$SD(\hat{\theta})$	$E\{se2(\hat{\theta})\}$	95% Coverage using se1
Structural parameters						
$\gamma 01$	1.00	0.99	0.17	0.19	0.18	95.0
$\gamma 10$	0.50	0.50	0.06	0.07	0.09	95.0
$\tau 00$	1.00	0.97	0.20	0.18	0.16	89.0
$var(\xi_{.j})$	0.43	0.43	0.08	0.09	0.07	89.0
Measurement parameters						
a_{x1}	0.80	0.80	0.07	0.06	0.07	98.0
a_{x2}	1.00	1.01	0.10	0.09	0.09	91.0
a_{x3}	1.20	1.22	0.12	0.10	0.11	92.0
a_{x4}	1.40	1.40	0.12	0.10	0.13	84.0
a_{x5}	1.60	1.60	0.15	0.13	0.15	73.0
a_{y1}	0.80	0.80	0.07	0.07	0.06	95.0
a_{y2}	1.00	1.01	0.07	0.07	0.07	94.0
a_{y3}	1.20	1.21	0.10	0.09	0.09	86.0
a_{y4}	1.40	1.39	0.10	0.09	0.10	89.0
a_{y5}	1.60	1.61	0.10	0.13	0.13	74.0
c_{x1}	0.80	0.80	0.14	0.08	0.06	94.0
c_{x2}	0.00	0.00	0.07	0.09	0.05	95.0
c_{x3}	-1.20	-1.22	0.09	0.12	0.08	91.0
c_{x4}	0.70	0.69	0.12	0.11	0.07	89.0
c_{x5}	-0.80	-0.80	0.12	0.15	0.08	89.0
c_{y1}	0.80	0.81	0.08	0.09	0.06	87.0
c_{y2}	0.00	0.01	0.11	0.11	0.06	78.0
c_{y3}	-1.20	-1.20	0.13	0.13	0.08	75.0
c_{y4}	0.70	0.71	0.15	0.15	0.07	62.0
c_{y5}	-0.80	-0.79	0.14	0.18	0.08	59.0
Efficiency						
		35~40min		90~120min		

Note. θ = Generating values; E($\hat{\theta}$) = mean of point estimates; E{se1($\hat{\theta}$)} = mean of recursively approximated standard error estimates (67 converged replications); E{se2($\hat{\theta}$)} = mean of post-convergence approximated standard errors; SD($\hat{\theta}$) = Standard deviation of point estimates; 95% Coverage using se1: Percentage coverage rate of generating value using post-convergence approximated standard errors; a = item slope parameters; c = item threshold parameters.

Recursively approximated standard errors are closer to the Monte Carlo standard deviations of item parameter estimates than the post-convergence approximated standard errors. More specifically, the most prominent differences are found in the standard errors of intercept parameters in that post-convergence approximated standard errors for item intercept parameters are underestimated.

Results: Cross-level interaction model. The generating values and the corresponding estimates from analyzing the first simulated data set using different algorithms are summarized in Table 3. Unlike the composition effect model results, the number of quadrature points for the EM algorithm makes some noticeable differences in the mean point estimates as well as the standard errors.

Table 3

Generating values and estimates for a cross-level interaction model (N=2,000, ng=100, np=20, 1st simulated data set)

		EM (5qp)		EM (8qp)		MH-RM	
	θ	E(θˆ)	E{se(θˆ)}	E(θˆ)	E{se(θˆ)}	E(θˆ)	E{se(θˆ)}
Structural parameters							
γ01	1.00	1.86	0.25	1.35	0.22	1.44	0.22
γ10	0.50	1.94	0.15	0.63	0.13	0.63	0.05
γ11	0.50	1.27	0.45	0.83	0.29	0.83	0.06
τ00	1.00	0.85	0.11	0.88	0.12	0.90	0.18
τ11	1.00	0.78	0.33	0.83	0.25	0.79	0.16
τ01	0.50	0.96	0.15	0.49	0.12	0.49	0.11
var(ξ.j)	0.43	0.40	0.02	0.39	0.05	0.39	0.07

		EM (5qp)		EM (8qp)		MH-RM	
	θ	E(θˆ)	E{se(θˆ)}	E(θˆ)	E{se(θˆ)}	E(θˆ)	E{se(θˆ)}
Measurement parameters							
a_{x1}	0.80	0.78	–	0.78	–	0.78	0.08
a_{x2}	1.00	1.40	0.14	0.96	0.14	0.96	0.07
a_{x3}	1.20	2.05	0.19	1.41	0.19	1.41	0.12
a_{x4}	1.40	2.37	0.21	1.62	0.21	1.63	0.18
a_{x5}	1.60	2.51	0.24	1.69	0.25	1.71	0.12
a_{y1}	0.80	0.79	0.00	0.79	0.00	0.79	0.05
a_{y2}	1.00	0.95	0.11	0.93	0.11	0.93	0.06
a_{y3}	1.20	1.17	0.11	1.15	0.12	1.16	0.07
a_{y4}	1.40	1.00	0.14	0.98	0.15	1.22	0.08
a_{y5}	1.60	1.43	0.18	1.40	0.19	1.51	0.09
c_{x1}	-0.80	-0.68	0.06	-0.73	0.07	-0.74	0.05
c_{x2}	0.00	0.10	0.08	0.10	0.08	0.09	0.05
c_{x3}	1.20	1.43	0.11	1.43	0.12	1.41	0.09
c_{x4}	-0.70	-0.52	0.11	-0.51	0.12	-0.53	0.08
c_{x5}	0.80	1.11	0.13	1.10	0.14	1.09	0.08
c_{y1}	-0.80	-0.72	0.09	-0.73	0.11	-0.73	0.06
c_{y2}	0.00	0.03	0.11	0.04	0.13	0.03	0.06
c_{y3}	1.20	1.26	0.14	1.26	0.16	1.26	0.08
c_{y4}	-0.70	-0.53	0.14	-0.52	0.16	-0.52	0.07
c_{y5}	0.80	0.96	0.17	0.96	0.20	0.96	0.08
Efficiency							
8 processors		15 min		100 min		60min	
1 processor		40 min		4hour 40 min			

Note. θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E\{se(\hat{\theta})\}$ = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameter; c = item threshold parameter; qp = number of quadrature points used in estimation. Mplus does not allow standardized factor identification option; therefore, anchoring the first factor loading option was used to estimate the model and the results are transformed to make the estimate comparable. The differences are particularly prominent in the structural parameters and the slopes of predictor-side indicators, as within-level variance estimates of the predictor were different across the number of quadrature points being used. However, the results from MH-RM algorithm are closer to the 8-quadrature-points results, indicating that reducing the number of quadrature points for a higher dimensional model is not desirable.

Efficiency of the MH-RM algorithm compared to the EM algorithm was more prominent for this cross-level interaction model, even as it is still in R. Using Mplus, even with 8 processors, the estimation took more than 1 hour and 30 minutes, while it took similar or even shorter time for the MH-RM algorithm implemented in R. When 1 processor was used, it took about 4 to 5 hours to yield a result using Mplus. This difference is remarkable considering that R does not have support for multi-processors.

For further analysis, more simulated data sets were analyzed by applying the MH-RM algorithm, and the generating values and corresponding estimates are summarized in Table 4. The largest relative bias of the parameter estimates for both measurement and structural parts is less than 10%. Means of standard error estimates and Monte Carlo standard deviations of point estimates are reasonably compatible; however, underestimation of standard errors for threshold estimates was consistent, indicating that the post-convergence approximation approach can be chosen for efficiency reasons, but with a cost in accuracy.

