STRUCTURAL EQUATION MODELING
2026, VOL. 33, NO. 3
ISSN: 1070-5511 | E-ISSN: 1532-8007

Estimating Multilevel Structural Equation Models with Random Slopes with Laplace and Variational Approximations

Steffen Nestler

University of Münster

ABSTRACT

Multilevel structural equation models (MSEMs) are an important statistical approach to analyze hierarchically nested data. While Bayesian methods are commonly used to estimate the MSEM parameters, maximum likelihood (ML) approaches are less often employed because of the computational challenges in the required numerical integration. Building on Rockwood's reformulation of the MSEM, we investigate two computationally efficient approximation methods to the likelihood function—the Laplace approximation (LA) and the extended variational approximation (EVA)—by comparing their performance against Gauss-Hermite (GH) quadrature and a Bayesian approach in two simulation studies. Results demonstrate that LA and EVA provide accurate parameter estimates with substantially shorter computation times than GH, especially for a more complex model. Furthermore, LA and EVA showed almost no bias, good convergence rates, and appropriate coverage, particularly with larger sample sizes. Altogether, these findings suggest that LA and EVA are promising approaches for ML estimation of MSEMs with random slopes.

KEYWORDS

Maximum likelihood estimation; multilevel model; multilevel structural equation model; structural equation model

Multilevel structural equation models (MSEMs) have become an important methodological tool in psychological research. They are increasingly used by applied researchers to analyze hierarchically structured data, where, for example, students are nested within schools or patients within clinics. Also, they are often employed to analyze (intensive) longitudinal data (Asparouhov, Hamaker, & Muthén, 2018; Bolger & Laurenceau, 2013; Nestler & Humberg, 2024). The popularity of the MSEM stems from their flexibility to model effects at both the individual and the group level, or the time-point and the person level, while appropriately accounting for dependencies in the data (e.g., Goldstein & McDonald, 1988; Liang & Bentler, 2004; Muthén, 1989).

To estimate the MSEM parameters, Bayesian approaches are typically employed, as they allow the modeling of complex models that include both random intercepts and random slopes. Maximum likelihood (ML) estimators represent an alternative, but so far they can only efficiently estimate MSEMs with random intercepts (Asparouhov & Muthén, 2019; Rosseel, 2021). Specifically, although ML algorithms have been developed for models with random slopes, these often rely on EM algorithms that become computationally very demanding or practically infeasible for complex models—particularly when these models contain many random slopes in addition to the random intercepts (e.g., Liang & Bentler, 2004).

However, in a recent article, Rockwood (2020) showed how MSEMs with random slopes can be reformulated such that ML estimation no longer requires numerical integration over *all* specified random effects (i.e., random intercepts and slopes), but only over the random slopes of the latent variables at level 1. To approximate the remaining integrals, Rockwood proposed to use Gauss-Hermite quadrature (GH). While this method is feasible for simple models that include only one or two random slopes, it can become inefficient for models with multiple random slopes, even in this reformulation, because the computational burden grows exponentially with the number of random effects to be integrated (Skrondal & Rabe-Hesketh, 2004).

In this article, we therefore investigate two alternative approaches for integral approximation: the Laplace approximation (LA, e.g., Pinheiro & Bates, 1995; Tuerlinckx et al., 2006) and the extended variational approximation (EVA, e.g., Korhonen et al., 2023; Ormerod & Wand, 2012). Both methods have never been examined in the case of the MSEM with random slopes before, to the best of our knowledge, although they promise better scalability even for models with multiple random slopes. In the following, we first introduce the MSEM as described in Rockwood (2020). Second, we show how the MSEM parameters can be estimated. We will show that the optimization of the (marginal) log-likelihood functions involves complex integral approximations, and third, we will describe GH and the two approximation methods that are the focus of our research. Fourth, we present the results from two simulation studies designed to investigate the performance of the two ML

approaches. Finally, we summarize the results of the simulation studies and we outline some questions for future methodological research.

Multilevel Structural Equation Model (MSEM)

Each MSEM consists of a measurement model, a level-1 structural model and a structural model on level 2 (see Asparouhov & Muthén, 2019; Goldstein & McDonald, 1988; Rockwood, 2020). First, the measurement model[1] is given by:

$$\mathbf{y}_{ij} = \boldsymbol{\nu}_j + \Lambda^B \boldsymbol{\alpha}_j + \Lambda^W \boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{yij} \tag{1}$$

where $\boldsymbol{\nu}_j$ is an intercept vector of dimension p and Λ^B is a factor loading matrix of dimension $p \times m_B$ for the m_B between-level latent factors in $\boldsymbol{\alpha}_j$. Λ^W is a factor loading matrix of dimension $p \times m_W$ for the m_W-dimensional within-level latent factors $\boldsymbol{\eta}_{ij}$, and $\boldsymbol{\varepsilon}_{yij}$ is a $p \times 1$ vector of residuals. The latter are assumed to follow a multivariate normal distribution with zero means and covariance matrix Θ.

Second, the level-1 structural model is defined by

$$\boldsymbol{\eta}_{ij} = \mathbf{B}_j^W \boldsymbol{\eta}_{ij} + \Gamma_j^W \mathbf{x}_{ij} + \boldsymbol{\zeta}_{ij} \tag{2}$$

Here, $\mathbf{B}_j^W$ is a $m_W \times m_W$ coefficient matrix capturing relationships among the within-level latent variables, and Γ_j^W is a $m_W \times q$ matrix of regression coefficient for the q observed within-level covariates $\mathbf{x}_{ij}$. Finally, $\boldsymbol{\zeta}_{ij}$ are level-1 disturbance terms that are assumed to follow a multivariate normal distribution with zero mean and covariance matrix Ψ.

When we rearrange Equation 2 for $\boldsymbol{\eta}_{ij}$ and substitute the result of the rearrangement into Equation 1, we obtain the *reduced form* for $\mathbf{y}_{ij}$:

$$\mathbf{y}_{ij} = \boldsymbol{\nu}_j + \Lambda^B \boldsymbol{\alpha}_j + \Lambda^W (\mathbf{I} - \mathbf{B}_j^W)^{-1} \Gamma_j^W \mathbf{x}_{ij} + \Lambda^W (\mathbf{I} - \mathbf{B}_j^W)^{-1} \boldsymbol{\zeta}_{ij} + \boldsymbol{\varepsilon}_{yij} \tag{3}$$

As indicated in Equation 3, the elements in $\boldsymbol{\nu}_j$, $\boldsymbol{\alpha}_j$, Γ_j^W, and $\mathbf{B}_j^W$ may differ between the level-2 units. When we assume that the number of these elements is r and collect the respective terms in the vector $\boldsymbol{\eta}_j$, we can define the structural model on level 2:

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

where $\boldsymbol{\mu}$, $\mathbf{B}^B$, Γ^B are $r \times 1$, $r \times r$, and $r \times s$ vectors and matrices, respectively, of fixed effect parameters and $\boldsymbol{\zeta}_j$ is a $r \times 1$ vector of level-2 disturbance terms. Again, the latter terms are assumed to be normally distributed with zero expectation and covariance matrix $\boldsymbol{\Omega}$.

To illustrate the MSEM, we use the latent covariate model that was used as the data-generating model in the simulation study presented in Rockwood (2020, see also Asparouhov & Muthén, 2019). The model assumes, see Figure 1 for a path diagram, that two level-1 variables were assessed and that the linear relationships between the latent intercepts of the two variables are of interest. In addition, it is assumed that the slope, that results when the latent level-1 part of x is regressed on the latent level-1 part of z, also differs between the level-2 units. In terms of the equations of the MSEM, the measurement model and the level-1 structural model are given by

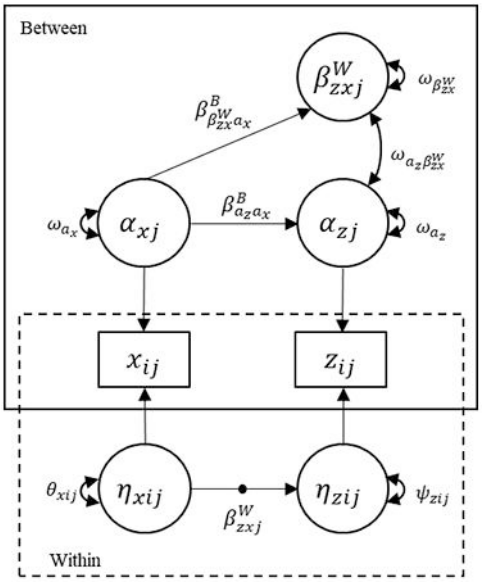

Figure 1. Path model for the illustrative MSEM example.

