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

Authors:
Ji Seung Yang
Li Cai

Journal publication date:
2014

Published in:
Journal of Educational and Behavioral Statistics, 39(6), 550–582

IES grant information:
Grant number R305D140046
Funded by National Center for Education Research (NCER)

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

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

Ji Seung Yang's dissertation research was supported by a dissertation grant from the Society of Multivariate Experimental Psychology. This project was also partially supported by an IES statistical methodology research grant (R305D100039). Li Cai's research is additionally supported by IES grant R305D140046 and NIDA grants R01DA026943 and R01DA030466. The views expressed here belong to the authors and do not reflect the views or policies of the funding agencies. We thank Drs. Michael Seltzer, Sandra Graham, and Steve Reise for thoughtful feedback.

Address all correspondence to: Ji Seung Yang, Department of Human Development and Quantitative Methodology, 1225 Benjamin Building, University of Maryland, College Park, MD 20742-1115. Tel: 301.405.6073 FAX: 301.314.9245 Email: jsyang@umd.edu

Authors
JI SEUNG YANG is an assistant professor of Measurement, Statistics and Evaluation (EDMS) Program in the Department of Human Development and Quantitative Methodology at University of Maryland-College Park; 1225 Benjamin Building, University of Maryland, College Park, MD 20742-1115; e-mail: jsyang@umd.edu. Her research interests include the development of statistical models for handling measurement error in predictor and outcome variables in multilevel settings.
LI CAI is a professor of Advanced Quantitative Methodology in the Department of Education at University of California-Los Angeles; Box 951521, 2022A Moore Hall, Los Angeles, CA 900951521; e-mail: lcai@ucla.edu. His research interests include the development, integration, and evaluation of innovative latent variable models that have wide-ranging applications in educational, psychological, and health-related domains of study.

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

Abstract

The main purpose of this study is to improve estimation efficiency in obtaining maximum marginal likelihood estimates of contextual effects in the framework of 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 estimates and standard errors efficiently. 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 contextual effects than the traditional approach. As an empirical illustration, data from the Programme for International Student Assessment (PISA; OECD, 2000) were analyzed.

Keywords: contextual effect, multilevel modeling, latent variable modeling, multilevel latent variable modeling

1 Introduction

In social science research, a contextual effect is traditionally defined as the difference between two coefficients in a hierarchial 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 group-level. A representative application of this kind of contextual effect in education was discussed in Raudenbush and Bryk (2002) using a subset of High School and Beyond (HS&B) data. 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 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. Though hierarchical linear modeling opened the door to estimating contextual effects, there have been two unresolved problems. 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, Ch. 3).

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). Lüdtke et al. (2008) proposed a multilevel latent variable modeling framework for contextual analysis. Lüdtke et al. (2008)'s simulation study is noteworthy in that the study 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 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). However, the type of manifest variables is limited to continuous only in Lüdtke et al.'s (2008) study.

Marsh et al. (2009) conducted another noted study using multilevel latent variable modeling for contextual effect analysis. Marsh and colleagues compared several contextual modeling options related to "big fish-little-pond effect (BFLPE)" estimation using an empirical data set in which academic achievement (predictor) and academic self-concept (outcome) were measured by, respectively, three and four continuous manifest variables. Among the tested models, a multilevel latent variable model 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 all the manifest variables, rather than using summed or averaged scores at both individual- and group-level. The study also illustrated how the nonlinear multilevel latent variable modeling approach can provide flexibility in modeling by including random slopes, latent (within-level or cross-level) interactions, and latent quadratic effects. In both Lüdtke et al.'s (2008) and Marsh et al.'s (2009) studies, they utilized continuous manifest variables, while the current study considers categorical indicators (item-level data) for all latent variables in the model.

While theoretically desirable, nonlinear multilevel latent variable modeling poses significant computational difficulties. Standard approaches such as numerical integration (e.g., adaptive quadrature) based EM or Markov chain Monte Carlo (MCMC, e.g., Gibbs Sampling) based estimation methods have important limitations that make them less practical for routine use. With respect to numerical integration, its computational bur-

den increases exponentially when the dimensionality of latent variable space is high, as is the case with the current nonlinear multilevel latent variable model. On the other hand, while MCMC is entirely free from the curse of multidimensionality, it is not immune from issues that include advanced tuning requirements, specification of priors, and convergence analysis for complex models. Lüdtke et al. (2011) also reported the occurrence of unstable estimates. The model has difficulty converging when small sample size is combined with low intraclass correlation coefficient (ICC) in predictors, and also when there are substantial amount of missing observations in the manifest variables.

The main objective of this study is to develop a more efficient and stable 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). This study significantly extends the applications of MH-RM algorithm to the case of multilevel modeling. Prior research using MH-RM is limited to single level applications, e.g., exploratory and confirmatory item factor analysis (Cai, 2010a, 2010b), latent regression modeling (von Davier & Sinharay, 2010), and item response theory modeling with non-normal latent variables (Monroe & Cai, 2014).

Computational efficiency and parameter recovery were assessed in a comparison with an implementation of EM algorithm using adaptive Gauss-Hermite quadrature (Mplus; Muthén & Muthén, 2008). Another objective was to find, through a simulation study, the extent to which measurement error and sampling error can influence contextual effect estimates under different conditions. The results can provide practical rationales for the application of computationally demanding nonlinear multilevel latent variable models. The last objective of this study was to provide an empirical illustration of estimating contextual effects by applying nonlinear multilevel latent variable models to empirical data that contain complex measurement structures and unbalanced data. A subset of data from Programme for International Student Assessment (PISA; Adams & Wu, 2002) were analyzed to illustrate a contextual effect model.

2 Contextual Effects in a Nonlinear Multilevel Latent Variable Model

The particular contextual effect of interest in the current study is one that occurs when a group-level characteristic is measured by individual-level variables, and the individual-level variables are in turn measured by categorical manifest variables. This study considers a contextual effect as a compositional effect that captures the influence of contextual variables on individual-level outcomes, controlling for the effect of the individual-level predictor.

2.1 Structural Models

In traditional HLM, a compositional effect β_c can be defined as follows:

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

In Equation (1), Y_{ij} and X_{ij} denote the outcome and predictor values of individual i in level-2 unit j, respectively. For the level-1 equation, the predictor values are centered around the group means $\bar{X}_{.j}$. For the level-2 model, the predictor values are centered around the grand mean $\bar{X}_{..}$.

In typical educational research settings, Y_{ij} and X_{ij} can be constructed by summing or averaging item scores from self-reports or other instruments. 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. In this particular definition of a *contextual effect* as a compositional effect, the slope γ_{10} is the same across the level-2 units (a fixed effect).

In a nonlinear multilevel latent variable model, the predictors and outcomes become latent variables that are denoted as η_{ij} and ξ_{ij}. Those latent variables are connected to manifest variables through measurement models. For notational simplicity, latent

individual deviations from latent group means can be defined as $\delta_{ij} = \xi_{ij} - \xi_{.j}$, and group mean deviations from the latent grand mean can be defined as $\delta_{.j} = \xi_{.j} - \xi_{..}$. Then the latent variable counterpart to Equation (1) is:

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

Note that we have centered the latent level-1 predictor values around the group means, and the latent level-2 predictor values around the grant mean, maintaining comparability with Equation (1). Similarly, 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.

