MAXIMUM LIKELIHOOD ESTIMATION OF MULTILEVEL STRUCTURAL EQUATION MODELS WITH RANDOM SLOPES FOR LATENT COVARIATES

Nicholas J. Rockwood

loma linda university

This is a post-peer-review, pre-copyedit version of an article accepted for publication in *Psychometrika*. (DOI pending)

I thank the editor, associate editors, and reviewers, as well as Drs. Andrew Hayes, Paul De Boeck, Jolynn Pek, and Robert Cudeck for helpful comments and discussions that led to the improvement of this manuscript. A portion of this research was conducted at The Ohio State University.

Correspondence should be sent to Nicholas J. Rockwood, Division of Interdisciplinary Studies, School of Behavioral Health, Loma Linda University, 11065 Campus St., Loma Linda, CA 92350. Email: nrockwood@llu.edu
Website: njrockwood.com

MAXIMUM LIKELIHOOD ESTIMATION OF MULTILEVEL STRUCTURAL EQUATION MODELS WITH RANDOM SLOPES FOR LATENT COVARIATES

Abstract

A maximum likelihood estimation routine for two-level structural equation models with random slopes for latent covariates is presented. Because the likelihood function does not typically have a closed-form solution, numerical integration over the random effects is required. The routine relies upon a method proposed by du Toit and Cudeck (2009) for reformulating the likelihood function so that an often large subset of the random effects can be integrated analytically, reducing the computational burden of high-dimensional numerical integration. The method is demonstrated and assessed using a small-scale simulation study and an empirical example.

Key words: multilevel SEM, random effects, random slopes, maximum likelihood estimation

1. Introduction

Multilevel structural equation modeling (MSEM) is an emerging statistical framework for the analysis of hierarchical data, where the underlying units of analysis are nested within larger units, known as groups or clusters, a popular example being students nested within classrooms or schools. B. O. Muthén (1989) suggests that the development of the MSEM framework began decades ago with the analysis of multivariate random intercept models in the unpublished dissertation of Schmidt (1969). Since then, Goldstein and McDonald (1988), McDonald and Goldstein (1989), B. O. Muthén (1989), B. O. Muthén and Satorra (1989), Lee (1990), and McDonald (1993) have each proposed two-level structural equation models (SEMs). Liang and Bentler (2004) noted the similarities between each of the formulations presented, developed a general MSEM formulation that contained many of the previously proposed models as special cases, and constructed an efficient Expectation-Maximization (EM) algorithm which was implemented within EQS (Bentler, 2004). Similar models can be estimated using LISREL (Jöreskog & Sörbom, 1996) and the R (R Core Team, 2017) package lavaan (Rosseel, 2012).

The only types of level-2 latent variables that can be included in these models are latent factors and random intercepts, which allow the means of the modeled variables to randomly vary across groups or clusters. Yet random slopes, which allow the relationship between variables to randomly vary across clusters, are an important component of multilevel modeling (MLM). Frameworks for MSEMs with random slopes for observed covariates have been presented by Mehta and Neale (2005) and Rabe-Hesketh, Skrondal, and Pickles (2004), which are respectively implemented within the R package openMx (Neale et al., 2016) and the Stata (StataCorp, 2005) package gllamm (Rabe-Hesketh et al., 2004). Shin and Raudenbush (2010) have also presented a form of an MSEM that allows for random slopes for observed covariates. Although such models are quite flexible, one limitation is that random slopes may not be specified for latent variables. That is, random heterogeneity in the relationship among latent variables cannot be modeled, which is a severe limitation.

The modeling framework implemented within Mplus (L. K. Muthén & Muthén, 2017), on the other hand, allows for MSEMs with random slopes for observed and latent covariates. However, maximum likelihood (ML) estimation of such MSEMs is substantially more complex than for previously proposed MSEMs because the likelihood function does not have a closed-form solution. Therefore, the function must be approximated using numerical methods,

such as Gaussian quadrature (e.g., Pinheiro & Bates, 1995), which are computationally intensive. For example, referring to MSEMs with random slopes for latent covariates, Asparouhov and Muthén (2019a) state:

> Within the Mplus ML framework all the random effects will need to be numerically integrated and thus such an estimation is limited by the number of variables, covariates, and random effects. With more than 3 or 4 random effects the ML estimation based on numerical integration will be slow, less precise, and quite likely to lead to convergence problems.

As an alternative, the Mplus developers have adopted Bayesian estimation for such models, which avoids quadrature-based integration by sampling from the posterior distribution of the parameters. However, Bayesian estimation can also be computationally intensive, especially when the prior distributions and likelihood function are non-conjugate. Further, extra care must be taken to assess whether the sampling routine has converged onto the posterior distribution of the parameters, ensuring that the estimates are meaningful.

In this paper, it is demonstrated that ML estimation of MSEMs with random slopes for latent covariates does not require numerical integration over all of the random effects. Instead, a computational method proposed by du Toit and Cudeck (2009) in the context of nonlinear mixed effects models can be applied so that only a subset of the random effects within MSEMs, which tends to be fewer than 3 or 4 for most practical models, need to be numerically integrated. Thus, ML estimation of such models is not typically computationally impractical. In fact, it can be relatively fast and accurate.

The remainder of this paper is organized as follows. In Section 2, the general MSEM is introduced and the advantages of the MSEM framework are discussed. An ML estimation routine for the MSEM is presented in Section 3. Within the routine, the likelihood function is restructured so that only a subset of the random effects need to be numerically integrated. In Section 4, a small-scale simulation study is conducted to compare the convergence rate and estimates of the new estimation routine with those obtained using ML estimation in Mplus. An example model is then fit to data from the 2003 Program for International Student Assessment (PISA; OECD, 2003) in Section 5. The paper concludes with a discussion in Section 6.

2. Model

In this section, the general MSEM and the advantages of the MSEM framework are presented. Attention is restricted to two-level SEMs with (conditionally) normal response variables. Potential methods for extending the modeling framework to account for additional levels of analysis and other response distributions (e.g., Poisson, binomial) are discussed in Section 6.

Let $\mathbf{z}_j$ be a k-dimensional vector of cluster-level (i.e., level-2) observations for cluster j ($j = 1, \ldots, J$) and $\mathbf{y}_{ij}$ be a p-dimensional vector of individual-level (i.e., level-1) observation for unit i ($i = 1, \ldots, n_j$) nested within cluster j. The observations $(\mathbf{z}_j', \mathbf{y}_{ij}')'$ are modeled as

$$\begin{pmatrix} \mathbf{z}_j \\ \mathbf{y}_{ij} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\nu}^{(z)} \\ \boldsymbol{\nu}_j^{(y)} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\Lambda}_B^{(z)} \\ \boldsymbol{\Lambda}_B^{(y)} \end{pmatrix} \boldsymbol{\alpha}_j + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Lambda}_W \end{pmatrix} \boldsymbol{\eta}_{ij} + \begin{pmatrix} \boldsymbol{\epsilon}_{zj} \\ \boldsymbol{\epsilon}_{yij} \end{pmatrix}, \tag{1}$$

where $\boldsymbol{\nu}^{(z)}$ and $\boldsymbol{\nu}_j^{(y)}$ are k-dimensional and p-dimensional vectors of intercepts , $\boldsymbol{\Lambda}_B^{(z)}$ and $\boldsymbol{\Lambda}_B^{(y)}$ are $k \times m_B$ and $p \times m_B$ loading matrices for the m_B-dimensional vector of level-2 (i.e., between-cluster) latent factors $\boldsymbol{\alpha}_j$, $\boldsymbol{\Lambda}_W$ is a $p \times m_W$ loading matrix for the m_W-dimensional vector of level-1 (i.e., within-cluster) latent factors $\boldsymbol{\eta}_{ij}$, and $\boldsymbol{\epsilon}_{zj}$ and $\boldsymbol{\epsilon}_{yij}$ are k- and p-dimensional multivariate normal error vectors with means $\mathbf{0}$ and covariance matrices $\boldsymbol{\Theta}_B$ and $\boldsymbol{\Theta}_W$, respectively. It is assumed that the covariances between all combinations of $\boldsymbol{\alpha}_j$, $\boldsymbol{\eta}_{ij}$, $\boldsymbol{\epsilon}_{zj}$, and $\boldsymbol{\epsilon}_{yij}$ are $\mathbf{0}$.

Equation 1 is often termed the *measurement model*, as it relates the observed dependent variables to the latent variables. The within-cluster latent variables $\boldsymbol{\eta}_{ij}$ are modeled via the within-cluster structural model, which is defined as

$$\boldsymbol{\eta}_{ij} = \mathbf{B}_{Wj}\boldsymbol{\eta}_{ij} + \boldsymbol{\Gamma}_{Wj}\mathbf{x}_{ij} + \boldsymbol{\zeta}_{ij}, \tag{2}$$

where $\mathbf{B}_{Wj}$ is an $m_W \times m_W$ matrix relating the within-cluster latent variables to one another and $\boldsymbol{\Gamma}_{Wj}$ is an $m_W \times q$ matrix relating q observed level-1 covariates $\mathbf{x}_{ij}$ to level-1 latent variables. The disturbances $\boldsymbol{\zeta}_{ij}$ are multivariate normal random variables with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Psi}$. It is assumed that $\text{Cov}(\boldsymbol{\epsilon}_{yij}, \boldsymbol{\zeta}_{ij}) = \mathbf{0}$ for all combinations of i, j.

Plugging the equation for $\boldsymbol{\eta}_{ij}$ into Equation 1 results in the reduced form of the model for $\mathbf{y}_{ij}$:

$$\mathbf{y}_{ij} = \boldsymbol{\nu}_j^{(y)} + \boldsymbol{\Lambda}_B^{(y)}\boldsymbol{\alpha}_j + \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj})^{-1}\boldsymbol{\Gamma}_{Wj}\mathbf{x}_{ij} + \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj})^{-1}\boldsymbol{\zeta}_{ij} + \boldsymbol{\epsilon}_{yij}. \tag{3}$$

This model closely resembles the single-level SEM of B. Muthén (1984) except for the level-2 latent factors $\boldsymbol{\alpha}_j$ and the j subscripts for $\boldsymbol{\nu}_j^{(y)}$, $\mathbf{B}_{Wj}$, and $\boldsymbol{\Gamma}_{Wj}$, which imply that these parameters may vary across level-2 units. Suppose there are r total elements of $\boldsymbol{\nu}_j^{(y)}$, $\boldsymbol{\alpha}_j$, $\mathbf{B}_{Wj}$, and $\boldsymbol{\Gamma}_{Wj}$ that vary across clusters. All of these r elements can be combined into the vector $\boldsymbol{\eta}_j$, which differs from $\boldsymbol{\eta}_{ij}$, and modeled with the following between-cluster structural model:

$$\boldsymbol{\eta}_j = \boldsymbol{\mu} + \mathbf{B}_B\boldsymbol{\eta}_j + \boldsymbol{\Gamma}_B\mathbf{x}_j + \boldsymbol{\zeta}_j. \tag{4}$$

Here, $\boldsymbol{\zeta}_j$ contains r multivariate normal random effects with covariance matrix $\boldsymbol{\Omega}$. Further, $\boldsymbol{\mu}$, $\mathbf{B}_B$, and $\boldsymbol{\Gamma}_B$ are $r \times 1$, $r \times r$, and $r \times s$ vectors and matrices, respectively, containing fixed effects, and $\mathbf{x}_j$ contains s level-2 observed covariates.

As with single-level SEMs, the observed endogenous variables ($\mathbf{y}_{ij}$ and $\mathbf{z}_j$) can be regressed on observed and latent covariates at each level by constructing single-indicator latent variables and utilizing the $\mathbf{B}$ and $\boldsymbol{\Gamma}$ matrices at the corresponding level. This approached is discussed in the context of single-level SEMs by Bollen (1989) and others.

2.1. Advantages of the MSEM framework

The general MSEM framework has several advantages over traditional SEM and MLM frameworks. The main advantage over the traditional SEM framework is the ability to model dependence between lower-level observations due to clustering. Some advantages of the MSEM framework relative to the traditional MLM framework are the ability to easily fit multivariate MLMs and seamlessly integrate level-1 and level-2 response variables within a unified model.

Another notable advantage of the MSEM framework is the capability to specify measurement models (i.e., factor analysis models) using hierarchically structured data. In addition to accounting for random measurement error at both levels of analysis, multilevel factor analysis presents a more sophisticated framework for understanding level-2 constructs. Such level-2 latent factors may be reflective (i.e., climate, shared) constructs in which level-1 responses are viewed as indicators of some shared cluster-level variable (e.g., student ratings of their teacher's effectiveness), formative (i.e., contextual, configural) constructs in which the level-2 factor corresponds to the mean of the level-1 construct within a cluster (e.g., latent school means of students' interest in science), or both (Lüdtke et al., 2008; Marsh et al., 2012; Stapleton, Yang, & Hancock, 2016).