$$\begin{pmatrix} x_{ij} \\ z_{ij} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_{xj} \\ \alpha_{zj} \end{pmatrix} + \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{xij} \\ \eta_{zij} \end{pmatrix}$$
$$\begin{pmatrix} \eta_{xij} \\ \eta_{zij} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ \beta^W_{zxj} & 0 \end{pmatrix} \begin{pmatrix} \eta_{xij} \\ \eta_{zij} \end{pmatrix} + \begin{pmatrix} \zeta_{\eta_{xij}} \\ \zeta_{\eta_{zij}} \end{pmatrix} \tag{5}$$

Since three coefficients are allowed to differ between the level-2 units, the level-2 structural model is:

$$\begin{pmatrix} \alpha_{xj} \\ \alpha_{zj} \\ \beta^W_{zxj} \end{pmatrix} = \begin{pmatrix} \mu_{\alpha_x} \\ \mu_{\alpha_z} \\ \mu_{\beta^W_{zx}} \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ \beta^B_{\alpha_z\alpha_x} & 0 & 0 \\ \beta^B_{\beta^W_{zx}\alpha_x} & 0 & 0 \end{pmatrix} \begin{pmatrix} \alpha_{xj} \\ \alpha_{zj} \\ \beta^W_{zxj} \end{pmatrix} + \begin{pmatrix} \zeta_{\alpha_{xj}} \\ \zeta_{\alpha_{zj}} \\ \zeta_{\beta^W_{zxj}} \end{pmatrix} \tag{6}$$

Finally, the covariance matrices on level 1 are

$$\Theta = \begin{pmatrix} \theta_{xij} & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \Psi = \begin{pmatrix} 0 & 0 \\ 0 & \psi_{\eta_{zij}} \end{pmatrix} \tag{7}$$

and the covariance matrix on level 2 is

$$\Omega = \begin{pmatrix} \omega_{\alpha_x} & 0 & 0 \\ 0 & \omega_{\alpha_z} & \omega_{\alpha_z, \beta^W_{zx}} \\ 0 & \omega_{\beta^W_{zx}, \alpha_z} & \omega_{\beta^W_{zx}} \end{pmatrix}. \tag{8}$$

ML Estimation of the MSEM Parameters

To estimate the MSEM parameters with ML, we assume—in addition to the distributional assumptions already mentioned—independence among $\boldsymbol{\alpha}_j$, $\boldsymbol{\eta}_{ij}$, and $\boldsymbol{\varepsilon}_{yij}$ (i.e., all cross-covariances equal zero, see Asparouhov & Muthén, 2019;

[1] In Rockwood (2020) the MSEM also included a SEM for level-2 variables. We do not consider this part in our descriptions to simplify the exposition.

Goldstein & McDonald, 1988; Liang & Bentler, 2004). Additionally, the residual terms on the two levels are also assumed to be uncorrelated (i.e., $\mathrm{Cov}(\boldsymbol{\varepsilon}_{yij}, \boldsymbol{\zeta}_{ij}) = \mathbf{0}$).

Let ϑ contain the parameters of the model that we want to estimate, the likelihood function given the data $\boldsymbol{y}$ then is:

$$L(\vartheta|\boldsymbol{y}) = \prod_{j=1}^{J} L_j(\vartheta|\boldsymbol{y}_j) = \prod_{j=1}^{J} \left\{ \int_{\boldsymbol{\eta}_j} \left[\prod_{i=1}^{n_j} f(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_j, \vartheta) \right] f(\boldsymbol{\eta}_j|\vartheta) d\boldsymbol{\eta}_j \right\} \tag{9}$$

where the term within the larger brackets is the likelihood contribution of a level-2 unit j. Furthermore, the log-likelihood function is:

$$ll(\vartheta|\boldsymbol{y}) = \sum_{j=1}^{J} \log \left\{ \int_{\boldsymbol{\eta}_j} \left[\prod_{i=1}^{n_j} f(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_j, \vartheta) \right] f(\boldsymbol{\eta}_j|\vartheta) d\boldsymbol{\eta}_j \right\}. \tag{10}$$

From the model definition and the assumptions it follows that $f(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_j, \vartheta)$ is the density of a multivariate normal distribution whose expectation and covariance matrix is:

$$\begin{aligned} \boldsymbol{\mu}_{y_{ij}} &= \boldsymbol{\nu}_j + \Lambda^B \boldsymbol{\alpha}_j + \Lambda^W (\boldsymbol{I} - \boldsymbol{B}_j^W)^{-1} \Gamma_j^W \boldsymbol{x}_{ij} \\ \Sigma_{y_{ij}} &= \Lambda^W (\boldsymbol{I} - \boldsymbol{B}_j^W)^{-1} \Psi (\boldsymbol{I} - \boldsymbol{B}_j^W)^{-1,T} \Lambda^{W,T} + \Theta \end{aligned} \tag{11}$$

Furthermore, $\boldsymbol{\eta}_j$ is also multivariate normally distributed with:

$$\begin{aligned} \boldsymbol{\mu}_{\eta_j} &= (\boldsymbol{I} - \boldsymbol{B}^B)^{-1} (\boldsymbol{\mu} + \Gamma^B \boldsymbol{x}_j) \\ \Sigma_{\eta_j} &= (\boldsymbol{I} - \boldsymbol{B}^B)^{-1} \Omega (\boldsymbol{I} - \boldsymbol{B}^B)^{-1,T} \end{aligned} \tag{12}$$

Maximizing Equation 9 or Equation 10 is computationally demanding because one has to approximate the integral for each level-2 unit j. To reduce this computational burden, Rockwood (2020) rewrites the model by partitioning the vector $\boldsymbol{\eta}_j$ into random effects that enter the model in a linear and effects that enter them in a nonlinear way (for similar approaches see Du Toit & Du Toit, 2008; Nestler, 2020). The latter random effects are those that are multiplied with the level-1 latent variables (i.e., $\boldsymbol{\zeta}_{ij}$ in Equation 3) and thus appear in the mean as well as the covariance structure of the level-1 model. By contrast, linear are those random effects that are not multiplied with the level-1 latent variables and are only part of the mean structure.

Let $\boldsymbol{\xi}_j$ contain the linear effects of a specified MSEM, that is, the elements of $\boldsymbol{\nu}_j$, $\boldsymbol{\alpha}_j$, and Γ_j^W that differ between the level-2 units. The nonlinear effects are contained in $\boldsymbol{\beta}_j^W$ and they refer to elements in $\boldsymbol{B}_j^W$ that differ between the level-2 units. Since we assumed that the whole vector $\boldsymbol{\eta}_j$ is normally distributed, the vector of expectations and the covariance matrix can also be partitioned in terms of the linear and nonlinear effects:

$$\boldsymbol{\eta}_j = \begin{pmatrix} \boldsymbol{\xi}_j \\ \boldsymbol{\beta}_{Wj} \end{pmatrix} \sim \mathcal{N} \left[\begin{pmatrix} \mu_\xi \\ \mu_{\boldsymbol{\beta}_W} \end{pmatrix}, \begin{pmatrix} \Sigma_\xi & \Sigma_{\xi,\boldsymbol{\beta}_W} \\ \Sigma_{\xi,\boldsymbol{\beta}_W} & \Sigma_{\boldsymbol{\beta}_W} \end{pmatrix} \right] \tag{13}$$

Furthermore, the conditional distribution of $\boldsymbol{\xi}$ given $\boldsymbol{\beta}_j^W = \boldsymbol{b}$ is then also normal (Wasserman, 2004) with expectations and covariance matrix:

$$\begin{aligned} \boldsymbol{\mu}_{\xi|\boldsymbol{\beta}^W} &= \mu_\xi + \Sigma_{\xi,\boldsymbol{\beta}^W} \Sigma_{\boldsymbol{\beta}^W}^{-1} (\boldsymbol{b} - \mu_{\boldsymbol{\beta}^W}) \\ \Sigma_{\xi|\boldsymbol{\beta}^W} &= \Sigma_\xi - \Sigma_{\xi,\boldsymbol{\beta}^W} \Sigma_{\boldsymbol{\beta}^W}^{-1} \Sigma_{\xi,\boldsymbol{\beta}^W}^T \end{aligned} \tag{14}$$