For identification purposes, we impose the restriction of $\xi_{..} = 0$ to fix the location of the predictor latent variable in the model. This implies that $\delta_{.j} = \xi_{.j}$ and the level-1 latent predictor value is expressed as a group mean plus a deviation term $\xi_{ij} = \xi_{.j} + \delta_{ij}$. To identify the location of the outcome latent variable, we set the intercept γ_{00} to 0 as well. To identify the scale of the latent variables, we impose additional restrictions on δ_{ij} and r_{ij}. These are disturbance terms, so they should have zero means and as is customary in other item response theory modeling situations, we set their variances to unity, i.e., $var(\delta_{ij}) = 1$ and $\sigma^2 = 1$. This particular identification constraint leaves open the possibility to estimate the variance of $\xi_{.j}$, which will be denoted ψ, as well as the variance of u_{0j}, which is τ_{00}. We also make the regression model specification assumption that the deviation $\xi_{.j}$ and the random effect u_{0j} are statistically independent.

2.2 Measurement Models

The measurement models define the relationship between manifest variables and latent variables. For brevity, only the measurement models of the latent predictor variable

ξ_{ij} will be described in this section, since the measurement models for the latent outcome η_{ij} can be defined analogously.

When manifest variables are ordinal response variables with multiple categories (including 0-1 responses), as is often the case with instruments used in educational research, a logistic version of Samejima (1969)'s classical graded response model can be utilized. Let item l have K_l ordered categories. The conditional cumulative probability for a response in category $k \in \{0, 1, ..., K_l - 1\}$ and above are defined as follows:

$$\begin{aligned} T_0(\xi_{ij}) &= 1, \\ T_1(\xi_{ij}) &= \frac{1}{1 + \exp[-(c_{1,l} + a_l\xi_{ij})]}, \\ &\vdots \\ T_{K_l-1}(\xi_{ij}) &= \frac{1}{1 + \exp[-(c_{K_l-1,l} + a_l\xi_{ij})]}, \end{aligned} \tag{3}$$

where $c_{1,l}, \ldots, c_{K_l-1,l}$ represent a vector of $K_l - 1$ item intercept parameters, and a_l is the item slope. The category response probability is defined as the difference between two adjacent cumulative probabilities:

$$P_k(\xi_{ij}) = T_k(\xi_{ij}) - T_{k+1}(\xi_{ij}), \tag{4}$$

for $k \in \{0, 1, ..., K_l - 1\}$, where $T_{K_l}(\xi_{ij}) = 0$.

Let $X_{ijl} \in \{0, 1, ..., K_l - 1\}$ be a random variable representing the ith individual's response in the jth level-2 unit to the lth item, and let x_{ijl} be a realization of X_{ijl}. Conditional on ξ_{ij}, the distribution of X_{ijl} is multinomial with trial size 1 in K_l categories:

$$f_{\boldsymbol{\theta}}(x_{ijl}|\xi_{ij}) = \prod_{k=0}^{K_l-1} P_k(\xi_{ij})^{\chi_k(x_{ijl})}, \tag{5}$$

where $\chi_k(x_{ijl})$ is an indicator function which equals 1 if and only if x_{ijl} is equal to k, and

0 otherwise. Note that missing observations are handled naturally in this conditional multinomial formulation. If x_{ijl} is a missing data point, the indicator function is always 0, and hence only observed responses contribute to the measurement of ξ_{ij}.

The conditional density $f_{\boldsymbol{\theta}}(x_{ijl}|\xi_{ij})$ is indexed by $\boldsymbol{\theta}$, which is our generic notation for a vector of all free parameters in the model that includes the item intercepts and slopes, the fixed effects (γ_{01}, γ_{10}), and the variance components (τ_{00} and ψ). Let $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijL_x})'$ be a $L_x \times 1$ vector of item responses from individual i in level-2 unit j to the L_x items measuring ξ_{ij}. Invoking the critically important assumption of conditional independence of item responses given the latent variable, we may write

$$f_{\boldsymbol{\theta}}(\mathbf{x}_{ij}|\xi_{ij}) = \prod_{l=1}^{L_x} f_{\boldsymbol{\theta}}(x_{ijl}|\xi_{ij}) = f_{\boldsymbol{\theta}}(\mathbf{x}_{ij}|\xi_{.j}, \delta_{ij}), \tag{6}$$

where the last equality follows from the fact that $\xi_{ij} = \xi_{.j} + \delta_{ij}$.

2.3 Observed and Complete Data Likelihoods

Similar to the case of ξ_{ij}, let us consider the measurement of η_{ij}. Let L_y be the number of manifest variables for η_{ij}. Again under conditional independence, the conditional response probabilities factor into item response probabilities:

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}|\eta_{ij}) = \prod_{l=1}^{L_y} f_{\boldsymbol{\theta}}(y_{ijl}|\eta_{ij}), \tag{7}$$

where $\mathbf{y}_{ij}$ is the $L_y \times 1$ vector of item responses from individual i in level-2 unit j to the outcome measures. Recall from Equation (2) that

$$\eta_{ij} = \beta_{0j} + \beta_{1j}\delta_{ij} + r_{ij} = \gamma_{00} + \gamma_{01}\xi_{.j} + u_{0j} + \gamma_{10}\delta_{ij} + r_{ij}.$$

We note that given fixed effects, if we knew the random effect u_{0j}, the latent group mean $\xi_{.j}$, the latent deviation term δ_{ij}, and the equation disturbance term r_{ij}, η_{ij} would be completely determined. This implies that we may rewrite the conditional distribution of

$\mathbf{y}_{ij}$ as:

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}|\eta_{ij}) = f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}|\xi_{.j}, u_{0j}, \delta_{ij}, r_{ij}). \tag{8}$$

If we integrate r_{ij} out of Equation (8), we have left

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}|\xi_{.j}, u_{0j}, \delta_{ij}) = \int f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}|\xi_{.j}, u_{0j}, \delta_{ij}, r_{ij}) f(r_{ij}) d(r_{ij}), \tag{9}$$

where $f(r_{ij})$ is the density of a standard normal random variable, given preceding assumptions about the disturbance term. Bringing in results from Equation (6) and integrating out δ_{ij} yields a conditional density that depends only on the level-2 latent variables and random effects

$$f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, u_{0j}) = \int f_{\boldsymbol{\theta}}(\mathbf{x}_{ij}|\xi_{.j}, \delta_{ij}) f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}|\xi_{.j}, u_{0j}, \delta_{ij}) f(\delta_{ij}) d(\delta_{ij}), \tag{10}$$

where $f(\delta_{ij})$ is the density of a standard normal random variable. Equation (10) makes it clear that we assume, under correct model specification, the outcome measures ($\mathbf{y}_{ij}$) and predictor measures ($\mathbf{x}_{ij}$) are conditionally independent.

Let J and I_j stand for the number of level-2 units and number of individuals in level-2 unit j. Let $\mathbf{Y}_j = \{\mathbf{y}_{ij}\}_{i=1}^{I_j}$ and $\mathbf{X}_j = \{\mathbf{x}_{ij}\}_{i=1}^{I_j}$ represent the collected responses to the outcome manifest variables and predictor manifest variables, respectively, from all individuals in level-2 unit j. We now make the critical assumption of conditional independence again - that the individuals are independent conditionally on the level-2 latent variables/random effects $\xi_{.j}$ and u_{0j}. Thus the conditional joint density of $\mathbf{Y}_j$ and $\mathbf{X}_j$ becomes:

$$f_{\boldsymbol{\theta}}(\mathbf{Y}_j, \mathbf{X}_j|\xi_{.j}, u_{0j}) = \prod_{i=1}^{I_j} f_{\boldsymbol{\theta}}(\mathbf{y}_{ij}, \mathbf{x}_{ij}|\xi_{.j}, u_{0j}). \tag{11}$$

Integrating out the level-2 latent variables and random effects yields the marginal prob-

ability, wherein we have utilized the independence of $\xi_{.j}$ and u_{0j}:

$$f_{\boldsymbol{\theta}}(\mathbf{Y}_j, \mathbf{X}_j) = \int \int \prod_{i=1}^{I_j} f_{\boldsymbol{\theta}}(\mathbf{Y}_j, \mathbf{X}_j | \xi_{.j}, u_{0j}) f(\xi_{.j}) f(u_{0j}) d(\xi_{.j}) d(u_{0j}) \tag{12}$$

By this point we have integrated all latent variables and random effects out of the joint probabilities. We now make the routine multilevel modeling assumption that the level-2 units are the independent sampling units. Upon observing $\mathbf{Y}_j$ and $\mathbf{X}_j$ and treating them as fixed, the marginal (observed data) likelihood function for the entire sample is

$$L(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}) = \prod_{j=1}^{J} f_{\boldsymbol{\theta}}(\mathbf{Y}_j, \mathbf{X}_j), \tag{13}$$

where $\mathbf{Y} = \left\{\mathbf{Y}_j\right\}_{j=1}^{J}$ and $\mathbf{X} = \left\{\mathbf{X}_j\right\}_{j=1}^{J}$ collect together the full set of outcome and predictor observed variable responses, respectively. Directly maximizing this marginal likelihood function over $\boldsymbol{\theta}$ would lead to the maximum marginal likelihood estimator of the structural parameters.