Table 4

Generating values and estimates for a cross-level interaction model using MH-RM algorithm (N=2,000, ng=100, np=20, 26/50 converged)

	θ	$E(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$SD(\hat{\theta})$
Structural parameters				
γ_{01}	1.00	1.07	0.18	0.21
γ_{10}	0.50	0.55	0.07	0.14
γ_{11}	0.50	0.46	0.27	0.19
τ_{00}	1.00	1.06	0.29	0.17
τ_{11}	1.00	1.05	0.28	0.27
τ_{01}	0.50	0.50	0.15	0.12
$var(\xi_{.j})$	0.43	0.43	0.07	0.09

	θ	$E(\hat{\theta})$	$E\{se(\hat{\theta})\}$	$SD(\hat{\theta})$
Measurement parameters				
a_{x1}	0.80	0.78	0.08	0.06
a_{x2}	1.00	0.98	0.08	0.08
a_{x3}	1.20	1.23	0.11	0.09
a_{x4}	1.40	1.37	0.12	0.14
a_{x5}	1.60	1.59	0.18	0.12
a_{y1}	0.80	0.77	0.06	0.06
a_{y2}	1.00	0.97	0.07	0.06
a_{y3}	1.20	1.19	0.11	0.06
a_{y4}	1.40	1.37	0.12	0.14
a_{y5}	1.60	1.56	0.17	0.13
c_{x1}	-0.80	-0.77	0.06	0.09
c_{x2}	0.00	0.00	0.05	0.09
c_{x3}	1.20	1.21	0.08	0.12
c_{x4}	-0.70	-0.66	0.07	0.14
c_{x5}	0.80	0.78	0.08	0.14
c_{y1}	-0.80	-0.79	0.06	0.12
c_{y2}	0.00	0.00	0.06	0.15
c_{y3}	1.20	1.21	0.09	0.19
c_{y4}	-0.70	-0.67	0.08	0.23
c_{y5}	0.80	0.84	0.09	0.24
Efficiency				
60~90min				

Note. θ = Generating values; $E(\hat{\theta})$ = mean of point estimates; $E\{se(\hat{\theta})\}$ = mean of estimated SEs (post-convergence approximated SEs); a = item slope parameter; c = item threshold parameter.

Given the iteration conditions, only 26 of 50 replications converged within the specified number of iterations. For this condition, the cause of low convergence rate was mostly due to the approximation of observed data information matrix rather than point estimates themselves. Either allowing larger numbers of iterations or achieving more efficient approximation of the observed

data information matrix would help the convergence rate increase. As a trial, 1000 iterations was tried, and this could increase the convergence rate up to 78% for this condition.

Simulation Study 2: Comparison of Models

Methods. The second simulation study was conducted to examine how measurement error and sampling error may influence compositional effect and cross-level interaction estimates across different conditions with both a traditional HLM model and a latent variable model.

Simulation conditions. Data generation conditions varied with respect to compositional effect sizes (compositional effect of 0, 0.2 or 0.5 and cross level interaction of 0, 0.5 or 1), sampling conditions (*ng*=100, *n p*=20; *ng*=100, *n p*=5; *ng*=20, *n p*=20), ICC sizes (0.1 or 0.3), and measurement conditions (see, Table 5). 100 and 50 replications were attempted for the contextual effect model and cross-level interaction model, respectively.

Table 5

Conditions of measurement models and generating values for item parameters

Measurement Model 1		
Condition	Slope	Intercept
	ξ_{ij} indicators X1~X5 (2PL)	η_{ij} indicators Y1~Y5 (2PL)
X1, Y1	0.8	-1
X2, Y2	1.0	0
X3, Y3	1.2	1
X4, Y4	1.4	-0.5
X5, Y5	1.6	0.5
Measurement Model 2		
	ξij indicators X1~X5 (GR, K=5)	ηij indicators Y1~Y5 (GR, K=5)
X1, Y1	0.8	-1, 0, 1, 2
X2, Y2	1.0	-1, 0, 1, 2
X3, Y3	1.2	-1, 0, 1, 2
X4, Y4	1.4	-1, 0, 1, 2
X5, Y5	1.6	-1, 0, 1, 2

***Analysis*.** Each data set has three sets of parameter estimates: 1) estimates from analyzing the generating values of η_{ij} and ξ_{ij} with a traditional multilevel model, which is treated as the gold standard (denoted as *G*), 2) estimates obtained by applying latent variable model (denoted as *L*), and 3) the estimates from analyzing the observed summed scores with the manifest variable approach (denoted as *M*). All of the traditional HLM analyses were conducted using an R package *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2012).

***Statistics*.** To compare these three sets of estimates, three statistics are calculated: 1) the percentage bias of the estimate relative to the magnitude of generating value, 2) the observed coverage of the 95% confident interval (CI) for true value, and 3) the observed power to detect the effect of interest as significant.

It should be noted that the regression coefficient estimates from the observed summed score analysis using a traditional multilevel model are not on the same scales as those obtained using the latent variable approach. To make the coefficient estimates more comparable, the estimates from traditional model approach were standardized by multiplying the parameter estimates by the ratio of standard deviation of the predictor to the standard deviation of the outcome.

Results: Compositional effect model. Relative percentage bias in $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$ is summarized in Figure 2. First, with respect to measurement model 1, in which the generating values of η_{ij} and ξ_{ij} are analyzed, the bias of $\hat{\gamma}_{01}$ ranged from 1 to 15% across the sampling conditions. While latent variable modeling analysis resulted in similar magnitude of bias with the generating value analysis, traditional HLM resulted in substantial bias (from 30 to 70%) in both estimates (see, the gray bars in Figure 2). Therefore, small ICC conditions are problematic in general. When small ICC is combined with a small number of people per group, the bias gets worse. It is noteworthy that the bias in the compositional effect from the traditional model can be upward when the ICC is large and the contextual effect size is small (see, the last plot of Figure 3). Performance of the traditional model and the latent variable model in terms of estimating $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and compositional effect, is similar across measurement conditions (see, Figure 4), indicating the measurement model is a less influential source of bias in this study.

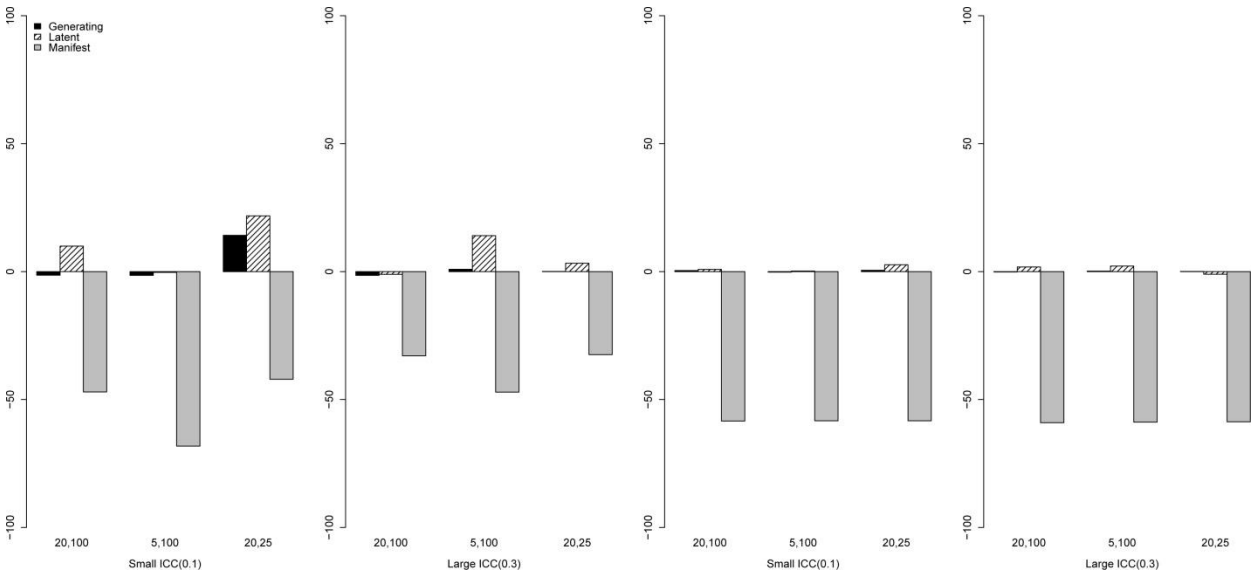

Figure 2. Relative Percentage Bias in $\hat{\gamma}_{01}$ (first two plots) and $\hat{\gamma}_{10}$ (last two plots), Large CE, MM 1.