The log-likelihood contribution of a level-2 unit j for the whole data vector of this level-2 unit $\boldsymbol{y}_j$ can then be reformulated by integrating the linear effects out. Specifically, we can rewrite the likelihood contribution of a level-2 unit j as

$$\begin{aligned} L_j(\vartheta|\boldsymbol{y}_j) &= \int_{\boldsymbol{\eta}_j} \left[\prod_{i=1}^{n_j} f(\boldsymbol{y}_{ij}|\boldsymbol{\eta}_j, \vartheta) \right] f(\boldsymbol{\eta}_j|\vartheta) d\boldsymbol{\eta}_j \\ &= \int_{\boldsymbol{\beta}_j^W} \left\{ \int_{\xi_j} \left[\prod_{i=1}^{n_j} f(\boldsymbol{y}_{ij}|\boldsymbol{\beta}_j^W, \xi_j, \vartheta) \right] f(\boldsymbol{\beta}_j^W, \xi_j|\vartheta) d\xi_j \right\} d\boldsymbol{\beta}_j^W \\ &= \int_{\boldsymbol{\beta}_j^W} \left\{ \int_{\xi_j} \left[\prod_{i=1}^{n_j} f(\boldsymbol{y}_{ij}|\boldsymbol{\beta}_j^W, \xi_j, \vartheta) \right] f(\xi_j|\boldsymbol{\beta}_j^W, \vartheta) d\xi_j \right\} f(\boldsymbol{\beta}_j^W|\vartheta) d\boldsymbol{\beta}_j^W \\ &= \int_{\boldsymbol{\beta}_j^W} \left\{ \int_{\xi_j} f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \xi_j, \vartheta) f(\xi_j|\boldsymbol{\beta}_j^W, \vartheta) d\xi_j \right\} f(\boldsymbol{\beta}_{Wj}|\vartheta) d\boldsymbol{\beta}_j^W \\ &= \int_{\boldsymbol{\beta}_j^W} f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta) f(\boldsymbol{\beta}_j^W|\vartheta) d\boldsymbol{\beta}_j^W \end{aligned} \tag{15}$$

where the product of the densities $f(\boldsymbol{y}_{ij}|\boldsymbol{\beta}_j^W, \boldsymbol{\xi}_j, \vartheta)$ can be replaced by the joint density of the whole data vector $f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta)$ in the third step, because the former are independent given the random effects. Furthermore, in the last step $\boldsymbol{\xi}_j$ can be integrated out, because they enter the means of $\boldsymbol{y}_j$ only. $f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta)$ is the density of a normal distribution whose expectation and covariance matrix contain the elements of $\boldsymbol{\mu}_{\xi|\boldsymbol{\beta}^W}$ and $\Sigma_{\xi|\boldsymbol{\beta}^W}$. Specifically, let $\boldsymbol{\nu}_j^*$, $\boldsymbol{\alpha}_j^*$, and Γ_j^{W*} denote the original vectors with their values replaced by the respective elements in $\boldsymbol{\mu}_{\xi|\boldsymbol{\beta}^W}$. A single entry in $\mu_{\boldsymbol{y}_j}$ is then

$$\boldsymbol{\mu}_{\boldsymbol{y}_{ij}} = \boldsymbol{\nu}_j^* + \Lambda^B \boldsymbol{\alpha}_j^* + \Lambda^W (\boldsymbol{I} - \boldsymbol{B}_j^W)^{-1} \Gamma_j^{*W} \boldsymbol{x}_{ij} \tag{16}$$

and the covariance matrix $\Sigma_{\boldsymbol{y}_j}$ is

$$\Sigma_{\boldsymbol{y}_j} = \boldsymbol{Q}_j^* \Sigma_{\xi|\boldsymbol{\beta}^W} \boldsymbol{Q}_j^{*T} + \boldsymbol{I}_{n_j} \otimes \Sigma_{\boldsymbol{y}_{ij}} \tag{17}$$

where $\boldsymbol{Q}_j^*$ is a design matrix that contains the elements of $\boldsymbol{\nu}_j$, $\boldsymbol{\alpha}_j$, and Γ_j^W that differ between the level-2 units for a specified model with the values replaced by their conditional expectations. Finally, $\boldsymbol{y}_j$ is the whole data vector of level-2 unit j; its distribution is thus of much higher dimension than the distribution of a single vector $\boldsymbol{y}_{ij}$. Nevertheless, integrating the linear effects out effectively reduces the dimension of the integration; in fact, the integration now only depends on those weights or slopes of the latent variables that differ between the level-2 units (i.e., the elements in $\boldsymbol{B}_j^W$). It is this dimensionality reduction that substantially improves computational efficiency. For instance, Rockwood (2020) found that the EM algorithm implemented in Mplus, in which the integral is integrated over all random effects (see Liang & Bentler, 2004), required approximately six hours to estimate the parameters of the illustrative MSEM (i.e., the latent covariate model with seven random effects), whereas the method integrating out linear effects converged in only seven minutes.

Approximating the Integral

The foregoing expositions showed that the marginal log-likelihood function for parameter estimation is:

$$ll(\vartheta|\boldsymbol{y}) = \sum_{j=1}^{J} \log \left\{ \int_{\boldsymbol{\beta}_j^W} f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta) f(\boldsymbol{\beta}_j^W|\vartheta) d\boldsymbol{\beta}_j^W \right\} \tag{18}$$

To maximize this function, the integral has to be approximated. Different approaches can be used for this. Here, we focus on GH-quadrature, the LA, and EVA.

GH Quadrature

Rockwood (2020) suggested to use (adaptive) GH-quadrature to approximate the integrals in Equation 18. The basic idea behind GH quadrature (see Pinheiro & Bates, 1995; Rabe-Hesketh, Skrondal, & Pickles, 2002; Tuerlinckx et al., 2006) is to generate P vectors of GH nodes $\boldsymbol{t}_p$ and weights $\boldsymbol{w}_p$ that have the same dimensionality as $\boldsymbol{\beta}_j^W$ (say m). For each of the node vectors, one then computes $f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W = \boldsymbol{t}_p, \vartheta)$ and the integral of a *single* level-2 unit j is then approximated with a weighted sum

$$L_j^{GH}(\vartheta) = \sum_{p_1=1}^{P} \cdots \sum_{p_m=1}^{P} f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W = \boldsymbol{t}_p, \vartheta)\boldsymbol{w}_{p_1} \cdots \boldsymbol{w}_{p_m}. \tag{19}$$

Depending on the number of nodes, the approximation of the integral using GH quadrature can be very accurate. However, this is also the disadvantage of this method, as the number of nodes P increases exponentially with the dimensionality of $\boldsymbol{\beta}_j^W$. For example, when $P = 10$ and when there are two random effects, $f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W = \boldsymbol{t}_p, \vartheta)$ needs to be computed for each of the $10^2 = 100$ vectors *and* each level-2 unit j. Thus, when the dimensionality of $\boldsymbol{\beta}_j^W$ is large, the computational burden associated with approximating the integral is too large to be feasible, at least when one is interested in an accurate integral approximation.

Laplace Approximation

The idea of the Laplace approximation is to replace the function under the integral with a function for which a closed-form expression is available (Tuerlinckx et al., 2006). To derive this function, one assumes that the modes of $\boldsymbol{\beta}_j^W$ of the logarithm of the joint distribution are available for each level-2 unit j:

$$f(\boldsymbol{\beta}_j^W) = \log\left[f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta)\right] + \log\left[f(\boldsymbol{\beta}_j^W|\vartheta)\right] \tag{20}$$

Using a second-order Taylor-approximation at $\hat{\boldsymbol{\beta}}_j^W$, the integrand $f(\boldsymbol{\beta}_j^W)$ can be expressed as

$$f(\boldsymbol{\beta}_j^W) \approx f(\hat{\boldsymbol{\beta}}_j^W) - \frac{1}{2}(\boldsymbol{\beta}_j^W - \hat{\boldsymbol{\beta}}_j^W)'\boldsymbol{H}_{\hat{\beta}^W}(\boldsymbol{\beta}_j^W - \hat{\boldsymbol{\beta}}_j^W) \tag{21}$$

where $\boldsymbol{H}_{\hat{\beta}^W}$ is the covariance matrix of the estimated modes. The log-likelihood function can then be approximated with

$$ll_j^{LA}(\vartheta) = \frac{q}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{H}_{\hat{\beta}^W}| + \log\left[f(\boldsymbol{y}_j|\hat{\boldsymbol{\beta}}_j^W, \vartheta)f(\hat{\boldsymbol{\beta}}_j^W|\vartheta)\right], \tag{22}$$

because the second term in Equation 21 is the kernel of a multivariate normal distribution that integrates to one.