A final advantage of the MSEM framework, which has been the recent focus of many methodologists, is the ability to *latent-center* level-1 covariates. When a level-1 covariate varies at both levels of analysis, the relationship between the covariate and the response variable may differ at each level (Cronbach et al., 1976). Further, the covariate may have a different meaning at each level of analysis. Thus, it is recommended to decompose the covariate into additive and orthogonal components that vary only at the within- and between- levels, respectively. The relationship between each of these components and the response variable can then be modeled.

Within the MLM framework, this is almost always performed using an observed variable decomposition, where the observed cluster means of the covariate is a predictor on the between- level and the observed cluster-centered covariate is used as a predictor on the within-level (see, e.g., Enders & Tofighi, 2007). However, Lüdtke et al. (2008) demonstrated that, when the construct is reflective or the sampling ratio of lower-level units is small, the observed cluster means may contain measurement error due to sampling variability and this error may result in a biased estimate of the between-group effect if the true within- and between- effects differ. This bias is larger when the predictor has a lower intraclass correlation coefficient (ICC) and within-cluster sample sizes are smaller, as the cluster means are estimated less reliably.

Alternatively, the MSEM framework allows for a latent decomposition (i.e., latent centering) of the predictor, where the between- component is modeled using the latent cluster means and the within- component is modeled using the original predictor centered around the latent cluster means. Because the means are treated as latent, uncertainty in the true cluster means is directly modeled and the bias in the between-cluster effect is eliminated (Lüdtke et al., 2008).

3. Parameter estimation

The difficulty of ML parameter estimation for MSEMs with random slopes for latent covariates can be demonstrated using the latent covariate model in which the within-cluster effect of x_{ij} on y_{ij} is random. This model can be formulated within the general MSEM framework defined by Equations 1-4,

$$\begin{pmatrix} x_{ij} \\ y_{ij} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{x_j} \\ \alpha_{y_j} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{x_{ij}} \\ \eta_{y_{ij}} \end{pmatrix} \tag{5}$$

$$\begin{pmatrix} \eta_{x_{ij}} \\ \eta_{y_{ij}} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta_{Wj} & 0 \end{pmatrix} \begin{pmatrix} \eta_{x_{ij}} \\ \eta_{y_{ij}} \end{pmatrix} + \begin{pmatrix} \zeta_{x_{ij}} \\ \zeta_{y_{ij}} \end{pmatrix} \tag{6}$$

$$\begin{pmatrix} \alpha_{x_j} \\ \alpha_{y_j} \\ \beta_{Wj} \end{pmatrix} = \begin{pmatrix} \mu_{\alpha_x} \\ \mu_{\alpha_y} \\ \mu_{\beta_{Wj}} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ \beta_B & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{x_j} \\ \alpha_{y_j} \\ \beta_{W,j} \end{pmatrix} + \begin{pmatrix} \zeta_{\alpha_{x_j}} \\ \zeta_{\alpha_{y_j}} \\ \zeta_{\beta_{Wj}} \end{pmatrix}, \tag{7}$$

which treats β_{Wj} as a random variable that is modeled at the between-cluster level. In scalar form, it can be seen that x_{ij} is decomposed into latent between-cluster (α_{x_j}) and within-cluster ($\eta_{x_{ij}}$) components,

$$x_{ij} = \alpha_{x_j} + \eta_{x_{ij}}. \tag{8}$$

The outcome variable y_{ij}, which is also decomposed into between-cluster (α_{y_j}) and within-cluster ($\eta_{y_{ij}}$) variables, is then regressed on each of these components and the slope for $\eta_{x_{ij}}$ is random,

$$y_{ij} = \underbrace{\mu_{\alpha_y} + \beta_B \alpha_{x_j} + \zeta_{\alpha_{y_j}}}_{\alpha_{y_j}} + \underbrace{\beta_{Wj}\eta_{x_{ij}} + \zeta_{y_{ij}}}_{\eta_{y_{ij}}}. \tag{9}$$

Modeling β_{Wj} as random results in an interaction between the level-2 latent variable β_{Wj} and level-1 latent variable $\eta_{x_{ij}}$. This latent variable interaction, which drastically increases the computational complexity of the model, is not possible within many of the MSEM frameworks discussed previously, such as those of Liang and Bentler (2004), Rabe-Hesketh et al. (2004), and Mehta and Neale (2005).

When β_{Wj} is fixed, then y_{ij} is a linear function of the normally distributed random variables α_{x_j}, $\zeta_{\alpha_{y_j}}$, $\eta_{x_{ij}}$, and $\zeta_{y_{ij}}$. As a result, the marginal distribution of y_{ij} is also normal and so the likelihood function can be computed in closed-form. But when β_{Wj} is random, the marginal distribution of y_{ij} is not normal and cannot be computed in closed-form. Instead, it must be numerically approximated. For example, in Mplus an (accelerated) EM algorithm is implemented in which the full r-dimensional integral is approximated using quadrature within the E-step (i.e., α_{x_j}, $\zeta_{\alpha_{y_j}}$, and $\eta_{x_{ij}}$ are all numerically integrated). Since the computational burden of quadrature-based numerical integration increases exponentially as a function of the dimension of integration, the types of models that can practically be fit using ML in Mplus are restricted to those with relatively small r.

However, as will be demonstrated here, r-dimensional numerical integration is not needed for the MSEM. Instead, a method for reducing the dimension of numerical integration introduced in the context of nonlinear mixed models can be applied. du Toit and Cudeck (2009) demonstrated that nonlinear mixed models with normally distributed random effects and conditionally normal response variables may contain random effects that enter the

function both linearly and nonlinearly. By conditioning on the nonlinear random effects, they were able to integrate the linear random effects out of the likelihood function analytically. This is straightforward as the sum of normally distributed random variables is also normally distributed and, by definition, the data are modeled as a linear function of the linear random effects.

An application of this method was presented by Cudeck, Harring, and du Toit (2009) for a specific single-level SEM in which two latent factors interact. By conditioning on one of the factors, the remaining factors enter the model linearly. Thus, only one dimension of numerical integration was required for the ML estimation approach. This method, which has not yet been implemented for the estimation of MSEMs, is relied upon here.

A ML estimation routine for the general MSEM defined in Section 2 is presented in this section. After defining the likelihood function, the model is slightly reparameterized to help distinguish between linear and nonlinear random effects. Next, the nonlinear random effects are conditioned upon so that the conditional distribution of the linear random effects, as well as the conditional distribution of the data marginalized over the linear random effects, can be derived. After simplifying the computations for this conditional distribution, Gaussian quadrature is described for numerically integrating over the nonlinear random effects and the computation of standard errors is discussed.

3.1. Likelihood

Letting $\mathbf{y}_j = (\mathbf{y}_{1j}', \mathbf{y}_{2j}', \ldots, \mathbf{y}_{n_j,j}')'$ and $\boldsymbol{\vartheta}$ contain all of the free and non-redundant parameters, the marginal likelihood for $\mathbf{d}_j = (\mathbf{z}_j', \mathbf{y}_j')'$ is

$$\mathcal{L}_j(\boldsymbol{\vartheta}|\mathbf{d}_j) = \int_{\boldsymbol{\eta}_j} \left[\prod_{i=1}^{n_j} f(\mathbf{y}_{ij}|\boldsymbol{\eta}_j, \boldsymbol{\vartheta}) \right] f(\mathbf{z}_j|\boldsymbol{\eta}_j, \boldsymbol{\vartheta}) f(\boldsymbol{\eta}_j|\boldsymbol{\vartheta}) d\boldsymbol{\eta}_j, \tag{10}$$

where

$$\begin{gathered} \mathbf{y}_{ij}|\boldsymbol{\eta}_j \sim \mathcal{N}(\boldsymbol{\mu}_{y_{ij}}, \boldsymbol{\Sigma}_W), \quad \mathbf{z}_j|\boldsymbol{\eta}_j \sim \mathcal{N}(\boldsymbol{\nu}^{(z)} + \boldsymbol{\Lambda}_B^{(z)}\boldsymbol{\alpha}_j, \boldsymbol{\Theta}_B), \\ \text{and} \quad \boldsymbol{\eta}_j \sim \mathcal{N}\big((\mathbf{I} - \mathbf{B}_B)^{-1}(\boldsymbol{\mu} + \boldsymbol{\Gamma}_B\mathbf{x}_j), \boldsymbol{\Sigma}_\eta\big), \end{gathered} \tag{11}$$

with

$$\boldsymbol{\mu}_{\mathbf{y}_{ij}} = \boldsymbol{\nu}_j^{(y)} + \boldsymbol{\Lambda}_B^{(y)}\boldsymbol{\alpha}_j + \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj})^{-1}\boldsymbol{\Gamma}_{Wj}\mathbf{x}_{ij}, \tag{12}$$

$$\boldsymbol{\Sigma}_W = \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj})^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B}_{Wj})^{-1'}\boldsymbol{\Lambda}_W' + \boldsymbol{\Theta}_W, \quad \text{and} \quad \boldsymbol{\Sigma}_\eta = (\mathbf{I} - \mathbf{B}_B)^{-1}\boldsymbol{\Omega}(\mathbf{I} - \mathbf{B}_B)^{-1'}.$$

Note that within this formulation, the level-1 latent factors $\boldsymbol{\eta}_{ij}$ have already been marginalized out of the likelihood function, but the r-dimensional integration over the level-2 random effects remains. However, by relying on the method proposed by du Toit and Cudeck (2009), the level-2 random effects can be partitioned into those that enter linearly and those that enter nonlinear. By conditioning on the nonlinear random effect, the linear random effect can be integrated out of the likelihood function analytically.

3.2. Linear vs. nonlinear random effects

Letting $\boldsymbol{\gamma}_{Wj} = \text{vec}(\boldsymbol{\Gamma}_{Wj})$, $\boldsymbol{\Gamma}_{Wj}\mathbf{x}_{ij}$ can be rewritten as

$$\boldsymbol{\Gamma}_{Wj}\mathbf{x}_{ij} = (\mathbf{x}'_{ij} \otimes \mathbf{I}_{m_W})\boldsymbol{\gamma}_{Wj}, \tag{13}$$

where $\otimes$ denotes the Kronecker product and the vec$(\cdot)$ operator stacks the columns of the corresponding matrix. Further, $\boldsymbol{\nu}_j^{(z)}$ can be constructed from $\boldsymbol{\nu}^{(z)}$ and $\boldsymbol{\epsilon}_{zj}$, such that $\boldsymbol{\nu}_j^{(z)} = \boldsymbol{\nu}^{(z)} + \boldsymbol{\epsilon}_{zj}$. Consequently, all residual random variation in $\mathbf{z}_j$ is modeled via $\boldsymbol{\nu}_j^{(z)}$, which is now included within $\boldsymbol{\eta}_j$, and $\boldsymbol{\Omega}$ is expanded to include the elements within $\boldsymbol{\Theta}_B$. This step is not necessary, per se, but it will simplify later computations. After this slight reparameterization, the reduced-form for the general MSEM can be written as:

$$\begin{pmatrix} \mathbf{z}_j \\ \mathbf{y}_{ij} \end{pmatrix} = \begin{pmatrix} \mathbf{G} \\ \mathbf{Q}_{ij} \end{pmatrix} \boldsymbol{\xi}_j + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj})^{-1} \end{pmatrix} \boldsymbol{\zeta}_{ij} + \begin{pmatrix} \mathbf{0} \\ \boldsymbol{\epsilon}_{yij} \end{pmatrix} \tag{14}$$

where

$$\begin{pmatrix} \mathbf{G} \\ \mathbf{Q}_{ij} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \boldsymbol{\Lambda}_B^{(z)} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \boldsymbol{\Lambda}_B^{(y)} & \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj})^{-1}(\mathbf{x}'_{ij} \otimes \mathbf{I}_{m_W}) \end{pmatrix}, \tag{15}$$

and $\boldsymbol{\xi}_j = (\boldsymbol{\nu}_j^{(z)'}, \boldsymbol{\nu}_j^{(y)'}, \boldsymbol{\alpha}'_j, \boldsymbol{\gamma}'_{Wj})'$. Only a subset of $\boldsymbol{\xi}_j$ will vary across clusters. These elements that vary across clusters can be denoted using a tilde ("$\sim$"). That is, let $\tilde{\boldsymbol{\nu}}_j$, $\tilde{\boldsymbol{\alpha}}_j$, and $\tilde{\boldsymbol{\gamma}}_{Wj}$ denote the elements of $\boldsymbol{\nu}_j$, $\boldsymbol{\alpha}_j$, and $\boldsymbol{\gamma}_{Wj}$ that vary across clusters j, respectively, where $\boldsymbol{\nu}_j = (\boldsymbol{\nu}_j^{(z)'}, \boldsymbol{\nu}_j^{(y)'})'$. These elements can be combined, in the same order as previously listed, into the vector $\tilde{\boldsymbol{\xi}}_j$. Let $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{Q}}_{ij}$ denote the columns of $\mathbf{G}$ and $\mathbf{Q}_{ij}$ corresponding to the elements of $\tilde{\boldsymbol{\xi}}_j$. That is, if $\tilde{\boldsymbol{\xi}}_j$ consists of the first, third, and seventh elements of $\boldsymbol{\xi}_j$, then $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{Q}}_{ij}$ contains the first, third, and seventh columns of $\mathbf{G}$ and $\mathbf{Q}_{ij}$, respectively. The elements of $\boldsymbol{\beta}_{Wj} = \text{vec}(\mathbf{B}_{Wj})$ that vary across clusters can similarly be denoted as $\tilde{\boldsymbol{\beta}}_{Wj}$. Thus, the level-2 random effects are $\boldsymbol{\eta}_j = (\tilde{\boldsymbol{\xi}}'_j, \tilde{\boldsymbol{\beta}}'_{Wj})'$.