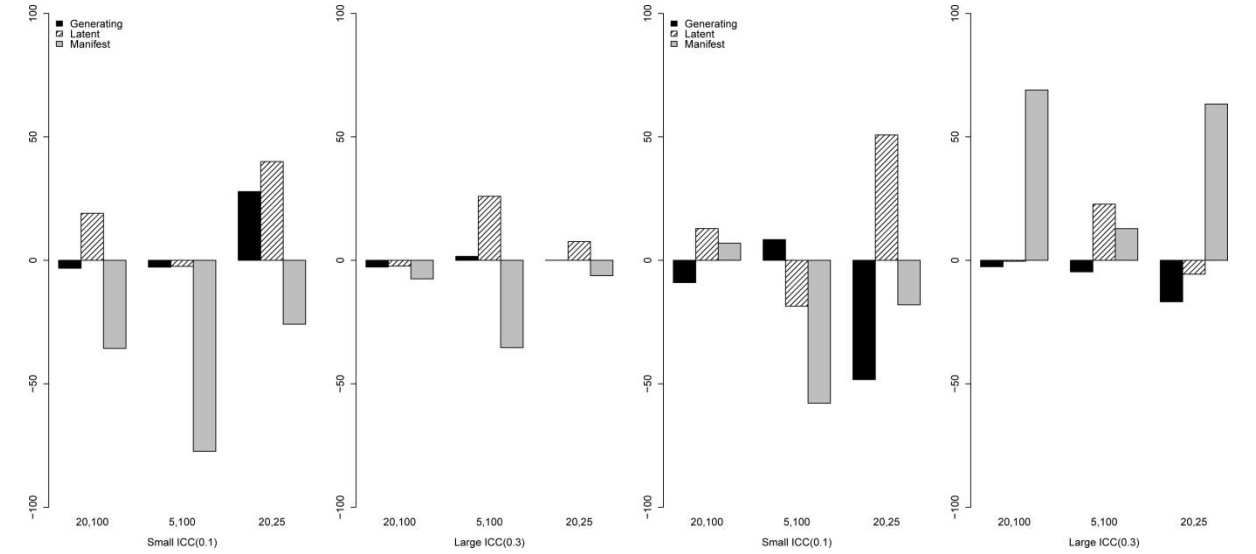

Figure 3. Relative Percentage Bias in $\hat{\gamma}_{01} - \hat{\gamma}_{10}$, Large (first two plots) and Small CE (last two plots), MM 1.

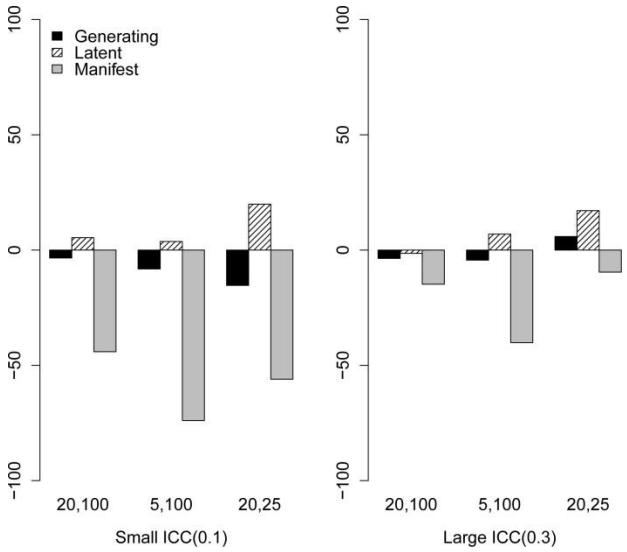

Figure 4. Relative Percentage Bias in $\hat{\gamma}_{01} - \hat{\gamma}_{10}$, Large CE, MM 2.

Second, to examine the performance of standard errors, the 95% CI coverage rate for the true compositional effect was calculated across simulated data conditions and models. Results are summarized in Figure 5. When generating values are analyzed, the coverage rates of contextual effect across sample conditions are generally as close to 95%, except for the cases where ICC is small and the number of group sampled is small. In this case the coverage rate can be low as 85%. The coverage rates of the latent variable model were also similar to those from generating value analysis, ranging from 88% to 98% for measurement model 1 and 2. When more item parameters need to be estimated, the sample is associated with a small ICC, and a small number of groups are sampled, the coverage rate can be as low as about 79%. Traditional model performance in terms of coverage rate for the contextual effect can be very problematic when both the number of people per group and ICC are small, in that the coverage can be as low as 7%.

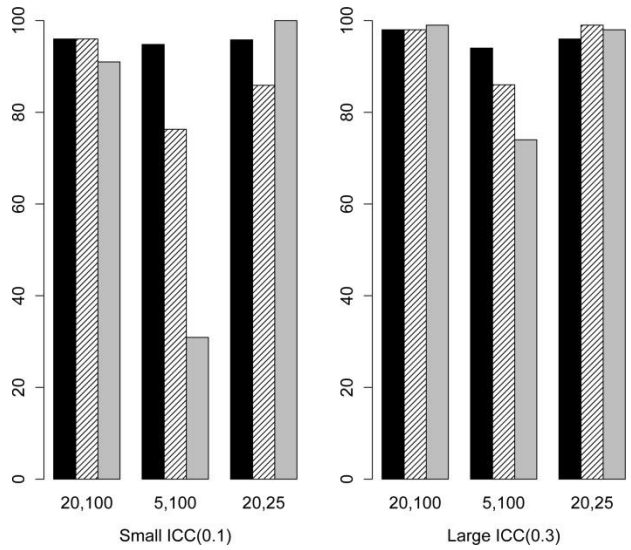

Figure 5. 95% Coverage of $\hat{\gamma}_{01} - \hat{\gamma}_{10}$, Large CE, MM 1.

Third, Figure 6 shows empirical Type I error rates of models across data conditions. Generating a value analysis model yields acceptable Type I error rates of .05 to .07 across sampling conditions. The latent variable model is similar, except for the cases when the number of people per group is small. When the number of people per group is small and ICC is small, Type I error increases to .14, indicating that it is more likely to conclude that there is a significant contextual effect than other approaches. For a traditional model, the Type I error rate inflation is huge—up to .57 when ICC is large and the number of people per group is large. Under the conditions when small ICC combines with a small number of group or a small number of people per group, the Type I error of the traditional model remains at a proper level.