The derivation of the Laplace approximation presumed that the modes and their covariance matrix are known for each level-2 unit. However, since the modes are unknown, the Laplace approximation can be implemented in at least two ways: The modes are (a) estimated given the current parameter estimates and the data, and then (b), the model parameters are estimated by maximizing Equation 22 given the current modes and the data. This *two-step* procedure is repeated until the algorithm has converged. It can be shown that this implementation of the Laplace approximation is identical to an adaptive GH approach with one node vector (see Pinheiro & Bates, 1995). In the second type of implementation, the modes and the parameters are estimated at the same time (i.e., by maximizing Equation 22). This approach has the benefit that it takes the dependency between the modes and the parameters into account, and in our experience, this implementation is more stable and yields more accurate results than the first implementation. Importantly, the optimization of Equation 22 should be faster than optimizing Equation 18, because the function does not contain any integrals and because analytical expressions for the gradient of the last term in the approximation can be derived.

Variational Approximation

The third approach considered here consists of optimizing a *lower* bound of the density in Equation 18 that is computationally simpler. The lower bound can be obtained by assuming that a 'variational' density $q(\boldsymbol{\beta}_j^W|\boldsymbol{a}_j, \boldsymbol{A}_j)$ is available for the random effects $\boldsymbol{\beta}_j^W$. Let $\boldsymbol{a}_j$ be the expectation of this density and $\boldsymbol{A}_j$ the covariance matrix, the lower bound is then (Korhonen et al., 2023; Ormerod & Wand, 2012):

$$\begin{aligned} ll_j(\vartheta|\boldsymbol{y}_j) &\geq \int_{\boldsymbol{\beta}_j^W} \log\left(\frac{f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta)f(\boldsymbol{\beta}_j^W|\vartheta)}{q(\boldsymbol{\beta}_j^W|\boldsymbol{a}_j, \boldsymbol{A}_j)}\right) q(\boldsymbol{\beta}_j^W|\boldsymbol{a}_j, \boldsymbol{A}_j)d\boldsymbol{\beta}_j^W \\ &= \mathbb{E}_q\left[\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta\right] + \mathbb{E}_q\left[\log(f(\boldsymbol{\beta}_j^W|\vartheta)\right] + \mathbb{H}_q\left[q(\boldsymbol{\beta}_j^W|\boldsymbol{a}_j, \boldsymbol{A}_j)\right] \end{aligned} \tag{23}$$

where we use $\mathbb{E}_q$ to highlight that the expected value refers to the variational density and $\mathbb{H}_q[q(\boldsymbol{\beta}_j^W|\boldsymbol{a}_j, \boldsymbol{A}_j)]$ is the entropy of this density.

A standard assumption in the literature is that the variational density is a multivariate normal distribution (Ormerod & Wand, 2010, 2012). The lower bound then is

$$\begin{aligned} ll_j(\vartheta|\boldsymbol{y}_j) \geq\ & \mathbb{E}_q\left[\log(f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta))\right] \\ & + \frac{1}{2}\left[\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}^W}^{-1}\boldsymbol{A}_j| - \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}^W}^{-1}(\boldsymbol{A}_j + (\boldsymbol{a}_j - \boldsymbol{\mu}_{\boldsymbol{\beta}^W})(\boldsymbol{a}_j - \boldsymbol{\mu}_{\boldsymbol{\beta}^W})^T) + \frac{m}{2}\right] \end{aligned} \tag{24}$$

where tr is the trace of a matrix. In the last step, one needs to find the expectation of the data log-likelihood given the variational density q. There are a few models for which analytical solutions exist for this expectation, but the MSEM does not belong to these functions. Therefore, we follow the approach outlined in Korhonen et al. (2023) and approximate $\log(f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta))$ using a second-order Taylor approximation around $\boldsymbol{a}_j$. Inserting this approximation into Equation 24 results in

$$\begin{aligned} ll_j(\vartheta|\boldsymbol{y}_j) \geq\ & \log\left[f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W = \boldsymbol{a}_j, \vartheta)\right] + \frac{1}{2}\mathrm{tr}\left[\boldsymbol{H}(\boldsymbol{a}_j, \boldsymbol{\beta}_j^W)\boldsymbol{A}_j\right] \\ & + \frac{1}{2}\left[\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}^W}^{-1}\boldsymbol{A}_j| - \mathrm{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}^W}^{-1}(\boldsymbol{A}_j + (\boldsymbol{a}_j - \boldsymbol{\mu}_{\boldsymbol{\beta}^W})(\boldsymbol{a}_j - \boldsymbol{\mu}_{\boldsymbol{\beta}^W})^T) + \frac{m}{2}\right] \end{aligned} \tag{25}$$

where $\boldsymbol{H}(\boldsymbol{a}_j, \boldsymbol{\beta}_j^W)$ is the second derivative of $\log\left[f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W, \vartheta)\right]$ after $\boldsymbol{\beta}_j^W$ with $\boldsymbol{a}_j$ inserted.

The lower bound in Equation 25 could already be used for parameter estimation. However, a drawback of using this function in optimization is that not only the model parameters contained in ϑ but also the parameters of the J variational densities have to be estimated. Thus, when the number of elements in $\boldsymbol{\beta}_j^W$ is m, $m \cdot J$ variational expectations and $0.5 \cdot m \cdot (m-1) \cdot J$ variational covariance parameters have to be additionally estimated. To simplify the optimization problem further, we use that each $\boldsymbol{A}_j$ has a closed form optimum. This optimum is

$$\boldsymbol{A}_j^* = \left[\Sigma_{\boldsymbol{\beta}^W}^{-1} - \boldsymbol{H}(\boldsymbol{a}_j, \boldsymbol{\beta}_j^W)\right]^{-1} \tag{26}$$

Substituting this expression into Equation 25 results in

$$ll_j^{EVA}(\vartheta) = \log\left(f(\boldsymbol{y}_j|\boldsymbol{\beta}_j^W = \boldsymbol{a}_j, \vartheta)\right) - \frac{1}{2}\log|\boldsymbol{A}_j^*| - \frac{1}{2}\mathrm{tr}\left[\Sigma_{\boldsymbol{\beta}^W}^{-1}(\boldsymbol{a}_j - \boldsymbol{\mu}_{\boldsymbol{\beta}^W})(\boldsymbol{a}_j - \boldsymbol{\mu}_{\boldsymbol{\beta}^W})^T\right] \tag{27}$$

and that is the approximation that we used in the simulations.

Implementation in Software

We implemented all three approximations to the likelihood and log-likelihood function, respectively, within `R` (R Core Team, 2025) and C++ via the `TMB` package (Kristensen et al., 2016). The `TMB` package computes the gradient of the to-be-optimized function via automatic differentiation. The negative of the log-likelihood function was then minimized within `R` using the `nlminb` function. Note that we computed the value of the conditional distribution of the data given the random effects using the simplifications derived in Du Toit & Du Toit (2008) and suggested in Rockwood (2020, see also Rosseel, 2021). Furthermore, we do not directly optimize the elements of $\Sigma_{\boldsymbol{\beta}^W}$ and Θ, but rather the elements of their Cholesky factors, because this secures that the variance parameters remain positive. The standard errors of the elements in $\Sigma_{\boldsymbol{\beta}^W}$ and Θ are then obtained using the multivariate delta theorem (Wasserman, 2004). The standard errors of all other model parameters were computed at the final parameter based on the observed information matrix. This matrix is the inverse of the negative Hessian matrix that we also obtained using automatic differentiation. For EVA, we computed the Hessian for the vector of to-be estimated model parameters and the vector of the variational mean parameters. The standard errors were then obtained by considering the respective *subblock* in the observed information matrix (see Korhonen et al., 2023).

Study 1

We conducted two simulation studies to examine the performance of the three approximation methods. In the first simulation, the population model was the latent covariate model that we employed to illustrate the MSEM and that was used in Rockwood (2020) to examine the performance of the GH quadrature approach. Similar to Rockwood (2020), we varied the number of level-2 and level-1 units and the ICC of the predictor. Furthermore, since the parameters of MSEMs are often estimated with a Bayesian approach implemented in Mplus, we also compared the three ML approaches to a Bayesian estimator in both simulation studies.

