Correction for Item Response Theory Latent Trait Measurement Error in Linear Mixed Effects Models

Chun Wang[1], Gongjun Xu[2], Xue Zhang[3]

1. University of Washington
2. University of Michigan
3. Northeast Normal University

Correspondence concerning this manuscript should be addressed to Chun Wang at:

312E Miller Hall
Measurement & Statistics
College of Education
Box 353600
Seattle, WA 98195-3600
e-mail: wang4066@uw.edu
phone: 206-616-6306

Acknowledgement: This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305D170042 (or R305D160010) awarded to the University of Washington (University of Minnesota originally). The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Citation: Wang, C., Xu, G., & Zhang, X. (2019). Correction for item response theory latent trait measurement error in linear mixed effects models. *Psychometrika*, *84*, 673-700.

Correction for Item Response Theory Latent Trait Measurement Error in Linear Mixed Effects Models

Abstract

When latent variables are used as outcomes in regression analysis, a common approach that is used to solve the ignored measurement error issue is to take a multilevel perspective on item response modeling (IRT). Although recent computational advancement allow efficient and accurate estimation of multilevel IRT models, we argue that a two-stage divide-and-conquer strategy still has its unique advantages. Within the two-stage framework, three methods that take into account heteroscedastic measurement errors of the dependent variable in stage II analysis are introduced, they are the closed-form marginal MLE (MMLE), the Expectation Maximization (EM) algorithm, and the moment estimation method. They are compared to the naïve two-stage estimation and the one-stage MCMC estimation. A simulation study is conducted to compare the five methods in terms of model parameter recovery and their standard error estimation. The pros and cons of each method are also discussed to provide guidelines for practitioners. Finally, a real data example is given to illustrate the applications of various methods using the National Educational Longitudinal Survey data (NELS 88).

It is not uncommon to have latent variables as dependent variables in regression analysis. For instance, the item response theory (IRT) scaled θ score is often used as an outcome measure to make high-stakes decisions such as evaluating performance of individual teachers or schools. However, there exist potential errors in estimating the latent θ scores (or any other latent variables from factor analysis perspective), and ignoring the measurement errors will adversely bias the subsequent statistical inferences (Fox & Glas, 2001, 2003). In particular, measurement error can diminish the statistical power of impact studies, yield inconsistent or biased estimates of model parameters (Lu, Thomas, & Zumbo, 2005), and weaken the ability of researchers to identify relationships among different variables affecting outcomes (Rabe-Hesketh & Skrondal, 2004). The consequence can be especially severe when the sample size is small, the hierarchical structure is sparsely populated, or when the number of items is small (e.g., Zwinderman, 1991).

When measurement error follows a normal distribution with a constant variance, correcting for the error can be easily handled via reliability adjustment (e.g., Bollen, 1989; Hsiao, Kwok, & Lai, 2018). The main challenge of having IRT θ score as dependent variable is that the measurement error in $\hat{\theta}$ is heteroscedastic with its variance depending on true θ. With the growing computational power nowadays, a recommended approach to address the measurement error challenge is to use an integrated multilevel IRT model (Adams et al., 1997; Fox & Glas, 2001, 2003; Kamata, 2001; Pastor & Beretvas, 2006; Wang, Kohli, & Henn, 2016) such that all model parameters are estimated simultaneously. This unified one-stage approach incorporates the standard errors of the latent trait estimates into the total variance of the model, avoiding the possible bias when using the estimated θ as the dependent variable in subsequent analysis.

Despite the statistical appeal of the one-stage approach, we advocate that a "divide-and-conquer" two-stage approach has its practical advantages. In the two-stage approach, an

appropriate measurement model is first fitted to the data, and the resulting $\hat{\theta}$ scores are used in subsequent analysis. This idea is in the same spirit as "factor score regression" proposed decades ago (Skrondal & Laake, 2001). The benefit of this approach includes *clearer definition of factors*, *convenience for secondary data analysis*, *convenience for model calibration and fit evaluation*, and *avoidance of improper solutions*. Indeed, it is known that unless an adequate number of good indicators of each latent factor are available, improper solutions (a.k.a., Heywood cases, negative variance estimates) can occur. Anderson and Gerbing (1984) found that with correct models, their simulation showed 24.9% of replications had improper solutions. With improper solutions, test statistics no longer have their assumed distributions, and consequently statistical inference and model evaluation become difficult (e.g., Stoel, Garre, Dolan, & van den Wittenboer, 2006).

Moreover, it has been known that partial misspecification in a model causes large bias in the estimates of other free parameters in structural equation modeling (SEM). In the presence of misspecification, a one-step approach will suffer from *interpretational confounding* (Burt, 1973, 1976), which refers to the inconsistency between the empirical meaning assigned to an unobserved construct and the a priori meaning of the construct. The potential for interpretation confounding is minimized when the two-step estimation approach is employed (Anderson & Gerbing, 1988). Furthermore, the specification errors in particular parts of an integrated model can be isolated by using the separate estimation approach.

Another compelling argument in support of two-stage estimation is the *convenience for secondary data analysis*. In a large-scale survey such as NAEP or NELS88, usually hundreds of test items and educational, demographic, and attitudinal variables are included, such that droves of descriptive statistics, multiple regression analyses, and SEM models might be entertained. In

this case, neither carrying out all of these analyses nor providing sufficient statistics for them is feasible. Oftentimes, these survey data provide either item parameters, or estimated θ's along with their standard errors. Hence, the methods introduced in this paper will come in handy to handle secondary data analysis with limited available information.

In this paper, we investigate different methods of addressing the measurement error challenge within a two-stage framework. These methods will be compared to the naïve two-stage method and an integrative one-stage Markov chain Monte Carlo (MCMC) method (Fox & Glas, 2001, 2003; Wang & Nydick, 2015) in a simulation study. We intend to show that the proposed two stage methods outperform the naïve method and they produce comparable results to the MCMC method.

1. Literature Review

With the advent and popularity of Item Response Theory (IRT), the IRT-based scaled scores (i.e., θ) has been widely used as an indicator of different latent traits, such as academic achievement in education. Hence, θ is treated as a dependent variable in various statistical analysis, including simple descriptive statistics (Fan, Chen, & Matsumoto, 1997), two sample t-test (Jeynes,1999), multiple regression (Goldhaber & Brewer, 1997; Nussbaum, Hamilton, & Snow, 1997), analysis of variance (ANOVA, Cohen, Bottge, & Wells, 2001), linear mixed models (Hill, Rowan, & Ball, 2005), hierarchical linear modeling (Bacharach, Baumeister, & Furr, 2003), and latent growth curve modeling (Fraine, Damme, & Onghena, 2007). In all these cited studies, θ scores were first obtained from separate IRT model fitting, and then they were used as variables in different statistical models as if they were "true" values without measurement errors. Complications arise, however, if the latent θ scores were estimated with non-ignorable measurement errors.

If a linear test when a fixed number of items is given to students, the resulting measurement error (or standard error, SE) typically follows a bowl shape with SE being smaller when the true latent trait is in the middle (e.g., Kolen, Hanson, & Brennan, 1992; Wang, 2015) of the θ scale. When an adaptive test is given to students, the resulting SE is more of a uniform shape (e.g., Thompson & Weiss, 2011; van der Linden & Glas, 2010). The differential SE, depending on the true θ level and test mode, complicates the treatment of measurement error issue in the subsequent statistical analysis.

There are quite a few studies that have accounted for the measurement errors in $\hat{\theta}$ assuming a constant measurement error term. In other words, simple measurement error models precipitate corrections to estimate "true" variances and correlations from their "observed" counterparts. For instance, Hong and Yu (2007) analyzed the Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K) data using a multivariate hierarchical model to study the relationship between early-grade retention and children's reading and math learning. Let Y_{tij} denote child *i*'s T-score [1]in school *j* in Year *t*, then the level-1 model in their analysis is generically expressed as

$$Y_{tij} = T_{tij} + e_{tij}, \quad e_{tij} \sim N(0, \sigma_t^2). \tag{1}$$

The test reliability was then used to compute the error variance σ_t^2 in each year. Although correctly accounting for measurement error improves the estimation precision, this treatment overlooks the fact that the measurement error of IRT θ scores is not constant across the θ scale. A statistically sound approach that follows through the assumption of IRT is to let $e_{tij} \sim N(0, \sigma_{tij}^2)$, however, the relaxation of the common variance assumption in Equation (1)

[1] The T-score is a standardized score, which was in fact a transformation of an IRT θ score.

imposes computational complexity to the model. The objective of this paper, therefore, is to investigate methods for addressing challenging measurement error issues in the two-stage approach. We need to acknowledge that this paper only focuses on the measurement errors occurred on the dependent variables, whereas there is extensive literature on dealing with measurement errors in covariates (i.e., independent variables). Methods for the latter scenario may include the method-of-moment (Carroll, et al., 2006; Fuller, 2006), simulation-extrapolation (Carroll et al., 2006; Devanarayan & Stefanski, 2002), and latent regression (Bianconcini & Cagnone, 2012; Bollen, 1989; Skrondal & Rabe-Hesketh, 2004). For a comparison of methods, please refer to Lockwood and McCaffrey (2014).

The rest of the paper is organized as follows. First, we will introduce the multilevel model that is considered throughout the study. In other applications, both the measurement model and the structural model can take other forms as long as the latter is a linear mixed effects model, and all methods introduced in the paper still apply. Second, four different methods are introduced within the two-stage framework, including a naïve method. Then, a simulation study is designed to evaluate and compare the performance of different methods, followed by a real data example. A discussion is given in the end that summarizes the pros and cons of each method.

2. Models

The model is comprised of two main levels, the measurement model and structural model. In this paper, we will focus specifically on the linear mixed effects model (LME) as the structural model in stage II inference. In particular, we will base the discussion on the scenario of longitudinal assessment, i.e., modeling individual and group level growth trajectories of student latent abilities over time via the latent growth curve model (LGC). Because the LGC model

belongs to the family of LME models, the methods introduced in this paper can be easily applied in all specific types of LME models for different nested structures.

At the measurement model level, the three-parameter logistic (3PL) model (Baker & Kim, 2004) is considered. The probability for a correct response y_{ijt} at time t (t = 1,..., T) for item j (j = 1,..., J) and person i (i = 1,..., N) can be written as

$$P(y_{ijt}=1\left|\theta_{it},a_{jt},b_{jt},c_{jt}\right.)=c_{jt}+(1-c_{jt})\frac{1}{1+\exp\left[-D(a_{jt}\theta_{it}-b_{jt})\right]}, \quad (2)$$

where D is a scaling constant that usually set to be 1.7. a_{jt}, b_{jt}, and c_{jt} are the discrimination, difficulty, and pseudo-guessing parameter of item j at time t , and θ_{it} is the ability of person i at time t . In longitudinal assessment, although the item parameters could differ across time (i.e., the subscript t is embedded for item parameters in (2)), anchor items need to be in place to link the scale across years (e.g., Wang, Kohli, Henn, 2016).

In the structural model level, we have a LME model with $\boldsymbol{\theta}_i$ as dependent variables written as follows

$$\boldsymbol{\theta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}_i + \boldsymbol{e}_i \ . \quad (3)$$

Considering the LGC model as a special case of (3), if assuming a unidimensional θ_i is measured per time point, then both $\boldsymbol{\theta}_i$ and $\boldsymbol{e}_i$ are T-by-1 vectors. $\boldsymbol{X}_i$ and $\boldsymbol{Z}$ are the T-by-p and T-by-q design matrices, and $\boldsymbol{\beta}$ and $\boldsymbol{u}_i$ are p-by-1 and q-by-1 vectors denoting fixed and random effects respectively. T is the total number of time points. In a more general case, $\boldsymbol{Z}$ can also differ across individuals ($\boldsymbol{Z}_i$).