An obvious computational limitation to the direct marginal likelihood approach is the integration involved in arriving at the observed data likelihood. All of the integrals must be approximated numerically, which can be computationally challenging. An alternative stance is to treat the random effects and latent variables r_{ij}, δ_{ij}, $\xi_{.j}$, and u_{0j} as missing data. This leads to a missing data formulation of the latent variable model. Had the missing data been observed, the complete data likelihood function can be written as

$$L(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}, \mathbf{Z}) = \prod_{j=1}^{J} \left[\prod_{i=1}^{I_j} f_{\boldsymbol{\theta}}(\mathbf{y}_{ij} | \xi_{.j}, u_{0j}, \delta_{ij}, r_{ij}) f_{\boldsymbol{\theta}}(\mathbf{x}_{ij} | \xi_{.j}, \delta_{ij}) f(\delta_{ij}) f(r_{ij}) \right] f_{\boldsymbol{\theta}}(u_{0j}) f_{\boldsymbol{\theta}}(\xi_{.j}), \tag{14}$$

where $\mathbf{Z}$ collects together all the level-1 random effects/latent variables $\left\{ \left\{ r_{ij}, \delta_{ij} \right\}_{i=1}^{I_j} \right\}_{j=1}^{J}$ as well as those at level-2 $\left\{ u_{0j}, \xi_{.j} \right\}_{j=1}^{J}$. In other words, $\mathbf{Z}$ represents the "missing data."

This missing data formulation prompts us to consider an alternative estimation ap-

proach that eschews numerical integration. In particular, the missing data may be "filled in" by drawing imputations from their posterior predictive distribution $f(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta})$. Note that in our case the posterior predictive distribution is proportional to the complete data likelihood, greatly facilitating the use of MCMC sampling methods to draw from the posterior. The imputations lead to complete data sets, and the complete data likelihood function is much easier to handle than the observed data likelihood function due to its completely factored form. Instead of directly solving the observed data optimization problem, a sequence of complete data optimizations can iteratively improve the parameters estimates until convergence.

3 Metropolis-Hastings Robbins-Monro Algorithm for Contextual Models

3.1 Metropolis-Hastings Robbins-Monro Algorithm

The MH-RM algorithm was initially proposed by Cai (2008) for nonlinear latent structure analysis with a comprehensive measurement model, and the application of the algorithm has been expanded to other measurement and statistical models (e.g. Cai, 2010a; Cai, 2010b; Monroe & Cai, 2014). The MH-RM algorithm combines the Metropolis-Hastings (MH; Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953) algorithm and the Robbins-Monro (RM; Robbins & Monro, 1951) stochastic approximation algorithm.

Utilizing the missing data formation of the latent variable model, the random effects and latent variables are treated as missing data. Once the missing data are "filled in" by the MH sampler, complete data likelihoods can be optimized iteratively. Because imputation noise is introduced in the MH step, the RM algorithm is used to filer out the noise. Let the parameter estimate at iteration t be denoted $\boldsymbol{\theta}^{(t)}$, the $(t+1)$th iteration of the MH-RM algorithm consists of three steps: Stochastic Imputation, Stochastic Approximation, and Robbins-Monro Update.

Step 1. Stochastic Imputation

Draw M_t sets of missing data, which are the random effects and latent variables, from a

Markov chain that has the posterior predictive distribution of missing data $f(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}^{(t)})$ as the target. Then, M_t sets of complete data are formed as follows:

$$\left\{\mathbf{Y}, \mathbf{X}, \mathbf{Z}_m^{(t+1)}; m = 1, ..., M_t\right\}. \tag{15}$$

Step 2. Stochastic Approximation

Let

$$\mathbf{s}(\boldsymbol{\theta}^{(t)}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}_m^{(t+1)}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}^{(t)}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}_m^{(t+1)}) \tag{16}$$

denote the gradient vector of the complete data log-likelihood function, evaluated at the current parameter value $\boldsymbol{\theta}^{(t)}$ and missing data imputation $\mathbf{Z}_m^{(t+1)}$. We first compute the sample average of gradients of the complete data log-likelihood:

$$\tilde{\mathbf{s}}_{t+1} = \frac{1}{M_t} \sum_{j=1}^{M_t} \mathbf{s}(\boldsymbol{\theta}^{(t)}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}_m^{(t+1)}). \tag{17}$$

By Fisher's Identity (Fisher, 1925), the conditional expectation of the complete data gradient vector over the posterior distribution of the missing data is the same as the gradient vector of the observed data log-likelihood, under mild regularity conditions, i.e.,

$$\frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}) = \int \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathbf{Y}, \mathbf{X}, \mathbf{Z}) f(\mathbf{Z}|\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}) d\mathbf{Z}. \tag{18}$$

In other words, though noise-corrupted, $\tilde{\mathbf{s}}_{t+1}$ gives the direction of likelihood ascent because it is a Monte Carlo approximation of the conditional expected complete data gradient vector (the right hand side of Equation 18), which is also an approximation of the observed data gradient vector (the left hand side of Equation 18).

Step 3. Robbins-Monro Update

To improve stability and speed, we also compute a recursive approximation of the con-

ditional expectation of the information matrix of the complete data log-likelihood:

$$\mathbf{\Gamma}_{t+1} = \mathbf{\Gamma}_t + \epsilon_t \left[\frac{1}{M_t} \sum_{m=1}^{M_t} \mathbf{H}(\boldsymbol{\theta}^{(t)} | \mathbf{Y}, \mathbf{X}, \mathbf{Z}_m^{(t+1)}) - \mathbf{\Gamma}_t \right], \tag{19}$$

where

$$\mathbf{H}(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}, \mathbf{Z}) = -\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \log L(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X}, \mathbf{Z})$$

is the complete data information matrix, i.e., the negative second derivative matrix of the complete data log-likelihood. Updated parameters are computed recursively:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \epsilon_t (\mathbf{\Gamma}_{t+1}^{-1} \tilde{\mathbf{s}}_{t+1}), \tag{20}$$

where $\{\epsilon_t; t \geq 0\}$ is a sequence of gain constants (to be elaborated).

The iterations are started from initial values $\boldsymbol{\theta}^{(0)}$ and a positive definite matrix $\mathbf{\Gamma}_0$. They can be terminated when the changes in parameter estimates are sufficiently small. As a practical method for convergence check, Cai (2008) proposed to monitor a "window" of the largest absolute differences between two adjacent iterations. Cai (2008) suggested 3 as a reasonable width of the window to be monitored in practice. Cai (2008) showed that the MH-RM iterations converge to a local maximum of the observed data likelihood $L(\boldsymbol{\theta} | \mathbf{Y}, \mathbf{X})$ with probability one as t increases without bounds.