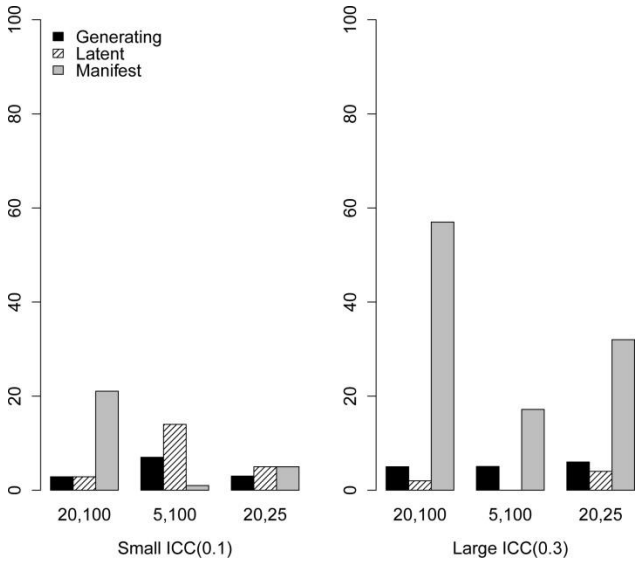

Figure 6. Empirical Type I error rates, MM 1.

When a compositional effect is large (see Figure 7), generating value analysis yields power of about .85 when ICC is large and the number of groups is large. When ICC is small, the power decreases to as low as .35 even with favorable sampling conditions. The lowest power (.15) is found when ICC is small and the number of groups is small. The patterns are similar for the latent variable model, but the latent variable model yields a slightly higher percentage of significant compositional effects in this condition. While the traditional model can yield a very high percentage of significant compositional effects when the ICC is large and the number of people per group is large, the power decreases remarkably when both ICC and the number of people per group or the number of groups are small.

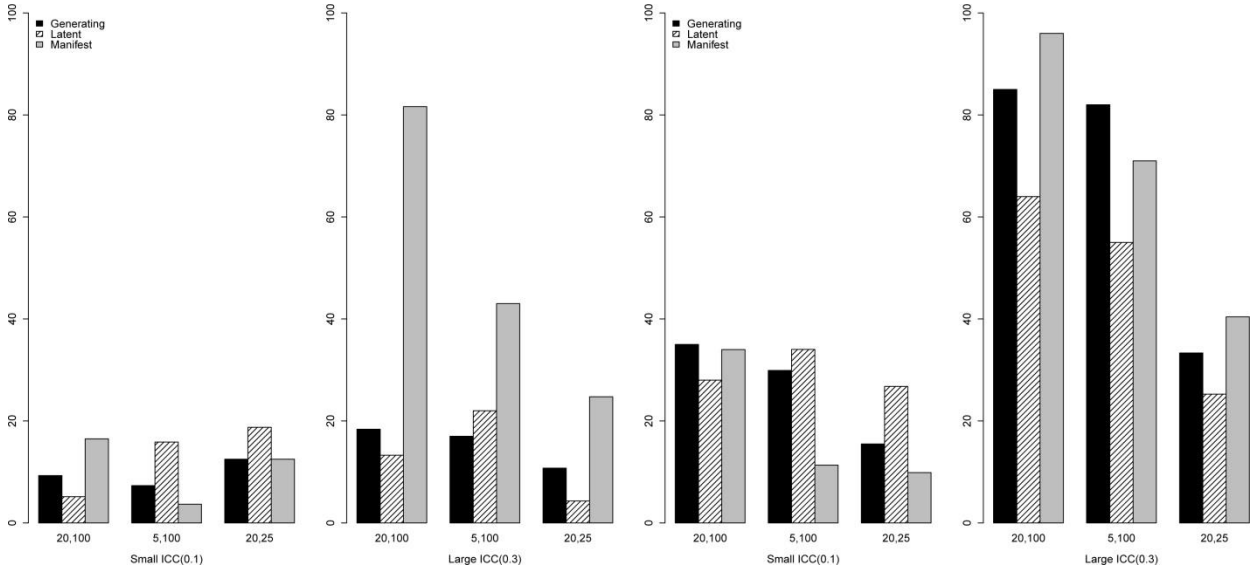

Figure 7. Percentage of significant compositional effect, Small (first two plots) and Large CE (last two plots), MM 1.

Results: Cross-level interaction model. The relative percentage bias in $\hat{\gamma}_{11}$ across simulated data conditions is summarized in Figure 8. First, when generating values are analyzed, bias can be as small as about 2% when the sampling condition is favorable and ICC is large enough. However, the bias can be as large as about 40% even when generating values are analyzed when the ICC is small and the number of groups sampled is 25. While the traditional approach yields more than 75% underestimation across conditions and reached almost 100% when a small ICC is combined with limited sample conditions, the bias in $\hat{\gamma}_{11}$ from the latent variable model analysis was smaller than that from the manifest variable model analysis.

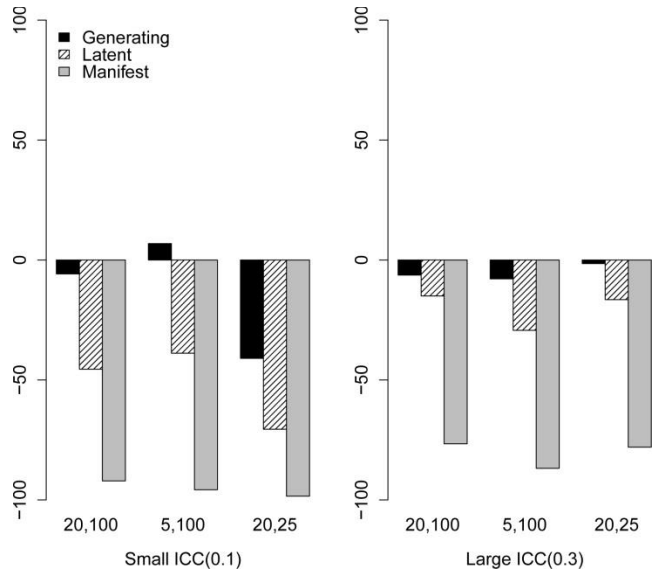

Figure 8. Relative Percentage Bias in $\hat{\gamma}_{11}$, Small CLI, MM 1.

Coverage rates for true cross-level interaction effects using 95% confidence intervals are reported in Figure 9. When generating values were analyzed, 95% confidence intervals covered the true cross-level interaction 81 to 100% of the time. When the latent variable model was applied, the coverage rates ranged from 12 to 87% depending on sampling conditions. When the number of sampled groups was small, the confidence intervals hardly captured the true values, even with the latent variable modeling approach. However, these coverage rates were still much higher than those from the traditional model approach. As bias in estimates was big and the standard error estimates were small in the traditional model approach, it was extremely rare to observe that confidence intervals actually covered the true value. Most of the coverage rates were 0.

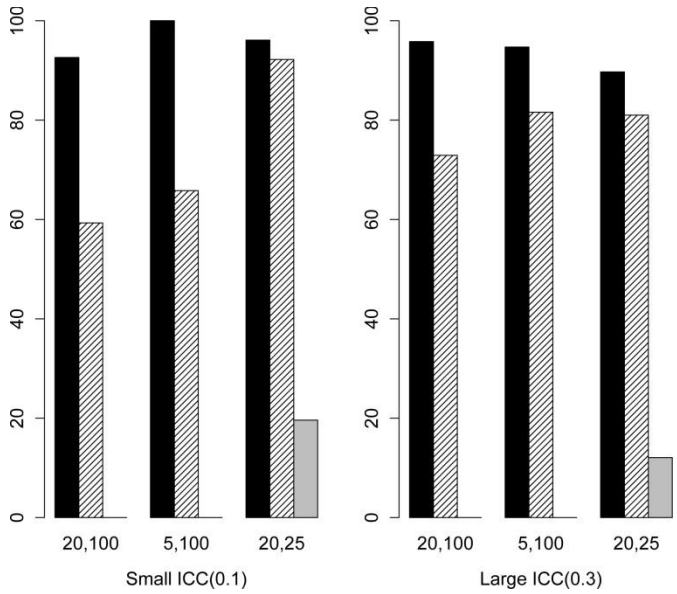

Figure 9. 95% coverage rates of $\hat{\gamma}_{11}$, Small CLI, MM 1.