For the rest of the paper, we consider a simplest linear growth pattern, i.e., $\boldsymbol{X} \equiv \boldsymbol{Z} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{bmatrix}$. But the methods discussed can be easily generalized to the conditions when $\boldsymbol{X}$ and $\boldsymbol{Z}$ differ. For instance, if one is interested in the treatment effect, and let g_i denote the observed covariate of treatment, with $g_i = 1$ indicating person *i* belongs to the treatment group, and 0 otherwise. Then $\boldsymbol{X}_i$ is updated as $\boldsymbol{X}_i = (\boldsymbol{Z}, g_i \times \mathbf{1}_{1\times 4})$ whereas $\boldsymbol{Z}$ stays the same. Similarly, if one is interested in the treatment by time interaction, then $\boldsymbol{X}_i = (\boldsymbol{Z}, g_i \times [0, 1, \dots, T-1]^{\mathrm{t}})$ where the superscript "t" denotes the transpose throughout the paper.

The random effects, $\boldsymbol{u}_i$, are typically assumed to follow multivariate normal distribution,

$$\boldsymbol{u}_i = \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim MVN\left(\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_u = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}\right),$$

and for simplicity, we assume an independent error structure, i.e., $e_{it} \sim N(0, \sigma^2)$.

If a multivariate latent trait (i.e., *D* dimensions) is measured at each time point, let $\boldsymbol{\theta}_i = [\theta_{i11}, \dots, \theta_{i1T}, \dots, \theta_{iD1}, \dots, \theta_{iDT}]^t$ with the first *T* elements refer to the latent trait at dimension 1 across *T* time points, Equation (3) still holds. But $\boldsymbol{X}$ becomes a (*D*×*T*) -by-(*D*×*2*) matrix taking the form of $\boldsymbol{I}_D \otimes \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & T-1 \end{bmatrix}$, where $\boldsymbol{I}_D$ is an identity matrix of size *D*-by-*D*, and $\otimes$ is the kronecker product. $\boldsymbol{\beta} = (\beta_{01}, \dots, \beta_{0D}, \beta_{11}, \dots, \beta_{1D})^t$ and $\boldsymbol{u}_i = (u_{i01}, \dots, u_{i0D}, u_{i11}, \dots, u_{i1D})^t$ both become (*D*×*2*) -by-1 vectors of fixed and random effects respectively.

3. Model Estimation

3.1 Unified One-stage Estimation

To estimate the multilevel IRT model simultaneously, the current available estimation methods include, but are not limited to, the generalized linear and nonlinear methodologies described in De Boeck and Wilson (2004), the generalized linear latent and mixed model framework of Skrondal and Rabe-Hesketh (2004), Bayesian methodology of Lee and Song (2003) including the Gibbs sampler and Markov chain Monte Carlo (MCMC, Fox & Glas, 2001, 2003; Fox, 2010). These methods are suitable for a general family of models allowing linear/nonlinear relations among normal latent variables and a variety of indicator types (e.g., ordinal, binary).

Among them, the first two approaches require numerical integration and calculation of the likelihood, which becomes computationally prohibitive or even impossible when the model is complex or the number of variables is large. Rabe-Hesketh and Skrondal (2008) admitted that "*estimation can be quite slow, especially if there are several random effects*". The Bayesian approach requires careful selection of prior distributions for each parameter, which might not come naturally for researchers who are unfamiliar with Bayesian methods. Other methods that supposedly alleviate the high-dimensional challenge (von Davier & Sinharay, 2007) include: adaptive Gaussian quadrature (Pinheiro & Bates, 1995), limited-information weighted least squares (WLS), and graphical models approach (Rijmen, Vansteelandt, & De Boeck, 2008). All of these methods have proven to work well in respective studies. Even so, a divide-and-conquer two-stage estimation approach still has its own advantages (e.g., reasons presented at the beginning) and it is the main focus of this paper. Given the flexibility MCMC offers to deal with the 3PL model, we will use it as a comparison to the two-stage estimation methods.

3.2 Two-stage Estimation

Let $\boldsymbol{\Psi} = (\boldsymbol{\beta}, \sum_u, \sigma^2)$ denote the set of structural parameters of interest, and let $\boldsymbol{\vartheta}_M = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ denote the set of item parameters pertaining only to the measurement part of the integrated model (Skrondal & Kuha, 2012). Throughout this paper, we assume the item parameters $\boldsymbol{\vartheta}_M = (\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c})$ are known to alleviate any propagation of errors (such as sampling error) from item parameter calibration. For readers who are concerned about item calibration errors, please refer to the method proposed in Liu and Yang (2018), namely, the Bootstrap-calibrated interval estimation approach.

Within the divide-and-conquer two-stage estimation scheme, because the latent outcome variable $\boldsymbol{\theta}_i$ (for person *i*) is measured with error, instead of observing $\boldsymbol{\theta}_i$, one only observes $\widehat{\boldsymbol{\theta}}_i$ from stage-one IRT calibration, and

$$\widehat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \tag{4}$$

where $\boldsymbol{\varepsilon}_i$ is the vector of measurement errors with a mean 0 and covariance matrix, $\boldsymbol{\Sigma}_{\theta_i}$. $\boldsymbol{\Sigma}_{\theta_i}$ is also known as the error covariance matrix, the magnitude of which depends on many factors, including (1) test information at $\boldsymbol{\theta}_i$, which also depends on whether the test is delivered via linear mode or adaptive mode; and (2) IRT model data fit. In the first stage, both $\widehat{\boldsymbol{\theta}}_i$ and $\hat{\boldsymbol{\varepsilon}}_i$ are estimated. Either the Maximum a Posteriori (MAP) or the Expected a Posteriori (EAP) is used to obtain the point estimate of $\widehat{\boldsymbol{\theta}}_i$ along with the error covariance matrix estimate, $\widehat{\boldsymbol{\Sigma}}_{\theta_i}$, for each person separately. Chang and Staut (1993) have shown that when test length is sufficiently long and when MLE is used, ε_i will follow normal distribution with mean 0 and variance proportional to the inverse of the Fisher information evaluated at true θ_i, i.e., $\sum_{\theta_i} \approx I^{-1}(\theta_i)$. Their results can be generalized to multidimensional $\boldsymbol{\theta}$'s and to MAP (e.g., Wang, 2015). Even though true θ_i is

unknown in practice, we have $\hat{\Sigma}_{\theta_i} \approx \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\theta}}_i)$ by plugging in $\widehat{\boldsymbol{\theta}}_i$ instead of $\boldsymbol{\theta}_i$. That is, using $\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\theta}}_i)$ as a proxy to the error variance of $\widehat{\boldsymbol{\theta}}_i$ is still viable as long as $\widehat{\boldsymbol{\theta}}_i$ is close to the true value (e.g., Koedel, Leatherman, & Parons, 2012; Shang, 2012). Although Lockwood and McCaffrey (2014) argued that $E\left[\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\theta}}_i)\right]$ is likely an overestimate of $E[\boldsymbol{I}^{-1}(\boldsymbol{\theta}_i)]$, and such a positive bias can lead to systematic errors in measurement error correction based on test reliability, this bias is no longer problematic in our methods because we treat each $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\boldsymbol{\Sigma}}_{\theta_i}$ individually, and we do not need a reliability estimate from $E\left[\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\theta}}_i)\right]$ to correct for measurement error.

Given the linear mixed effects model defined in Equation (3), the likelihood of both random and fixed effects is therefore

$$\prod_{i=1}^{N}\left[\phi(\boldsymbol{\theta}_i; \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}_i, \sigma^2\boldsymbol{I}_T)\right]\phi(\boldsymbol{u}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_u), \tag{5}$$

where N denotes sample size and $\phi(.)$ denotes the multivariate normal density. The likelihood in Equation (5) assumes that the random effect follows a multivariate normal distribution with a covariance matrix of $\sum_u$. A non-normal distribution of the random effect is also allowed if needed. Maximum likelihood estimation proceeds with integrating out the random effects first, leading to a marginal likelihood of

$$\prod_{i=1}^{N}\int\left[\phi(\boldsymbol{\theta}_i; \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}_i, \sigma^2\boldsymbol{I}_T)\right]\phi(\boldsymbol{u}_i; \boldsymbol{0}, \boldsymbol{\Sigma}_u)d\boldsymbol{u}\ , \tag{6}$$

which needs to be maximized to find the solution of $\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \hat{\boldsymbol{\Sigma}}_u$. Then the individual coefficient $\boldsymbol{u}_i$ will be predicted via the best linear unbiased predictor (BLUP).

Combining the linear mixed effects model in Equation (3) with the measurement error model in (4), the likelihood in Equation (5) is updated as

$$L(\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{u}) = \prod_{i=1}^{N}\left[\phi(\boldsymbol{\theta}_i;\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}_i,\sigma^2\boldsymbol{I}_T)\right]\phi(\boldsymbol{u}_i;\mathbf{0},\boldsymbol{\Sigma}_u)\varphi(\hat{\boldsymbol{\theta}}_i;\boldsymbol{\theta}_i,\hat{\boldsymbol{\Sigma}}_{\theta_i}), \quad (7)$$

in which case both random coefficient $\boldsymbol{u}_i$ and latent factors $\boldsymbol{\theta}_i$ need to be integrated to obtain the marginal likelihood. In Equation (7), $\varphi(.)$ denotes the density of the measurement error distribution. Therefore, the joint log-likelihood of $\boldsymbol{\Psi}$= ($\boldsymbol{\beta}$, $\sum_u$, σ^2) based off (7) is written as

$$l(\boldsymbol{\psi},\boldsymbol{\theta},\boldsymbol{u}) = \sum_{i=1}^{N} l(\boldsymbol{\psi},\boldsymbol{\theta}_i,\boldsymbol{u}_i) = \sum_{i=1}^{N}\left\{\log\left[\phi(\boldsymbol{\theta}_i;\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u}_i,\sigma^2\boldsymbol{I}_T)\right] + \log\phi(\boldsymbol{u}_i;\mathbf{0},\boldsymbol{\Sigma}_u) + \log\varphi(\hat{\boldsymbol{\theta}}_i;\boldsymbol{\theta}_i,\hat{\boldsymbol{\Sigma}}_{\theta_i})\right\}. \quad (8)$$

This equation will be used throughout the following explication.

We need to emphasize that the discussion hereafter is based on the assumption that the measurement error follows normal or multivariate normal distribution with error covariance matrix $\hat{\boldsymbol{\Sigma}}_{\theta_i}$. Diakow (2013) suggested using Warm (1989)'s weighted maximum likelihood in stage I along with a more precise version of the asymptotic standard error (Magis & Raiche, 2012). As the paper unfolds below, the non-normal measurement error distribution is also allowed in the method described in section 3.2.3. In fact, both methods provided in sections 3.2.2 and 3.2.3 are suitable for a level-1 variance-known problem (Raudenbush & Bryk, 2002, chapter 7), and our goal is to provide an accurate method for secondary data analysis that is convenient and understandable for applied research (Diakow, 2013).