The gain constant ϵ_t is a sequence of decreasing non-negative real numbers such that $\epsilon_t \in (0, 1]$, $\sum_{t=0}^{\infty} \epsilon_t = \infty$, and $\sum_{t=0}^{\infty} \epsilon_t^2 < \infty$. In practical implementations of MH-RM, the starting parameter values $\boldsymbol{\theta}^{(0)}$ are often sufficiently far away from the mode of the marginal likelihood that extra care must be taken with the gain constant sequence so that MH-RM does not terminate prematurely. We typically implement a 3-stage procedure wherein the first $M1$ iterations of MH-RM uses non-decreasing gain constants to quickly move the provisional estimates to a vicinity of the final solution. The next $M2$ iterations use the same non-decreasing gain constants but the estimates are averaged to start the

final MH-RM iterations with decreasing gain constants. For the last stage, the sequence of gain constants is taken to be $\epsilon_t = 0.1/(t+1)^{0.75}$ as some experimentation.

3.2 Approximating the Observed Information Matrix

One of the benefits of using the MH-RM algorithm is that the observed data information matrix can be 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. Utilizing Fishier's Identity, the gradient vector is approximated recursively,

$$\hat{\mathbf{s}}_{t+1} = \hat{\mathbf{s}}_t + \epsilon_t\{\tilde{\mathbf{s}}_{t+1} - \hat{\mathbf{s}}_t\}, \tag{21}$$

where $\tilde{\mathbf{s}}_t$ is defined as Equation (17). A Monte Carlo estimate of the conditional expectation of the complete data information matrix minus the conditional covariance of the complete data gradient vector is defined as follows:

$$\tilde{\mathbf{G}}_t = \frac{1}{M_t}\sum_{j=1}^{m_k}\left[\mathbf{H}(\boldsymbol{\theta}^{(t)}|\mathbf{Y},\mathbf{X},\mathbf{Z}_m^{(t+1)}) - \mathbf{s}(\boldsymbol{\theta}^{(t)}|\mathbf{Y},\mathbf{X},\mathbf{Z}_m^{(t+1)})[\mathbf{s}(\boldsymbol{\theta}^{(t)}|\mathbf{Y},\mathbf{X},\mathbf{Z}_m^{(t+1)})]'\right]. \tag{22}$$

A more stable estimate can be found by further recursive approximation:

$$\hat{\mathbf{G}}_{t+1} = \hat{\mathbf{G}}_t + \epsilon_t\{\tilde{\mathbf{G}}_{t+1} - \hat{\mathbf{G}}_t\}. \tag{23}$$

Finally, the observed information matrix is approximated as

$$\mathcal{I}_{t+1} = \hat{\mathbf{G}}_{t+1} + \hat{\mathbf{s}}_{t+1}\hat{\mathbf{s}}'_{t+1}. \tag{24}$$

Cai (2010a) discussed the rationale behind this approximation as a recursive application of Louis's (1982) formula. The main benefit is that the information matrix becomes a by-product of the MH-RM iterations. Another practical option for approximating the ob-

served information matrix is a direct application of Louis's (1982) formula, in which the gradient vector and the conditional expectations are approximated directly after converge of the MH-RM algorithm using additional Monte Carlo samples. In this study, standard errors obtained by the first method are called *recursively approximated standard errors* and those from the latter are called *post-convergence approximated standard errors*.

4 Simulation Studies

4.1 Simulation Study 1: Comparison of Estimation Algorithms

4.1.1 Methods

The first study examined parameter recovery and standard errors across two algorithms, MH-RM algorithm and an existing EM algorithm. The data-generating and fitted models followed Equation (2). 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 the graded model in Equation (3). For η_{ij}, the manifest variables are Y_1, Y_2, Y_3, Y_4, and Y_5. For ξ_{ij}, which is the sum the level-2 latent group mean and the deviation terms ($\xi_{.j} + \delta_{ij}$), the manifest variables are X_1, X_2, X_3, X_4, and X_5. The item parameters were the same across levels, representing cross-level measurement invariance.

We attempted 100 Monte Carlo replications. 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 based EM approach implemented in Mplus (Muthén & Muthén, 2010). The MH-RM algorithm's convergence criterion was 5.0×10^{-5}, and the maximum number of iterations for the first two stages of MH-RM with constant gains were $M1 = 100$ and $M2 = 500$. To calculated post-convergence approximated standard errors, 100 to 500 additional random samples were used. All replications converged within 600 MH-RM iterations with decreasing gains.

4.1.2 Results

The generating values and the corresponding estimates for the compositional effect 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 (qp=5 and qp=14). The default number of quadrature points is 15 in Mplus, but the computer cannot handle 15 quadrature points for this four-dimensional model. The maximum possible number of quadrature points was 14 for a compositional effect model. A smaller number of quadrature points (5) was tested to compare point estimates and standard errors. The fourth set of columns includes the corresponding point estimates and standard errors using the MH-RM algorithm.

The means of point estimates from different algorithms are generally very close to one another. For structural parameter estimates, 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 ψ. Standard errors are also very similar.

For measurement parameter estimates, both the means of point estimates and the standard errors were the same up to the second decimal place across different numbers of quadrature points. The largest difference in average point estimates between EM and MH-RM was 0.02, indicating that the two approaches yield highly similar estimates. However, 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 biggest difference in standard error estimates for measurement parameters between two algorithms was 0.13.

The natural logarithm of standard error estimates from EM algorithm, MH-RM algorithm (post-convergence approximated standard errors) are plotted against the natural logarithm of empirical standard deviations of point estimates across the Monte Carlo

replications in Figure 1. The estimates are clustered along the diagonal reference line, indicating that the 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.

With regards to computing time, when one processor was used for estimation, EM with 5 quadrature points generally required a small amount of time, while EM with 14 quadrature points generally required over an hour. The MH-RM algorithm required about 40 minutes. Note that MH-RM is implemented in R (an interpreted language) with explicit looping, while Mplus is written in FORTRAN (a compiled language). As an interpreted language is expected to be several orders of magnitude slower compared to a compiled language in terms of looping, a direct comparison is inappropriate. What we can safely conclude is that when ported into a compiled language, MH-RM is poised to be substantially more efficient.

To examine the performance of the MH-RM algorithm further, all 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 components. The Monte Carlo standard deviations of parameter estimates (column 5) are also similar to standard error estimates from both EM and MH-RM (column 4 and 6); the largest difference is 0.02. With respect to measurement parameters, the average item parameter estimates are very close to generating values.

However, we see that recursively approximated standard errors are generally 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, where post-convergence approximated standard errors for item intercept parameters are underestimated. Therefore, we find that recursively approximated standard errors perform better than post-

convergence approximated standard errors. With that said, a drawback of using recursively approximated standard errors is the requirement of a relatively larger number of main MH-RM iterations (at least 1000 in our experience) to reach a positive definite approximate observed information matrix. For this reason, post-convergence approximated standard errors are adopted for the remaining simulations in this study since this approach gives proper standard error estimates for structural parameters and can be faster.

Finally, 95% confidence intervals for each parameter were calculated. The post-convergence approximated standard errors were used to form these two-sided Wald-type confidence intervals. The percentages of intervals that cover the generating values are reported in the last column of Table 2. Based on the 100 replications performed, coverage of structural parameters appears well calibrated in general. For measurement parameters, the coverage rates tend to decrease as the magnitude of parameters becomes larger. Coverage rates are at the lowest for the more extreme intercept parameters due to their underestimated standard errors.

4.2 Simulation Study 2: Comparison of Models

The second simulation study was conducted to examine how measurement error and sampling error may influence compositional effect estimation across different conditions with both a traditional HLM and a multilevel latent variable model.

4.2.1 Simulation Conditions

A total of 30 data generating conditions were examined: 2 compositional effect sizes, $\times$ 3 sampling conditions $\times$ 2 ICC sizes $\times$ 2 measurement condition + 6 conditions for a model with no compositional effect.