Figure 10 shows observed percentage of significant cross-level interaction across different sampling conditions and analysis models. Results from the generating value analyses are encouraging in that power can be about .80 for both large and small cross-level interactions, as long as ICC is large enough and a sufficient number of groups is sampled. However, when a small number of groups is sampled, the power can be as low as .32 for a large cross-level interaction and .06 for a small cross-level interaction. The latent variable model approach can detect cross-level interaction better than the traditional modeling approach in that the percentages of significant cross-level interactions are higher in general than those from the traditional model analysis. However, when the cross-level interaction is large and the sampling condition is favorable with large ICC, the traditional model can detect the effect slightly more frequently than the latent variable modeling approach. However, it should be noted that the CI's do not cover the true value in this case, even though the traditional model can detect the existence of the cross-level interaction. It is notable that the power of the traditional model decreases dramatically when either ICC or the number of people per group is small.

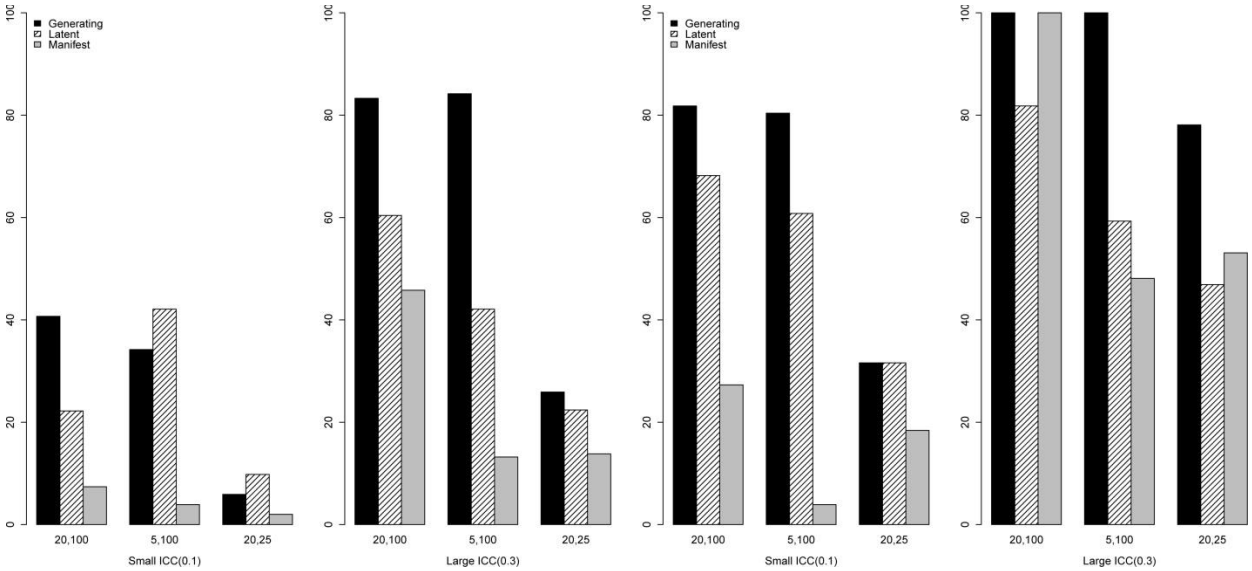

Figure 10. Percentage of significant compositional effect, Small (first two plots) and Large (last two plots) CLI, MM 1.

Empirical Applications

Compositional Effect Model: A "Big-fish-little-pond" Effect

Data. For this compositional effect analysis, a subset of PISA (2000; OECD, 2000) data were extracted and analyzed. A sample of students from the United States who worked on *reading literacy* booklets 8 and 9 was analyzed in this study for the purpose of illustration. These two booklets included 32 reading items (3 graded responses items with 3 categories and 29 dichotomously scored items) and there were 667 students from 141 schools. The number of students within a school ranged from 1 to 8, which is rather a small number of students per group. The outcome variable *self concept in reading* was measured by three items (CC02Q05, CC02Q09, and CC02Q23). Each item has a Likert-type scale, ranging from 1 (disagree) to 4 (agree).

Results. The structural parameter estimates from the multilevel latent variable model analysis (EM algorithm and the MH-RM algorithm) and traditional multilevel model analysis are reported in Table 6. In general, a positive and significant within-level coefficient $\hat{\gamma}_{10}$ is found across different models and algorithms. Between-level coefficient $\hat{\gamma}_{01}$ estimates were not significantly different from 0 when the multilevel latent model was applied, while the estimate was significantly different from 0 when the traditional multilevel was applied, due to the small standard error.

Table 6

Structural parameter estimates from PISA 2000 USA data analysis using the compositional effect model

	Latent variable model						Manifest variable model		
	MH-RM			EM			EM		
Parameter θ	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value
γ10	0.42	0.06	7.17	0.42	0.05	7.92	0.11	0.01	7.75
γ01	0.16	0.11	1.43	0.18	0.11	1.68	0.07	0.02	3.60
τ00	0.47	0.11	0.39	0.47	0.11	4.28	0.37	0.61 (*SD*)	190.31 (χ2)
var(ξ.j)	0.12	0.07	2.30	0.11	0.06	1.86	N/A	N/A	N/A
BFLPE	-0.27	0.13	-2.12	-0.24	0.12	-1.98	-0.04	0.02	-1.76
Computation time	1 hour 40 min M1=100, M2=300, M3=300 burn-in=5			1 hour 40 min 14qp, 1 processor					

Note. Reported standard errors for MH-RM algorithm are from recursively approximated observed data information. M1=Number of maximum iterations at initializing stage; M2=Number of maximum iterations at the constant gain stage; M3=Number of maximum iterations at the decreasing gain stage; qp=number of adaptive quadrature points.

The compositional effect "big-fish-little-pond" is calculated by subtracting $\hat{\gamma}_{10}$ from $\hat{\gamma}_{01}$. The direction of the compositional was negative as reported in previous research (Marsh et al., 2009). This indicates that two students who have the same levels of achievement can have different level of academic self-concept, depending on the group-level academic achievement. As the compositional effect is negative, the students who belong to a higher-level achievement group tend to have lower academic self-concept compared to students who belong to a lower-level achievement group. On the other hand, the students who belong to a lower-level achievement group tend to have higher academic self-concept compared to students who belong to a higher-level achievement group—just like a fish that feels big if the pond where it lives is small. However, in terms of the statistical significance of the compositional effect, the traditional model yields that the effect is not significantly different from 0. This result is consistent with what was found in the simulation study presented in Figure 7 in that the power of the latent variable model to detect a compositional effect is higher than that of the traditional model, when the data set is associated with a sufficiently large number of groups and a small number of students per group.

Cross-level Interaction Model: Co-operative Learning Preference and Reading Literacy

Data. For this cross-level interaction model analysis, a subset of PISA 2000 was extracted and analyzed. The data were collected in Korea, and those students who were administered

booklets 8 and 9 for reading literacy were used in this analysis. In the process of data cleaning, 4 reading items were dropped, since all item responses were zero. 29 item responses (3 graded responses and 26 dichotomously scored items) of 1,103 students in 143 schools were analyzed. These 29 items are the indicators for the latent predictor variable. The number of students within a school ranged from 1 to 8, which can be considered a small number of students per group. The outcome variable, *co-operative learning preference*, was measured by four items (CC02Q02, CC02Q08, CC02Q19, and CC02Q22). Each item has a Likert-type scale, ranging from 1 (disagree) to 4 (agree).