3.2.1 Method I: Marginalized MLE (MMLE)

When both $\phi(.)$ and $\varphi(.)$ in Equation (8) follow or can be well approximated by a normal distribution (or multivariate normal), it can be derived that the marginal likelihood of the combined model, after integrating out both random coefficient $\boldsymbol{u}_i$ and latent factors $\boldsymbol{\theta}_i$ in (8), has a closed form up to a certain constant (for detailed derivations, please see Appendix A)

expressed below. To be specific, given the joint likelihood in Equation (7), the marginal log-likelihood of the target model parameters can be shown to be

$$l(\boldsymbol{\psi}) = \log L(\boldsymbol{\psi}) \propto -N \log\left|\boldsymbol{\Sigma}_u\right| - \frac{N}{\sigma^2}\left\|\boldsymbol{X}_i\boldsymbol{\beta}\right\|^2 + \sum_{i=1}^{N}(\log\left|\boldsymbol{\Sigma}_{u_i}^*\right| - \log\left|\sigma^2\hat{\boldsymbol{\Sigma}}_{\theta_i}^{-1} + \boldsymbol{I}_T\right|)$$
$$+ \sum_{i=1}^{N}\left(\left\|(\hat{\boldsymbol{\Sigma}}_{\theta_i}^{-1} + \sigma^{-2}\boldsymbol{I}_T)^{-1/2}(\hat{\boldsymbol{\Sigma}}_{\theta_i}^{-1}\hat{\boldsymbol{\theta}}_i + \sigma^{-2}\boldsymbol{X}_i\boldsymbol{\beta})\right\|^2 + \left\|(\boldsymbol{\Sigma}_{u_i}^*)^{1/2}\boldsymbol{\mu}_{u_i}^*\right\|^2\right) \quad (9)$$

where

$$(\boldsymbol{\Sigma}_{u_i}^*)^{-1} = \boldsymbol{\Sigma}_u^{-1} + \sigma^{-2}\boldsymbol{Z}^t\boldsymbol{Z} - \sigma^{-4}\boldsymbol{Z}^t(\sigma^2\hat{\boldsymbol{\Sigma}}_{\theta_i}^{-1} + \boldsymbol{I}_T)^{-1}\boldsymbol{Z}, \text{ and} \quad (10)$$

$$\boldsymbol{\mu}_{u_i}^* = \sigma^{-2}\boldsymbol{Z}^t(\boldsymbol{X}_i\boldsymbol{\beta}) - \boldsymbol{Z}^t(\sigma^2\hat{\boldsymbol{\Sigma}}_{\theta_i}^{-1} + \boldsymbol{I}_T)^{-1}(\hat{\boldsymbol{\Sigma}}_{\theta_i}^{-1}\hat{\boldsymbol{\theta}}_i + \sigma^{-2}\boldsymbol{X}_i\boldsymbol{\beta}) \quad . \quad (11)$$

In above equations, $|\square|$ denotes the determinant of a matrix, and $\| \|^2$ denotes an inner product of a vector. The closed form marginal likelihood for the longitudinal MIRT model is also presented in the Appendix A.

The MMLE proceeds with maximizing the closed-form marginal log-likelihood in (9). The '`optim`' function in '`stats`' library of R is used for solving the maximization problem. This function provides general purpose optimization based on Nelder-Mead, quasi-Newton, and conjugate-gradient algorithms. It allows for user-specified box constraints on parameters. Instead of using the default Nelder-Mead method (Nelder & Mead, 1965) which tends to be slow, we choose to use "L-BGFS-B" method available in the function because our objective function in (9) is differentiable. In particular, BGFS is the quasi-Newton method proposed by Broyden (1970), Fletcher (1970), Goldfarb (1970), and Shanno (1970), which uses both function values and gradients to construct a surface to be optimized. L-BGFS-B is then an extension of BGFS (Byrd et al., 1995) that allows box constraints in which each variable is given a lower and/or upper bound as long as the initial values satisfy the constraints. In our application, the constraints

include -1000< β_0, β_1<1000, 0.001<σ_{u_0}, σ_{u_1}<5, -.99<ρ<0.99, and 0.001<σ^2<1000[2]. The initial values for all parameters are set at 0.1. Both the parameter point estimates and their standard errors are output from the function, with the former being the final estimates upon convergence, and the latter obtained from the Hessian matrix. In some extreme cases when Hessian matrix is not available, we use numeric differentiation available in the 'numDeriv' package instead.

3.2.2 Method II: Expectation-Maximization (EM)

In this section, we will describe an alternative method to resolve the challenge of high-dimensional integration involved in the marginal likelihood. It is complementary to Method I when the closed-form marginal likelihood is not available, or when the numeric optimization fails to converge properly.

In particular, when treating the random effects and latent variables, $\boldsymbol{u}_i$ and $\boldsymbol{\theta}_i$, as missing data, this method proceeds iteratively between the expectation (E) and maximization (M) steps until convergence. At the (m+1)th iteration, in the E-step, take the expectation of log-likelihood with respect to the posterior distribution of $\boldsymbol{u}_i$ and $\boldsymbol{\theta}_i$ as

$$E(\boldsymbol{\psi} \mid \hat{\boldsymbol{\psi}}^{(m)}) = \sum_{i=1}^{N} \int l(\boldsymbol{\psi}, \boldsymbol{\theta}_i, \boldsymbol{u}_i) f(\boldsymbol{\theta}_i, \boldsymbol{u}_i \mid \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Sigma}}_{\theta_i}, \hat{\boldsymbol{\psi}}^{(m)}) d\boldsymbol{u}_i d\boldsymbol{\theta}_i \,, \tag{12}$$

where $f(\boldsymbol{\theta}_i, \boldsymbol{u}_i \mid \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Sigma}}_{\theta_i}, \hat{\boldsymbol{\psi}}^{(m)})$ denotes the posterior distribution, and $l(\boldsymbol{\psi}, \boldsymbol{\theta}_i, \boldsymbol{u}_i)$ takes the form in Equation (8). The integration in (12) can be obtained easily when one samples directly from the posterior distribution, such that

$$E(\boldsymbol{\psi} \mid \hat{\boldsymbol{\psi}}^{(m)}) = \sum_{i=1}^{N} \left[\frac{1}{Q} \sum_{q=1}^{Q} l(\boldsymbol{\psi}, \boldsymbol{\theta}_i^q, \boldsymbol{u}_i^q) \right], \tag{13}$$

[2] Originally, $\boldsymbol{\Sigma}_u$ needs to be constrained to be non-negative definite. However, this is not a box-constraint that 'optim' function can handle. We therefore impose constraints on the variance and correlation terms.

where $(\boldsymbol{\theta}_i^q, \boldsymbol{u}_i^q)$ is the qth draw from the posterior distribution, and Q is the total number of Monte Carlo draws. This Monte Carlo based integration is appropriate even if the measurement error or random effects does not follow normal distributions, and hence we consider this approach more general than the MMLE method.

If both the measurement error and random effects are indeed normal, then the conditional expectation in (12) has a closed form which can be directly computed without resorting to numeric integration. That is, given $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\sum}_{\theta_i}^{-1}$ from Stage I estimation and $\hat{\boldsymbol{\psi}}^{(m)}$ from the mth EM cycle, the joint posterior distribution of $(\boldsymbol{\theta}_i, \boldsymbol{u}_i)$ follows a multivariate normal, with a variance of

$$\widehat{\sum}^{(m)} = \begin{bmatrix} (\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1} + \widehat{\sum}_{\theta_i}^{-1} & -(\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1}\boldsymbol{Z} \\ -\boldsymbol{Z}^t(\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1} & (\widehat{\sum}_u^{(m)})^{-1} + \boldsymbol{Z}^t(\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1}\boldsymbol{Z} \end{bmatrix}^{-1} \equiv \begin{bmatrix} \sum^{11} & \sum^{12} \\ \sum^{21} & \sum^{22} \end{bmatrix}^{-1}. \qquad (14)$$

and a mean of

$$\begin{bmatrix} \widehat{\boldsymbol{\mu}}_\theta^{(m)} \\ \widehat{\boldsymbol{\mu}}_u^{(m)} \end{bmatrix}$$

$$= \begin{bmatrix} (\sum^{11} - \sum^{12}(\sum^{22})^{-1}\sum^{21})^{-1}\left[\widehat{\sum}_{\theta_i}^{-1}\widehat{\boldsymbol{\theta}}_i + (\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1}\boldsymbol{X_i}\widehat{\boldsymbol{\beta}}^{(m)} + \sum^{12}(\sum^{22})^{-1}\boldsymbol{Z}^t(\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1}\boldsymbol{X_i}\widehat{\boldsymbol{\beta}}^{(m)}\right] \\ (\sum^{22} - \sum^{21}(\sum^{11})^{-1}\sum^{12})^{-1}\left[-\boldsymbol{Z}^t(\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1}\boldsymbol{X_i}\widehat{\boldsymbol{\beta}}^{(m)} - \sum^{21}(\sum^{11})^{-1}\left(\widehat{\sum}_{\theta_i}^{-1}\widehat{\boldsymbol{\theta}}_i + (\hat{\sigma}^{2(m)}\boldsymbol{I})^{-1}\boldsymbol{X_i}\widehat{\boldsymbol{\beta}}^{(m)}\right)\right] \end{bmatrix}$$

(15)

M-step proceeds with maximizing the conditional expectation in (12) with respect to $\boldsymbol{\psi}$. Given the form of $l(\boldsymbol{\psi}, \boldsymbol{\theta}_i, \boldsymbol{u}_i)$ in Equation (8), $\boldsymbol{\beta}$, $\sum_u$, σ^2 all have the closed form solution as follows, which greatly simplifies the maximization step,

$$\hat{\boldsymbol{\beta}}^{(m+1)} = \left(\sum_{i=1}^{N} \boldsymbol{X}_i^t \boldsymbol{X}_i\right)^{-1} \times \sum_{i=1}^{N} \boldsymbol{X}_i^t \left[E^{(m)}(\boldsymbol{\theta}_i)^t - \boldsymbol{Z}E^{(m)}(\boldsymbol{u}_i)\right], \qquad (16)$$

$$\hat{\sigma}^{2(m+1)} = \frac{\sum_{i=1}^{N} E^{(m)}\left((\boldsymbol{\theta}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(m+1)} - \boldsymbol{Z}\boldsymbol{u}_i)^t(\boldsymbol{\theta}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(m+1)} - \boldsymbol{Z}\boldsymbol{u}_i)\right)}{N \times T}, \tag{17}$$

$$\hat{\boldsymbol{\Sigma}}_u^{(m+1)} = \frac{\sum_{i=1}^{N} E^{(m)}(\boldsymbol{u}_i\boldsymbol{u}_i^t)}{N}. \tag{18}$$

The notation of $E^{(m)}$ indicates that, at the (m+1)th EM cycle, the expected values in (16)~(18) are obtained from the first and second moments of the posterior multivariate normal distribution $f(\boldsymbol{\theta}_i, \boldsymbol{u}_i \mid \hat{\boldsymbol{\theta}}_i, \hat{\boldsymbol{\Sigma}}_{\theta_i}, \hat{\boldsymbol{\psi}}^{(m)})$ with mean and variance specified in (14) and (15). Equation (17) adopts the expectation conditional maximization (ECM) idea in Meng and Rubin (1993) in that the closed-form solution for residual variance only exists conditioning on the updated parameter $\hat{\boldsymbol{\beta}}^{(m+1)}$. The ECM algorithm shares all the appealing convergence properties of EM.