Within this formulation, it can be seen that the observed data are modeled as a linear function of $\boldsymbol{\xi}_j$ (and, thus, $\tilde{\boldsymbol{\xi}}_j$). The random effects $\tilde{\boldsymbol{\beta}}_{Wj}$, on the other hand, enter the model nonlinearly as these elements are multiplied by the random vectors $\boldsymbol{\zeta}_{ij}$ and $\tilde{\boldsymbol{\gamma}}_{Wj}$, resulting in one or more latent variable interactions. However, once the nonlinear random effects are conditioned upon, the (conditional) likelihood of the data marginalized over the linear random effects $\tilde{\boldsymbol{\xi}}_j$ and $\boldsymbol{\zeta}_{ij}$ can be computed analytically. Consequently, only the random effects $\tilde{\boldsymbol{\beta}}_{Wj}$ will need to be numerically integrated within the likelihood function.

3.3. Conditional distribution of the linear random effects

Within the general MSEM, the distribution of $\boldsymbol{\eta}_j$ is multivariate normal:

$$\boldsymbol{\eta}_j \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\eta}}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}}), \tag{16}$$

where

$$\boldsymbol{\mu}_{\boldsymbol{\eta}} = (\mathbf{I} - \mathbf{B}_B)^{-1}(\boldsymbol{\mu} + \boldsymbol{\Gamma}_B \mathbf{x}_j), \tag{17}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = (\mathbf{I} - \mathbf{B}_B)^{-1}\boldsymbol{\Omega}(\mathbf{I} - \mathbf{B}_B)^{-1'}. \tag{18}$$

Partitioning $\boldsymbol{\eta}_j$ into the two subvectors $\tilde{\boldsymbol{\xi}}_j$ and $\tilde{\boldsymbol{\beta}}_{Wj}$ corresponds to partitioning $\boldsymbol{\eta}_j$ into the random effects that enter the model linearly and nonlinearly, respectively. The vector $\boldsymbol{\mu}_{\boldsymbol{\eta}}$ and matrix $\boldsymbol{\Sigma}_{\boldsymbol{\eta}}$ can also be partitioned accordingly:

$$\boldsymbol{\eta}_j = \begin{pmatrix} \tilde{\boldsymbol{\xi}}_j \\ \tilde{\boldsymbol{\beta}}_{Wj} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{\xi}} \\ \boldsymbol{\mu}_{\boldsymbol{\beta}_W} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\xi}} & \boldsymbol{\Sigma}_{\boldsymbol{\xi},\boldsymbol{\beta}_W} \\ \boldsymbol{\Sigma}_{\boldsymbol{\beta}_W,\boldsymbol{\xi}} & \boldsymbol{\Sigma}_{\boldsymbol{\beta}_W} \end{pmatrix} \right]. \tag{19}$$

The conditional distribution of $\tilde{\boldsymbol{\xi}}_j$ given $\tilde{\boldsymbol{\beta}}_{Wj} = \mathbf{b}$ then follows as

$$\tilde{\boldsymbol{\xi}}_j | \tilde{\boldsymbol{\beta}}_{Wj} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}, \boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}), \tag{20}$$

with

$$\boldsymbol{\mu}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W} = \boldsymbol{\mu}_{\boldsymbol{\xi}} + \boldsymbol{\Sigma}_{\boldsymbol{\xi},\boldsymbol{\beta}_W}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_W}^{-1}(\mathbf{b} - \boldsymbol{\mu}_{\boldsymbol{\beta}_W}), \tag{21}$$

and

$$\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W} = \boldsymbol{\Sigma}_{\boldsymbol{\xi}} - \boldsymbol{\Sigma}_{\boldsymbol{\xi},\boldsymbol{\beta}_W}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_W}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_W,\boldsymbol{\xi}}. \tag{22}$$

3.4. Conditional distribution of the response variables

The distribution of $\mathbf{d}_j = (\mathbf{z}_j', \mathbf{y}_j')'$ conditional on $\tilde{\boldsymbol{\beta}}_{Wj}$ can then be derived using the conditional distribution of $\tilde{\boldsymbol{\xi}}_j$. Letting $\boldsymbol{\nu}_j^*$, $\boldsymbol{\alpha}_j^*$, $\boldsymbol{\Gamma}_{Wj}^*$, $\mathbf{B}_{Wj}^*$, and $\tilde{\mathbf{Q}}_{ij}^*$ denote the original vectors and matrices with the linear random effects replaced by their corresponding conditional expectation in $\boldsymbol{\mu}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}$ and the nonlinear random effects replaced by the value for which they are conditioned on, $\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj}$ is multivariate normal with mean

$$\boldsymbol{\mu}_{\mathbf{d}_j} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{z}_j} \\ \boldsymbol{\mu}_{\mathbf{y}_j} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{z}_j} \\ \boldsymbol{\mu}_{\mathbf{y}_{1j}} \\ \boldsymbol{\mu}_{\mathbf{y}_{2j}} \\ \vdots \\ \boldsymbol{\mu}_{\mathbf{y}_{n_j,j}} \end{pmatrix}, \tag{23}$$

where

$$\boldsymbol{\mu}_{\mathbf{z}_j} = \boldsymbol{\nu}_j^{(z)*} + \boldsymbol{\Lambda}_B^{(z)}\boldsymbol{\alpha}_j^*, \tag{24}$$

and

$$\boldsymbol{\mu}_{\mathbf{y}_{ij}} = \boldsymbol{\nu}_j^{(y)*} + \boldsymbol{\Lambda}_B^{(y)}\boldsymbol{\alpha}_j^* + \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj}^*)^{-1}\boldsymbol{\Gamma}_{Wj}^*\mathbf{x}_{ij}. \tag{25}$$

The (conditional) covariance matrix of $\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj}$ is given as

$$\boldsymbol{\Sigma}_{\mathbf{d}_j} = \begin{pmatrix} \tilde{\mathbf{G}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{G}}' & \text{symmetric} \\ \tilde{\mathbf{Q}}_j^*\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{G}}' & \tilde{\mathbf{Q}}_j^*\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{Q}}_j^{*'} + \mathbf{I}_{n_j} \otimes \boldsymbol{\Sigma}_W^* \end{pmatrix} \tag{26}$$

where

$$\tilde{\mathbf{Q}}_j^* = \begin{pmatrix} \tilde{\mathbf{Q}}_{1j}^* \\ \tilde{\mathbf{Q}}_{2j}^* \\ \vdots \\ \tilde{\mathbf{Q}}_{n_j,j}^* \end{pmatrix}, \tag{27}$$

and $\boldsymbol{\Sigma}_W^* = \boldsymbol{\Lambda}_W(\mathbf{I} - \mathbf{B}_{Wj}^*)^{-1}\boldsymbol{\Psi}(\mathbf{I} - \mathbf{B}_{Wj}^*)^{-1'}\boldsymbol{\Lambda}_W' + \boldsymbol{\Theta}_W$.

Letting

$$\boldsymbol{\epsilon}_{\mathbf{d}_j} = \begin{pmatrix} \boldsymbol{\epsilon}_{\mathbf{z}_j} \\ \boldsymbol{\epsilon}_{\mathbf{y}_j} \end{pmatrix} = \begin{pmatrix} \mathbf{z}_j - \boldsymbol{\mu}_{\mathbf{z}_j} \\ \mathbf{y}_j - \boldsymbol{\mu}_{\mathbf{y}_j} \end{pmatrix}, \tag{28}$$

the conditional density of $\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj}$ can be written as:

$$f(\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj}) = (2\pi)^{-(pn_j+k)/2}|\boldsymbol{\Sigma}_{\mathbf{d}_j}|^{-1/2}\exp\left\{ -\frac{1}{2}\boldsymbol{\epsilon}_{\mathbf{d}_j}'\boldsymbol{\Sigma}_{\mathbf{d}_j}^{-1}\boldsymbol{\epsilon}_{\mathbf{d}_j} \right\}. \tag{29}$$

3.5. Simplifying the conditional distribution of $\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj}$

The matrix $\boldsymbol{\Sigma}_{\mathbf{d}_j}$ is often of a very high dimension $(pn_j + k)$ and the calculation of its determinant and inverse, which must be computed for each group j, is computationally intensive. Luckily the expressions for $|\boldsymbol{\Sigma}_{\mathbf{d}_j}|$ and $\boldsymbol{\epsilon}'_{\mathbf{d}_j}\boldsymbol{\Sigma}^{-1}_{\mathbf{d}_j}\boldsymbol{\epsilon}_{\mathbf{d}_j}$ can be simplified by following similar derivations provided by McDonald (1993) and du Toit and du Toit (2008, Appendix).

First, define:

$$\begin{gathered}
\mathbf{A}_j = \sum_i^{n_j} \mathbf{A}_{ij} = \sum_i^{n_j} \tilde{\mathbf{Q}}^{*\prime}_{ij}\boldsymbol{\Sigma}^{*-1}_W\tilde{\mathbf{Q}}^{*}_{ij}, \quad \mathbf{T}_j = (\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W} + \mathbf{A}_j)^{-1}, \\
\mathbf{C}_j = (\mathbf{I} - \mathbf{A}_j\mathbf{T}_j), \quad \mathbf{D}_j = (\mathbf{I} - \mathbf{A}_j\mathbf{T}_j)\mathbf{A}_j = \mathbf{C}_j\mathbf{A}_j, \\
\boldsymbol{\Sigma}_{zz.y} = \tilde{\mathbf{G}}(\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W} - \boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\mathbf{D}_j\boldsymbol{\Sigma}'_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W})\tilde{\mathbf{G}}', \\
\mathbf{E}_j = \tilde{\mathbf{G}}'\boldsymbol{\Sigma}^{-1}_{zz.y}\tilde{\mathbf{G}}, \quad \mathbf{F}_j = \mathbf{C}'_j\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\mathbf{E}_j\boldsymbol{\Sigma}'_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\mathbf{C}_j, \\
\mathbf{H}_j = \mathbf{F}_j - \mathbf{T}_j, \quad \text{and} \quad \mathbf{p}_j = \sum_{i=1}^{n_j}\mathbf{p}_{ij} = \sum_{i=1}^{n_j}\tilde{\mathbf{Q}}^{*\prime}_{ij}\boldsymbol{\Sigma}^{*-1}_W\boldsymbol{\epsilon}_{\mathbf{y}_{ij}}.
\end{gathered} \tag{30}$$

Using these expressions, the determinant of $\boldsymbol{\Sigma}_{\mathbf{d}_j}$ can be re-expressed as:

$$|\boldsymbol{\Sigma}_{\mathbf{d}_j}| = |\boldsymbol{\Sigma}^*_W|^{n_j}|\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}||\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W} + \mathbf{A}_j||\boldsymbol{\Sigma}_{zz.y}|, \tag{31}$$

and $\boldsymbol{\epsilon}'_{\mathbf{d}_j}\boldsymbol{\Sigma}^{-1}_{\mathbf{d}_j}\boldsymbol{\epsilon}_{\mathbf{d}_j}$ can be re-expressed as:

$$\begin{aligned}
\boldsymbol{\epsilon}'_{\mathbf{d}_j}\boldsymbol{\Sigma}^{-1}_{\mathbf{d}_j}\boldsymbol{\epsilon}_{\mathbf{d}_j} =& \operatorname{tr}\Big[\boldsymbol{\Sigma}^{*-1}_W\sum_{i=1}^{n_j}\boldsymbol{\epsilon}_{\mathbf{y}_{ij}}\boldsymbol{\epsilon}'_{\mathbf{y}_{ij}}\Big] + \mathbf{p}'_j\mathbf{H}_j\mathbf{p}_j \\
&- 2\mathbf{p}'_j\mathbf{C}'_j\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{G}}'\boldsymbol{\Sigma}^{-1}_{zz.y}\boldsymbol{\epsilon}_{\mathbf{z}_j} \\
&+ \boldsymbol{\epsilon}'_{\mathbf{z}_j}\boldsymbol{\Sigma}^{-1}_{zz.y}\boldsymbol{\epsilon}_{\mathbf{z}_j}.
\end{aligned} \tag{32}$$

Each of these simplifications require much smaller matrix inversions and determinants than computing Equation 29 directly. Specifically, the largest matrix inversion or determinant is of dimension $\max\{p, k, r_L\}$, where r_L is the number of linear level-2 random effects (i.e., the number of elements in $\tilde{\boldsymbol{\xi}}_j$).