First, two different sizes of compositional effect were considered in this study. The generating value of γ_{01} was 1.0. The generating value of γ_{10} was either 0.5 or 0.8, giving a compositional effect of 0.5 or 0.2, respectively. Second, the combination of large (ng=100, np=20) and small (ng=25, np=5) numbers of groups and individuals makes a

total of 4 different sampling conditions. However, the combination of 25 groups and group size of 5 leads to too small a total sample size (125), which is not entirely appropriate for the stable estimation of a high-dimensional latent variable model. Therefore, only three different sampling conditions were used for this simulation study. For latent predictor ICC levels, 0.1 and 0.3 were used to generate small- and a large-ICC conditions by manipulating ψ – the variance of $\xi_{.j}$. Finally, two different measurement structures were considered. The observed variables in the first condition were dichotomous and in the second condition, they were 5-category ordinal responses. The true item parameters are given in in Table 3. Additionally, data were generated from a model with no compositional effect ($\gamma_{01} = \gamma_{10}$) with the first measurement condition and analyzed to examine empirical Type I error rates for the compositional effect estimates with both the traditional model and the latent variable model. In each condition, 100 Monte Carlo replications were attempted.

4.2.2 Analysis

Because all simulated data sets have the true generating values of η_{ij} and ξ_{ij}, these values (true scores) can be analyzed using a traditional model. The resulting parameter estimates can be considered gold standard estimates that are influenced only by sampling fluctuations but not by measurement conditions. Therefore, 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 HLM, which is treated as the gold standard, 2) estimates obtained by applying latent variable model, and 3) the estimates from analyzing the observed summed scores of outcomes and predictors with the standard approach using manifest variables. All of the traditional HLM analyses were conducted using an R package *nlme* (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2012).

4.2.3 Evaluation Statistics

To compare these three sets of estimates, three statistics are calculated: 1) the percentage bias of the estimate relative to the magnitude of its generating value, 2) the observed

coverage rate of the 95% confident interval, and 3) the observed power to detect the compositional effect of interest as significant.

It should be noted that the regression coefficient estimates from the observed sum score analysis using a traditional multilevel model are not on the same scales as those obtained using the latent variable approach, which yields naturally standardized fixed effects coefficients due to the identification conditions discussed earlier. To make the coefficient estimates more comparable, the estimates from the traditional HLM approach were standardized by multiplying the parameter estimates by the ratio of standard deviation of the predictor to the standard deviation of the outcome.

4.2.4 Results

Convergence rates and mean computing time across generating data conditions are reported in Table 4. Only converged replications were used to calculate evaluation statistics. The worst cases of non-convergence occur when the number of level-2 units is low and the ICC is small. This is particularly true for the second measurement condition when a substantially larger number of item parameters for the multiple-categorical items must be estimated from the data.

Let us examine the first measurement condition where all items are dichotomously scored. Because a compositional effect estimate is defined as the difference between $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$, those two parameter estimates are examined together, along with the compositional effect estimate (the difference) itself. Relative percentage biases in $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$ are summarized in Figure 2. When the generating values of η_{ij} and ξ_{ij} were analyzed, the bias of $\hat{\gamma}_{01}$ ranged from 1% to 15% across the sampling conditions. Latent variable modeling resulted in a similar magnitude of bias. But traditional HLM resulted in more substantial bias in both $\hat{\gamma}_{01}$ and $\hat{\gamma}_{10}$ (from 30% to 70%) (see the gray bars in Figure 2).

The biases in the traditional HLM estimates of the regression coefficients lead to an interesting pattern of biases in the compositional effect estimate. The bias can be as small as 8% when the predictor ICC is large and the sampling condition favorable (more

individuals in each group), but the bias can be as large as 80% when the ICC is small and the group size is small (see Figure 3). It is also noteworthy that the bias in the compositional effect estimate from the traditional HLM model can also be positive when the ICC is large and the contextual effect size is small (see the last plot of Figure 3).

On the other hand, comparing the two plots in Figure 4 with the first two plots in Figure 3 reveals that the performance of the traditional HLM and the latent variable model in terms of estimating $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, as well as the compositional effect, is highly similar across the two measurement conditions. This indicates the measurement model is a less influential source of bias in this study.

To examine the standard error estimates, the coverage rates of the 95% confidence intervals for the true compositional effect were calculated. Results from the condition with large true compositional effect and the first measurement condition are summarized in Figure 5. When generating values are analyzed, the coverage rates across sampling conditions are generally close to 95%, except for the case where ICC is small and the number of groups is also small. In this case the coverage rate can be as low as 85%. The coverage rates based on the latent variable model parameter estimates were similar or slightly worse than those from generating value analysis. Coverage rates with traditional HLM estimates can be problematic when the number of individuals per group and the ICC are both low.

To examine how researchers can make different inferential decisions when they apply a traditional model and a latent variable model, the empirical Type I error rates are calculated for the conditions where the true data generating model has zero compositional effect. Figure 6 shows empirical Type I error rates across ICC and sampling conditions for the first measurement condition.

Generating value analysis yields Type I error rates of .05 to .07 across sampling conditions. The latent variable model maintains similar Type I error rate calibration, except for the cases when the number of individuals per group is small. For traditional HLM

analysis, Type I error rate inflation is dramatic. Only under the conditions when a small predictor ICC is coupled with a small number of group or a small number of individuals per group, does the traditional method maintains proper Type I error rate.

Turning to statistical power, 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 also large. When ICC is small, power decreases to .35 even with favorable sampling conditions. The lowest statistical power (.15) is found when predictor ICC is small and the number of groups is also small.

The patterns are similar for the latent variable analysis. But when ICC is small, and the number of individuals per group or the number of groups is small, latent variable modeling actually yields a slightly higher percentage of significant compositional effects. While the traditional HLM analysis yields a very high percentage of significant compositional effects when the ICC is large and the number of individuals per group is also large, the power decreases remarkably when ICC is small and when the sampling condition deteriorates (i.e., when the number of individuals per group or the number of groups is small). Also, the relatively high statistical power associated with the traditional HLM analysis is partially attributable to the inflated Type I error rates observed earlier - the test is liberal overall.

In summary, relative bias of $\hat{\gamma}_{01}$ are $\hat{\gamma}_{10}$ are large when the traditional HLM is applied. This is consistent with findings from previous research. However, the relative bias in the difference between the two coefficients (the compositional effect estimate) can sometimes be kept at bay, since both coefficients can be biased in the same direction. We note that the true compositional effect can be estimated with traditional methods when ICC is large and the sampling condition is favorable. However, Type I error rates are severely inflated under this very condition, when the true compositional effect is zero. Thus this model can frequently make the false claim that there is a significant compositional effect even when there is none.

On the other hand, biases in point estimates seems rather unavoidable when the sampling condition is not favorable (small group sizes and low sample size in general), and especially when ICC is also low. Even with generating true scores the estimates show some biases. However, the latent variable model tends to yield less biased estimates in general. When ICC is small and the number of individuals per group is small, the Type I error rate associated with the latent variable compositional effect estimate increases slightly, but the magnitude of the elevation is still much more preferable compared to the traditional HLM analysis. We also find that the main issue with the latent variable model approach in terms of sampling conditions is related more to small number of groups rather than to the number of individuals per group. As long as the number of groups sampled is sufficiently large, the performance of the latent variable modeling approach can be satisfactory.

Finally, we find that the measurement structure to be less influential in this study. It certainly may be due to the particularly set of item parameters chosen. The results from the second measurement condition, however, indicate that the estimation of too many item parameters with limited sample size can possibly undermine the performance of the latent variable modeling approach.