If the measurement model is the multidimensional IRT model with *D* dimensions, and if the residual error covariance matrix is still assumed to be diagonal, then the aforementioned EM algorithm only needs to be modified minimally. In particular, in the E-step, one simply needs to replace $\boldsymbol{I}_T$ with $\boldsymbol{I}_{DT}$, whereas $\boldsymbol{\Sigma}_u$ and $\boldsymbol{X}_i$ take the updated forms. In the M-step, at the (m+1)th iteration, the closed-form update for $\hat{\boldsymbol{\beta}}^{(m+1)}$ stays exactly the same as in (16). The update for $\hat{\sigma}^{2(m+1)}$ is modified as

$$\hat{\sigma}^{2(m+1)} = \frac{\sum_{i=1}^{N} E^{(m)}\left((\boldsymbol{\theta}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(m+1)} - \boldsymbol{Z}\boldsymbol{u}_i)^t(\boldsymbol{\theta}_i - \boldsymbol{X}_i\hat{\boldsymbol{\beta}}^{(m+1)} - \boldsymbol{Z}\boldsymbol{u}_i)\right)}{N \times D \times T}. \tag{19}$$

The standard error of the parameter estimates are obtained using the supplemented EM algorithm (Dempster, Laird, & Rubin, 1977; Cai, 2008). The principle idea is reiterated briefly as follows. The large sample error covariance matrix of MLE is known to be

$$\boldsymbol{V}(\hat{\boldsymbol{\psi}} \,|\, \boldsymbol{Y}) = \boldsymbol{I}^{-1}(\hat{\boldsymbol{\psi}} \,|\, \boldsymbol{Y}) = \boldsymbol{I}_c^{-1}(\hat{\boldsymbol{\psi}})[\mathbf{I}_d - \boldsymbol{\Delta}(\hat{\boldsymbol{\psi}})]^{-1}, \tag{20}$$

where $\boldsymbol{I}(\hat{\boldsymbol{\psi}}|\boldsymbol{Y})$ is the Fisher information matrix based on observed response data, $\boldsymbol{Y}$. $\boldsymbol{I}_c(\hat{\boldsymbol{\psi}})$ is the natural by-product of the E-step as it is simply the second derivative of Equation (12) with respect to all elements in $\boldsymbol{\Psi}$. $\Delta(\hat{\boldsymbol{\psi}})$ is the fraction of missing information, which can be obtained via numerical differentiation as

$$\Delta(\hat{\boldsymbol{\psi}}) = \frac{\partial M(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}}\Big|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}, \tag{21}$$

where $M(\boldsymbol{\psi})$ defines the vector-valued EM map as $\boldsymbol{\psi}^{(m+1)} = M(\boldsymbol{\psi}^{(m)})$. Upon convergence, $\hat{\boldsymbol{\psi}} = M(\hat{\boldsymbol{\psi}})$. For details regarding the calculation of $\Delta(\hat{\boldsymbol{\psi}})$ in general, please refer to Cai (2008) or Tian, Cai, and Xin (2013). We use a direct forward difference method (i.e., Eq. 8 and 9 in Tian et al., 2013) with a perturbation tuning parameter η=1. For details with respect to the specific form of $\boldsymbol{I}_c(\hat{\boldsymbol{\psi}})$, please see Appendix B.

3.2.3 Method III: Moment Estimation Method

If framing the estimation problem from a slightly different perspective, the linear mixed effects model in Equation (3) actually leads to the mean and covariance structure as follows,

$$\mu_\theta = E(\boldsymbol{\theta}_i) = \boldsymbol{X}_i\boldsymbol{\beta}\ ;\ \boldsymbol{\Sigma}_\theta = \boldsymbol{Z}\boldsymbol{\Sigma}_u\boldsymbol{Z}^t + \sigma^2\boldsymbol{I}_T. \tag{22}$$

It implies that to recover the structural parameters, $\boldsymbol{\Psi}$=($\boldsymbol{\beta}$, $\sum_u$, σ^2), only the $\hat{\boldsymbol{\mu}}_\theta$ and $\hat{\boldsymbol{\Sigma}}_\theta$ (i.e., estimated population mean and covariance of $\boldsymbol{\theta}$) need to be obtained in Stage I, rather than the individual point estimate of $\boldsymbol{\theta}_i$ and Σ_{θ_i}. This is consistent with the traditional wisdom in structural equation modeling (SEM), in which the inputs can be the mean and covariance matrix rather than the raw data. In our application, we assume θ_{it}'s follow multivariate normal in the population. When this assumption is satisfied, the mean and covariance contain all information (i.e., sufficient statistics), and when this assumption is violated, this method may still provide robust, consistent parameter estimates.

In stage I, the $\boldsymbol{\theta}_i$ and Σ_{θ_i} are estimated from raw response data via the EM algorithm (Mislevy, Beaton, Kaplan, & Sheehan, 1992). In particular, without imposing any particular growth pattern on latent traits over time, the full joint likelihood is

$$L\left(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta \mid \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\right) = \prod_{i,j,t} p_{ijt}^{y_{ijt}} \left(1 - p_{ijt}\right)^{1-y_{ijt}} \phi\left(\boldsymbol{\theta}_i, \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta\right), \tag{23}$$

where ϕ (·) again denotes multivariate normal density. Then in the E-step, for the (m+1)th cycle, the conditional expectation of $(\mu_{\boldsymbol{\theta}}, \sum_{\boldsymbol{\theta}})$ is

$$E(\log L(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta) \mid \hat{\boldsymbol{\mu}}_\theta^{(m)}, \hat{\boldsymbol{\Sigma}}_\theta^{(m)}) = \int l\left(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta \mid \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}\right) P(\boldsymbol{\theta} | \hat{\boldsymbol{\mu}}_\theta^{(m)}, \hat{\boldsymbol{\Sigma}}_\theta^{(m)}, \boldsymbol{y}) d\boldsymbol{\theta}\,, \tag{24}$$

where the integral can be obtained via Monte Carlo integration by drawing Q samples of $\boldsymbol{\theta}^q$'s from multivariate normal with mean $\hat{\boldsymbol{\mu}}_\theta^{(m)}$ and covariance $\hat{\boldsymbol{\Sigma}}_\theta^{(m)}$.

M-step follows with maximizing the conditional expectation in (24) with respect to $(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$, using the following closed-form expressions,

$$\begin{aligned} \hat{\boldsymbol{\mu}}_\theta^{(m+1)} &= \frac{1}{N} \sum_{i=1}^{N} \int \boldsymbol{\theta} P(\boldsymbol{\theta} \mid \hat{\boldsymbol{\mu}}_\theta^{(m)}, \hat{\boldsymbol{\Sigma}}_\theta^{(m)}, \boldsymbol{y}_i) d\boldsymbol{\theta} \\ \hat{\boldsymbol{\Sigma}}_\theta^{(m+1)} &= \frac{1}{N} \sum_{i=1}^{N} \int (\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}_\theta^{(m)})^T (\boldsymbol{\theta} - \hat{\boldsymbol{\mu}}_\theta^{(m)}) P(\boldsymbol{\theta} \mid \hat{\boldsymbol{\mu}}_\theta^{(m)}, \hat{\boldsymbol{\Sigma}}_\theta^{(m)}, \boldsymbol{y}_i) d\boldsymbol{\theta} \end{aligned} . \tag{25}$$

The estimators in (26) are maximum likelihood estimates of $(\mu_{\boldsymbol{\theta}}, \sum_{\boldsymbol{\theta}})$, consistent in sample size (i.e., N) regardless of test length (Mislevy et al., 1992).

In stage II, $\widehat{\boldsymbol{\Psi}}$ can be estimated using any off-the-shelf SEM packages, using $\hat{\boldsymbol{\mu}}_\theta$ and $\hat{\boldsymbol{\Sigma}}_\theta$ as input. An example is the R package 'lavaan' (Rosseel, 2012), from which the MLE estimates of $\widehat{\boldsymbol{\Psi}}$ are provided. Or in essence, the generalized least squares solution to $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{V}^{-1} \boldsymbol{\mu}_\theta$ and the MLE of $\widehat{\sum}_u$ and $\hat{\sigma}^2$ can be found based on the likelihood function

$$F = -\log\left|\boldsymbol{Z}\hat{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^T + \hat{\sigma}^2\boldsymbol{I}_T\right| - tr\left(\boldsymbol{\Sigma}_\theta(\boldsymbol{Z}\hat{\boldsymbol{\Sigma}}_u\boldsymbol{Z}^T + \hat{\sigma}^2\boldsymbol{I}_T)^{-1}\right), \quad (26)$$

where "tr" denotes the trace of a matrix. As there are no closed form solutions to (26), Newton-Raphson method is usually used (e.g., Lindstrom & Bates, 1988). The only input in (26) is $\hat{\boldsymbol{\Sigma}}_\theta$ from stage I. This method is extremely fast computationally. Because the individual latent score $\hat{\theta}_i$ is not needed in stage II estimation, the measurement error challenge vanishes.

Please note that this moment estimation method could also apply when $\boldsymbol{X}_i$ differs across individuals, i.e., when evaluating treatment effect is of interest. In this case, the sample mean of $\boldsymbol{X}_i$ along with $\hat{\mu}_\theta$ estimated from Stage I will be treated as the mean-structure input, whereas an expanded covariance matrix including $\hat{\boldsymbol{\Sigma}}_\theta$ as well as the covariance between $\boldsymbol{X}_i$ and $\boldsymbol{\theta}$ will be put into 'lavaan'. In this regard, Stage II estimation needs minimum update, whereas Stage I estimation (i.e., Equations 25 and 26) need to be updated accordingly.

In sum, the two-stage methods introduced in sections 3.2.1 and 3.2.2 rely on the assumption that $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\sum}_{\theta_i}$ are asymptotically unbiased. Whereas previous methods might suffer from such divide-and-conquer strategy due to finite sample bias in $\widehat{\boldsymbol{\theta}}_i$ and $\widehat{\sum}_{\theta_i}$, the third moment estimation method should be fine theoretically. One limitation of the method, however, is that sample size needs to be large enough to enable accurate (and consistent) recovery of $\hat{\boldsymbol{\mu}}_\theta$ and $\hat{\boldsymbol{\Sigma}}_\theta$ in stage I, especially $\hat{\boldsymbol{\Sigma}}_\theta$ has to be positive definite. The MMLE and EM methods, on the other hand, do not seem to be affected much by small sample size.

4. Simulation Study

Two simulation studies were conducted to evaluate the performance of five different methods, they are: (1) direct maximization of the closed-form marginal likelihood (MMLE),

(2) the EM algorithm, (3) the moment estimation method, (4) the naïve two-stage estimation, and (5) the one-stage MCMC estimation. The first simulation study focused on the unidimensional 3PL model as the measurement model, along with the latent growth curve model as the structural model; whereas the second simulation study focused on the two-dimensional compensatory IRT model along with the associative latent growth curve model. Throughout the simulation studies, all item parameters were fixed at known values to eliminate any potential contamination of item parameter estimation bias on the other targeted parameters. In addition, only dichotomous items were considered, but the 3PL and M3PL model could be easily replaced by the polytomous response models if needed.

4.1 Study I

4.1.1 Design

The fixed and manipulated factors in the study were drawn from the previous literature. Two factors were manipulated: examinee sample size (200 vs. 2,000), and covariance matrix of the random effects (Raudenbush & Liu, 2000; Ye, 2016). The 200 sample size is typically seen in psychology research whereas the 2,000 sample size is seen in education research. The medium and small covariance matrix of $\sum_u$ were set as follows (Raudenbush & Liu, 2000; Ye, 2016),

$$\begin{bmatrix} 0.2 & 0.05 \\ 0.05 & 0.1 \end{bmatrix} \text{(medium)}, \quad \begin{bmatrix} 0.1 & 0.025 \\ 0.025 & 0.05 \end{bmatrix} \text{(small)}.$$

The number of measurement waves was fixed at 4 (Khoo, Wes, Wu, & Kwok, 2006; Ye, 2016), and test length was fixed at 25, which is similar to the test length for science subject in NELS (National Educational Longitudinal Study).