Results. The structural parameter estimates from the multilevel latent variable model analysis (EM algorithm and the MH-RM algorithm) and traditional multilevel model analysis are reported in Table 7. In general, positive within- and between-level coefficients ($\hat{\gamma}_{10}$ and $\hat{\gamma}_{01}$) were found, indicating that the level of co-operative learning preference and reading literacy is positively associated. However, none of these were statistically significant when the MH-RM algorithm was applied, and only the between-level coefficient was significant at a $p < .05$ level when the EM algorithm was applied, which is also different from the traditional HLM analysis in that both coefficients are statistically different from 0 due to the small standard errors.

Table 7

Structural parameter estimates from PISA 2000 Korea data analysis using the cross-level interaction model

	Latent variable model						Manifest variable model		
	MH-RM			EM			EM		
Parameter θ	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value	$\hat{\theta}$	se($\hat{\theta}$)	*t*-value
$\gamma 10$	0.021	0.061	0.315	0.229 (0.018)	0.149	1.538	0.066	0.019	3.339
$\gamma 01$	0.045	0.068	0.739	0.233 (0.032)	0.009	26.972	0.041	0.016	2.618
$\gamma 11$	-0.088	0.062	-1.417	-0.364 (-0.050)	0.296	-1.232	-0.004	0.019	-1.363
$\tau 00$	0.021	0.005	4.556	0.002 (0.034)	0.000	3.918	0.353	0.594 (*SD*)	192.83 (χ^2)
$\tau 11$	0.073	0.015	4.709	1.744 (0.060)	0.615	2.837	0.005	0.070 (*SD*)	147.04 (χ^2)
$\tau 01$	-0.029	0.006	-4.517	-0.052 (-0.030)	0.016	-3.211	-0.023	0.598 (*SD*)	172.75 (χ^2)
$var(\xi_{.j})$	0.817	0.007	118.852	0.629 (0.830)	0.088	7.123	N/A	N/A	N/A
Computation time	18 hours M1=100, M2=1000, M3=1000 3000 for SE burn-in=5			8 hours 5qp,1processor Mstep iteration=5000 M convergence=0.00001					

Note. Reported standard errors for the MH-RM algorithm are obtained using the post-convergence approximated observed data information. Numbers in () are transformed point-estimates for comparison since different identification option was used form Mplus running. M1=Number of maximum iterations at initializing stage; M2=Number of maximum iterations at the constant gain stage; M3=Number of maximum iterations at the decreasing gain stage; qp=number of adaptive quadrature points.

The parameter estimate of interest that captures a cross-level interaction effect was $\hat{\gamma}_{11}$, which appears to be negative in this particular example across computational algorithms and models. The negative cross-level interaction can be interpreted as the relationship between co-operative learning preference and reading literacy is weaker in schools with higher achievement levels, indicating the slope of between two variables becomes less stiff as school-level achievement increases. If the negative cross-level interaction size is large enough, the direction of the relationship between the co-operative learning preference and reading literacy could be negative at schools where school-level reading literacy is very high. However, $\hat{\gamma}_{11}$ was not statistically different from 0 across models and computational algorithms.

With respect to computation, an 8 adaptive quadrature points estimation using Mplus did not converge, and only a 5-quadrature-point solution was available with some changes in default settings that are related to the M-step. When the MH-RM algorithm was applied, it took 18 hours to estimate, and a large number of samples (3,000) were used to calculate the observed data information.

Summary

This study is situated in the current streams of research (e.g., Goldstein & Browne, 2004; Goldstein, Bonnet, & Rocher, 2007; Kamata, Bauer, & Miyazaki, 2008) that try to develop a comprehensive unified model that benefits from both multilevel modeling and latent variable modeling by combining multidimensional IRT and factor analytic measurement modeling with the flexibility of nonlinear structural modeling in a multilevel setting. Considering that one of the most urgent needs in developing a unified model is an efficient estimation method, the current study contributes to nonlinear multilevel latent variable modeling by investigating an alternative estimation algorithm. The principles of the MH-RM algorithm and the previous study results (Cai, 2008) suggest that the algorithm can be more efficient than the existing algorithms when a model is associated with a large number of latent variables or random effects.

The main purpose of this study was to improve estimation efficiency in obtaining full-information maximum likelihood (FIML) estimates of contextual effects by adopting the Metropolis-Hastings Robbins-Monro algorithm (MH-RM; Cai, 2008, 2010a, 2010b). R programs (R Core Team, 2012) implementing the MH-RM algorithm were produced to fit nonlinear multilevel latent variable models. Computation efficiency and parameter recovery were assessed by comparing results with an EM algorithm that uses adaptive Gauss-Hermite quadrature for numerical integration. Results indicate that the MH-RM algorithm can obtain FIML estimates and their standard errors efficiently, and the efficiency of MH-RM was more prominent for a cross-level interaction model, which requires 5-dimensional integration. While using the EM algorithm with only 8 adaptive quadrature points required about 100 minutes to estimate a cross-level interaction model, the MH-RM algorithm required about 60 minutes to have similar results. Considering the difference between an interpreted language and a compiled language in which each algorithm is implemented, even more substantial improvement in efficiency is expected if the MH-RM algorithm is written in a compiled language in the future.

The second purpose of this study was to provide information about the performance of nonlinear multilevel latent variable modeling compared to traditional HLM through a simulation study with various sampling and measurement structure conditions. Results suggest that nonlinear multilevel latent variable modeling can more properly estimate and detect a contextual

effect than the traditional approach in most conditions. Substantial bias was found in the between-level coefficient in the compositional model and in the cross-level interaction coefficient when the traditional model is applied, Notably, when the intraclass correlation (ICC) and the number of individuals per group were both small, the bias can be more than 80%, and the CIs hardly capture the true values. This is because that when the ICC is small, the between-group variance is too small to be decomposed and estimated, indicating between-group variation is small and the characteristic of interest is homogenous across groups. When this issue is combined with a small number of groups or a small number of people per group, the condition exacerbates the difficulty in estimating between-group variance and yield difficulty in convergence and biased estimates.

Since the within-level coefficient is also underestimated in the traditional model analysis, the point estimate of a compositional effect can be unbiased when the ICC size and the number of level-1 units per level-2 unit are both large (e.g., ICC=0.3 and the number of level-1 units per level-2 =20). However, Type I error rates of the traditional model are substantially elevated (up to 60%) in this sampling condition, indicating that the compositional effect detected by the traditional model under desirable sampling conditions could be spurious. These unacceptable Type I error rates are caused by the small standard error of between-level regression coefficient in the traditional HLM. The standard error of the between-level coefficients in HLM is influenced by the variance of between-level coefficient estimate, which is the sum of parameter dispersion and error dispersion (Raudenbush & Bryk, 2002). As the error dispersion does not reflect measurement error in HLM, the variance of between-level coefficient estimate is underestimated and so is the standard error. In contrast, the latent variable approach yielded less biased estimates, and statistical inferences across sampling and the ICC size conditions were more consistent than those of the traditional model, as long as the number of groups is sufficiently large (25 was found to be too small).