Method

Data-Generating Model and Performance Measures

The data-generating model was a latent covariate model with two variables and a random slope, see Figure 1 for a path diagram. We used the same population parameter values as those used in the simulation study of Rockwood (2020). That is, the means and the matrix of path coefficients of the level-2 random effects $\boldsymbol{\eta}_j$ were set to:

$$\boldsymbol{\mu}_{\boldsymbol{\eta}_j} = \begin{pmatrix} 0 \\ 0 \\ -0.50 \end{pmatrix} \text{ and } \boldsymbol{B}^B = \begin{pmatrix} 0 & 0 & 0 \\ 0.50 & 0 & 0 \\ 0.25 & 0 & 0 \end{pmatrix}, \tag{28}$$

and the covariance matrices were set to

$$\Omega = \begin{pmatrix} \omega_x & 0 & 0 \\ 0 & 0.20 & 0.05 \\ 0 & 0.05 & 0.30 \end{pmatrix}, \Theta = \begin{pmatrix} 1-\omega_x & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \Psi = \begin{pmatrix} 0 & 0 \\ 0 & 0.80 \end{pmatrix}. \tag{29}$$

We varied the ICC of the predictor variable by setting ω_x to .05, .10, or .20. Finally, the number of level-2 units was either 100 or 300, and the number of level-1 units was either 10 or 25.

Data Generation and Analysis

We used `R` to generate 500 datasets in each of the $3 \times 2 \times 2 = 12$ simulation conditions. For each replication, we fitted the data using the three ML estimation approaches and the Bayesian approach. Since the model contains one random slope, we followed Rockwood (2020) and used 30 quadrature points in the GH approximation. For all three ML approaches we obtained starting values by fitting the random intercept version of the model with `lavaan` (i.e., without the random slope) because we believed that starting the optimization with suitable initial values corresponds to how one would proceed in standard (commercial) software. Bayesian estimates were obtained with Mplus (Muthén & Muthén 1998-2017) using the default noninformative prior specifications when estimating the model in each replication. Specifically, flat normal distributions were used as priors for the factor loadings, the latent mean parameters, and the path coefficients, and inverse Gamma or inverse Wishart distributions were specified for the variance and covariance parameters. We generated two chains of 15,000 samples

from the posterior distributions. The first 7,500 samples were discarded as burn-in for parameter estimation. Five prior practice runs in the condition with $J = 100$, $n_j = 10$, and ICC = .05 showed that this number of iterations and burn-in period resulted in an acceptable effective sample size for all parameters (the average effective sample size was above 400 for each estimate). For each replication, we obtained the posterior median of each marginal posterior distribution as an estimate of the specific parameter (this is the default output in Mplus).

To measure performance, we counted the number of replications in which the three ML algorithms converged. We also computed the median parameter estimate, the median absolute deviation as a robust measure of the standard deviation, and the coverage rate (CR). The CR was computed by calculating the 95% confidence interval (credibility interval) with the standard error (posterior standard deviation) of an estimate in each replication. When the actual parameter value was included in the interval, coverage was coded with 1 and it was coded with 0 when it was not included in the interval. Finally, we also report the mean and the standard deviation of the computation times for the three ML methods and the Bayesian approach.

Results and Discussion

We first analyzed the convergence rates of the different estimation approaches. For the three ML approaches, a replication was counted as not converged when the optimization algorithm reached the maximum number of allowed iterations and/or when undefined standard errors occurred in the final solution. For Bayes, a replication was counted as not converged, when the MCMC chains did not reach Mplus' default PSR stopping value. Overall, the number of nonconverged replications was very small: While all replications converged for Bayes, about 20 replications did not converge for the three ML approaches, and almost all of these occurred in the condition with $J = 100$, $n_j = 10$, and ICC = 0.05. Table 1 also displays the mean and standard deviation of the computation times. All approaches became slower the larger the number of level-2 units. For the three ML approaches, LA and EVA had shorter computation times than GH, and this difference was more pronounced the larger the number of participants. Finally, we note that the longer average computation times for Bayes resulted from our specification that 15,000 draws should be generated from the posterior distributions.

Table 2 presents the results of the simulation study for the two between-level effects (i.e., $\beta^B_{\alpha_z\alpha_x}$ and $\beta^B_{\beta^W_{zx}\alpha_x}$) and the variance of the random slope (i.e., $\omega_{\beta^W_{zx}}$). The results for the other parameters were very similar to the results of the three selected parameters. We provide the results for the other parameters in the OSF project accompanying this article (see https://osf.io/ncsz8). The OSF project also contains the results for the means and standard deviations of the parameter estimates, whose pattern was very similar to the pattern of the medians and the MADs. As can be seen in Table 2, there were almost no differences between the four approaches with regard to the performance measures. All four approaches yielded acceptable median estimates and CRs in all simulation conditions except when the ICC was small (although even in this case, the bias was rather small). However, the larger the number of level-2 and level-1 units, the better the estimates were even in these conditions. The MAD results showed that, independent of the estimation approach, the estimates were more variable when the number of level-2 and level-1 units was small, and the ICC was also small. However, with a larger number of level-2 units, all three approaches became more efficient.

Study 2

The results of study 1 provided encouraging first evidence that the LA and the EVA approach may provide reasonable estimates of the MSEM parameters. However, the latent covariate model is a rather simple model as it only contains one random slope. We therefore conducted a second simulation study with a more complex data-generating model containing three random slopes (in addition to six linear random effects). We expected that LA and EVA would also perform well for this model. Furthermore, given the more complex data-generating model, we also expected to observe

Table 1. No. Of nonconverged replications, average computation time, and standard deviation of computation time in study 1.

			GH			LA			EVA			Bayes		
J	n_j	ICC	Ncr	T	SD_T	Ncr	T	SD_T	Ncr	T	SD_T	Ncr	T	SD_T
100	10	0.05	0.03	2.90	0.49	0.03	1.40	0.18	0.03	1.22	0.20	0.00	25.6	0.51
		0.10	0.00	3.11	0.18	0.00	1.46	0.08	0.00	1.23	0.05	0.00	25.5	0.11
		0.20	0.00	3.05	0.32	0.00	1.51	0.12	0.00	1.30	0.07	0.00	25.6	0.13
	25	0.05	0.00	4.26	0.21	0.00	2.07	0.08	0.00	1.81	0.07	0.00	31.4	0.13
		0.10	0.00	4.47	0.27	0.00	2.09	0.09	0.00	1.79	0.07	0.00	31.5	0.13
		0.20	0.00	4.41	0.48	0.00	2.12	0.14	0.00	1.82	0.08	0.00	31.7	0.17
300	10	0.05	0.00	10.3	0.52	0.00	4.72	0.17	0.00	7.78	0.20	0.00	72.3	0.35
		0.10	0.00	10.6	0.61	0.00	4.71	0.20	0.00	7.69	0.21	0.00	72.2	0.20
		0.20	0.00	10.1	1.03	0.00	4.53	0.33	0.00	7.69	0.18	0.00	72.3	0.20
	25	0.05	0.00	14.2	0.66	0.00	7.10	0.21	0.00	11.1	0.21	0.00	89.9	0.27
		0.10	0.00	14.7	0.89	0.00	7.12	0.29	0.00	10.9	0.22	0.00	90.1	0.24
		0.20	0.00	14.2	1.49	0.00	6.87	0.47	0.00	10.9	0.22	0.00	90.5	0.49

Note. Ncr = Number of nonconverged replications as relative percentages, T = mean computation time in seconds. SD_T = standard deviation of computation times in seconds.

Table 2. Median parameter estimate (Md), median absolute deviation (MAD), and coverage rate (CR) for selected parameters depending on the number of level-2 units J, the number of level-1 units n_j, and the ICC in study 1.