In terms of fixed effects, the mean intercept was set at 0 (i.e., β_0=0), and mean slope was set at 0.15 (i.e., β_1=.15). Given the medium slope variance of .1 specified above, the mean slope

of .15 leads to a medium standardized effect size of .5 (see Raudenbush & Liu, 2000). Regarding the 3PL item parameters, *a*-parameters were drawn from Uniform (1.5, 2.5), *b*-parameters were drawn from Normal (0, 1) (Cai, 2010), and *c*-parameters were drawn from Uniform (0.1, 0.2). The scaling factor *D* was set at 1.7. Residual variance was $\sigma_e^2=\sigma^2 = .15$ (Kohli et al., 2015).

The details of the MCMC method including the priors are presented in the Appendix C. As shown in the Appendix C, conjugate priors are used whenever possible to enable direct Gibbs sampler. However, because we considered the logistic model throughout the paper, the Metropolis-Hastings algorithm is used to construct the Markov chains of certain parameters (i.e., $\boldsymbol{\theta}$). Otherwise, when the normal ogive model is considered, the efficiency of MCMC will be further improved.

In stage I estimation, a combination of maximum likelihood estimator (MLE) and maximum a posteriori (MAP) estimator was used. That is, MLE was considered first and if the absolute value of the estimate was larger than 3, then the estimation method switched to MAP with a normal prior $N(0, 5)$. The recovery of the structural model parameters is the focus of this report, including mean intercept (β_0), slope (β_1,), residual variance (σ^2), and covariance matrix of random effects ($\sum_u$). For these parameters, the average bias was computed as the mean of all bias estimates from all replications. Taking the mean intercept parameter as an example, the relative bias and RMSE were computed as $\frac{1}{R}\sum_{r=1}^{R}\frac{(\hat{\beta}_0^r-\beta_0)}{\beta_0}$ and $\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\beta}_0^r-\beta_0)^2}$. Here, *R* denotes the total number of replications, and $\hat{\beta}_0^r$ denotes the estimate from the *r*th replication. 50 replications were conducted per condition. In addition, the average estimated standard error for every parameter from each replication was computed, and the final mean values across replications were reported.

4.1.2 Results

Table 1 presents the bias and relative bias (in the parentheses) for the structural model parameters. Several trends can be spotted from this table. First and consistent with our expectation, the naïve two-stage method generated the largest bias for the residual variance σ^2, and in most cases, the largest bias for the elements in the random effects covariance matrix $\sum_u$ (e.g., τ_{00}). However, not all elements in $\sum_u$ suffered from high bias, which might be due to the unsystematic measurement errors across time (i.e., the measurement error is not in an explicit increasing or decreasing order). Second, both MMLE and EM method tended to perform well in most conditions by reducing the bias of σ^2 and elements in $\sum_u$. There is no appreciable difference between these two methods. Although the MMLE works with the closed-form marginal likelihood, hence it circumvents the numerical integration that subjects to Monte Carlo error, the optimization in the six dimensional space can still cause numeric error. On the other hand, the EM works either with Monte-Carlo based integration or closed-form integration in the E-step, but the closed-form solution in M-step avoids numeric optimization. Therefore, numeric approximations appear in different steps of these two methods, resulting in slight to no differences between them. Third, the moment estimation method generated the most accurate parameter recovery among all methods as this method does not depend on the assumption of normal measurement error. Hence, when the population distribution is assumed normal, this method is recommended. Unsurprisingly, the MCMC method also produced accurate parameter estimates, and in the cases when sample size is large, the best parameter estimates among all methods. It is only when the sample size is small and when the covariance matrix of random effects is small that MCMC yielded slightly higher bias in $\widehat{\sum}_u$. This could be explained by the known effect of "regression toward mean" for Bayesian estimates, and such an effect will diminish when sample size increases.

In terms of the manipulated factors, the true value of the random effects covariance matrix did not seem to affect the results much, so did not the sample size. The parameters (especially β_0, β_1, and σ^2) from the moment estimation method seemed to improve slightly with larger sample size, simply because the mean and covariance matrix of θ recovered better in stage I with a larger sample size. The other three methods treated each individual θ_i from stage I as a fallible estimate from its own measurement error model (i.e., Equation 5), so increasing sample size does not help reduce the measurement error. Overall, our observations of results are similar to Diakow (2013)'s conclusion where she used gllamm command (Rabe-Hesketh, Skrondal & Pickles, 2004) in Stata (StatCorp, 2011) with adaptive Gauss-Hermite quadrature method.

Table 1. Bias (and relative bias) of structural model parameters for IRT+LGC model

Covariance	N		MCMC	Moment Estimation	MMLE	EM	Naïve
Medium	200	β_0	.006	-.008	.018	.017	.017
		β_1	.006 (.041)	-.008 (-.051)	.003 (.019)	.002 (.013)	.007 (.047)
		σ^2	.008 (.056)	-.045 (-.302)	.065 (.433)	.063 (.420)	.141 (.937)
		τ_{00}	.015 (.074)	-.015 (-.076)	.006 (.032)	.011 (.053)	.028 (.140)
		τ_{01}	-.004 (-.083)	.003 (.060)	-.006 (-.120)	-.008 (-.163)	-.002 (-.048)
		τ_{11}	.019 (.194)	-.017 (-.168)	-.021 (-.213)	-.029 (-.201)	.021 (-.209)
	2000	β_0	.003	.001	.027	.025	.029
		β_1	.002 (.013)	-.006 (-.037)	.005 (.034)	.003 (.023)	.008 (.054)
		σ^2	.001 (.004)	-.022 (-.145)	.065 (.437)	.065 (.432)	.142 (.944)
		τ_{00}	.005 (.025)	-.007 (-.033)	.012 (.060)	.014 (.069)	.034 (.170)
		τ_{01}	-.000 (-.009)	.008 (.160)	-.005 (.096)	-.006 (-.114)	-.002 (-.030)
		τ_{11}	.003 (.032)	-.022 (-.217)	-.022 (-.218)	.021 (-.213)	-.021 (-.213)
Small	200	β_0	.009	-.009	.010	.009	.007
		β_1	-.003 (-.021)	-.002 (-.013)	.009 (.061)	.008 (.056)	.016 (.105)
		σ^2	.001 (.006)	-.036 (-.237)	.050 (.334)	.044 (.296)	.123 (.820)
		τ_{00}	.030 (.296)	-.003 (-.032)	-.004 (-.039)	.011 (.110)	.006 (.060)
		τ_{01}	-.014 (-.564)	.000 (.010)	.003 (.124)	-.004 (-.177)	.007 (.270)
		τ_{11}	.020 (.405)	-.009 (-.189)	-.003 (-.053)	.001 (.026)	-.001 (-.027)
	2000	β_0	-.002	-.007	.011	.011	.008
		β_1	.000 (.002)	-.002 (-.015)	.008 (.058)	.008 (.057)	.015 (.099)
		σ^2	-.003 (-.017)	.021 (-.142)	.047 (.312)	.044 (.295)	.119 (.794)
		τ_{00}	.007 (.068)	-.007 (-.073)	-.006 (-.064)	.000 (.003)	.003 (.033)
		τ_{01}	-.003 (-.120)	.004 (.144)	-.000 (.004)	.001 (.037)	.008 (.313)
		τ_{11}	.003 (.070)	-.008 (-.159)	.004 (.070)	.001 (.027)	.001 (.028)

Note: The relative bias for the mean intercept (i.e., β_0) is not reported because the true value is 0.

On a separate note, because the accurate estimation of $\hat{\theta}_i$ and $\hat{\Sigma}_{\theta_i}$ are pivotal to the success of the proposed MMLE and EM methods, Tables 2 and 3 present $\hat{\theta}_i$ and $\hat{\Sigma}_{\theta_i}$ recovery results. Note that for Table 3, the bias of the measurement error estimate is computed as $\sqrt{I^{-1}(\hat{\boldsymbol{\theta}}_i)}$-$\sqrt{I^{-1}(\boldsymbol{\theta}_i)}$ for person *i* where $\boldsymbol{\theta}_i$ is the true value for person *i*. Then the average bias is computed across all individuals, and finally the medium value is obtained across replications. The medium is used instead of mean because there are a couple of outliers that may severely inflate the bias. As shown in Table 2, MCMC produced the smallest absolute bias and RMSE simply because it uses information from all time points. The estimation precision from MLE/MAP is also acceptable. A clear trend is that the RMSE is evidently larger at later time points, which is due to the way we simulated item parameters, resulting in a lack of "difficult" items. Regarding the recovery of the measurement error, $\hat{\Sigma}_{\theta_i}$, Table 3 shows that on average, there is about 10% bias. Therefore, it is expected that if Warm's WLE and bias-corrected measurement error computation is used (Diakow, 2013, Wang, 2015), the improvement of MMLE and EM over naïve method should be more salient.

Table 2. Average bias and RMSE of $\boldsymbol{\theta}$ estimates for the UIRT+LGC model

		Bias				**RMSE**			
		Small covariance		Medium covariance		Small covariance		Medium covariance	
		MCMC	MLE/MAP	MCMC	MLE/MAP	MCMC	MLE/MAP	MCMC	MLE/MAP
200	θ_1	-.001	.011	-.001	-.001	.186	.242	.201	.265
	θ_2	-.003	.015	-.003	.014	.194	.260	.220	.284
	θ_3	.007	.019	.007	.014	.222	.293	.279	.338
	θ_4	.005	.019	.011	-.024	.276	.329	.370	.439
2000	θ_1	.001	.005	-.000	.007	.187	.246	.200	.263
	θ_2	.001	.017	-.000	.017	.196	.266	.220	.294
	θ_3	.000	.026	-.000	.016	.225	.297	.277	.338
	θ_4	.002	.017	.002	-.024	.280	.333	.366	.444

Table 3. Bias (and relative bias) of $\hat{\Sigma}_{\theta_i}$ for the UIRT+LGC model

N		**200**		**2000**
Covariance	Small	Medium	Small	Medium
$\left(\hat{\Sigma}_{\boldsymbol{\theta_i}}\right)_{\mathbf{11}} \equiv \hat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_1}}$	.024 (.087)	.033 (.117)	.029 (.106)	.041 (.142)
$\left(\hat{\Sigma}_{\boldsymbol{\theta_i}}\right)_{\mathbf{22}} \equiv \hat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_2}}$	.031 (.106)	.036 (.134)	.037 (.135)	.038 (.147)
$\left(\hat{\Sigma}_{\boldsymbol{\theta_i}}\right)_{\mathbf{33}} \equiv \hat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_3}}$	.037 (.135)	-.010 (.126)	.039 (.156)	-.013 (.146)
$\left(\hat{\Sigma}_{\boldsymbol{\theta_i}}\right)_{\mathbf{44}} \equiv \hat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_4}}$	.016 (.141)	-.450 (.092)	.013 (.153)	-.675 (.094)

Table 4 presents the average standard error (SE) of all structural model parameters for different methods under different conditions. Consistent with our expectation, the naïve method generated higher SE for all parameters compared to MMLE and EM methods under all conditions. The SEs from MCMC was also slightly high because they contained Monte Carlo sampling error by nature. Again, the level of covariance (i.e., Σ_u) did not affect the magnitude of SE much, and EM yielded slightly lower standard error than MMLE, but the difference is marginal. The moment estimation method generated slightly higher SE because it did not take into account all individual information in stage I but rather only used mean and covariance estimates, hence "limited" information. For all methods, SE dropped when sample size increased.