5 Empirical Application: "Big-fish-little-pond" Effect

5.1 Data

For this compositional effect demonstration, a subset of publicly available data from The Programme for International Student Assessment (PISA 2000; OECD, 2000) were extracted and analyzed. PISA is a large international comparative survey. A large amount of student and school level information that covering cognitive and affective domains was collected with a complex sampling scheme.

Though 42 countries participated in the data collection, a sample of students from the US was analyzed in this study for the purpose of illustration only. Originally, a total of 129 reading items were administered to estimate country level reading literacy using

a balanced incomplete block design. However, for simplicity, only booklets 8 and 9 were used for this analysis. These two booklets included 33 reading items, but 1 item was dropped prior to our analysis because all item responses to this item were scored as incorrect, which meant that the item contributes no information. Therefore, the analysis data set contained responses to 32 reading items (3 ordinal items with 3 categories each and 29 dichotomous items) from 667 students nested within 141 schools in the US. The number of students within a school ranged from 1 to 8 in this analysis data set. The outcome variable is the students' *self concept in reading*. It was measured by three items (CC02Q05, CC02Q09, and CC02Q23). Each item has a Likert-type scale, ranging from 1 (disagree) to 4 (agree).

5.2 Results

The structural parameter estimates from the multilevel latent variable analysis (EM algorithm and the MH-RM algorithm) and traditional HLM analysis are summarized in Table 5. In general, a positive and significant within-school coefficient $\hat{\gamma}_{10}$ is found across different models and algorithms. The between-school coefficient estimate ($\hat{\gamma}_{01}$) was not statistically significantly different from 0 when the multilevel latent model was applied (with EM or MH-RM algorithm), while the estimate was significantly different from 0 when the traditional HLM was applied.

The compositional "big-fish-little-pond" effect is calculated by subtracting $\hat{\gamma}_{10}$ from $\hat{\gamma}_{01}$. The direction of the compositional was negative. This is consistent with reports from previous research (Marsh et al., 2009). It indicates that two students who have the same levels of reading achievement can have different level of academic self-concept, depending on school-level academic achievement. As the compositional effect is negative, the students who attends a higher achieving school tend to have lower academic self-concept when compared with a student who attends a lower achieving school. On the other hand, a student who belongs to a lower achieving school is expected to have higher academic self-concept when compared with a student who belongs to a higher

achieving school – just like a fish that feels big if the pond in which it lives is small.

However, in terms of the statistical significance of the compositional effect, the effect is not significantly different from 0 if we use the estimates and standard errors from the traditional HLM; but if we use the latent variable estimates, the compositional effect is significant. This result is consistent with what we found via the simulation study in that the power of the latent variable model to detect a compositional effect is higher than that of the traditional method, when the data set is associated with a sufficiently large number of schools and a small number of students per school.

Finally, the item parameter estimates from the MH-RM algorithm are plotted against those from the EM algorithm in Figure 8. As can be seen, the estimates are very close. Standard errors of the item parameters exhibited a similar pattern as found previously (see Figure 9), confirming that the post-convergence approximation method yields slightly smaller standard errors, while the recursive approximation tends to yield larger standard errors.

6 Conclusion

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

The primary purpose of this study was to improve estimation efficiency in obtaining

maximum likelihood estimates of contextual effects by adopting the MH-RM algorithm (Cai, 2008, 2010a, 2010b). R programs 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. Results indicate that the MH-RM algorithm can obtain maximum likelihood estimates and their standard errors efficiently. Considering the difference between an interpreted language (R) and a compiled language (FORTRAN) in which EM is implemented, substantial improvement in efficiency is expected if the MH-RM estimation code is ported to a compiled language in the future.

The second purpose of this study was to provide information about the performance of the nonlinear multilevel latent variable model in comparison to traditional HLM through a simulation study that covers various sampling and measurement 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. Type I error rates of the compositional effect estimate from the traditional model can also be substantially elevated whereas latent variable modeling leads to more proper Type I error rate calibration.

The third purpose of this study was to provide an empirical illustration using a subset of data extracted from PISA (Adams & Wu, 2002). A negative compositional effect was found for the relationship between reading literacy and academic self-concept, supporting the results from previous studies, on the "Big-fish-little-pond" effect (e.g. Marsh et al., 2009). The compositional effect was statistically significant at the .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.

This study is limited several important ways. The latent variable model itself contains a series of strong specification and distributional assumptions. These assumptions require careful checking in empirical settings because the violations of these assumptions

can lead to substantial unknown estimation biases. The simulation study only examined a limited set of conditions with fixed item and structural parameters. The data generating and fitted models in the simulation study also do not contain any model specification error. More complex structural models should also be considered. In future research, an obvious extension of the model discussed here is one that includes cross-level interactions in latent variables.

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 & M. 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 Algo-

rithm. *Journal of the Royal Statistical Society*, *44*(2), 226-233.

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.

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.

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.

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, 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: Muthèn & Muthèn.

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.

von Davier, M., & Sinharay, S. (2010). Stochastic approximation for latent regression item response models. *Journal of Educational and Behavioral Statistics*, *35*(2), 174-193.

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

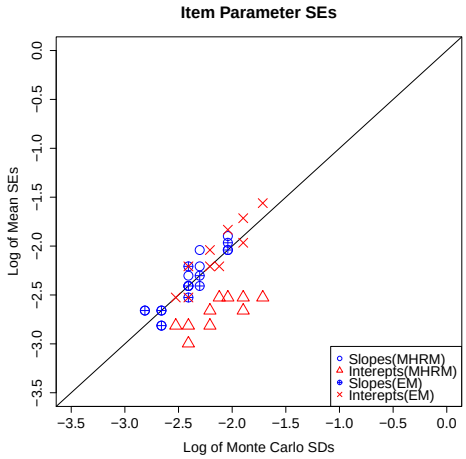

Figure 1: Comparisons of standard errors for item parameters.

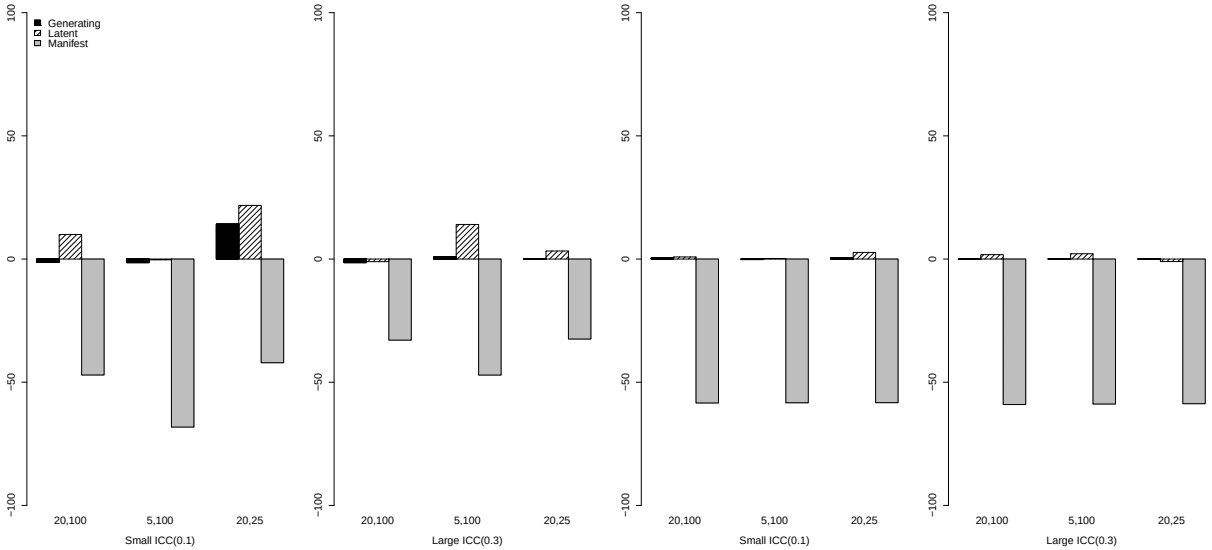

Figure 2: Relative percentage bias in $\hat{\gamma}_{01}$ (first two plots) and $\hat{\gamma}_{10}$ (last two plots), large true compositional effect, measurement condition 1, by the sampling conditions (number of individuals in each group, and number of groups).