			Md				Mad				CR			
J	n_j	ICC	GH	LA	EVA	Bayes	GH	LA	EVA	Bayes	GH	LA	EVA	Bayes
Level-2 Path Coefficients $\beta^B_{a_z a_x}$ *(True Value = .50)*														
100	10	.05	.531	.540	.538	.593	.543	.547	.548	.636	.948	.942	.945	.964
		.10	.522	.526	.523	.479	.292	.290	.293	.283	.945	.945	.945	.936
		.20	.508	.508	.507	.476	.170	.170	.170	.168	.953	.953	.953	.958
	25	.05	.511	.511	.511	.504	.336	.337	.337	.320	.967	.967	.967	.970
		.10	.510	.510	.510	.503	.201	.200	.200	.196	.940	.940	.940	.944
		.20	.511	.511	.511	.505	.131	.131	.131	.132	.940	.940	.940	.954
300	10	.05	.532	.534	.533	.455	.293	.295	.295	.265	.958	.960	.960	.938
		.10	.509	.511	.510	.458	.155	.156	.156	.150	.955	.955	.955	.942
		.20	.506	.505	.505	.482	.095	.094	.095	.090	.927	.928	.928	.928
	25	.05	.494	.495	.495	.485	.197	.198	.198	.192	.933	.935	.935	.938
		.10	.508	.509	.508	.497	.116	.116	.116	.118	.943	.943	.943	.934
		.20	.504	.504	.504	.501	.076	.076	.076	.074	.953	.953	.953	.954
Level-2 Path Coefficients $\beta^B_{\beta^W_{zx} a_x}$ *(True Value = .25)*														
100	10	.05	.263	.257	.259	.291	.514	.520	.528	.568	.971	.966	.967	.946
		.10	.241	.241	.242	.243	.284	.288	.288	.291	.953	.953	.952	.948
		.20	.250	.251	.251	.243	.170	.170	.173	.175	.952	.950	.950	.960
	25	.05	.257	.255	.255	.260	.388	.390	.390	.397	.952	.952	.952	.950
		.10	.260	.258	.257	.245	.214	.214	.213	.219	.943	.945	.945	.948
		.20	.249	.249	.249	.247	.146	.144	.144	.149	.945	.943	.943	.948
300	10	.05	.256	.253	.254	.261	.276	.281	.282	.277	.958	.958	.958	.960
		.10	.250	.249	.249	.245	.171	.172	.174	.170	.957	.953	.953	.954
		.20	.246	.244	.245	.240	.102	.102	.103	.100	.945	.945	.945	.960
	25	.05	.248	.246	.246	.255	.198	.197	.197	.187	.948	.948	.948	.954
		.10	.249	.248	.248	.246	.129	.129	.129	.126	.973	.973	.973	.970
		.20	.257	.257	.257	.258	.080	.079	.079	.079	.957	.957	.958	.960
Variance Random Slope $\omega_{\beta_{Wzx}}$ *(True Value = .30)*														
100	10	.05	.279	.277	.279	.297	.063	.063	.062	.065	.910	.906	.909	.948
		.10	.286	.285	.285	.313	.057	.057	.057	.061	.932	.933	.932	.944
		.20	.283	.283	.283	.313	.060	.060	.060	.067	.918	.918	.918	.950
	25	.05	.286	.287	.287	.309	.045	.045	.045	.049	.925	.935	.935	.956
		.10	.288	.289	.289	.312	.046	.045	.045	.046	.925	.933	.933	.942
		.20	.289	.289	.289	.311	.047	.047	.047	.050	.918	.922	.922	.954
300	10	.05	.296	.296	.296	.304	.033	.033	.033	.033	.948	.948	.948	.954
		.10	.296	.296	.296	.305	.034	.034	.034	.034	.958	.958	.958	.960
		.20	.297	.297	.297	.307	.034	.034	.034	.036	.932	.928	.928	.940
	25	.05	.298	.297	.297	.306	.027	.028	.028	.030	.938	.940	.940	.938
		.10	.297	.297	.297	.303	.029	.029	.029	.030	.940	.938	.938	.952
		.20	.298	.298	.298	.306	.027	.027	.027	.027	.937	.938	.938	.956

larger differences between the three ML approaches in the average computation time.

Method

Data-Generating Model and Simulation Conditions

A path diagram of the data-generating model is provided in Figure 2. The model contains five observed variables. At level 1, three variables load on one latent factor (i.e., η_{vij}). This factor mediates the linear relationships between the two other variables. All path coefficients of the level-1 mediation model were allowed to differ between the level-2 units. The latent factor at level 2 is measured by the random intercepts of those indicators that load on the factor at level 1. Again, this factor is the mediator of the relationship of the intercepts of the two other variables. Finally, the differences between the random slopes were assumed to be correlated but unrelated to the differences in the intercepts.

The factor loadings of the latent mediator were set to the same values for the two levels. However, we did not assume invariance during estimation. The loading of the second indicator was set to 0.80, and it was set to 0.60 for the third indicator (the loading of the first indicator is fixed to 1 during estimation). The vector of (latent) means was set to $\boldsymbol{\mu} = (\mu_{v_1}, \mu_{v_2}, \mu_{v_3}, \mu_{\alpha_x}, 0, \mu_{\alpha_z}, \mu_{\beta^W_{vx}}, \mu_{\beta^W_{zv}}, \mu_{\beta^W_{zx}})^T = (0, 0, 0, 0, 0, 0, 0.20, 0.05, 0.25)^T$. The nonzero elements in B^B, that is, the path coefficients of the level-2 mediation model, were $\beta^B_{vx} = 0.3$, $\beta^B_{zv} = 0.05$, and $\beta^B_{zx} = 0.2$. Furthermore, the nonzero (diagonal) elements of the residual covariance matrix $\boldsymbol{\Theta}$ On level 1 were $\theta_{v1ij} = \theta_{v2ij} = \theta_{v3ij} = 0.1$ and the variance terms of the latent factors at level 1 were $\psi_{\eta_{xij}} = 0.90$, $\psi_{\eta_{vij}} = 0.40$, and $\psi_{\eta_{zij}} = 0.25$. Finally, the covariance matrix on level 2 was set to

$$\boldsymbol{\Omega} = \begin{pmatrix} \omega_{v_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \omega_{v_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \omega_{v_3} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \omega_{\alpha_x} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \omega_{\alpha_v} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \omega_{\alpha_z} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \omega_{\beta^W_{vx}} & \omega_{\beta^W_{vx},\beta^W_{zv}} & \omega_{\beta^W_{vx},\beta^W_{zx}} \\ 0 & 0 & 0 & 0 & 0 & 0 & \omega_{\beta^W_{vx},\beta^W_{zv}} & \omega_{\beta^W_{zv}} & \omega_{\beta^W_{zv},\beta^W_{zx}} \\ 0 & 0 & 0 & 0 & 0 & 0 & \omega_{\beta^W_{vx},\beta^W_{zx}} & \omega_{\beta^W_{zv},\beta^W_{zx}} & \omega_{\beta^W_{zx}} \end{pmatrix}$$

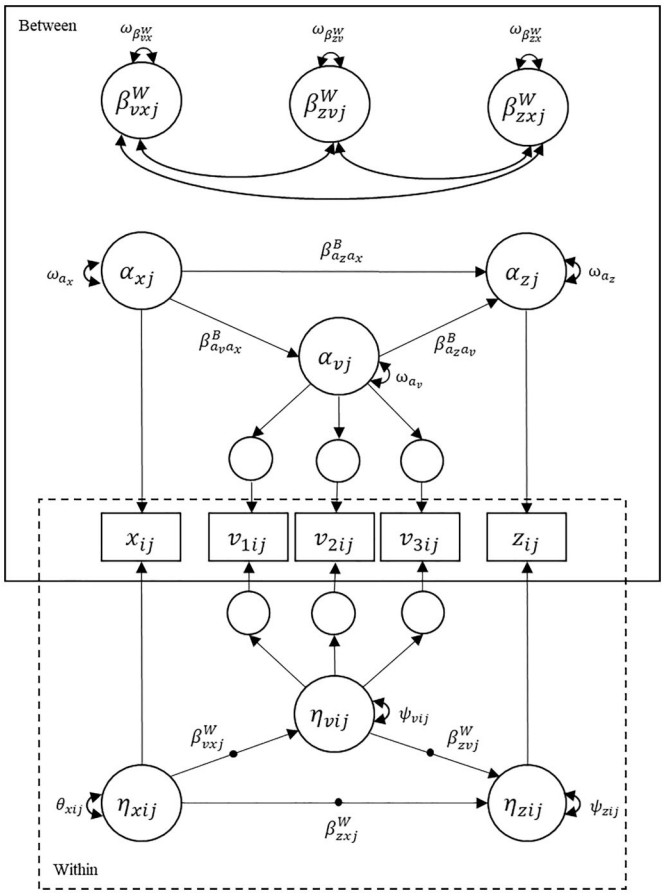

Figure 2. Path model for the MSEM used in simulation Study 2.

$$= \begin{pmatrix} 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0.1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.05 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.052 & 0.028 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.052 & 0.15 & 0.012 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0.028 & 0.012 & 0.1 \end{pmatrix} \quad (30)$$

implying correlations between the random slopes of 0.3 (0.052), 0.2 (0.028), and 0.1 (0.012). Of note, given the data-generating model's variance terms, the ICC of all observed variables is about 0.10.

Data Generation and Analysis

We used `R` to generate 500 datasets in each of the four simulation conditions. In each replication, we first estimated the random intercept version of the model in `lavaan` (Rosseel, 2012) to obtain reasonable starting values. Then, we fitted the data using the three ML estimation approaches and the Bayesian approach. Since the model contains three random slopes, using 30 quadrature points in the GH approximation would not be feasible because this would result in 27,000 node vectors per level-2 unit. In the literature it is sometimes suggested to use at least three to five points per dimension (see Verbeke & Molenberghs, 2005),

Table 3. No. Of nonconverged replications, average computation time, and standard deviation of computation time in study 2.