Table 4. Estimated standard error for structural model parameters for UIRT+LGC model

		Medium covariance					Small covariance				
N		MCMC	Moment Estimation	MMLE	EM	Naïve	MCMC	Moment Estimation	MMLE	EM	Naïve
200	β_0	.043	.036	.027	.018	.045	.037	.030	.022	.017	.037
	β_1	.030	.023	.016	.010	.025	.024	.018	.014	.010	.022
	σ^2	.016	.007	.010	.006	.018	.014	.008	.009	.005	.017
	τ_{00}	.038	.026	.023	.017	.043	.026	.019	.016	.008	.030
	τ_{01}	.017	.012	.012	.007	.017	.011	.008	.010	.004	.013
	τ_{11}	.019	.011	.008	.007	.013	.011	.007	.005	.004	.010
2000	β_0	.013	.012	.009	.006	.014	.011	.010	.007	.006	.012
	β_1	.009	.007	.005	.003	.008	.007	.006	.004	.003	.007
	σ^2	.005	.003	.003	.002	.005	.005	.003	.003	.002	.005
	τ_{00}	.012	.009	.007	.005	.014	.008	.006	.005	.002	.010
	τ_{01}	.006	.004	.004	.002	.005	.004	.003	.003	.001	.004
	τ_{11}	.006	.003	.003	.002	.004	.003	.002	.002	.001	.003

4.2 Study II

4.2.1 Design

In this second simulation study, the two-dimensional simple-structure IRT model was used. The test length was fixed at 40 at each measurement wave, hence there were 20 items loading on each dimension. The item parameters per domain were simulated the same as in Study I. The only difference is, the mean of the difficulty parameter increased over time, which were taken to be the average of the mean θ from the two dimensions at the corresponding time point. This way, the items tend to align better with θ as the respective time points. The number of measurement waves were also fixed at 4, and the fixed effects were set at $\beta = [0, 0, 0.15, 0.15]$. Here the first two elements refer to the mean intercepts and the last two elements refer to the mean slopes. Residual variance was fixed at $\sigma^2 = 0.15$ for simplicity. Given that the size of the random effects covariance matrix did not affect the results much from study I, we decided to fix the covariance matrix as

$$\sum_u = \begin{bmatrix} \sum_{u_0} & 0 \\ 0 & \sum_{u_1} \end{bmatrix} = \begin{bmatrix} \sigma^2_{u_{01}} & \sigma_{u_{01}u_{02}} & & \\ \sigma_{u_{01}u_{02}} & \sigma^2_{u_{02}} & & \\ & & \sigma^2_{u_{11}} & \sigma_{u_{11}u_{12}} \\ & & \sigma_{u_{11}u_{12}} & \sigma^2_{u_{12}} \end{bmatrix} = \begin{bmatrix} .2 & .1 & 0 & 0 \\ .1 & .2 & 0 & 0 \\ 0 & 0 & .2 & .1 \\ 0 & 0 & .1 & .2 \end{bmatrix}.$$

As shown above, the intercepts and slopes were uncorrelated, whereas the two intercepts correlated and the two slopes correlated. This simplification resulted in a reduction of the total number of parameters, which, to some extent, benefited the MMLE method. This is because in MMLE, larger number of parameters means searching in a high-dimensional space. The EM method was not affected, however, because of the closed-form solution in both the E-step and the M-step. But similar constraints were still added in the EM estimation to make a fair comparison.

4.2.2 Results

Table 5 presents the bias and relative bias (in the parenthesis) of the structural model parameters. First of all, as expected, the MCMC method produced the most accurate parameter estimates for all parameters under both conditions. Second, consistent with the findings from the previous simulation study, all methods produced acceptable fixed parameter estimates, and the bias for β_{01} and β_{02} are second smallest for the moment estimation method. This may be because, with slightly shorter test length (20 per dimension vs. 25 from study I), the individual $\hat{\theta}$ and its SE may be prone to larger error, whereas the population mean and covariance estimates are less affected. However, the difference is not salient. The naïve method again yielded the largest positive bias for residual variance (σ^2) and intercept variance ($\sigma^2_{u_{01}}$ and $\sigma^2_{u_{02}}$). The moment estimation method, on the other hand, resulted in slightly large negative bias for residual variance but it generated accurate slope variance estimates. In contrast, the other three methods resulted in slightly negative bias for slope variance, and naïve method even outperformed the

other two by a little margin. These results match both Diakow (2013) and Verhelst (2010), who found that in the hierarchical linear modeling, "within-cluster variance is overestimated by the naïve method while between-cluster variance is recovered".

Table 5. Bias (and relative bias) of structural model parameters for MIRT+LGC model

N		MCMC	Moment Estimation	MMLE	EM	Naïve
200	β_{01}	.004	-.019	.053	.049	.033
	β_{02}	.001	-.004	.066	.062	.055
	β_{11}	.001 (.008)	.002 (.011)	.003 (.022)	-.002 (-.012)	.002 (.015)
	β_{12}	.000 (.001)	-.003 (-.022)	-.001 (-.006)	-.006 (-.039)	-.007 (-.045)
	σ^2	.008 (.053)	-.083 (-.550)	.011 (-.062)	.011 (.075)	.212 (1.413)
	$\sigma^2_{u_{01}}$	.019 (.095)	-.021 (-.104)	-.012 (-.062)	-.011 (-.055)	.033 (.164)
	$\sigma_{u_{01}u_{02}}$	-.010 (-.103)	-.008 (-.078)	-.002 (-.018)	-.002 (-.016)	.016 (.164)
	$\sigma^2_{u_{02}}$	.020 (.098)	-.022 (-.111)	-.006 (-.033)	-.007 (-.035)	.024 (.118)
	$\sigma^2_{u_{11}}$	.012 (.058)	-.008 (-.041)	-.059 (-.298)	-.059 (-.295)	-.053 (-.262)
	$\sigma_{u_{11}u_{12}}$	-.001 (-.011)	.001 (.013)	-.029 (-.285)	-.028 (-.284)	-.027 (-.271)
	$\sigma^2_{u_{12}}$	.003 (.016)	-.008 (-.041)	-.056 (-.278)	-.056 (-.278)	-.054 (-.272)
2000	β_{01}	-.001	-.012	.063	.060	.040
	β_{02}	-.000	-.0088	.062	.059	.040
	β_{11}	.001 (.006)	-.004(-.026)	-.004(-.026)	-.008(-.055)	-.003(-.018)
	β_{12}	.001 (.005)	-.003(-.021)	.001(.004)	-.004(-.025)	-.000(-.000)
	σ^2	.002 (.011)	-.088(-.585)	.013(.085)	.013(.085)	.212(1.409)
	$\sigma^2_{u_{01}}$	.003 (.014)	-.016(-.080)	-.004(-.022)	-.004(-.021)	.027(.134)
	$\sigma_{u_{01}u_{02}}$	-.003 (-.029)	-.008(-.079)	-.003(-.030)	-.003(-.025)	.014(.137)
	$\sigma^2_{u_{02}}$	.002 (.010)	-.016(-.083)	-.006(-.032)	-.006(-.032)	.036(.179)
	$\sigma^2_{u_{11}}$	.001 (.007)	-.016(-.079)	-.059(-.299)	-.059(-.298)	-.056(-.278)
	$\sigma_{u_{11}u_{12}}$	.001 (.011)	-.006(-.064)	-.031(-.310)	-.031(-.309)	-.029(-.289)
	$\sigma^2_{u_{12}}$	.002 (.008)	-.017(-.085)	-.058(-.294)	-.058(-.293)	-.054(-.270)

Tables 6 and 7 present the recovery of $\boldsymbol{\theta}_i$ and $\boldsymbol{\Sigma}_{\boldsymbol{\theta}_i}$ respectively. In general, the MCMC produced more accurate $\widehat{\boldsymbol{\theta}}_i$ estimates than the MLE/MAP method unsurprisingly. The RMSE increases slightly at a later time also due to lack of suitable items for the certain range of θ. As to the recovery of the measurement error, while the relative bias is around 10% for the first three time points, which is similar to the results in Table 5, the relative bias drops considerably for the last time point and the bias itself increases dramatically. This is again because of the mismatch

between the item difficulties and θ at time 4. From the LGC model where true θ's were simulated, the ranges of θ are (-2.5, 2.5), (-2.5, 3), (-3, 4), and (-4, 6) for the four time points respectively. However, the variance of item difficulty was fixed at 1 across all time points, so there were not enough items with extreme difficulty levels for extreme θ's at time 4. It is anticipated that both the RMSE in Table 6 and the measurement error bias will decrease if items with wider difficulty levels are added.

Table 6. Average bias and RMSE of $\boldsymbol{\theta}$ estimates for the MIRT+LGC model

	Bias				**RMSE**			
		N=200		N=2000		N=200		N=2000
	MCMC	MLE/MAP	MCMC	MLE/MAP	MCMC	MLE/MAP	MCMC	MLE/MAP
$\boldsymbol{\theta_{11}}$	-.002	.033	-.000	.042	.315	.443	.323	.448
$\boldsymbol{\theta_{21}}$	.005	.044	-.000	.039	.326	.452	.318	.454
$\boldsymbol{\theta_{12}}$	.001	.046	-.000	.038	.309	.493	.317	.483
$\boldsymbol{\theta_{22}}$	.002	.042	-.000	.035	.313	.605	.311	.595
$\boldsymbol{\theta_{13}}$	-.001	.052	-.000	.039	.367	.447	.371	.449
$\boldsymbol{\theta_{23}}$	.001	.046	.000	.036	.370	.457	.367	.453
$\boldsymbol{\theta_{14}}$	.000	.041	-.000	.042	.479	.480	.484	.484
$\boldsymbol{\theta_{24}}$	.000	.022	-.000	.040	.484	.591	.478	.589

Table 7. Bias (and relative bias) of $\hat{\Sigma}_{\theta_i}$ for the MIRT+LGC model

	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{11}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{12}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{13}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{14}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{21}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{22}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{23}}}$	$\widehat{\boldsymbol{\sigma}}_{\boldsymbol{\theta_{24}}}$
N=200	.037	.026	.023	.031	-.091	-.084	-10.643	-6.122
	(.113)	(.081)	(.118)	(.101)	(.074)	(.083)	(-.001)	(-.003)
N=2000	.034	.042	.022	.025	-.236	-.175	14.367	-6.822
	(.108)	(.120)	(.105)	(.098)	(.082)	(.089)	(-.008)	(.001)

Table 8 presented the estimated standard error for structural parameters. Overall, the results are consistent with the previous findings that the naïve method generated somewhat larger standard error because "the biased estimates of the variance components might affect the estimated standard errors of the regression coefficients" (Diakow, 2013).