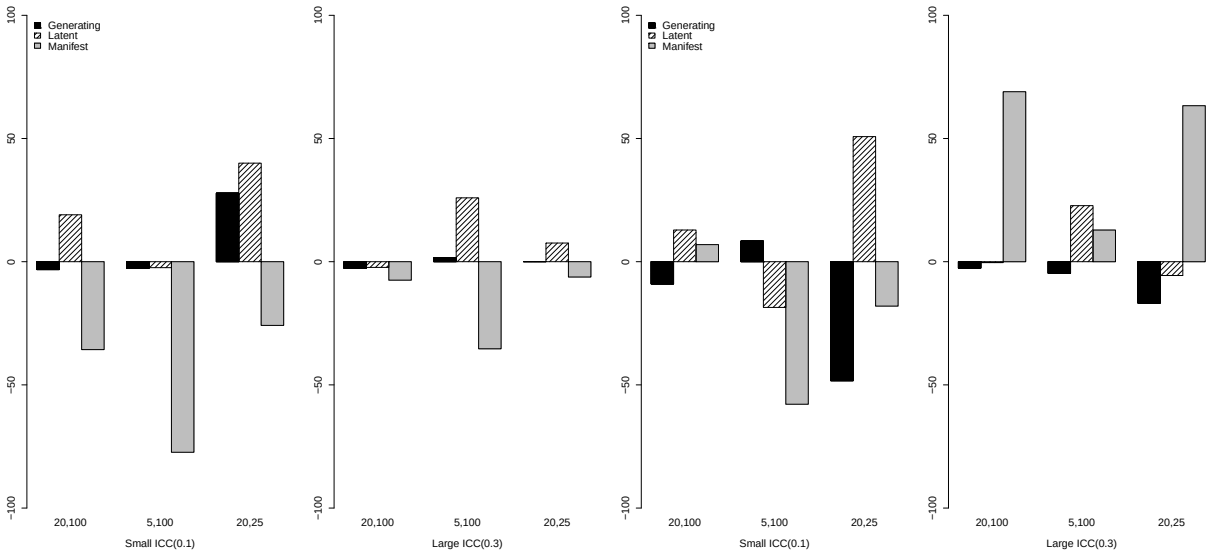

Figure 3: Relative percentage bias in compositional effect estimate $\hat{\gamma}_{01} - \hat{\gamma}_{10}$, large true compositional effect (first two plots) and small true compositional effect (last two plots), first measurement condition, by the sampling conditions (number of individuals in each group, and number of groups).

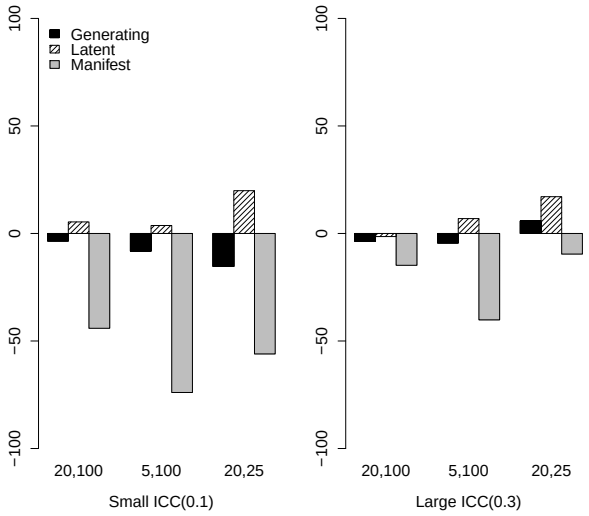

Figure 4: Relative percentage bias in compositional effect estimate $\hat{\gamma}_{01} - \hat{\gamma}_{10}$, large true compositional effect, second measurement condition, by the sampling conditions (number of individuals in each group, and number of groups).

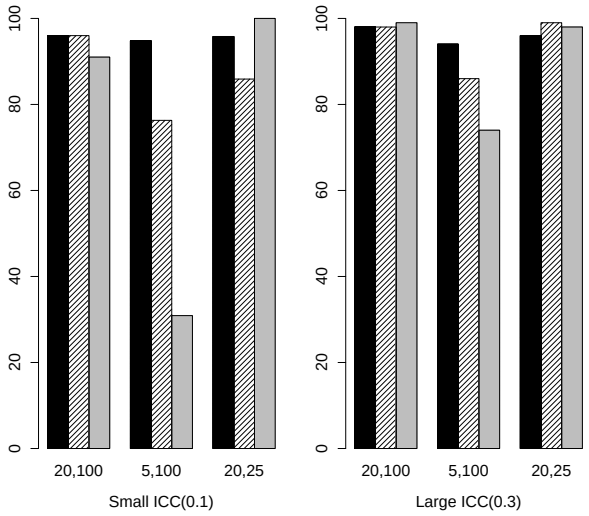

Figure 5: 95% compositional effect estimate confidence interval coverage rate, large true compositional effect, first measurement condition, by the sampling conditions (number of individuals in each group, and number of groups).

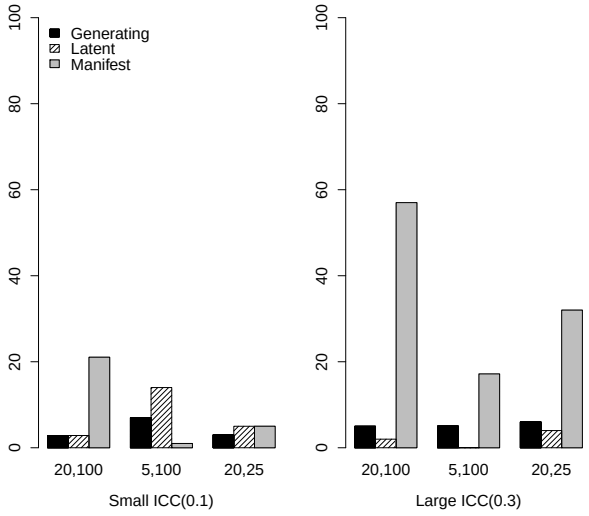

Figure 6: Empirical Type I error rates for the compositional effect estimate, first measurement condition.

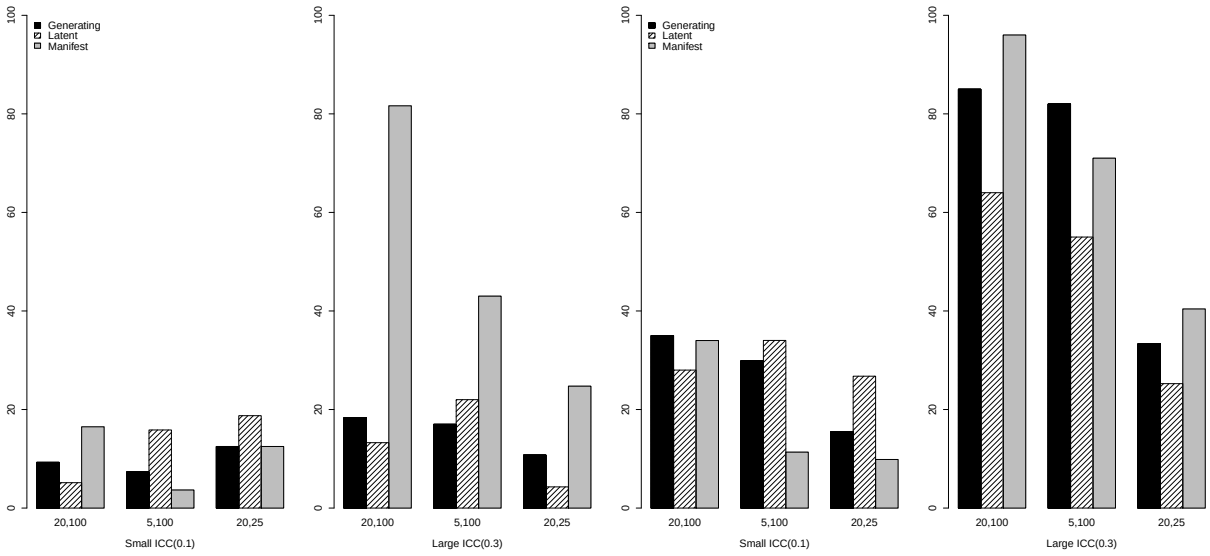

Figure 7: Percentage of significant compositional effect (estimated power), small true compositional effect (first two plots) and large true compositional effect (last two plots), first measurement condition.