3.5.1. Further simplifications for special cases of the model

Various special cases of the MSEM lend themselves to even further simplified computations of the (conditional) likelihood. For example, when the model only includes

observed level-1 covariates with fixed effects, such that $\tilde{\boldsymbol{\xi}}_j$ contains no elements of $\boldsymbol{\Gamma}_{Wj}$ (i.e., $\boldsymbol{\Gamma}_{Wj} = \boldsymbol{\Gamma}_W$ for all j), the matrix $\tilde{\mathbf{Q}}^*_{ij}$ no longer varies across units (i.e., $\tilde{\mathbf{Q}}^*_{ij} = \tilde{\mathbf{Q}}^*$ for all ij). Consequently, the conditional density takes a strikingly similar structural form as the random intercept MSEM densities of McDonald and Goldstein (1989), McDonald (1993), and B. O. Muthén (1989), and so it can be further simplified using an adaptation of their derivations. Specifically, after defining

$$\begin{gathered}
\boldsymbol{\Sigma}_{zz} = \tilde{\mathbf{G}}\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{G}}', \quad \boldsymbol{\Sigma}_{yz} = \tilde{\mathbf{Q}}^*\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{G}}', \\
\boldsymbol{\Sigma}_{yy.z} = \tilde{\mathbf{Q}}^*\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{Q}}^{*\prime} - \boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{yz}', \\
\boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}^*_W + n_j\boldsymbol{\Sigma}_{yy.z}, \quad \text{and} \quad \bar{\boldsymbol{\epsilon}}_{\mathbf{y}_j} = n_j^{-1}\sum_{i=1}^{n_j}\boldsymbol{\epsilon}_{\mathbf{y}_{ij}},
\end{gathered} \tag{33}$$

the log of $f(\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj})$ can be reformulated as

$$\begin{aligned}
\log\{f(\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj})\} = &-\frac{1}{2}\Big\{(pn_j + k)\log(2\pi) + \log|\boldsymbol{\Sigma}_{zz}| + (n_j - 1)\log|\boldsymbol{\Sigma}^*_W| + \log|\boldsymbol{\Sigma}_j| \\
&+ \text{tr}\{[\boldsymbol{\Sigma}_{zz}^{-1} + n_j\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{yz}'\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_{yz}\boldsymbol{\Sigma}_{zz}^{-1}]\boldsymbol{\epsilon}_{\mathbf{z}_j}\boldsymbol{\epsilon}_{\mathbf{z}_j}'\} - 2n_j\text{tr}\{\boldsymbol{\Sigma}_{zz}^{-1}\boldsymbol{\Sigma}_{yz}'\boldsymbol{\Sigma}_j^{-1}\bar{\boldsymbol{\epsilon}}_{\mathbf{y}_j}\boldsymbol{\epsilon}_{\mathbf{z}_j}'\} \\
&+ \text{tr}\Big[\ \boldsymbol{\Sigma}_W^{*-1}\sum_{j}^{n_j}\boldsymbol{\epsilon}_{\mathbf{y}_{ij}}\boldsymbol{\epsilon}_{\mathbf{y}_{ij}}'\Big] - n_j\text{tr}\{[\boldsymbol{\Sigma}_W^{*-1} - \boldsymbol{\Sigma}_j^{-1}]\bar{\boldsymbol{\epsilon}}_{\mathbf{y}_j}\bar{\boldsymbol{\epsilon}}_{\mathbf{y}_j}'\}\Big\}.
\end{aligned} \tag{34}$$

3.6. Gaussian quadrature

The likelihood function for cluster j now requires an r_{NL}-dimensional integration:

$$\mathcal{L}_j(\boldsymbol{\vartheta}|\mathbf{d}_j) = \int f(\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj})f(\tilde{\boldsymbol{\beta}}_{Wj})d\tilde{\boldsymbol{\beta}}_{Wj}, \tag{35}$$

where r_{NL} is the dimension of $\tilde{\boldsymbol{\beta}}_{Wj}$ and $\boldsymbol{\vartheta}$ contains all freely estimated and non-redundant model parameters. The intractable integral can be approximated using Gaussian quadrature. For a multidimensional integral where $\tilde{\boldsymbol{\beta}}_{Wj}$ has mean $\boldsymbol{\mu}_{\boldsymbol{\beta}_W}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_W}$, the appropriate nodes $\mathbf{t}_q = (t_{q_1}, \ldots, t_{q_{r_{NL}}})'$ and weights $\mathbf{w}_q = (w_{q_1}, \ldots, w_{q_{r_{NL}}})'$ are:

$$\mathbf{t}_q = \boldsymbol{\mu}_{\boldsymbol{\beta}_W} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}_W}^{1/2}\mathbf{t}^*_q, \quad \text{and} \quad \mathbf{w}_q = \frac{1}{\sqrt{\pi}}\mathbf{w}^*_q, \tag{36}$$

where $\mathbf{t}^*_q$ and $\mathbf{w}^*_q$ are standard Gaussian quadrature nodes and weights, respectively, and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_W}^{1/2}$ is the Cholesky decomposition of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_W}$. Thus, using Q nodes per dimension, Equation 35 is approximated as

$$\mathcal{L}_j(\boldsymbol{\vartheta}|\mathbf{d}_j) = \sum_{q_1=1}^{Q}\cdots\sum_{q_{r_{NL}}=1}^{Q} f(\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj} = \mathbf{t}_q)w_{q_1}\ldots w_{q_{r_{NL}}}. \tag{37}$$

For models with large ICCs and multidimensional integrals, *adaptive* Gaussian quadrature, where the nodes are centered and scaled adaptively for each cluster, can provide a more accurate approximation to the log-likelihood with fewer nodes. The details of this approach are discussed by Pinheiro and Bates (1995) and Rabe-Hesketh, Skrondal, and Pickles (2002), but are omitted here for space-saving purposes.

3.7. Standard errors

Maximization of the log-likelihood function can then be carried out using any general purpose optimization algorithm. An advantage of direct maximization of the likelihood function, relative to an indirect maximization via the EM algorithm, is that standard errors for the parameters can be obtained more easily. Let

$$\mathcal{I}(\boldsymbol{\vartheta}) = -\frac{\partial^2 l(\boldsymbol{\vartheta}|\mathbf{d})}{\partial \boldsymbol{\vartheta} \partial \boldsymbol{\vartheta}'} \tag{38}$$

denote the observed information matrix, which is the negative of the Hessian matrix of the log-likelihood function. The standard errors for the estimated parameters can be estimated as

$$se(\hat{\vartheta}_j) = [\mathcal{I}(\hat{\boldsymbol{\vartheta}})^{-1}]_{jj}^{\frac{1}{2}}, \tag{39}$$

where $\hat{\vartheta}_j$ is the jth parameter estimate and $\mathcal{I}(\hat{\boldsymbol{\vartheta}})$ is the observed information matrix evaluated at the vector of parameter estimates, $\hat{\boldsymbol{\vartheta}}$.

3.8. Implementation within R/C++

For the following simulation study and example analysis, the developed routine was implemented within R (R Core Team, 2017), C++, and Rcpp (Eddelbuettel, 2013). Specifically, the likelihood function was coded in C++ and maximized within R using the nlminb algorithm in the optimx package (Nash, 2014). The maximization algorithm uses first-order derivatives (i.e., the gradient vector). Rather than relying on numerical derivatives, which can be imprecise and computationally costly, especially for models with many parameters, the derivatives were computed using automatic differentiation (Griewank & Walther, 2008) as implemented within the C++ Stan Math Library (Carpenter et al., 2015) via the Rstan package (Stan Development Team, 2016). The standard errors were computed at the final parameter estimates by taking the numerical derivatives of the gradient vector

computed using automatic differentiation. A gentle introduction to automatic differentiation for psychometrics research is provided by Cudeck (2005), and Carpenter et al. (2015) document its implementation within Stan. The R, C++, and Stan source code used for fitting the model (and other example models) is available at https://osf.io/pxz5s/. Ideally, these files, along with this paper, will prompt statistical software developers to implement the proposed estimation routine within existing packages and software.

4. Simulation

In this section, the performance of the newly proposed ML estimation routine is assessed and compared to the ML routine implemented within Mplus. In addition to determining whether the parameters can be recovered adequately, interest lies in whether the new routine can reduce some of the non-convergence issues that can be prevalent when estimating MSEMs using ML in Mplus.

The latent covariate model with a random slope is used for this simulation and the between component of the predictor x_{ij} predicts both the outcome y_{ij} and the random slope β_{Wj}. The full model in scalar form is

$$x_{ij} = \alpha_{x_j} + \eta_{x_{ij}} \tag{40}$$

$$y_{ij} = \alpha_{y_j} + \beta_{Wj}\eta_{x_{ij}} + \epsilon_{y_{ij}} \tag{41}$$

$$\alpha_{y_j} = \mu_{\alpha_y} + \beta_B\alpha_{x_j} + \zeta_{y_j} \tag{42}$$

$$\beta_{Wj} = \beta_{0W} + \beta_{1W}\alpha_{x_j} + \zeta_{\beta_{Wj}} \tag{43}$$

$$\eta_{x_{ij}} \sim \mathcal{N}(0, \psi_x), \quad \epsilon_{y_{ij}} \sim \mathcal{N}(0, \theta_y), \quad \alpha_{x_j} \sim \mathcal{N}(\mu_x, \omega_x), \tag{44}$$

$$\begin{pmatrix} \zeta_{y_j} \\ \zeta_{\beta_{Wj}} \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \omega_y & \omega_{y,\beta_W} \\ \omega_{\beta_W,y} & \omega_{\beta_W} \end{pmatrix} \right]. \tag{45}$$

This model was recently used by Asparouhov and Muthén (2019a) who demonstrated that Mplus' Bayesian estimation can properly recover the true parameters for their simulation condition ($\text{ICC}_x = .5, J = 500, n_j = 15$). The authors did not estimate the model using ML, but had difficulty with convergence of a slightly more complex model when smaller level-1 and level-2 sample sizes were used (Asparouhov & Muthén, 2019b, Model A2).

For this simulation, three factors that have been shown to impact estimation performance of MSEMs were varied. These include the number of clusters ($J = 100, 200, 400$), the number

of level-1 units within each cluster ($n = 5, 10, 20$), and the ICC of the predictor ($\text{ICC}_x = .05, .1, .2, .3$). The main parameters of interest are the structural parameters (β_B, β_{0W}, β_{1W}) and the random slope variance (ω_{β_W}). In the simulation, these parameters were specified to be $\beta_B = 0.50$, $\beta_{0W} = -0.50$, $\beta_{1W} = 0.25$, and $\omega_{\beta_W} = 0.30$. The values for the simulation conditions and parameters were chosen because they are in the range of what is typical in behavioral science research and what has been specified in other simulation studies for MSEMs (e.g., Lüdtke et al., 2008).

Using R, one-hundred datasets were simulated from each of the $3 \times 3 \times 4 = 36$ simulation conditions, resulting in a total of 3600 datasets. Each of the datasets were fit using the estimation routine proposed here within R/C++ (see previous section for implementation details) and using the ML estimation routine in Mplus (L. K. Muthén & Muthén, 2017) via the MplusAutomation R package (Hallquist & Wiley, 2018). All of the code needed to generate the data and fit the models is provided at https://osf.io/pxz5s/.

ML estimation in Mplus requires four dimensions of numerical integration for this model. The maximum of 11 adaptive quadrature nodes per dimension (due to computer memory constraints) was used in the simulation study. Regarding model specification, the predictor x_{ij} was explicitly decomposed using the procedure outlined by Preacher, Zhang, and Zyphur (2016), which is currently the only method for ML estimation of the parameters using Mplus. Using the ML procedure developed here, only one dimension of numerical integration is needed, corresponding to the random slope β_{Wj}. For unidimensional numerical integration, it can often be more efficient to use nonadaptive quadrature with more nodes than adaptive quadrature with fewer nodes. Consequently, 30 nonadaptive quadrature nodes were used.

4.1. Convergence

Using Mplus, a replication was deemed to have not converged if the output contained an error beginning with "The model estimation did not terminate normally...". Using the R/C++ implementation of the new method, there were no specific errors that occurred during estimation in any of the replications. However, to ensure that the residual covariance matrices remained positive definite during estimation, (residual) variances were constrained to be greater than or equal to .0001 and the model was parameterized so (residual) correlations were constrained to be within -.995 and .995. Therefore, the estimation was deemed to have not

properly converged if any of the final parameter estimates were equal to their bound constraint. Although this creates a slight contrast in the definition of nonconvergence for each method, both definitions are ultimately equal from a practical standpoint in that they both indicate that a particular set of parameter estimates are inadmissible.