Table 8. Estimated standard error of structural model parameter

	N=200					**N=2000**				
	MCMC	Moment	MMLE	EM	Naïve	MCMC	Moment	MMLE	EM	Naïve

		Estimation					Estimation			
$\boldsymbol{\beta_{01}}$	.047	.034	.032	.046	.049	.015	.011	.010	.015	.016
$\boldsymbol{\beta_{02}}$	.048	.033	.032	.046	.049	.015	.011	.010	.015	.016
$\boldsymbol{\beta_{11}}$	.038	.032	.023	.032	.033	.012	.010	.007	.010	.010
$\boldsymbol{\beta_{12}}$	.038	.032	.023	.033	.033	.012	.010	.007	.010	.010
$\boldsymbol{\sigma^2}$	.016	.003	.011	.008	.018	.005	.001	.004	.003	.006
$\boldsymbol{\sigma^2_{u_{01}}}$	.041	.023	.027	.014	.050	.013	.007	.009	.004	.016
$\boldsymbol{\sigma_{u_{01}u_{02}}}$	.029	.017	.019	.008	.035	.009	.006	.006	.003	.011
$\boldsymbol{\sigma^2_{u_{02}}}$	.042	.023	.027	.014	.049	.013	.007	.009	.004	.016
$\boldsymbol{\sigma^2_{u_{11}}}$	.031	.021	.013	.009	.022	.009	.011	.004	.003	.007
$\boldsymbol{\sigma_{u_{11}u_{12}}}$	.021	.016	.009	.006	.016	.006	.011	.003	.002	.005
$\boldsymbol{\sigma^2_{u_{12}}}$	.030	.021	.014	.009	.022	.009	.010	.004	.003	.007

5. A Real Data Illustration

In this section, we briefly compared the performance of five methods using a real data example from the National Educational Longitudinal Study 88 (NELS 88). A nationally representative sample of approximately 24,500 students were tracked via multiple cognitive batteries from 8th to 12th grade (the first three studies) in years 1988, 1990, and 1992. The science subject data were used in this section. The sample size was 7,282 after initial data cleaning, and we used list-wise deletion to eliminate the effect of missing data[3]. The data contains binary responses to 25 items in each year. The true item parameters were obtained from NELS 88 psychometrics report (https://nces.ed.gov/pubs95/95382.pdf). The mean discrimination parameters were 0.85, 0.95, and 0.95 for the three measurement occasions, with the standard deviation of 0.29, 0.30, and 0.30 respectively. The mean and standard deviation of difficulty parameters were (-0.28, 0.10, 0.22) and (0.90, 0.71, 0.96) respectively. The mean and standard deviation of guessing parameters were (0.20, 0.19, 0.18) and (0.14, 0.13, 0.12) respectively. In

[3] We used the list-wise deletion because we wanted to create a complete data set for illustration. Our intention was to evaluate the performance of different methods without possible interference of missing data. Because we used the NELS provided item parameters and because our structural model is simple, the possible bias introduced by list-wise deletion may be ignored.

stage I analysis, the unidimensional 3PL model was considered, both the MLE estimation for individual ability ($\hat{\theta}^{MLE}$) and the EM algorithm for population mean and covariance were obtained. The estimated population mean and covariance were $\hat{\mu}$ = (-.43, .08, .28) and $\hat{\Sigma} = \begin{bmatrix} 0.92 & 0.77 & 0.77 \\ 0.77 & 0.96 & 0.84 \\ 0.77 & 0.84 & 0.96 \end{bmatrix}$, whereas the sample mean and covariance estimates from $\hat{\theta}^{MLE}$ were $\hat{\mu}$=(-.43, .09, .28) and $\hat{\Sigma} = \begin{bmatrix} 1.25 & 0.89 & 0.86 \\ 0.89 & 1.31 & 0.98 \\ 0.86 & 0.98 & 1.26 \end{bmatrix}$. The two means are close, whereas the sample variances were larger.

Table 5 presents the parameter estimates and their standard error (in the parenthesis) for the five different methods. As reflected in Table 5, the fixed effects estimates from different methods were close. The naïve method, as expected, resulted in largest residual variance and intercept variance estimates. Both MMLE and EM tended to yield smaller variance estimates, which are consistent with the findings in Diakow (2013). This is because the random variances in the data can actually be decomposed as measurement error, randomness across individuals (random effects), and randomness within individuals (i.e., residual error). By actively incorporating the measurement error term in the model, the other two variances were reduced.

Also of note is that the estimated measurement error obtained in stage 1 for extreme $\hat{\theta}^{MLE}$ (i.e., close to -3 or 3) could be over 1 (in particular for measurement waves 2 and 3) due to lack of information in the tests for students with extreme abilities. In this case, the imprecision in the estimated measurement error could adversely affect the parameter estimates in the MMLE and EM methods (Diakow, 2013).

Table 5. Parameter estimates and their standard error (in the parenthesis) for NLES 88 Science data

	MCMC	Moment Estimation	MMLE	EM	Naïve
β_0	-.349(.012)	-.376(.011)	-.324(.001)	-.339(.009)	-.371(.013)
β_1	.351(.005)	.353(.004)	.350(.002)	.350(.003)	.354(.005)
σ^2	.054(.003)	.145(.002)	.078(.001)	.076(.003)	.315(.005)
τ_{00}	.828(.019)	.775(.015)	.599(.010)	.596(.011)	.906(.020)
τ_{01}	-.006(.006)	.004(.004)	-.006(.004)	.007(.0004)	-.002(.006)
τ_{11}	.029(.003)	.014(.002)	.008(.002)	.0231(.0002)	.032(.004)

6. Discussion

In this paper, we considered three model estimation methods for (secondary) data analysis when the outcome variable in a linear mixed effects model is latent and therefore measured with error. All of them fall within the scheme of two-stage estimation that embraces the advantages of "divide-and-conquer" strategy. Such advantages include convenience for model calibration and fit evaluation, avoidance of improper solutions, and convenience of secondary data analysis. The last aspect is especially appealing from a practical perspective because oftentimes, the raw response data is considered restricted-use data and not publicly available, whereas $\hat{\theta}$ (or certain linear transformation of it) with its SE are publicly available.

The three methods explored in the study overcome the limitation of the naïve two-stage estimation that ignores the measurement errors in latent trait estimates ($\hat{\theta}$) when treating them as dependent variables. It is known that ignoring the measurement error in $\hat{\theta}$ when $\hat{\theta}$ is treated as a dependent variable still yields a consistent and unbiased estimate of fixed effects (i.e., $\boldsymbol{\beta}$), but the standard error of $\boldsymbol{\beta}$ will be inflated, and the random effects estimates (i.e., $\widehat{\boldsymbol{\Sigma}}_u$) as well as residual variances will be distorted. For the MMLE and EM methods, the point estimate $\hat{\theta}_i$ and its corresponding measurement error for each student per time point are obtained in stage I measurement model calibration. And these two pieces of information become the key input for

stage II estimation. The moment estimation method, on the other hand, only needs population estimates of the mean and covariance matrix from stage I as input.

To elaborate, the MMLE method builds upon the assumption of (multivariate) normal measurement errors such that the marginal joint likelihood of the model parameters can be written in a closed form. This closed form marginal likelihood is then directly maximized to obtain parameter estimates. Neither the known challenge of curse of dimensionality (i.e., numerical approximation of a high-dimensional integration) nor the lengthy sampling iterations is an issue any more. Comparing to MMLE, the EM method has greater flexibility because it no longer requires the (multivariate) normal measurement error, which may not always be satisfied in practice especially when there are few items per dimension. Although in this paper and in the simulation studies, we still assume the measurement error of $\hat{\theta}$ follows normal/multivariate normal just to check the feasibility of the algorithm, it can be modified to incorporate non-normal measurement error cases.

The modification can be established based on the importance sampling idea. The critical piece to facilitate the entire importance sampling machinery is the change-of-measure sampling distribution, $H(\boldsymbol{\theta}_i, \boldsymbol{u}_i)$. Regardless of whether or not the multivariate normality assumption is satisfied, $H(\boldsymbol{\theta}_i, \boldsymbol{u}_i)$ can take the form of joint multivariate normal because it serves as a close approximation to the actual (and sometimes complicated) joint distribution of $(\boldsymbol{\theta}_i, \boldsymbol{u}_i)$. Moreover, the random values drawn from the sampling distribution of $H(\boldsymbol{\theta}_i, \boldsymbol{u}_i)$ are all independent, as opposed to the correlated draws from Gibbs or Metropolis-Hastings sampler in MCMC. The form of $H(\boldsymbol{\theta}_i, \boldsymbol{u}_i)$ can be derived based on the results from Stage I, and drawing samples from multivariate normal distribution is very easy, hence the numerical approximation to the expectation in EM becomes quite straightforward.

The proposed MMLE and EM are based on the measurement error model that is essentially a random-effects meta-regression (Raudenbush & Bryk, 1985; Raudenbush & Bryk, 2002, chapter 7), and it is in the broader framework for considering second-stage estimates in the presence of heteroscedasticity (Buonaccorsi, 1996). In particular, Buonaccorsi (1996) derived unbiased estimates of the structural model parameters (i.e., Σ_u) under different specific forms of heteroscedasticity. Because the conditional standard error of measurement from the 3PL model is a nonlinear function of both item parameters and θ, Buonaccorsi's (1996) derived results may not directly apply. However, the take-away message is the analytic results hold under the assumption of conditionally unbiased estimators and conditionally unbiased standard errors in stage I. Therefore, it is of paramount importance to obtain reliable $\hat{\theta}$ estimates in stage I. Diakow (2013) suggested using weighted maximum likelihood (WLE, Warm, 1989), and it is promising to check in the future for both unidimensional models and multidimensional models (Wang, 2015).

The plausible value multiple-imputation method is another method of addressing measurement error issues in large-scale educational statistical inference. The statistical theory of this method is that, as long as the plausible values are constructed from the results of a comprehensive extensive marginal analysis, population characteristics can be estimated accurately without attempting to produce accurate point estimates for individual students (Sirotnik & Wellington, 1977; Mislevy, et al., 1992). Because most imputation procedures available in standard statistical software packages (e.g., SAS, Stata, and SPSS) assume that observations are independent, research on imputation strategies in the context of linear mixed effects models (or multilevel models) is still limited. From a theoretical perspective, using a multilevel model at the imputation stage is recommended to ensure congeniality between the imputation model and the model used by the analyst (Meng, 1994; Drechsler, 2015). Several

researches have demonstrated plausible values drawn from a simplified model without accounting for higher level dependency yielded substantial bias for random effects and negligible bias for fixed effects in secondary analysis (Monseur & Adams, 2009; Diakow, 2010; Drechsler, 2015). Future research could compare the proposed methods with the plausible value approach.

The two methods considered in the paper (MMLE and EM) account for the potentially non-constant error variance in the dependent variable by including a measurement error model with heteroskedastic variance at the lowest level of the multilevel model. We consider these two methods convenient and useful alternative to the well-studied multiple imputation method. One profound advantage of the proposed methods is that is does not require a correct conditioning model, which is required in the multiple imputation method. This is important because it is almost infeasible to find, and to sample from, a correct conditioning model that is exhaustive of all possible nesting structures and secondary analyses are impossible to predict. However, these two proposed methods do rely on the precision of $\hat{\theta}$ and its SE estimates.

In this paper, we provide technical details for the three two-stage methods for interested readers to replicate and extend our study for other types of linear or nonlinear mixed effects models. The source code of all methods will also made available to readers upon request. On the other hand, the combined model (e.g., Equation 7) could potentially be fitted using off-the-shelf specialized software packages that can handle heteroskedastic variance at the lowest level, such as the gllamm command (Rabe-Hesketh, et al., 2004) in Stata (StatCorp, 2011) and HLM (Raudenbush, Bryk & Congdon, 2004).

