Received: 5 July 2021 | Accepted: 28 September 2022

DOI: 10.1111/bmsp.12294

ARTICLE

Multilevel SEM with random slopes in discrete data using the pairwise maximum likelihood

Maria T. Barendse[1,2] | Yves Rosseel[3]

[1]Oral Public Health Department, Academic Centre for Dentistry, Amsterdam, Netherlands

[2]Language and Genetics Department, Max Planck Institute, Nijmegen, Netherlands

[3]Department of Data Analysis, Ghent University, Ghent, Belgium

Correspondence
Maria T. Barendse, Oral Public Health Department, Academic Centre for Dentistry.
Email: mariska.barendse@acta.nl

Abstract

Pairwise maximum likelihood (PML) estimation is a promising method for multilevel models with discrete responses. Multilevel models take into account that units within a cluster tend to be more alike than units from different clusters. The pairwise likelihood is then obtained as the product of bivariate likelihoods for all within-cluster pairs of units and items. In this study, we investigate the PML estimation method with computationally intensive multilevel random intercept and random slope structural equation models (SEM) in discrete data. In pursuing this, we first reconsidered the general 'wide format' (WF) approach for SEM models and then extend the WF approach with random slopes. In a small simulation study we the determine accuracy and efficiency of the PML estimation method by varying the sample size (250, 500, 1000, 2000), response scales (two-point, four-point), and data-generating model (mediation model with three random slopes, factor model with one and two random slopes). Overall, results show that the PML estimation method is capable of estimating computationally intensive random intercept and random slopes multilevel models in the SEM framework with discrete data and many (six or more) latent variables with satisfactory accuracy and efficiency. However, the condition with 250 clusters combined with a two-point response scale shows more bias.

KEYWORDS

discrete data, multilevel models, pairwise maximum likelihood, random slopes

INTRODUCTION

Structural equation modelling (SEM) is an effective modelling framework used for measuring latent variables through indicators and studying the associations between the latent variables and/or observed variables (see Bollen, 1989; Kline, 2015). The generality of SEM is reflected in the number of applications in the social sciences (e.g., Guo et al., 2008; Kline, 2015; MacCallum & Austin, 2000; Xiong et al., 2015). SEM was introduced for continuous data and later extended to handle more complex data. Here, we will focus on the pairwise maximum likelihood (PML) estimation method for SEM to deal with two data complexities: discrete responses (i.e., binary coding or three- or four-point scales) and multilevel structures.

To estimate discrete data in the context of SEM, the PML estimation method was introduced by Jöreskog and Moustaki (2001). With PML estimation the product of bivariate (and sometimes univariate) likelihoods is calculated (see Jöreskog & Moustaki, 2001). Outside the SEM context, PML is part of a broader framework of composite maximum likelihood estimators (see Varin, 2008; Varin et al., 2011). Katsikatsou et al. (2012) showed that PML estimation produced low biases within the SEM framework. Compared to other frequentist SEM estimation methods for discrete data, the most prominent advantage of the PML estimation method is the possibility of computing models with a large number of latent variables. The marginal maximum likelihood (MML; see Bock & Aitkin, 1981) estimation method calculates a full likelihood and cannot estimate models with too many latent variables. Alternatively, the (weighted) least squares estimation method (Browne, 1984; Muthén et al., 1997) can be used to estimate discrete data. This estimation method also uses bivariate and univariate information and is in that respect comparable to the PML estimation method.

Estimating models with both discrete data and a multilevel structure complicates the estimation. In multilevel data, lower-level units are selected within higher-level units (i.e., clusters). As a consequence, units within a cluster tend to be more alike than units from different clusters. This dependency in the data introduces an additional source of variation that needs to be taken into account in analysing data (Hox et al., 2017). Multilevel data in SEM are usually analysed with the 'long format' (LF) approach (see McDonald & Goldstein, 1989; Muthén, 1990), where each row corresponds to the data of a single unit and multiple rows constitute a single cluster. In the 'wide format' (WF) or 'multivariate' approach, each data row is independent and corresponds to a single cluster (see Barendse & Rosseel, 2020, for a multilevel model framework with a random intercept). However, the number of columns can increase substantially with large clusters in the WF approach. The WF approach is therefore particularly useful for data with relatively small cluster sizes (e.g., Koomen et al., 2007; Lau et al., 2015; Mahlke et al., 2016; Moorman, 2016; NLSAH, 2005).