The convergence rates for Mplus' ML estimation and the ML estimation routine developed here are displayed in Table 1. Overall, the ML estimation routine within Mplus had substantially more convergence issues than the ML estimation method introduced here. For example, out of all 3600 replications, the new method failed to converge only 27 times (rate $< .01$), whereas the Mplus ML routine failed to converge 790 times (rate = .22). The most problematic conditions were those with a level-1 sample size of $n_j = 5$. Using the new method, the impact of a small level-1 sample size could be somewhat offset by a larger level-2 sample size or ICC_x, as the routine only resulted in nonconvergence when the level-2 sample size and ICC_x were both small. This was not the case, however, with the Mplus ML routine. For example, even when $J = 400$ and $\text{ICC}_x = .30$, the Mplus ML routine failed to converge 62% of the replications when $n_j = 5$.

[Table 1 about here.]

4.2. Parameter estimates

The bias and root mean squared error (RMSE) for a subset of the parameter estimates obtained using the newly introduced ML routine and the ML routine implemented within Mplus are displayed in Tables 2 and 3, respectively. The bias and RMSE were computed using only the estimates from the converged models. To save space, only the bias and RMSE for β_{0W}, β_B, β_{1W}, and ω_{β_W} are displayed, as these correspond to the parameters that are typically of the most interest substantively.

[Table 2 about here.]

[Table 3 about here.]

The estimates of β_{0W} are generally unbiased using both methods under all 36 simulation conditions. Both estimation routines also recover ω_{β_W} quite well, though it is slightly underestimated in conditions with small level-1 and level-2 samples sizes and a low ICC_x. This

is not surprising as ML estimation is known to produce underestimated level-2 variance components when level-2 sample sizes are small (Maas & Hox, 2005).

When estimated using the ML routine developed here, the effects of the between-cluster latent component of x_{ij} (α_x) on the between-cluster latent component of y_{ij} and the random slope β_{Wj} have seemingly non-negligible positive bias in the same conditions (i.e., small level-1 and level-2 sample sizes and low ICC_x). As one or more of these factors increase, the bias decreases. For example, when $J = 200$, $n_j = 5$, and $\text{ICC}_x = .05$, the estimated bias of β_B is .25. However, the bias drops to .14 when $J = 400$, .05 when $\text{ICC}_x = .10$, and .05 when $n_j = 10$.

Although the bias of β_B and β_{1W} when the model was fit using ML in Mplus tends to also be somewhat larger in these conditions, it does not appear to change as systematically as a function of the simulation conditions. For example, referring back to when $J = 200$, $n_j = 5$, and $\text{ICC}_x = .05$, the estimated bias of β_B is $-.05$ using Mplus. The bias actually increases (in absolute value) when any one of the three simulation factors increase (e.g., moving from $J = 200$ to $J = 400$ or from $n_j = 5$ to $n_j = 10$).

Overall, the ML estimation routine developed here performed better than the Mplus ML routine in terms of convergence rate and bias, though there were some simulation conditions (e.g., when $n_j = 5$ and $\text{ICC}_x = .05$) when the Mplus routine was less biased. In total, the new estimation routine produced less biased estimates than Mplus in 78% of the 4 (parameters) $\times$ 36 (simulation conditions) = 144 cells. Further, the RMSE using the new method was less than that from Mplus in 79% of the cells. Thus, it appears that the newly developed routine may be able to both improve the convergence rate and the reliability of the estimates for such models (and simulation conditions).

5. Example

In this section, a more elaborate model is fit to data from the 2003 Program for International Student Assessment (PISA), a large-scale study designed to test the knowledge of 15-year-old students in mathematics, reading, and science (OECD, 2003). The students are the level-1 units and schools are the level-2 units. After describing the data and specifying the model, which highlights many of the advantages of the MSEM framework, the parameter estimates obtained using the newly proposed ML estimation routine are presented and compared to those obtained using ML estimation in Mplus.

5.1. Data

For this analysis, data from a total of 9,729 students nested within 359 schools in Spain are used. The dataset contains a mixture of item-level responses, as well as some scale scores. The following student-level variables are used in this analysis: students' perception of teacher support (SPTS, ICC = 0.12), four items measuring students' enjoyment related to mathematics (ENJ_1 to ENJ_4, ICCs from 0.03 to 0.04), and students' mathematics achievement score (MATH, ICC = 0.18). Because these variables contain both within-school and between-school variability, they can be modeled at both levels of analysis. Additionally, the following school-level variables are used: quality of educational resources (QUAL) and three items measuring teacher enthusiasm (ENTH_1 to ENTH_3).

5.2. Model

At the within-school level, interest lies in the relationship between students' perception of teacher support and mathematics achievement via enjoyment of mathematics. Further, the relationship between enjoyment and achievement is expected to vary across schools. Variability of this slope is modeled at the between-school level as a function of school-level student's perception of teacher support and quality of educational resources. Additional effects at the between-school level include the between- counterparts of those modeled at the within-school level, and the effects of teacher enthusiasm on perceptions of support, enjoyment, and mathematics achievement. Lastly, quality of educational resources is included as a covariate for all between-school effects.

Formally, the measurement model is specified as

$$\begin{pmatrix} \text{ENTH}_{1j} \\ \text{ENTH}_{2j} \\ \text{ENTH}_{3j} \\ \text{SPTS}_{ij} \\ \text{ENJ}_{1ij} \\ \text{ENJ}_{2ij} \\ \text{ENJ}_{3ij} \\ \text{ENJ}_{4ij} \\ \text{MATH}_{ij} \end{pmatrix} = \begin{pmatrix} \nu_{1j} \\ \nu_{2j} \\ \nu_{3j} \\ \nu_4 \\ \nu_5 \\ \nu_{6j} \\ \nu_{7j} \\ \nu_{8j} \\ 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ \lambda_{B21} & 0 & 0 & 0 \\ \lambda_{B31} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \lambda_{B63} & 0 \\ 0 & 0 & \lambda_{B73} & 0 \\ 0 & 0 & \lambda_{B83} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{1j} \\ \alpha_{2j} \\ \alpha_{3j} \\ \alpha_{4j} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{W62} & 0 \\ 0 & \lambda_{W72} & 0 \\ 0 & \lambda_{W82} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{1ij} \\ \eta_{2ij} \\ \eta_{3ij} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \epsilon_{5ij} \\ \epsilon_{6ij} \\ \epsilon_{7ij} \\ \epsilon_{8ij} \\ 0 \end{pmatrix}, \tag{46}$$

and the level-1 structural model is

$$\begin{pmatrix} \eta_{1ij} \\ \eta_{2ij} \\ \eta_{3ij} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ \beta_{W21} & 0 & 0 \\ \beta_{W31} & \beta_{W32j} & 0 \end{pmatrix} \begin{pmatrix} \eta_{1ij} \\ \eta_{2ij} \\ \eta_{3ij} \end{pmatrix} + \begin{pmatrix} \zeta_{1ij} \\ \zeta_{2ij} \\ \zeta_{3ij} \end{pmatrix}. \tag{47}$$

For this analysis, the following cross-level measurement invariance constraints are placed on the loading parameters for the enjoyment factors: $\lambda_{B63} = \lambda_{W62}$, $\lambda_{B73} = \lambda_{W72}$, and $\lambda_{B83} = \lambda_{W82}$. Thus the between-school enjoyment factor is specified as a configural construct (see, e.g. Stapleton & Johnson, 2019), which represents the latent school means on the student-level enjoyment construct.

The vector of $r = 11$ level-2 random effects $\boldsymbol{\eta}_j$ includes all elements of the measurement model and level-1 structural model that contain (only) a j subscript. These random effects are modeled via the level-2 structural model:

$$\begin{pmatrix} \nu_{1j} \\ \nu_{2j} \\ \nu_{3j} \\ \nu_{6j} \\ \nu_{7j} \\ \nu_{8j} \\ \alpha_{1j} \\ \alpha_{2j} \\ \alpha_{3j} \\ \alpha_{4j} \\ \beta_{W32j} \end{pmatrix} = \begin{pmatrix} \mu_{\nu_1} \\ \mu_{\nu_2} \\ \mu_{\nu_3} \\ \mu_{\nu_6} \\ \mu_{\nu_7} \\ \mu_{\nu_8} \\ 0 \\ 0 \\ 0 \\ \mu_{\alpha_4} \\ \mu_{\beta_{W32}} \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & 0 & 0 & 0 & 0 \\ 0 & \dots & 0 & \beta_{B87} & 0 & 0 & 0 \\ 0 & \dots & 0 & \beta_{B97} & \beta_{B98} & 0 & 0 \\ 0 & \dots & 0 & \beta_{B10,7} & \beta_{B10,8} & \beta_{B10,9} & 0 \\ 0 & \dots & 0 & \beta_{B11,7} & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \nu_{1j} \\ \nu_{2j} \\ \nu_{3j} \\ \nu_{6j} \\ \nu_{7j} \\ \nu_{8j} \\ \alpha_{1j} \\ \alpha_{2j} \\ \alpha_{3j} \\ \alpha_{4j} \\ \beta_{W32j} \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \gamma_{B7} \\ \gamma_{B8} \\ \gamma_{B9} \\ \gamma_{B10} \\ \gamma_{B11} \end{pmatrix} \text{QUAL}_j + \begin{pmatrix} \zeta_{\nu_{1j}} \\ \zeta_{\nu_{2j}} \\ \zeta_{\nu_{3j}} \\ \zeta_{\nu_{6j}} \\ \zeta_{\nu_{7j}} \\ \zeta_{\nu_{8j}} \\ \zeta_{\alpha_{1j}} \\ \zeta_{\alpha_{2j}} \\ \zeta_{\alpha_{3j}} \\ \zeta_{\alpha_{4j}} \\ \zeta_{\beta_{W32j}} \end{pmatrix}. \tag{48}$$

Except for the freely estimated residual covariance ($\omega_{\alpha_4,\beta_{W32}}$) between the school-level math achievement variable (α_4) and the random slope (β_{W32}), all residuals at both levels are constrained to have covariances of zero. Since there is very little remaining between-school variability within the first enjoyment item (ENJ_1) after accounting for the enjoyment factor, the residual variance for this item at the between-school level is constrained to 0. Thus, ν_5 is fixed, rather than random.

The random slope from within-school enjoyment to math achievement is modeled as a function of school-level perceptions of support and quality of educational resources, so the effect is conditional and the interpretation of the intercept (i.e., $\mu_{\beta_{W32}}$) depends on the centering of support and resources. Therefore, when support is decomposed into latent within- and between- components, the between- level component is modeled as a sum of an intercept and a between- latent variable with mean 0, rather than an intercept fixed to 0 and a between-

latent variable with an estimated mean. Further, quality of educational resources was mean centered. Consequently, the conditional effect $\mu_{\beta_{W32}}$ is interpreted as the expected within-school effect of enjoyment on math for a school with average school-level perceptions of support and quality of educational resources.

This model specification results in a total of 48 freely estimated parameters (9 at the within-school level, 36 at the between-school level, and 3 factor loadings constrained to be equal at the within- and between-school levels). A path digram corresponding to the model is displayed in Figure 1. The diagram follows similarly to that used within the Mplus User Guide (L. K. Muthén & Muthén, 2017) and within publications utilizing MSEM (see, e.g., Preacher, Zyphur, & Zhang, 2010; Preacher et al., 2016). Observed variables are represented using rectangles and latent variables (i.e., factors and random effects) are represented using circles. The outer solid and dotted rectangles correspond to the unit at which the included variables vary. Here, the solid rectangle corresponds to between-school variability and the dotted rectangle corresponds to within-school variability. The observed student perception, enjoyment, and math achievement variables are within both rectangles, as they contain both within- and between-school variability, whereas the observed teacher enthusiasm variables only contain between-school variability. Lastly, the solid black circle on the path from η_{2ij} to η_{3ij} indicates that this slope is a random variable. It varies across schools, and so is represented as a latent variable (labeled β_{W32j}) at the between-school level.

[Figure 1 about here.]

Interestingly, this model utilizes all four of the advantages of the MSEM framework discussed in Section 2.1. First, the model is multivariate in that many response variables are simultaneously modeled. Second, factor models are used at the within- and between-school levels to model student enjoyment and a factor model is used at the between-school level to model teacher enthusiasm. Thus, random measurement error and unique variability for the individual items are separated from the latent factors. Third, level-2 response variables (the teacher enthusiasm items) are modeled. Lastly, a latent decomposition (i.e., latent centering) is used for students' perception of teacher support. By using the latent school-level means rather than the observed means, sampling error is reduced from the between-school component which ultimately reduces potential bias in the corresponding estimated between-school effects.