		GH			LA			EVA			Bayes		
J	n_j	Ncr	T	SD_T	Ncr	T	SD_T	Ncr	T	SD_T	Ncr	T	SD_T
100	10	0.01	138.7	15.9	0.14	31.1	3.9	0.08	41.9	1.8	0.00	222.1	0.7
	25	0.01	195.6	14.4	0.04	39.7	3.2	0.03	54.3	1.9	0.00	173.3	0.3
300	10	0.00	449.8	31.3	0.07	101.7	11.2	0.04	291.0	5.7	0.00	365.0	1.0
	25	0.00	603.8	57.1	0.02	123.7	13.3	0.01	367.8	5.6	0.00	445.1	1.1

Note. Ncr = Number of nonconverged replications in per cent, T = mean computation time in seconds. SD_T = standard deviation of computation times in seconds.

we therefore decided to use five points per dimension in the simulation. To obtain the Bayesian estimates, we again used Mplus (Muthén & Muthén 1998-2017) with the same non-informative prior distributions as in Study 1. The chain length was again set to 15,000 samples, with the first 7,500 samples discarded as burn-in. Similar to Study 1, prior practice runs—in the $J = 100$ and $n_j = 10$ condition—showed that this chain length resulted in an average effective sample size greater than 400 for each parameter. Finally, we used the posterior medians as parameter estimates and examined the same performance measures as in Study 1.

Results and Discussion

We first analyzed the number of nonconverged replications (using the same approach as in Study 1) and the mean and standard deviation of the computation times (see Table 3). For Bayes, all replications converged. For the ML approaches, we found that GH had the fewest convergence problems, with nonconvergence occurring in only 1% of the replications when the number of level-2 units was 100. The number of nonconverged replications was larger for LA and particularly pronounced when the number of level-2 units was 100 and the number of level-1 units was 10. However, the convergence rate improved substantially with larger sample sizes. EVA fell in between, with nonconvergence rates ranging from smaller to larger values depending on the number of level-2 and level-1 units. With regard to the computation times, we found that LA was the fastest method in all four simulation conditions. GH was substantially slower than both LA and EVA, requiring about 3-5 times longer compared with the other two ML approaches. Interestingly, when we looked at the time to convergence only, we found that EVA was faster than LA (in the order of the conditions in Table 3, the averages were 12.5, 18.5, 47.2, and 70.7 seconds); that is, EVA was slower than LA because we computed the Hessian matrix across both the parameters *and* the expectations of the variational densities. Finally, Bayes was slower than LA and EVA, but faster than GH. However, this result may partly be due to our specification that 15,000 samples should be generated from the posterior distributions.

Preliminary analyses showed that the differences between the four approaches in estimating the parameters of the level-1 model were negligible. The complete results, including the means and standard deviations of the estimates (whose pattern was similar to that of the medians and

Table 4. Median parameter estimate (Md), median absolute deviation (MAD), and coverage rate (CR) for selected parameters depending on the number of level-2 units J and the number of level-1 units n_j in study 2.

	Par.	GH	LA	EVA	Bayes	GH	LA	EVA	Bayes	GH	LA	EVA	Bayes	GH	LA	EVA	Bayes
Md	λ_2^B	.809	.809	.811	.836	.791	.791	.794	.814	.799	.802	.802	.819	.800	.803	.802	.806
	β_{zx}^B	.193	.196	.198	.198	.202	.196	.197	.205	.204	.206	.206	.207	.202	.202	.203	.202
	$\mu_{\beta_{zx}^W}$	.250	.248	.248	.248	.259	.248	.248	.249	.252	.250	.250	.249	.253	.250	.250	.250
	ω_{v_1}	.096	.095	.095	.096	.095	.095	.095	.100	.098	.098	.099	.100	.098	.099	.098	.100
	$\omega_{\beta_{zx}^W}$	.119	.090	.092	.111	.109	.096	.097	.107	.112	.094	.093	.103	.102	.098	.099	.103
	$\rho_{\beta_{zv}^W\beta_{zx}^W}$	.205	.103	.110	.083	.222	.105	.104	.099	.230	.105	.109	.099	.241	.103	.102	.099
MAD	λ_2^B	.280	.282	.282	.259	.229	.229	.234	.228	.153	.152	.153	.152	.141	.141	.141	.141
	β_{zx}^B	.151	.150	.151	.159	.126	.125	.125	.122	.093	.094	.094	.093	.062	.063	.064	.064
	$\mu_{\beta_{zx}^W}$	.050	.043	.043	.043	.056	.037	.037	.039	.029	.026	.027	.027	.034	.022	.022	.022
	ω_{v_1}	.034	.034	.034	.025	.032	.032	.032	.019	.020	.020	.020	.016	.018	.018	.018	.010
	$\omega_{\beta_{zx}^W}$	.036	.029	.029	.033	.026	.018	.018	.021	.019	.014	.014	.015	.016	.010	.010	.010
	$\rho_{\beta_{zv}^W\beta_{zx}^W}$	.244	.156	.156	.148	.228	.123	.121	.116	.142	.101	.101	.098	.151	.068	.069	.069
CR	λ_2^B	.926	.925	.923	.947	.938	.938	.939	.972	.937	.934	.943	.938	.952	.956	.953	.964
	β_{zx}^B	.944	.952	.949	.964	.925	.910	.915	.936	.930	.929	.937	.938	.950	.954	.955	.956
	$\mu_{\beta_{zx}^W}$	.889	.950	.951	.954	.649	.934	.934	.946	.907	.939	.932	.938	.698	.937	.938	.936
	ω_{v_1}	.949	.950	.949	.956	.945	.950	.952	.928	.965	.964	.967	.940	.963	.968	.966	.950
	$\omega_{\beta_{zx}^W}$	.886	.882	.887	.930	.705	.920	.920	.926	.898	.905	.923	.964	.728	.935	.936	.942
	$\rho_{\beta_{zv}^W\beta_{zx}^W}$	.705	.954	.956	.962	.392	.946	.945	.948	.637	.946	.950	.956	.388	.949	.949	.954

Note. True parameter values were $\lambda_{v2}^B = .80$, $\beta_{zx}^B = .20$, $\mu_{\beta_{zx}^W} = .25$, $\omega_{v1} = .10$, $\omega_{\beta_{zx}^W} = .10$, and $\rho_{\beta_{zv}^W\beta_{zx}^W} = .10$.

MADs), can be downloaded from the OSF project accompanying this manuscript. Therefore, we present the results for a selection of parameters of the level-2 model in Table 4. Turning to bias first, we see that all approaches provided almost unbiased estimates of the factor loadings, the path coefficients at level 2, and mean parameters at level 2 (i.e., elements in Λ^B, $\boldsymbol{B}^B$, and $\boldsymbol{\mu}$). When there was a bias, it became smaller with an increasing number of level-2 units. Similar results were obtained for the MAD of these parameters, with less variable estimates for all approaches as the sample size increased. For the variance and covariance parameters (i.e., the last three parameters in each block of Table 4), large biases occurred for GH whereas estimates from LA, EVA, and Bayes were again almost unbiased. For all estimation methods, the MAD decreased as the sample size increased. However, this decrease in variability probably led to the confidence intervals for GH narrowing around the wrong value, which may explain the poor results for the CR of GH. For LA and EVA, the CR was close to nominal for most parameters. The exception was $\omega_{\beta_{zx}^W}$, where undercoverage was observed when the number of level-2 units was small. Here, Bayesian credibility intervals were closer to the nominal value.

General Discussion

In this article we investigated the performance of LA and EVA with regard to the estimation of the parameters of an MSEM with random slopes. We conducted two simulation studies that differed in the complexity of the data-generating model. In both studies, we manipulated the number of level-2 units and the number of level-1 units. Furthermore, in the first simulation, we also examined the impact of the ICC on estimation performance. In the first simulation, the population model was a latent covariate model with a single random slope. The results showed that there were no substantial differences in computation time between the three ML approaches, and all four approaches (the three ML methods and the Bayesian approach) performed similarly well in terms of bias, variability, and coverage. For all approaches, we found that the lower the ICC, the worse was the performance for some parameters, especially at level 2. This result is consistent with other simulation research showing that level-2 parameters are more difficult to estimate when the proportion of between-cluster variance is small (e.g., Maas & Hox, 2005).