The third purpose of this study was to provide empirical illustrations using two subsets of data extracted from PISA (Adams & Wu, 2002). A negative compositional effect was found from the U.S. data in terms of the relationship between reading literacy and self-concept about reading, supporting the results from previous studies, which is called “Big-fish-little-pond” effect (e.g., Marsh et al., 2009). The compositional effect was statistically significant at $p < .05$ level when the nonlinear multilevel latent variable model was applied. On the other hand, the traditional HLM approach could not detect a statistically significant effect. It is because the power of HLM substantially decreases when the numbers of people per group are small and this subset of data was the case. With respect to a cross-level interaction model, the relation between reading literacy and co-operative learning preference was examined, using a subset of PISA data

collected in Korea. A negative, but not statistically significant, cross-level interaction was found between reading literacy and co-operative learning preference. The nonlinear multilevel latent variable model and the traditional HLM approach yielded similar results in that the cross-level interaction estimates were not statistically different from zero in both results.

Unlike the results from the simulation study, the results of empirical applications were not dramatically different in model comparison-wise. One possible explanation is that predictor variable reading literacy is measured by a large number of well-developed items for these empirical applications, and accordingly, the summed scores are very reliable. However, in other circumstances where less reliable measures (e.g., affective domain measures or teacher instructional variables) are used as predictors or where even a smaller number of people per group are sampled, it is expected to observe more substantial differences between the results from a nonlinear multilevel latent variable model and a traditional HLM. In addition, these two models also can yield divergent statistical inferences even when there are a sufficient size of ICC and a large number of people per group due the substantial elevation of Type I error rates when the traditional HLM is applied. Therefore, a wide range of further empirical applications should be followed, and the improved estimation efficiency, by adopting an MH-RM algorithm for the nonlinear multilevel latent variable models, can contribute to further applications by making the nonlinear multilevel latent variable modeling framework more practical in routine use.

References

Adams, R., & Wu, M. (2002). *PISA 2000 technical report*. Paris, France: Organization for Economic Cooperation and Development.

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.

Fisher, R. A. (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, *22*, 700-725.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*(3), 252-286.

Goldstein, H., & Browne, W. (2004). Multilevel factor analysis models for continuous and discrete data. In Maydeu-Olivares & M. J. J. (Eds.), *Contemporary Psychometrics* (p. 7270-7274). Mahwah, NJ: Lawrence Erlbarum Associates.

Hastings, W. K. (1970). Monte carlo simulation methods using markov chains and their applications. *Biometrika*, *57*, 97-109.

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). *Multilevel modeling of educational data*. In A. A. OConnell & McCoach, D. B. (Eds.), (pp. 345–388). Charlotte, NC: Information Age Publishing.

Lee, V. E., & Bryk, A. (1989). A multilevel model of the social distribution of educational achievement. *Sociology of Education*, *62*, 172-192.

Louis, T. A. (1982). Fiding the Observed Information Matrix when Using the EM Algorithm. *Journal of the Royal Statistical Society*, *44*(2), 226-233.

Lüdtke, O., Marsh, H., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*, 203-229.

Lüdtke, O., Marsh, H., Robitzsch, A., & Trautwein, U. (2011). A 2 x 2 taxonomy of multilevel latent covariate models: Accuracy and bias trade-offs in full and partial error-correction models. *Psychological Methods*, *16*(4), 444-467.

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. e. a. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, *44*, 764-802.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equations of state space calculations by fast computing machines. *Journal of Chemical Physics, 21,* 1087-1092.

Muthén, L. K., & Muthén, B. O. (2008). *Mplus 5.0* [Computer software]. Los Angeles, CA.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus user's guide* (Sixth ed.). Los Angeles, CA: Muthe` n & Muthe` n.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2012). *nlme: Linear and nonlinear mixed effects models* [Computer software manual]. (R package version 3.1-104).

R Core Team. (2012). R: *A language and environment for statistical computing* [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org (ISBN 3-900051-07-0).

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59,* 1–17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*(4), 307-335.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics, 22,* 400-407.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monographs, 17.*

Spearman, C. (1904). "General intelligence" objectively determined and measured. *American Journal of Psychology, 15,* 201-293.

Willms, J. D. (1986). Social class segregation and its relationship to pupils' examination results in Scotland. *American Sociological Review, 55,* 224-241.

Appendix A: observed and complete data likelihoods

The conditional density for x_{ijl} follows a multinomial with trial size 1 in K_l categories:

$$f_\theta(x_{ijl}|\xi_{ij}) = \prod_{k=0}^{K_l-1} P_\theta(x_{ijl} = k|\xi_{ij})^{\chi_k(x_{ijl})}, \tag{1}$$

where χ_k is an indicator function in which χ_k is 1 if $x_{ijl} = k$, or 0 otherwise. As ξ_{ij} is measured by $\mathbf{x}_{ij}$, η_{ij} is measured by $\mathbf{y}_{ij}$, the conditional density of $\mathbf{y}_{ij}$ is written as:

$$f_\theta(\mathbf{y}_{ij}|\eta_{ij}) = f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j, r_{ij}), \tag{2}$$

If we integrate r_{ij} out of Equation (2),

$$\int f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j) f_\theta(r_{ij}) d(r_{ij}) = f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j), \tag{3}$$

where $f_\theta(r_{ij})$ is the density of a normal distribution $N(0, \sigma^2)$. For identification purpose, σ^2 is fixed at 1 in this study, which makes $f_\theta(r_{ij})$ the density of a standard normal random variable. Integrating out ξ_{ij} yields

$$\begin{aligned} & f_\theta(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, \boldsymbol{\beta}_j) \\ &= \int f_\theta(\mathbf{x}_{ij}|\xi_{ij}) f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j) f(\xi_{ij}) d(\xi_{ij}) \end{aligned} \tag{4}$$

When J and I_j stand for the number of groups and number of individuals in group j, the conditional joint density of $\mathbf{y}_{.j}$ and $\mathbf{x}_{.j}$ for group j is the multiplication of the conditional joint densities for $\mathbf{y}_{ij}$ and $\mathbf{x}_{ij}$ in the same group as can be seen in the following equation:

$$f_\theta(\mathbf{y}_{.j}, \mathbf{x}_{.j}|\xi_{.j}, \boldsymbol{\beta}_j) = \prod_{i=1}^{I_j} f_\theta(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, \boldsymbol{\beta}_j) \tag{5}$$

Integrating out level-2 latent variable and random coefficients $\xi_{.j}$ and $\boldsymbol{\beta}_j$ yields

$$f_\theta(\mathbf{y}_{.j}, \mathbf{x}_{.j}) = \int \prod_{i=1}^{I_j} f_\theta(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, \boldsymbol{\beta}_j) f(\xi_{.j}) f(\boldsymbol{\beta}_j) d(\xi_{.j}) d(\boldsymbol{\beta}_j) \tag{6}$$

In this manner, one can integrate all latent variables and random coefficients out of the model to get a marginal distribution from which the parameters can be estimated. Treating η_{ij}, ξ_{ij}, $\xi_{.j}$, $\boldsymbol{\beta}_j$ and r_{ij} as missing data, the complete data likelihood, when J and I_j stand for the number of groups and number of individuals in group j, is:

$$\prod_{j=1}^{J} \Big[\prod_{i=1}^{I_j} f_\theta(\mathbf{y}_{ij}|\xi_{ij}, \xi_{.j}, \boldsymbol{\beta}_j, r_{ij}) f_\theta(\mathbf{x}_{ij}|\xi_{ij}) f_\theta(\xi_{ij}) f_\theta(r_{ij}) \Big] \times f_\theta(\boldsymbol{\beta}_j) f_\theta(\xi_{.j}) \tag{7}$$

where $f_\theta(\mathbf{x}_{ij}|\xi_{ij}) = \prod_{l=1}^{L_x} f_\theta(x_{ijl}|\xi_{ij})$ and $f_\theta(\mathbf{y}_{ij}|\boldsymbol{\xi}_{ij}, \boldsymbol{\xi}_{.j}, \boldsymbol{\beta}_j) = \prod_{l=1}^{L_y} f_\theta(y_{ijl}|\boldsymbol{\xi}_{ij}, \boldsymbol{\xi}_{.j}, \boldsymbol{\beta}_j)$. L_x and L_y are the number of manifest variables for ξ_{ij} and η_{ij}, respectively.