5.3. Parameter estimates

The parameter estimates obtained using the new estimation routine for the within- and between-school effects for the model are displayed in Tables 4 and 5, respectively. At the within-school level, there is a positive effect of students' perception of teacher support on enjoyment of mathematics ($\beta_{W21} = 0.17$), but a negative effect on mathematics performance ($\beta_{W31} = -0.12$) when controlling for enjoyment. The effect of enjoyment on performance, holding support constant, was modeled as a function of school-level support, the schools' quality of education resources, and a random school-specific residual. For schools with average support and resources, the effect of enjoyment on mathematics achievement is positive ($\mu_{\beta_{W32}} = 0.67$). Holding resources constant, the effect of enjoyment is stronger for schools with higher average perceptions of support ($\beta_{B11,8} = 0.17$). There is not enough evidence to suggest that the relationship is dependent on quality of educational resources ($\gamma_{B11} = 0.04$), but there is additional between-school variability in the effect not accounted for ($\omega_{\beta_{W32}} = 0.06$).

[Table 4 about here.]

At the between-school level, quality of educational resources is a positive predictor of teacher enthusiasm ($\gamma_{B7} = 0.11$) and math achievement ($\gamma_{B10} = 0.17$), but not students' perception of teacher support ($\gamma_{B8} = -0.01$) or student enjoyment of mathematics ($\gamma_{B9} = -0.01$). Teacher enthusiasm is positively associated with students' perception of teacher support ($\beta_{B87} = 0.14$), but not student enjoyment of mathematics ($\beta_{B97} = 0.02$). Teacher support is, on the other hand, a positive predictor of enjoyment ($\beta_{B98} = 0.15$). Finally, teacher enthusiasm ($\beta_{B10,7} = 0.26$) and student enjoyment ($\beta_{B10,9} = 2.42$) are positive predictors of mathematics achievement, whereas students perception of teacher support is a negative predictor ($\beta_{B10,8} = -0.70$).

[Table 5 about here.]

5.4. Comparison with Mplus

The parameter estimates obtained using ML estimation in Mplus are also displayed in Tables 4–5. In general, the estimates are nearly identical to those obtained using the newly proposed estimation routine, though there are some slightly larger (though not substantive)

differences in some of the estimated between-school structural parameters. There is also a slight difference between the log-likelihood using the new method (−69657.64) and the Mplus routine (−69656.34). Because both methods utilize ML estimation, the difference can be attributable to how well the quadrature method approximates the log-likelihood (as well as minor differences in convergence settings, etc.).

In Mplus, the model required eight dimensions of numerical integration and so three adaptive quadrature nodes per dimension (the maximum due to computer memory constraints) were used to approximate the log-likelihood function. In contrast, the new method only required one dimension of numerical integration, corresponding to the nonlinear random effect β_{W32j}. As such, seven nonadaptive quadrature nodes were used to obtain the parameter estimates. Refitting the model using the new method with 25 nonadaptive nodes resulted in parameter estimates and a log-likelihood equivalent to at least two decimal places, indicating that the approximation using the new method is perhaps more accurate than the approximation from Mplus' ML routine. Although the slight deviation between the log-likelihoods obtained using the two methods results in negligible differences in parameter estimates, it may be more influential when comparing competing models using a likelihood ratio test.

The most dramatic difference between the newly proposed ML estimation routine and ML estimation in Mplus is the estimation time. Within Mplus, it took several hours to obtain parameter estimates. With the new routine, however, parameter estimates could be obtained in under seven minutes. Thus, although both estimation routines converged and produced nearly equivalent results, the newly proposed method is substantially faster due to a reduction in the dimension of numerical integration from eight dimensions to one dimension. The difference in estimation time can become especially prominent when multiple models must be fit, such as when comparing competing models or conducting simulation studies.

6. Discussion

A ML estimation routine for two-level SEMs with random slopes for observed and latent covariates was proposed. The routine relies on a reformulation of the likelihood function so that some of the random effects can be integrated analytically. A brief simulation study demonstrated that the new method can recover the true parameters within most of the

simulated conditions and also reduce some of the convergence issues that have plagued ML estimation of such models. Further, a complex MSEM was fit to a real data set, which highlighted some ways in which MSEMs can be used to address interesting substantive research questions. In this section, alternative computational methods are discussed and limitations are addressed. The paper concludes with recommended future directions for the estimation of MSEMs and the assessment of the resulting parameter estimates.

6.1. Alternative computational methods

Two major decisions were made in the construction of the ML estimation routine. These include the decision to either directly or indirectly maximize the likelihood function and the method used for reducing the dimension of numerical integration.

The major computational hurdle within the estimation routine presented within this paper is the need for numerical integration. This computation hurdle is also present via indirect maximization methods. For example, the intractable integral is contained within the E-step of the EM algorithm. Although direct methods were used here, it is also possible to restructure the E-step integral to aid in numerical integration in a similar method to what was used in the direct maximization algorithm. Thus, the major contribution of this paper, which demonstrates a method for reducing the dimension of numerical integration required for the general MSEM model, is also applicable to indirect maximization methods in addition to the specific estimation routine presented here.

The method used for reducing the dimension of numerical integration is essentially an extension of the method presented by du Toit and Cudeck (2009) to the MSEM context. The restructuring of the likelihood function relied upon the (conditional) conjugacy of the distributions of the linear random effects and data. An alternative method for restructuring the likelihood function for latent variable models with intractable integrals has been suggested by Rijmen (2009), who generalized a method proposed by Gibbons and Hedeker (1992). This method relies upon graph theory to determine the dependencies of the random effects and, depending on their structure, reformulates the likelihood so that the dimension of integration is reduced. This method has been particularly useful for multidimensional item response theory models (Gibbons & Hedeker, 1992; Rijmen, 2009; Cai, 2010; Rijmen, 2010). The graph theory methodology may be useful for extending the MSEM to account for other types of data

and designs, such as those discussed in the next section.

6.2. Limitations

The MSEM presented, as well as the proposed estimation routine, are not without their limitations. First, although no distributional assumptions are made about any observed covariates in the model (i.e., $\mathbf{x}_{ij}$ and $\mathbf{x}_j$), all response variables ($\mathbf{y}_{ij}$ and $\mathbf{z}_j$) are assumed to be normally distributed conditional on the random effects, which are also assumed to be normally distributed. Extending the model to account for non-normal data is certainly possible, but adds to the computational complexity of estimating the model parameters. As discussed, the proposed estimation routine relies upon the conjugacy between the normally distributed random effects and conditionally normally distributed response variables to reformulate the likelihood function so that many of the integrals can be computed analytically. This reformulation is not possible for response variables that depend on normally distributed latent variables, but are not conditionally normal. Therefore, the vector of random effects that must be numerically integrated would need to be expanded to include those that are predictors of categorical and count variables. Depending on the specific model, the dimension of numerical integration required may be too high to practically estimate the model parameters using ML.

Second, this paper focused on the general two-level SEM. Although some three-level growth models are possible to fit within the framework, the model discussed here does not easily generalize to other types of three-level models. Yet, data collected on higher order structures are fairly common in psychological, educational, and health science research. The two-level SEM could be extended to allow for additional levels of nesting, as is allowed within the GLLAMM framework (Rabe-Hesketh et al., 2004) and Mplus (L. K. Muthén & Muthén, 2017). In such a circumstance the random effects that need to be numerically integrated may be nested and so the method of numerically integrating nested random effects discussed by Rabe-Hesketh, Skrondal, and Pickles (2005) could be used.

Finally, the approach as documented in Section 3 did not account for missing data. However, it is relatively straightforward to extend the likelihood calculation to account for models in which the data can be assumed to be missing at random. The necessary extension is sketched in the Appendix.

6.3. Future directions

There are a few programming extensions that could be implemented to speed up the estimation routine. For example, analytically computing the derivatives of the likelihood function with respect to the parameters can reduce the estimation time. Without providing analytic derivatives, the derivatives are either approximated using numerical methods or computed using automatic differentiation. Computing the derivatives analytically is typically faster than either of these two approaches. Another method for speeding up estimation includes the use of parallel processing. The likelihood function can easily be broken into independent components (e.g., the likelihood contribution from each cluster) that can be computed in parallel.

In the modeling context, the performance of these complex MSEMs are still relatively unknown. The estimation routine presented allows for a much broader range of potential models to be estimated than what was currently possible using ML, but the simulation study presented was relatively basic compared to the possible models that can be fit. The next task is better understanding the conditions for which the parameter estimates are "good" and, more importantly, when they are not. As demonstrated here and elsewhere, the performance of such estimates likely depend on many things including, but not limited to, the level-1 and level-2 sample sizes, the ICCs, the amount of measurement error, and the complexity of the model. The effects of each of these components can, and should, be assessed via more comprehensive simulation studies. That is, now that it is practical to obtain ML parameter estimates for such complex MSEMs, these estimates should fully be assessed and compared to other estimation methods (e.g., Bayesian estimation).

References

Asparouhov, T., & Muthén, B. (2019a). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142.

Asparouhov, T., & Muthén, B. (2019b). Latent variable interactions using maximum-likelihood and bayesian estimation for single- and two-level models. *Mplus Web Notes: No. 23*.

Bentler, P. M. (2004). *Eqs 6: Structural equations program manual*. Multivariate software.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley.

Cai, L. (2010). A two-tier full-information item factor analysis model with applications. *Psychometrika*, *75*(4), 581–612.

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., & Betancourt, M. (2015). The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164*.

Cronbach, L. J., et al. (1976). Research on classrooms and schools: Formulation of questions, design and analysis.

Cudeck, R. (2005). Fitting psychometric models with methods based on automatic differentiation. *Psychometrika*, *70*(4), 599–617.

Cudeck, R., Harring, J. R., & du Toit, S. H. (2009). Marginal maximum likelihood estimation of a latent variable model with interaction. *Journal of Educational and Behavioral Statistics*, *34*(1), 131–144.

du Toit, S. H., & Cudeck, R. (2009). Estimation of the nonlinear random coefficient model when some random effects are separable. *Psychometrika*, *74*(1), 65–82.

du Toit, S. H., & du Toit, M. (2008). Multilevel structural equation modeling. In *Handbook of multilevel analysis* (pp. 435–478). Springer.

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York: Springer. (ISBN 978-1-4614-6867-7) doi: 10.1007/978-1-4614-6868-4

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*(3), 423–436.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*(4), 455–467.

Griewank, A., & Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation* (Vol. 105). Siam.

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(4), 621–638.

Jöreskog, K. G., & Sörbom, D. (1996). *Lisrel 8: User's reference guide*. Scientific Software International.

Lee, S.-Y. (1990). Multilevel analysis of structural equation models. *Biometrika*, *77*(4), 763–772.

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*(1), 101–122.

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229.

Maas, C. J., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92.

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, *47*(2), 106–124.

McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika*, *58*(4), 575–585.

McDonald, R. P., & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, *42*(2), 215–232.

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods*, *10*(3), 259–284.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*(4), 557–585.

Muthén, B. O., & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In *Multilevel analysis of educational data* (pp. 87–99). Elsevier.

Muthén, L. K., & Muthén, B. (2017). Mplus version 8 [Computer software manual]. Los Angeles, CA: Muthén & Muthén.

Nash, J. C. (2014). On best practice optimization methods in R. *Journal of Statistical Software*, *60*(2), 1–14. Retrieved from `http://www.jstatsoft.org/v60/i02/`

Neale, M. C., Hunter, M. D., Pritikin, J. N., Zahery, M., Brick, T. R., Kirkpatrick, R. M., ... Boker, S. M. (2016). OpenMx 2.0: Extended structural equation and statistical modeling. *Psychometrika*, *81*(2), 535-549.

OECD. (2003). *Programme for International Student Assessment 2003.* Retrieved from `http://www.oecd.org/pisa/data/database-pisa2003.htm`

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.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*(2), 189–205.

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209-233.

R Core Team. (2017). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `https://www.R-project.org/`

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

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*(2), 167–190. doi: 10.1007/bf02295939

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2005). Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics*, *128*(2), 301–323.

Rijmen, F. (2009). *An efficient EM algorithm for multidimensional IRT models: Full*

information maximum likelihood estimation in limited time (Tech. Rep.). Princeton, NJ: ETS Research Report (RR0903).

Rijmen, F. (2010). Formal relations and an empirical comparison among the bi-factor, the testlet, and a second-order multidimensional irt model. *Journal of Educational Measurement*, *47*(3), 361–372.

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling and more. version 0.5–12 (BETA). *Journal of Statistical Software*, *48*(2), 1–36.

Schmidt, W. H. (1969). *Covariance structure analysis of the multivariate random effects model* (Unpublished doctoral dissertation). University of Chicago, Department of Education.

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*(1), 26–53.