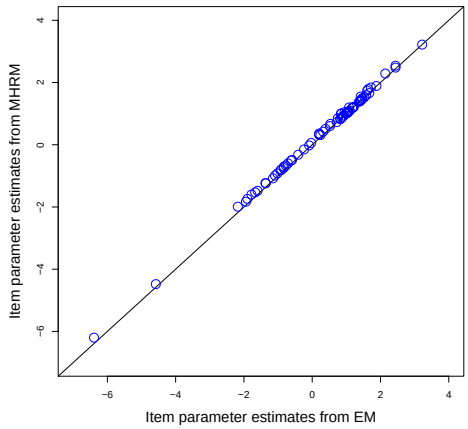

Figure 8: Item parameter estimates based on the EM and MH-RM algorithms for the PISA 2000 USA data analysis.

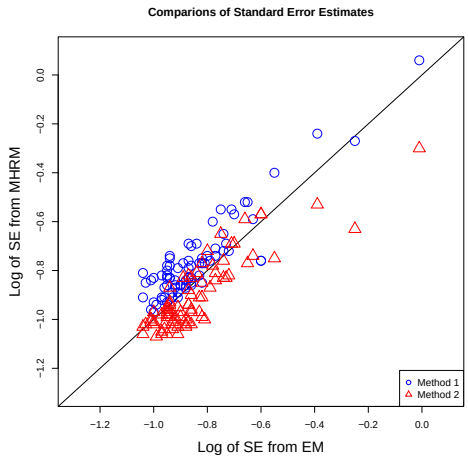

Figure 9: Standard errors of item parameters based on the EM and MH-RM algorithms for PISA 2000 USA data analysis. Method 1 uses recursively approximated standard errors. Method 2 uses post-convergence approximated standard errors.

Table 1: Generating values, EM estimates, and MH-RM estimates for a compositional effect model

Structural Parameters							
		EM (5qp)		EM (14qp)		MH-RM	
	θ	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	E($\hat{\theta}$)	E{se($\hat{\theta}$)}	E($\hat{\theta}$)	E{se($\hat{\theta}$)}
γ_{01}	1.00	1.02	0.19	1.01	0.19	1.00	0.18
γ_{10}	0.50	0.52	0.05	0.51	0.05	0.52	0.09
τ_{00}	1.00	0.90	0.16	0.91	0.17	0.93	0.16
ψ	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_{y1}	-0.80	-0.80	0.11	-0.80	0.11	-0.81	0.06
c_{y2}	0.00	0.01	0.13	0.01	0.13	0.00	0.05
c_{y3}	1.20	1.19	0.16	1.19	0.16	1.18	0.08
c_{y4}	-0.70	-0.74	0.18	-0.74	0.18	-0.75	0.07
c_{y5}	0.80	0.79	0.21	0.79	0.21	0.78	0.08
Computational Efficiency							
one processor		5∼7 min		60∼100min		35∼40min	

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 parameters; c = item threshold parameters.

Table 2: Generating values and MH-RM estimates for a compositional effect model

	θ	E($\hat{\theta}$)	E{se1($\hat{\theta}$)}	SD($\hat{\theta}$)	E{se2($\hat{\theta}$)}	95% CI Coverage Using se1
Structural Parameters						
γ_{01}	1.00	0.99	0.17	0.19	0.18	95.0
γ_{10}	0.50	0.50	0.06	0.07	0.09	95.0
τ_{00}	1.00	0.97	0.20	0.18	0.16	89.0
ψ	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
Computational 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; E{se2($\hat{\theta}$)} = mean of post-convergence approximated standard errors; SD($\hat{\boldsymbol{\theta}}$) = Monte Carlo standard deviation of point estimates; 95% confidence interval coverage rate using post-convergence approximated standard errors; a = item slope parameters; c = item threshold parameters.

Table 3: Conditions of measurement models and generating values for item parameters

Measurement Model 1 (MM 1)		
Condition	ξ_{ij} indicators X1∼X5 ($K_l = 2$)	η_{ij} indicators Y1∼Y5 ($K_l = 2$)
	slope (a_l)	intercept (c_l)
X1, Y1	0.8	-1.0
X2, Y2	1.0	0.0
X3, Y3	1.2	1.0
X4, Y4	1.4	-0.5
X5, Y5	1.6	0.5
Measurement Model 2 (MM 2)		
Condition	ξ_{ij} indicators X1∼X5 ($K_l = 5$)	η_{ij} indicators Y1∼Y5 ($K_l = 5$)
	slope (a)	intercepts ($c_{1,l}, c_{2,l}, c_{3,l}, c_{4,l}$)
X1, Y1	0.8	-1.0, 0.0, 1.0, 2.0
X2, Y2	1.0	-1.0, 0.0, 1.0, 2.0
X3, Y3	1.2	-1.0, 0.0, 1.0, 2.0
X4, Y4	1.4	-1.0, 0.0, 1.0, 2.0
X5, Y5	1.6	-1.0, 0.0, 1.0, 2.0

Table 4: Percentage of converged solution and average time per replication (in seconds)

	Large Compositional Effect = 0.5			
	np=20		np=5	
ng=100	MM1	MM2	MM1	MM2
ICC=0.1	100(2781)	89(4911)	97(972)	81(1593)
ICC=0.3	100(2657)	95(5301)	100(955)	95(1613)
ng=25	MM1	MM2	MM1	MM2
ICC=0.1	98(1046)	92(1522)	N/A	
ICC=0.3	99(865)	93(1524)		
	Small Compositional Effect = 0.2			
	np=20		np=5	
ng=100	MM1	MM2	MM1	MM2
ICC=0.1	97(2937)	91(5165)	95(1021)	92(1588)
ICC=0.3	98(1785)	92(4910)	100(1046)	91(1593)
ng=25	MM1	MM2	MM1	MM2
ICC=0.1	95(919)	78(1521)	N/A	
ICC=0.3	93(915)	95(1519)		

Note. MM1 = Measurement model 1; MM2 = Measurement model 2; ng = number of groups; np = number of individuals per group.

Table 5: Structural parameter estimates from PISA 2000 USA data analysis

	Multilevel latent variable model						Manifest variable HLM		
	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
γ_{10}	0.42	0.06	7.17	0.42	0.05	7.92	0.11	0.01	7.75
γ_{01}	0.16	0.11	1.43	0.18	0.11	1.68	0.07	0.02	3.60
τ_{00}	0.47	0.11	0.39	0.47	0.11	4.28	0.37	–	190.31$*$
ψ	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

Note. Reported standard errors for MH-RM algorithm are from recursively approximated observed data information matrix. $*$ The HLM software program produces a χ^2 test for the variance component τ_{00}.