Appendix B: First and second order derivatives of the complete data models

Latent Structure Models

Denote the expected value and covariance matrix of $\boldsymbol{\eta}$ by $\boldsymbol{\mu}$ and Σ. When $\boldsymbol{\mu}$ and Σ contain parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$ respectively, the complete data log-likelihood function can be written as,

$$l = -\frac{1}{2}[\boldsymbol{\eta} - \boldsymbol{\mu}(\boldsymbol{\theta})]'[\Sigma(\boldsymbol{\tau})]^{-1}[\boldsymbol{\eta} - \boldsymbol{\mu}(\boldsymbol{\theta})] - \frac{1}{2}log|\Sigma(\boldsymbol{\tau})| - \frac{1}{2}Nlog2\pi. \tag{8}$$

Then the first derivative of l with respect to the parameter vector $\boldsymbol{\theta}$ is

$$\frac{\partial l}{\partial \boldsymbol{\theta}} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\Sigma(\boldsymbol{\tau})^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu}(\boldsymbol{\theta})). \tag{9}$$

The first derivative of l with respect to a parameter τ_k is

$$\frac{\partial l}{\partial \tau_k} = -\frac{1}{2}\left[tr(\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}) - (\boldsymbol{\eta} - \boldsymbol{\mu})'\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1}(\boldsymbol{\eta} - \boldsymbol{\mu})\right]. \tag{10}$$

The second derivative of l with respect to the parameter vector $\boldsymbol{\theta}$ is

$$\frac{\partial^2 l}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = -\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}}\Sigma^{-1}\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\theta}'} + \left\{(\boldsymbol{\eta} - \boldsymbol{\mu})'\Sigma^{-1}\frac{\partial^2 \boldsymbol{\mu}}{\partial \theta_i \partial \boldsymbol{\theta}'}\right\}. \tag{11}$$

The second derivative of l with respect to parameters τ_k and τ_s is

$$\begin{aligned}\frac{\partial^2 l}{\partial \tau_s \partial \tau_k} = -\frac{1}{2}\Big\{ & tr\Big(\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_s}\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \tau_s \partial \tau_k}\Big) \\ & + (\boldsymbol{\eta} - \boldsymbol{\mu})'\Big[(-1)\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_s}\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1} + \Sigma^{-1}\frac{\partial^2 \Sigma}{\partial \tau_s \partial \tau_k}\Sigma^{-1} \\ & - \Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_k}\Sigma^{-1}\frac{\partial \Sigma}{\partial \tau_s}\Sigma^{-1}\Big](\boldsymbol{\eta} - \boldsymbol{\mu})\Big\}.\end{aligned} \tag{12}$$

Graded Responses

For the manifest variables that have more than two categories, Equation (**??**) can be redefined as follows, suppressing subscripts:

$$\begin{aligned} T_0 &= 1, \\ T_1 &= \frac{1}{1 + \exp[-(b_{1,l} + a\xi)]}, \\ T_2 &= \frac{1}{1 + \exp[-(b_{2,l} + a\xi)]}, \\ &\vdots \\ T_{K-1} &= \frac{1}{1 + \exp[-(b_{K_l - 1,l} + a\xi)]}, \\ T_K &= 0 \end{aligned}$$

The cumulative response probability for a category k is defined as $P_k = T_k - T_{k+1}$. Taking the log of the likelihood function of the complete data model yields the following equation,

$$l = \sum_{k=0}^{K-1} \chi_k(x) log P_k = \sum_{k=0}^{K-1} \chi_k(x) log(T_k - T_{k+1}), \tag{13}$$

where x is the response to a graded item with K categories. The first derivatives of the complete data model log-likelihood are

$$\begin{aligned}
\frac{\partial l}{\partial b_k} &= \frac{\partial}{\partial b_k}(\chi_{k-1}(x) log(T_{k-1} - T_k) + \chi_k(x) log(T_k - T_{k+1})) \\
&= -(\frac{\chi_{k-1}(x)}{T_{k-1} - T_k} - \frac{\chi_k(x)}{T_k - T_{k+1}})\frac{\partial T_k}{\partial b_k} \\
\frac{\partial l}{\partial a} &= \sum_{k=0}^{K-1} \frac{\chi_k(x)}{T_k - T_{k+1}}(\frac{T_k}{\partial a} - \frac{T_{k+1}}{\partial a}),
\end{aligned}$$

where

$$\frac{\partial T_k}{\partial b_k} = T_k(1 - T_k), \frac{\partial T_k}{\partial a} = T_k(1 - T_k)\xi.$$

The second derivatives are given by

$$\begin{aligned}
\frac{\partial^2 l}{\partial b_k^2} &= -(\frac{\chi_{k-1}(x)}{(T_{k-1} - T_k)^2} + \frac{\chi_k(x)}{(T_k - T_{k+1})^2})(\frac{\partial T_k}{\partial b_k})^2 \\
&\quad -(\frac{\chi_{k-1}(x)}{T_{k-1} - T_k} - \frac{\chi_k(x)}{T_k - T_{k+1}})(\frac{\partial}{\partial b_k}\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial b_{k-1}\partial b_k} &= \frac{\chi_{k-1}(x)}{(T_{k-1} - T_k)^2}(\frac{\partial T_{k-1}}{\partial b_{k-1}})(\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial b_{k+1}\partial b_k} &= \frac{\chi_k(x)}{(T_{k+1} - T_k)^2}(\frac{\partial T_{k+1}}{\partial b_{k+1}})(\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial a \partial b_k} &= -\frac{\chi_k(x)}{(T_{k+1} - T_k)^2}(\frac{\partial T_k}{\partial b_k})(\frac{\partial T_k}{\partial a} - \frac{\partial T_{k+1}}{\partial a}) \\
&\quad +\frac{\chi_{k-1}(x)}{(T_{k-1} - T_k)^2}(\frac{\partial T_k}{\partial b_k})(\frac{\partial T_{k-1}}{\partial a} - \frac{\partial T_k}{\partial a}) \\
&\quad -(\frac{\chi_{k-1}(x)}{T_{k-1} - T_k} - \frac{\chi_k(x)}{T_k - T_{k+1}})(\frac{\partial}{\partial a}\frac{\partial T_k}{\partial b_k}) \\
\frac{\partial^2 l}{\partial a \partial a'} &= \sum_{k=0}^{K-1}\{-\frac{\chi_k(x)}{(T_k - T_{k+1})^2}(\frac{\partial T_k}{\partial a} - \frac{\partial T_{k+1}}{\partial a})(\frac{\partial T_k}{\partial a'} - \frac{\partial T_{k+1}}{\partial a'}) \\
&\quad +\frac{\chi_k(x)}{T_k - T_{k+1}}(\frac{\partial}{\partial a}\frac{\partial T_k}{\partial a'} - \frac{\partial}{\partial a}\frac{\partial T_{k+1}}{\partial a'})\},
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial}{\partial b_k}\frac{\partial T_k}{\partial b_k} &= T_k(1-T_k)(1-2T_k) \\
\frac{\partial}{\partial a}\frac{\partial T_k}{\partial b_k} &= T_k(1-T_k)(1-2T_k)\xi \\
\frac{\partial}{\partial a}\frac{\partial T_k}{\partial a} &= T_k(1-T_k)(1-2T_k)\xi\xi'.
\end{aligned}$$