Stan Development Team. (2016). *RStan: The R interface to Stan.* Retrieved from `http://mc-stan.org/` (R package version 2.14.1)

Stapleton, L. M., & Johnson, T. L. (2019). Models to examine the validity of cluster-level factor structure using individual-level data. *Advances in Methods and Practices in Psychological Science*, *2*(3), 312–329.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520.

StataCorp. (2005). *Stata statistical software: Release 15.* College Station, TX: StataCorp LLC.

Appendix

In this section, methods for adapting the estimation routine to allow for data missing at random are sketched. As in Section 2, suppose there are k level-2 variables $\mathbf{z}_j$ and p level-1 variables $\mathbf{y}_{ij}$. However, now consider that one or more elements within these vectors for a given j or ij may be missing. Suppose cluster j has k_j non-missing elements of $\mathbf{z}_j$ and individual i in cluster j has p_{ij} non-missing elements in $\mathbf{y}_{ij}$.

Define $\mathbf{K}_j$ $(k_j \times k)$ and $\mathbf{M}_{ij}$ $(p_{ij} \times p)$ to be zero-one matrices that select the non-missing elements of $\mathbf{z}_j$ and $\mathbf{y}_{ij}$, respectively. For example, suppose k is 4 and cluster j' is missing the

third element of $\mathbf{z}_{j'}$, so that

$$\mathbf{K}_{j'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{A1}$$

can be used to select the non-missing subset of $\mathbf{z}_{j'}$:

$$\mathbf{z}^*_{j'} = \mathbf{K}_{j'}\mathbf{z}_{j'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} z_{1j'} \\ z_{2j'} \\ - \\ z_{4j'} \end{pmatrix} = \begin{pmatrix} z_{1j'} \\ z_{2j'} \\ z_{4j'} \end{pmatrix}. \tag{A2}$$

The matrix $\mathbf{M}_{ij}$ performs the same role as $\mathbf{K}_j$, except it is used to select non-missing elements of $\mathbf{y}_{ij}$ rather than $\mathbf{z}_j$. Thus, $\mathbf{z}^*_j = \mathbf{K}_j\mathbf{z}_j$ will be used in place of $\mathbf{z}_j$ and $\mathbf{y}^*_j = \mathbf{M}_{ij}\mathbf{y}_{ij}$ will be used in place of $\mathbf{y}_{ij}$.

By premultiplying some of the other model matrices within the likelihood calculation by $\mathbf{K}_j$ or $\mathbf{M}_{ij}$, the estimation routine can be adapted to account for the missing elements within $\mathbf{z}_j$ and $\mathbf{y}_{ij}$. Specifically, for Equations 24-25, replace $\boldsymbol{\mu}_{\mathbf{z}_j}$ and $\boldsymbol{\mu}_{\mathbf{y}_{ij}}$ with $\mathbf{K}_j\boldsymbol{\mu}_{\mathbf{z}_j}$ and $\mathbf{M}_{ij}\boldsymbol{\mu}_{\mathbf{y}_{ij}}$, respectively. Within Equations 26–30 replace $\tilde{\mathbf{G}}$ and $\tilde{\mathbf{Q}}^*_{ij}$ with $\mathbf{K}_j\tilde{\mathbf{G}}$ and $\mathbf{M}_{ij}\tilde{\mathbf{Q}}^*_{ij}$, and replace $\boldsymbol{\Sigma}^*_W$ with $\boldsymbol{\Sigma}^*_{Wij} = \mathbf{M}_{ij}\boldsymbol{\Sigma}^*_W\mathbf{M}'_{ij}$. Lastly, replace $\mathbf{I}_{n_j} \otimes \boldsymbol{\Sigma}^*_W$ in Equation 26 with

$$\bigoplus_{i=1}^{n_j} \boldsymbol{\Sigma}^*_{Wij}, \tag{A3}$$

where $\oplus$ is the direct sum. Using these replacements, the simplified expressions for $|\boldsymbol{\Sigma}_{\mathbf{d}_j}|$ and $\boldsymbol{\epsilon}'_{\mathbf{d}_j}\boldsymbol{\Sigma}^{-1}_{\mathbf{d}_j}\boldsymbol{\epsilon}_{\mathbf{d}_j}$ in the new conditional log-likelihood

$$f(\mathbf{d}_j|\tilde{\boldsymbol{\beta}}_{Wj}) = (2\pi)^{-(\sum_i p_{ij}+k_j)/2}|\boldsymbol{\Sigma}_{\mathbf{d}_j}|^{-1/2}\exp\left\{ -\frac{1}{2}\boldsymbol{\epsilon}'_{\mathbf{d}_j}\boldsymbol{\Sigma}^{-1}_{\mathbf{d}_j}\boldsymbol{\epsilon}_{\mathbf{d}_j} \right\} \tag{A4}$$

are

$$|\boldsymbol{\Sigma}_{\mathbf{d}_j}| = \left\{ \prod_{i=1}^{n_j} |\boldsymbol{\Sigma}^*_{Wij}| \right\} |\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}||\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W} + \mathbf{A}_j||\boldsymbol{\Sigma}_{zz.y}|, \tag{A5}$$

and

$$\begin{aligned} \boldsymbol{\epsilon}'_{\mathbf{d}_j}\boldsymbol{\Sigma}^{-1}_{\mathbf{d}_j}\boldsymbol{\epsilon}_{\mathbf{d}_j} = &\sum_{i=1}^{n_j} \boldsymbol{\epsilon}'_{\mathbf{y}_{ij}}\boldsymbol{\Sigma}^{*-1}_{Wij}\boldsymbol{\epsilon}_{\mathbf{y}_{ij}} + \mathbf{p}'_j\mathbf{H}_j\mathbf{p}_j \\ &- 2\mathbf{p}'_j\mathbf{C}'_j\boldsymbol{\Sigma}_{\boldsymbol{\xi}\bullet\boldsymbol{\beta}_W}\tilde{\mathbf{G}}'\boldsymbol{\Sigma}^{-1}_{zz.y}\boldsymbol{\epsilon}_{\mathbf{z}_j} \\ &+ \boldsymbol{\epsilon}'_{\mathbf{z}_j}\boldsymbol{\Sigma}^{-1}_{zz.y}\boldsymbol{\epsilon}_{\mathbf{z}_j}. \end{aligned} \tag{A6}$$

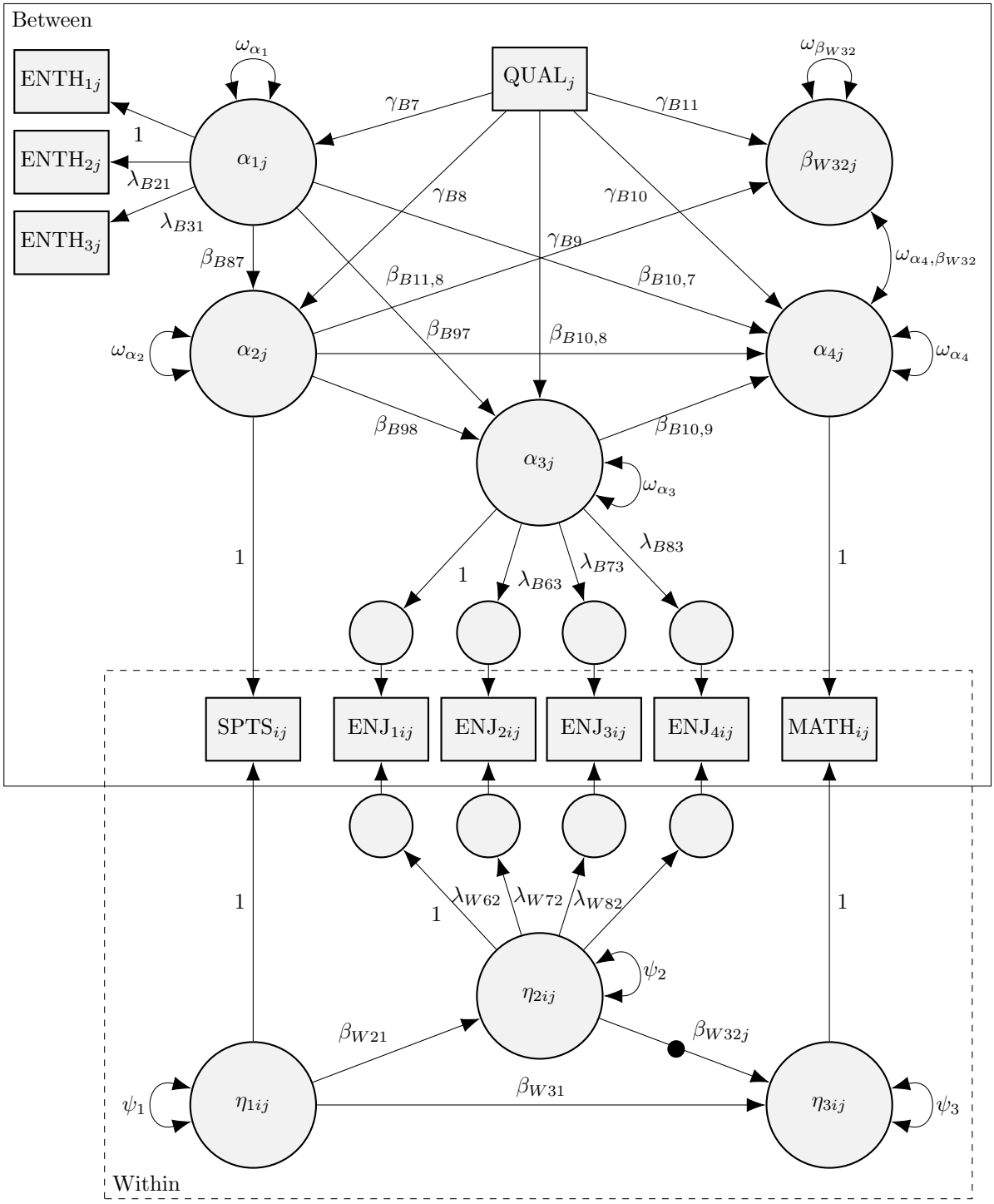

FIGURE 1.
A path diagram corresponding to the PISA model. To reduce clutter, item-specific level-1 and level-2 residual variances, as well as item intercepts and factor means, have been omitted.

		New Method				Mplus			
J	n_j	.05	.10	.20	.30	.05	.10	.20	.30
100	5	13	4	–	–	41	39	28	37
	10	2	–	–	–	10	9	5	9
	25	–	–	–	–	10	11	5	15
200	5	8	–	–	–	64	47	34	58
	10	–	–	–	–	14	14	4	9
	25	–	–	–	–	12	10	13	14
400	5	–	–	–	–	69	38	42	62
	10	–	–	–	–	11	10	3	6
	25	–	–	–	–	10	5	4	18

TABLE 1.

Percent of replications in the latent covariate simulation for which the model did not converge. Each column corresponds to a different ICC_x (.05, .10, .20, .30). A dash indicates that all reps in the corresponding condition properly converged.