The second simulation study examined a more complex model with three random slopes, resulting in a higher-dimensional integration problem. For this model, GH required a substantial reduction in the number of quadrature points per dimension to remain computationally feasible. Even with this reduction, GH took considerably longer than LA, EVA, and Bayes to converge. Furthermore, the use of fewer quadrature points resulted in larger biases for GH compared to LA and EVA, especially for the covariance parameters at level 2 (see Nestler & Erdfelder, 2023, for similar results). In contrast, and similar to the Bayesian approach, both LA and EVA provided parameter estimates with similar or smaller bias and variability, while requiring substantially less computation time. EVA took somewhat longer than LA, but this difference was primarily attributable to the computation of standard errors rather than to the optimization procedure itself. However, there were also more nonconverged replications for both approaches, although the overall number was still small. Finally, and independent of the concrete approach, we observed that larger numbers of level-1 and level-2 units were associated with reduced bias, more stable estimates, and improved CRs.

Altogether, our simulation studies provide encouraging evidence for the utility of LA and EVA in estimating complex MSEMs with random slopes. Nevertheless, several questions remain for future methodological research. First, alternative approaches for computing standard errors within the EVA framework should be further investigated. In our implementation, we computed standard errors by calculating the Hessian across the whole parameter vector, and this increased the overall computation time for EVA to a substantial extent. In a recent article, Grønberg, Ormerod, & Clemmensen (2025) proposed to adapt the method

proposed in Louis (1982) for the EM algorithm to the VA approximation in the context of a cross-random effects model for ordinal data. Whether this approach is feasible in the case of the MSEM, is an interesting question for future simulation research.

Second, future research should examine the performance of the two ML methods for even more complex models. For example, some applications require models in which factor loadings (Asparouhov, Hamaker, & Muthén, 2018) or residual variance terms (e.g., Nestler & Blozis, 2024) are allowed to vary randomly across level-2 units, representing another source of heterogeneity in the measurement model. Such specifications lead to additional random effects that enter the model nonlinearly, further increasing the dimensionality of the integration problem. It would be interesting to investigate whether LA and EVA continue to provide accurate estimates and remain computationally efficient in such contexts, or whether their performance diminishes relative to alternative ML methods and a Bayesian approach.

Third, other integral approximation methods should also be investigated. In our study, we compared LA and EVA with GH, but it might also be interesting to compare them to Quasi Monte-Carlo approximation (Crowther, 2017) or importance sampling (Kuk, 1999). Both methods have the advantage that the points used to approximate the integral do not increase with the size of the random effect vector. Furthermore, another promising approach might be composite likelihood estimation (Varin, Reid, & Firth, 2011). Here, the full likelihood is approximated by a product of lower-dimensional marginal or conditional likelihoods. This method has been successfully applied in various contexts where the full likelihood is intractable. Moreover, we assumed here that all observed variables are continuous. A challenging question for future research is to extend the MSEM to ordinal variables—or a mix of continuous and ordinal variables—and to develop an ML estimation routine for this type of model (McNeish, Somers, & Savord, 2024; Nestler, 2013).

Finally, when the two methods investigated in this article are intended to be used in applied research, they must be implemented in accessible statistical software. Currently, ML estimation of MSEM with random slopes is available in only a limited number of (commercial) software packages, and an EM algorithm is the most widely implemented approach. Our findings suggest that LA and EVA could provide valuable alternatives. However, to be used, they will need to be integrated into user-friendly software with intuitive model specification syntax and robust numerical implementations.

To conclude, this article demonstrated that the LA and EVA are promising approaches to estimate the parameters of the MSEM with random slopes using ML. Both methods provide accurate parameter estimates while offering substantial computational advantages for complex models (in comparison to the standard approach). As MSEMs will be increasingly used in psychological research, efficient estimation methods will become more important to allow applied researchers to analyze their real-world data in a suitable time.

References

Asparouhov, T., Hamaker, E. L., & Muthén, B. (2018). Dynamic structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 359–388. https://doi.org/10.1080/10705511.2017.1406803

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 119–142. https://doi.org/10.1080/10705511.2018.1511375

Bolger, N., & Laurenceau, J.-P. (2013). *Intensive longitudinal methods: An introduction to diary and experience sampling research*. Guilford Press.

Crowther, M. J. (2017). *Extended multivariate generalised linear and non-linear mixed effects models*.

Du Toit, S. H., & Du Toit, M. (2008). Multilevel structural equation modeling. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis*. (p. 435–478). Springer.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika*, *53*, 455–467. https://doi.org/10.1007/bf02294400

Grønberg, M., Ormerod, J. T., & Clemmensen, L. K. H. (2025). Gaussian variational approximation for ordinal data with crossed random effects. *Journal of Computational and Graphical Statistics*, *0*, 1–12. https://doi.org/10.1080/10618600.2025.2505018

Korhonen, P., Hui, F. K. C., Niku, J., & Taskinen, S. (2023). Fast and universal estimation of latent variable models using extended variational approximations. *Statistics and Computing*, *33*, 26. https://doi.org/10.1007/s11222-022-10189-w

Kristensen, K., Nielsen, A., Berg, C. W., Skaug, H., & Bell, B. M. (2016). TMB: Automatic differentiation and Laplace approximation. *Journal of Statistical Software*, *70*, 1–21. https://doi.org/10.18637/jss.v070.i05

Kuk, A. Y. C. (1999). Laplace importance sampling for generalized linear mixed models. *Journal of Statistical Computation and Simulation*, *63*, 143–158. https://doi.org/10.1080/00949659908548521

Liang, J., & Bentler, P. M. (2004). An EM algorithm for fitting two-level structural equation models. *Psychometrika*, *69*, 101–122. https://doi.org/10.1007/bf02295842

Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *44*, 226–233. https://doi.org/10.1111/j.2517-6161.1982.tb01203.x

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92. https://doi.org/10.1027/1614-2241.1.3.86

McNeish, D. M., Somers, J. A., & Savord, A. (2024). Dynamic structural equation models with binary and ordinal outcomes in mplus. *Behavior Research Methods*, *56*, 1506–1532. https://doi.org/10.3758/s13428-023-02107-3

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, *54*, 557–585. https://doi.org/10.1007/bf02296397

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus user's guide*. (8th edition). Muthén & Muthén.

Nestler, S. (2013). A monte carlo study comparing PIV, ULS, and DWLS in the estimation of dichotomous confirmatory factor analysis. *The British Journal of Mathematical and Statistical Psychology*, *66*, 127–143. https://doi.org/10.1111/j.2044-8317.2012.02044.x

Nestler, S. (2020). Modeling interindividual differences in latent within-person variation: The confirmatory factor level variability model. *The British Journal of Mathematical and Statistical Psychology*, *73*, 452–473. https://doi.org/10.1111/bmsp.12196

Nestler, S., & Blozis, S. A. (2024). A latent variable mixed-effects location scale model that also considers between-person differences in

the autocorrelation. *Statistics in Medicine*, *43*, 89–101. https://doi.org/10.1002/sim.9943

Nestler, S., & Erdfelder, E. (2023). Random effects multinomial processing tree models: A maximum likelihood approach. *Psychometrika*, *88*, 809–829. https://doi.org/10.1007/s11336-023-09921-w

Nestler, S., & Humberg, S. (2024). Univariate autoregressive structural equation models as mixed-effects models. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*, 357–366. https://doi.org/10.1080/10705511.2023.2212865

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*, 140–153. https://doi.org/10.1198/tast.2010.09058

Ormerod, J. T., & Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *21*, 2–17. https://doi.org/10.1198/jcgs.2011.09118

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*, 12–35. https://doi.org/10.1080/10618600.1995.10474663

R Core Team. (2025). 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: Promoting Communications on Statistics and Stata*, *2*, 1–21. https://doi.org/10.1177/1536867x0200200101

Rockwood, N. J. (2020). Maximum likelihood estimation of multilevel structural equation models with random slopes for latent covariates. *Psychometrika*, *85*, 275–300. https://doi.org/10.1007/s11336-020-09702-9

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y. (2021). Evaluating the observed log-likelihood function in two-level structural equation modeling with missing data: From formulas to R code. *Psych*, *3*, 197–232. https://doi.org/10.3390/psych3020017

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling*. Chapman and Hall, CRC Press.

Tuerlinckx, F., Rijmen, F., Verbeke, G., & De Boeck, P. (2006). Statistical inference in generalized linear mixed models: A review. *The British Journal of Mathematical and Statistical Psychology*, *59*, 225–255. https://doi.org/10.1348/000711005X79857

Varin, C., Reid, N., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*, 5–42.

Verbeke, G., & Molenberghs, G. (2005). *Models for discrete longitudinal data*. Springer.

Wasserman, L. (2004). *All of statistics: A concise course in statistical inference*. Springer.