There are two limitations of the study that worth mentioning. First, the IRT item parameters are assumed known throughout the study. If in case the calibration sample size is small that the sampling error can no longer be ignored, the Bootstrap–calibrated interval

estimates for θ (Liu & Yang, 2018) could be applied in stage I of the proposed two-stage framework. Second, while we focused only on the model parameters' point estimates and standard error estimates, future studies could go one step further to evaluate the power of detecting significant covariates (Ye, 2015). For that purpose, the simulation design will focus on manipulating the effect size of the covariate (treatment effect) and the amount of measurement error (which could be manipulated by test length).

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika, 49*(2), 155-173.

Anderson, J. C., & Gerbing, D. W. (1988). Structural equation modeling in practice: A review and recommended two-step approach. *Psychological Bulletin*, *103*, 411-423.

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics, 22*(1), 47-76.

Bentler, P. M. (1978). The interdependence of theory, methodology, and empirical data: Causal modeling as an approach to construct validation. *Longitudinal Drug Research*, 267-302.

Bacharach, V. R., Baumeister, A. A., & Furr, R. M. (2003). Racial and gender science achievement gaps in secondary education. *The Journal of Genetic Psychology*, *164*(1), 115–26.

Bianconcini, S., & Cagnone, S. (2012). A general multivariate latent growth model with applications to student achievement. *Journal of Educational and Behavioral Statistics*, *37*, 339-364.

Bollen, K. A. (1989). Structural equations with latent variables. John Wiley and Sons, Inc., New York.

Broyden, C.G. (1970). The convergence of a class of double-rank minimization algorithms 1. general considerations. *IMA Journal of Applied Mathematics*, *6*, 76.

Buonaccorsi, J. P. (1996). Measurement error in the response in the general linear model. *Journal of the American Statistical Association, 91*(434), 633-642.

Burt, R. S. (1973). Confirmatory factor-analytic structures and the theory construction process. *Sociological Methods & Research, 2*(2), 131-190.

Burt, R. S. (1976). Interpretational confounding of unobserved variables in structural equation models. *Sociological Methods & Research, 5*(1), 3-52.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM J. Scientific Computing*, 16, 1190–1208.

Cai, L. (2008). SEM of another flavour: Two new applications of the supplemented EM algorithm. *British Journal of Mathematical and Statistical Psychology, 61*(2), 309-329.

Carroll, R., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective* (2nd ed.). London, England: Chapman and Hall.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37-52.

Cohen, A. S., Bottge, B. A., & Wells, C. S. (2001). Using Item Response Theory to Assess Effects of Mathematics Instruction in Special Populations. *Exceptional Children*, *68*(1), 23–44. doi:10.1177/001440290106800102

De Boeck, P., & Wilson, M. (2004). *A framework for item response models*: Springer.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society. Series B (methodological)*, 1-38.

Devanarayan, V., & Stefanski, L. (2002). Empirical simulation extrapolation for measurement error models with replicate measurements. *Statistics & Probability Letters*, *59*, 219-225.
Diakow, R. (2010). *The use of plausible values in multilevel modeling*. Unpublished masters thesis, University of California, Berkeley.
Diakow, R. P. (2013). *Improving explanatory inferences from assessments*. Unpublished doctoral dissertation. University of California-Berkley.
Drechsler, J. (2015). Multiple imputation of multilevel missing data—rigor versus simplicity. *Journal of Educational and Behavioral Statistics, 40*(1), 69-95.
Fan, X., Chen, M., & Matsumoto, A. R. (1997). Gender differences in mathematics achievement: Findings from the National Education Longitudinal Study of 1988. *Journal of Experimental Education*, *65*(3), 229–242.
Fletcher, R. (1970). A new approach to variable metric algorithms. *The Computer Journal*, *13*, 317.
Fox, J.-P. (2010). Bayesian item response theory modeling: theory and applications. New York: Springer.
Fox, J.-P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.
Fox, J.-P., & Glas, C. A. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika, 68*(2), 169-191.
Fraine, B. De, Damme, J. Van, & Onghena, P. (2007). A longitudinal analysis of gender differences in academic self-concept and language achievement: A multivariate multilevel latent growth approach. *Contemporary Educational Psychology*, *32*(1), 132–150.
Fuller, W. (2006). *Measurement error models* (2nd ed.). New York, NY: John Wiley.
Goldfarb, D. (1970). A family of variable metric updates derived by variational means. Math. Comp., *24*, 23–26.
Goldhaber, D. D., & Brewer, D. J. (1997). Why Don't Schools and Teachers Seem to Matter? Assessing the Impact of Unobservables on Educational Productivity. *The Journal of Human Resources*, *32*(3), 505–523.
Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement. *American Educational Research Journal*, *42*(2), 371–406.
Hong, G., & Yu, B. (2007). Early-grade retention and children's reading and math learning in elementary years. *Educational Evaluation and Policy Analysis*, *29*, 239-261.
Hsiao, Y., Kwok, O., & Lai, M. (2018). Evaluation of two methods for modeling measurement errors when testing interaction effects with observed composite scores. Educational and Psychological Measurement, 78, 181-202.
Jeynes, W. H. (1999). Effects of Remarriage Following Divorce on the Academic Achievement of Children. *Journal of Youth and Adolescence*, *28*(3), 385–393. doi:10.1023/A:1021641112640
Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 79-93.
Khoo, S., West, S., Wu, W., & Kwok, O. (2006). Longitudinal methods. In M. Eid & E. Diener (Eds.), *Handbook of psychological measurement: A multimethod perspective* (pp. 301–317). Washington, DC: APA.

Koedel, C., Leatherman, R., & Parsons, E. (2012). Test measurement error and inference from value-added models. *The B. E. Journal of Economic Analysis and Policy*, *12*, 1-37.
Kohli, N., Hughes, J., Wang, C., Zopluoglu, C., & Davison, M. L. (2015). Fitting a linear–linear piecewise growth mixture model with unknown knots: A comparison of two common approaches to inference. *Psychological Methods, 20*(2), 259.
Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement*, 285-307.
Lindstrom, M. J. & Bates, D. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated measure data. *Journal of the American Statistical Association*, *83*, 1014-1022.
Liu, Y., & Yang, J. (2018). Bootstrap-calibrated interval estimates for latent variable scores in item response theory. *Psychometrika*, *83*, 333-354.
Lockwood, L. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANCOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, *39*, 22-52.
Lu, I. R., Thomas, D. R., & Zumbo, B. D. (2005). Embedding IRT in structural equation models: A comparison with regression based on IRT scores. *Structural Equation Modeling, 12*(2), 263-277.
Magis, D., & Raiche, G. (2012). On the relationships between Jeffrey's model and weighted likelihood estimation of ability under logistic IRT models. *Psychometrika*, *77*, 163-169.
Meng, X. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, *10*, 538-573.
Meng, X., & Rubin, D. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*, 267-278.
Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association, 80*(392), 993-997.
Mislevy, R. J. (1987). Recent developments in item response theory with implications for teacher certification. *Review of research in education*, 239-275.
Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133-161.
Mislevy, R. J., & Sheehan, K. M. (1989). Information matrices in latent-variable models. *Journal of Educational and Behavioral Statistics, 14*(4), 335-350.
Monseur, C. & Adams, R. J. (2009). Plausible values: How to deal with their limitations. *Journal of Applied Measurement*, 10(3), 320–334.
Nelder, J. A. and Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal* 7, 308–313.
Nussbaum, E., Hamilton, L., & Snow, R. (1997). Enhancing the validity and usefulness of large-scale educational assessment: IV.NELS:88 Science achievement to 12th grade. *American Educational Research Journal*, *34*, 151-173.
Pastor, D. A. (2003). The use of multilevel item response theory modeling in applied research: an illustration. *Applied Measurement in Education*, *16*, 223-243.
Pastor, D. A., & Beretvas, N. S. (2006). Longitudinal Rasch modeling in the context of psychotherapy outcomes assessment. *Applied Psychological Measurement*, *30*, 100-120.

Pinheiro, J. C., & Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of computational and Graphical Statistics, 4*(1), 12-35.
R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.
Rabe-Hesketh, S., & Skrondal, A. (2008). *Multilevel and longitudinal modeling using Stata*: STATA press.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM manual*. University of California, Berkeley: Berkeley Electronic Press.
Raudenbush, S. W. & Bryk, A. S. (1985). Empirical Bayes Meta-Analysis. *Journal of Educational and Behavioral Statistics*, *10*, 75-98.
Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). *HLM 6 for Windows* (Computer software). Lincolnwood, IL: Scientific Software International.
Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological methods, 5*(2), 199.
Rijmen, F., Vansteelandt, K., & De Boeck, P. (2008). Latent class models for diary method data: Parameter estimation by local computations. *Psychometrika, 73*(2), 167-182.
Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*. doi: 10.18637/jss.v048.i02
Shang, Y. (2012). Measurement error adjustment using the SIMEX method: An application to student growth percentiles. *Journal of Educational Measurement*, *49*, 446-465.
Shanno, D. F. (1970). Conditioning of quasi-Newton methods for function minimization. *Math. Comp.*, *24*, 647–656.
Sirotnik, K., & Wellington, R. (1977). Incidence sampling: an integrated theory for "matrix sampling". *Journal of Educational Measurement*, *14*, 343-399.
Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*(4), 563-575.
Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models*: CRC Press.
StataCorp (2011). *Stata Statistical Software: Release 12*. College Station, TX: StataCorp LP.
Stoel, R. D., Garre, F. G., Dolan, C., & Van Den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*(4), 439.
Thompson, N., & Weiss, D. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment*, *Research & Evaluation*, 16 (1). Available online: http://pareonline.net/getvn.asp?v=16&n=1
Tian, W., Cai, L., Thissen, D., & Xin, T. (2013). Numerical differentiation methods for computing error covariance matrices in item response theory modeling: An evaluation and a new proposal. *Educational and Psychological Measurement, 73*(3), 412-439.
van der Linden, W.J. & Glas, C.A.W. (Eds.) (2010). *Elements of adaptive testing*, (Statistics for Social and Behavioral Sciences Series). New York: Springer.
Verhelst, N. (2010). IRT models: Parameter estimation, statistical testing and application in EER. In B. P. Creemers, L. Kyriakides, & P. Sammons (Eds.), *Methodological advances in educational effectiveness research* (pp. 183-218). New York: Routledge.

von Davier, M., & Sinharay, S. (2007). An importance sampling EM algorithm for latent regression models. *Journal of Educational and Behavioral Statistics, 32*(3), 233-251.
Wang, C. (2015). On latent trait estimation in multidimensional compensatory item response models. *Psychometrika*, *80*, 428-449.
Wang, C., Kohli, N., & Henn, L. (2016). A second-order longitudinal model for binary outcomes: Item response theory versus structural equation modeling. *Structural Equation Modeling*: *A Multidisciplinary Journal*, *23*, 455-465.
Wang, C., & Nydick, S. (2015). Comparing Two Algorithms for Calibrating the Restricted Non-Compensatory Multidimensional IRT Model. *Applied Psychological Measurement* , *39*, 119-134.
Warm, T.A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427–450.
Ye, F. (2016). Latent growth curve analysis with dichotomous items: Comparing four approaches. *British Journal of Mathematical and Statistical Psychology*, *69*, 43-61.
Yves Rosseel (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36. URL http://www.jstatsoft.org/v48/i02/
Zwinderman, A. H. (1991). A generalized Rasch model for manifest predictors. *Psychometrika, 56*(4), 589-600.