J	n_j	ICC_x	β_{0W}	β_B	β_{1W}	ω_{β_W}
100	5	0.05	-0.02 (0.12)	0.34 (1.67)	0.32 (1.71)	-0.05 (0.10)
		0.10	-0.01 (0.08)	0.19 (0.59)	0.06 (0.51)	-0.02 (0.09)
		0.20	-0.01 (0.09)	0.05 (0.25)	0.00 (0.26)	-0.02 (0.09)
		0.30	-0.01 (0.08)	0.01 (0.13)	-0.01 (0.18)	-0.03 (0.10)
	10	0.05	0.00 (0.07)	0.24 (0.80)	0.04 (0.69)	-0.02 (0.06)
		0.10	0.00 (0.06)	0.03 (0.28)	-0.02 (0.27)	-0.02 (0.06)
		0.20	0.01 (0.07)	0.02 (0.16)	-0.03 (0.17)	-0.01 (0.05)
		0.30	-0.01 (0.08)	0.01 (0.13)	0.00 (0.12)	-0.02 (0.07)
	25	0.05	0.00 (0.06)	0.03 (0.37)	-0.09 (0.41)	0.00 (0.05)
		0.10	0.01 (0.06)	0.00 (0.20)	0.00 (0.23)	-0.02 (0.05)
		0.20	0.00 (0.06)	0.01 (0.13)	0.00 (0.13)	-0.01 (0.05)
		0.30	0.00 (0.06)	0.01 (0.10)	0.00 (0.11)	-0.01 (0.05)
200	5	0.05	0.00 (0.06)	0.25 (1.01)	0.10 (0.79)	-0.02 (0.06)
		0.10	0.00 (0.05)	0.05 (0.35)	0.00 (0.25)	-0.01 (0.06)
		0.20	0.00 (0.06)	0.00 (0.14)	-0.01 (0.16)	-0.01 (0.06)
		0.30	-0.01 (0.07)	0.01 (0.10)	-0.03 (0.13)	-0.01 (0.08)
	10	0.05	0.01 (0.05)	0.05 (0.43)	0.06 (0.36)	0.00 (0.04)
		0.10	0.00 (0.05)	0.05 (0.20)	0.05 (0.18)	-0.01 (0.04)
		0.20	0.00 (0.05)	0.01 (0.12)	0.00 (0.14)	0.00 (0.04)
		0.30	0.00 (0.05)	0.01 (0.08)	0.02 (0.10)	-0.01 (0.04)
	25	0.05	0.00 (0.04)	0.04 (0.22)	0.04 (0.23)	0.00 (0.04)
		0.10	0.00 (0.04)	0.00 (0.15)	-0.01 (0.16)	-0.01 (0.03)
		0.20	0.00 (0.04)	0.00 (0.09)	-0.02 (0.10)	0.00 (0.04)
		0.30	0.00 (0.04)	0.00 (0.06)	-0.01 (0.07)	0.00 (0.03)
400	5	0.05	0.00 (0.04)	0.14 (0.49)	0.04 (0.47)	-0.01 (0.04)
		0.10	0.00 (0.04)	0.00 (0.22)	0.03 (0.19)	0.00 (0.04)
		0.20	0.00 (0.04)	0.01 (0.12)	0.00 (0.12)	0.00 (0.04)
		0.30	0.00 (0.04)	0.00 (0.07)	-0.01 (0.09)	0.00 (0.05)
	10	0.05	0.00 (0.03)	0.02 (0.23)	-0.02 (0.24)	0.00 (0.03)
		0.10	0.00 (0.04)	0.02 (0.15)	0.02 (0.13)	0.00 (0.03)
		0.20	0.00 (0.03)	0.00 (0.08)	-0.01 (0.08)	0.00 (0.03)
		0.30	0.00 (0.04)	0.01 (0.07)	0.00 (0.07)	0.00 (0.04)
	25	0.05	0.00 (0.03)	0.01 (0.13)	-0.02 (0.17)	0.00 (0.03)
		0.10	0.00 (0.03)	-0.01 (0.09)	0.01 (0.11)	-0.01 (0.03)
		0.20	0.00 (0.03)	0.00 (0.06)	0.00 (0.07)	0.00 (0.03)
		0.30	0.00 (0.03)	0.00 (0.04)	-0.01 (0.06)	0.00 (0.03)

TABLE 2.
Bias (RMSE) using the new ML estimation routine. Population parameters are $\beta_{0W} = -0.50$, $\beta_B = 0.50$, $\beta_{1W} = 0.25$, and $\omega_{\beta_W} = 0.30$.

J	n_j	ICC_x	β_{0W}	β_B	β_{1W}	ω_{β_W}
100	5	0.05	-0.02 (0.10)	0.00 (1.36)	0.28 (1.20)	-0.04 (0.11)
		0.10	0.00 (0.09)	-0.07 (0.47)	0.17 (0.54)	-0.03 (0.10)
		0.20	0.01 (0.09)	-0.06 (0.23)	0.05 (0.28)	0.01 (0.14)
		0.30	-0.02 (0.07)	0.03 (0.14)	0.09 (0.18)	-0.04 (0.17)
	10	0.05	0.00 (0.08)	0.11 (1.00)	0.20 (0.96)	-0.02 (0.07)
		0.10	0.00 (0.06)	0.01 (0.26)	0.05 (0.32)	-0.02 (0.06)
		0.20	0.00 (0.07)	0.03 (0.16)	0.00 (0.19)	-0.01 (0.06)
		0.30	-0.02 (0.09)	0.02 (0.13)	0.03 (0.13)	-0.01 (0.07)
	25	0.05	-0.01 (0.06)	0.13 (0.52)	-0.05 (0.50)	0.00 (0.05)
		0.10	0.00 (0.06)	0.04 (0.23)	0.03 (0.25)	-0.01 (0.05)
		0.20	-0.01 (0.06)	0.03 (0.14)	0.02 (0.14)	0.00 (0.06)
		0.30	-0.01 (0.06)	0.03 (0.11)	0.02 (0.12)	-0.01 (0.05)
200	5	0.05	0.01 (0.06)	-0.05 (1.13)	0.40 (1.56)	-0.03 (0.08)
		0.10	0.01 (0.06)	-0.09 (0.27)	0.19 (0.32)	-0.04 (0.11)
		0.20	0.03 (0.07)	-0.18 (0.26)	0.04 (0.19)	0.03 (0.11)
		0.30	-0.01 (0.08)	0.01 (0.12)	0.07 (0.13)	-0.04 (0.09)
	10	0.05	0.01 (0.05)	-0.15 (0.30)	0.16 (0.42)	0.00 (0.04)
		0.10	0.00 (0.05)	0.07 (0.21)	0.15 (0.26)	-0.01 (0.05)
		0.20	-0.01 (0.05)	0.03 (0.12)	0.04 (0.15)	-0.01 (0.04)
		0.30	-0.01 (0.05)	0.03 (0.09)	0.03 (0.10)	0.00 (0.04)
	25	0.05	0.00 (0.04)	0.12 (0.29)	0.07 (0.27)	0.00 (0.04)
		0.10	-0.01 (0.04)	0.05 (0.18)	0.02 (0.19)	-0.01 (0.03)
		0.20	0.00 (0.04)	0.00 (0.12)	0.00 (0.11)	0.01 (0.04)
		0.30	-0.01 (0.04)	0.02 (0.07)	0.00 (0.07)	0.01 (0.04)
400	5	0.05	0.01 (0.04)	-0.30 (0.37)	0.12 (0.36)	-0.01 (0.04)
		0.10	0.01 (0.04)	-0.15 (0.25)	0.22 (0.28)	-0.06 (0.07)
		0.20	0.02 (0.04)	-0.14 (0.23)	0.08 (0.14)	0.03 (0.11)
		0.30	0.00 (0.04)	0.01 (0.06)	0.07 (0.09)	-0.02 (0.06)
	10	0.05	0.00 (0.03)	-0.10 (0.21)	0.08 (0.32)	0.00 (0.03)
		0.10	-0.01 (0.04)	0.06 (0.16)	0.13 (0.20)	0.00 (0.03)
		0.20	-0.01 (0.03)	0.01 (0.10)	0.03 (0.09)	-0.01 (0.03)
		0.30	-0.01 (0.04)	0.02 (0.07)	0.01 (0.07)	0.00 (0.04)
	25	0.05	-0.01 (0.03)	0.08 (0.18)	0.02 (0.21)	0.00 (0.03)
		0.10	-0.01 (0.03)	0.02 (0.11)	0.05 (0.13)	0.00 (0.03)
		0.20	-0.01 (0.03)	0.01 (0.09)	0.02 (0.08)	0.01 (0.04)
		0.30	-0.01 (0.03)	0.02 (0.05)	0.00 (0.06)	0.01 (0.03)

TABLE 3.

Bias (RMSE) using the ML estimation routine implemented within Mplus. Population parameters are $\beta_{0W} = -0.50$, $\beta_B = 0.50$, $\beta_{1W} = 0.25$, and $\omega_{\beta_W} = 0.30$.

Effect	Parameter	New (SE)	Mplus (SE)
SPTS → ENJ	β_{W21}	0.172 (0.008)	0.173 (0.008)
SPTS → MATH	β_{W31}	−0.116 (0.017)	−0.117 (0.017)
ENJ → MATH	β_{W32j}	–	–
ENJ_2 Loading	λ_{W62}	0.899 (0.010)	0.899 (0.010)
ENJ_3 Loading	λ_{W72}	1.152 (0.012)	1.153 (0.012)
ENJ_4 Loading	λ_{W82}	0.874 (0.011)	0.875 (0.011)
$\text{Var}(\text{ENJ}_1)$	θ_5	0.251 (0.005)	0.251 (0.005)
$\text{Var}(\text{ENJ}_2)$	θ_6	0.211 (0.004)	0.211 (0.004)
$\text{Var}(\text{ENJ}_3)$	θ_7	0.161 (0.004)	0.160 (0.004)
$\text{Var}(\text{ENJ}_4)$	θ_8	0.311 (0.005)	0.311 (0.005)
Var(SPTS)	ψ_1	0.908 (0.013)	0.908 (0.013)
Var(ENJ)	ψ_2	0.441 (0.010)	0.440 (0.010)
Var(MATH)	ψ_3	2.135 (0.032)	2.135 (0.032)
Log-likelihood		−69657.64	−69656.34
Parameters		48	48
Dim. of Integration		1	8
Estimation Time		6.76 minutes	5.43 hours

TABLE 4.

Parameter estimates (standard errors) for the within-school effects of the PISA example obtained using the newly proposed ML estimation routine and the ML estimation routine implemented within Mplus.

Effect	Parameter	New (SE)	Mplus (SE)
ENTH → SPTS	β_{B87}	0.143 (0.063)	0.148 (0.066)
ENTH → ENJ	β_{B97}	0.020 (0.029)	0.019 (0.029)
ENTH → MATH	$\beta_{B10,7}$	0.261 (0.116)	0.267 (0.117)
SPTS → ENJ	β_{B98}	0.152 (0.032)	0.157 (0.033)
SPTS → MATH	$\beta_{B10,8}$	−0.695 (0.146)	−0.739 (0.158)
SPTS → β_{W32j}	$\beta_{B11,8}$	0.169 (0.087)	0.167 (0.091)
ENJ → MATH	$\beta_{B10,9}$	2.422 (0.506)	2.517 (0.536)
QUAL → ENTH	γ_{B7}	0.114 (0.024)	0.144 (0.025)
QUAL → SPTS	γ_{B8}	−0.010 (0.022)	−0.011 (0.023)
QUAL → ENJ	γ_{B9}	−0.005 (0.010)	−0.005 (0.010)
QUAL → MATH	γ_{B10}	0.168 (0.040)	0.168 (0.041)
QUAL → β_{W32j}	γ_{B11}	0.035 (0.027)	0.035 (0.027)
ENTH_2 Loading	λ_{B21}	0.989 (0.101)	0.989 (0.101)
ENTH_3 Loading	λ_{B31}	0.629 (0.082)	0.629 (0.082)
ENJ_2 Loading	λ_{B63}	0.899 (0.010)	0.899 (0.010)
ENJ_3 Loading	λ_{B73}	1.152 (0.012)	1.153 (0.012)
ENJ_4 Loading	λ_{B83}	0.874 (0.011)	0.875 (0.011)
Var(ENTH_1)	ω_{ν_1}	0.129 (0.017)	0.128 (0.017)
Var(ENTH_2)	ω_{ν_2}	0.078 (0.016)	0.078 (0.016)
Var(ENTH_3)	ω_{ν_3}	0.223 (0.018)	0.223 (0.018)
Var(ENJ_2)	ω_{ν_6}	0.004 (0.001)	0.004 (0.001)
Var(ENJ_3)	ω_{ν_7}	0.003 (0.001)	0.003 (0.001)
Var(ENJ_4)	ω_{ν_8}	0.007 (0.002)	0.007 (0.002)
Var(ENTH)	ω_{α_1}	0.152 (0.023)	0.152 (0.023)
Var(SPTS)	ω_{α_2}	0.122 (0.012)	0.122 (0.013)
Var(ENJ)	ω_{α_3}	0.015 (0.003)	0.014 (0.003)
Var(MATH)	ω_{α_4}	0.366 (0.040)	0.362 (0.040)
Var(β_{W32j})	$\omega_{\beta_{W32}}$	0.063 (0.020)	0.062 (0.020)
Cov(MATH,β_{W32j})	$\omega_{\alpha_4,\beta_{W32}}$	−0.014 (0.019)	−0.014 (0.019)
Intercept(ENTH_1)	μ_{ν_1}	2.931 (0.028)	2.931 (0.028)
Intercept(ENTH_2)	μ_{ν_2}	3.093 (0.025)	3.092 (0.025)
Intercept(ENTH_3)	μ_{ν_3}	3.134 (0.028)	3.134 (0.028)
Intercept(SPTS)	ν_4	−0.088 (0.021)	−0.088 (0.021)
Intercept(ENJ_1)	ν_5	2.061 (0.011)	2.061 (0.011)
Intercept(ENJ_2)	ν_6	1.873 (0.011)	1.872 (0.010)
Intercept(ENJ_3)	ν_7	2.172 (0.013)	2.171 (0.012)
Intercept(ENJ_4)	ν_8	2.557 (0.011)	2.557 (0.011)
Intercept(MATH)	μ_{α_4}	4.916 (0.040)	4.913 (0.039)
Intercept(β_{W32j})	$\mu_{\beta_{W32}}$	0.665 (0.029)	0.666 (0.028)

TABLE 5.

Parameter estimates (standard errors) for the between-school effects of the PISA example obtained using the newly proposed ML estimation routine and the ML estimation routine implemented within Mplus.