This paper combines the challenges of multilevel structures including random slopes and discrete data using the PML estimation method. The pairwise likelihood in multilevel data is obtained as the product of bivariate likelihoods for within-cluster pairs of units and items. So far, the PML estimation method with random intercepts and random slopes has only been used for generalized mixed models for binary data (see Bellio & Varin, 2005; Cho & Rabe-Hesketh, 2011; Renard et al., 2004; Tibaldi et al., 2007). Random slopes show up in multilevel models if the effect of covariates is allowed to vary across clusters. To deal with the random slopes, casewise estimation is applied. Casewise estimation blurs away most of the differences between the WF and LF approaches. The aim of this study is to estimate complex multilevel random intercept and random slope SEM with at least six latent variables in the WF approach that cannot be estimated with other frequentist estimation methods. Computationally, these models cannot be estimated with intensive numerical (e.g., adaptive Gauss–Hermite quadrature) multilevel marginal maximum likelihood estimation methods as the number of latent variables that has to be integrated out is too high (MML; see Hedeker & Gibbons, 1994). In addition, these models cannot be estimated with multilevel weighted least squares (WLS; see Asparouhov & Muthén, 2007), as it does not allow for the casewise estimation that is essential for models with random slopes.

The paper is organized as follows. We first describe the general SEM framework and the PML estimation method for discrete data. Then we introduce a new procedure to analyse data in the WF approach in multilevel data with random slopes and show how PML estimation can be used within this framework if the data are discrete. Next, we perform a small simulation study using two complex multilevel models (i.e., a mediation model and a factor model) with random slopes to evaluate the PML estimation method in terms of the accuracy of the parameter estimates. Finally, we discuss the usefulness of the PML estimation method in multilevel data.

General SEM framework

Structural equation modelling can estimate a wide range of models. It can be considered as a combination of regression or path analyses and factor analysis. The general model for unit *i* with continuous data can be described by a measurement model, that relates measured variables to latent variables, and a structural model, that relates latent variables to one another:

$$\mathbf{y}_{\mathrm{i}} = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta}_{\mathrm{i}} + \boldsymbol{\epsilon}_{\mathrm{i}}, \tag{1}$$

$$\boldsymbol{\eta}_{\mathrm{i}} = \boldsymbol{\alpha} + \mathbf{B}\boldsymbol{\eta}_{\mathrm{i}} + \boldsymbol{\zeta}_{\mathrm{i}}, \tag{2}$$

where $\boldsymbol{y}_{\mathrm{i}}$ is a p-dimensional vector of observed variables for unit i, $\boldsymbol{\nu}$ is a p-dimensional vector of intercepts, $\boldsymbol{\Lambda}$ is a $p \times l$ matrix of factor loadings relating the observed variables to the l-dimensional latent variables $\boldsymbol{\eta}_{\mathrm{i}}$, $\boldsymbol{\epsilon}_{\mathrm{i}}$ is a p-dimensional vector of measurement errors or residuals, $\boldsymbol{\alpha}$ is an l-dimensional vector of latent factor means and intercepts, $\boldsymbol{B}$ is an $l \times l$ matrix of regression coefficients among the latent factors, and $\boldsymbol{\zeta}_{\mathrm{i}}$ is an l-dimensional vector of residuals for unit i. Assuming $\mathrm{Cov}(\boldsymbol{\epsilon}, \boldsymbol{\zeta}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\eta}, \boldsymbol{\epsilon}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\epsilon}_{\mathrm{i}}) = \mathbf{0}$, $\mathrm{E}(\boldsymbol{\zeta}) = \mathbf{0}$, $\mathrm{diag}(\boldsymbol{B}) = \mathbf{0}$, and that $(\boldsymbol{I} - \boldsymbol{B})$ is invertible, where $\boldsymbol{I}$ is an $l \times l$ identity matrix, we can find expressions for the $p \times p$ covariance matrix $\boldsymbol{\Sigma}_{\mathrm{i}}$ and the l-dimensional vector of the mean structure $\boldsymbol{\mu}_{\mathrm{i}}$ (Equation 4) of $\boldsymbol{y}_{\mathrm{i}}$:

$$\boldsymbol{\Sigma}(\boldsymbol{\theta})_{\mathrm{i}} = \boldsymbol{\Lambda}(\boldsymbol{I} - \mathbf{B})^{-1}\Psi(\boldsymbol{I} - \mathbf{B})^{-1T}\boldsymbol{\Lambda}^{T} + \Theta, \tag{3}$$

$$\boldsymbol{\mu}(\boldsymbol{\theta})_{\mathrm{i}} = \boldsymbol{\nu} + \boldsymbol{\Lambda}(\boldsymbol{I} - \mathbf{B})^{-1}\boldsymbol{\alpha}, \tag{4}$$

where the variances and covariances of $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$ are denoted by $\boldsymbol{\Psi}$ and Θ, respectively. The model parameter vector $\boldsymbol{\theta}$ includes the free parameters in $\boldsymbol{\Lambda}$, $\boldsymbol{B}$, Θ, $\boldsymbol{\Psi}$, $\boldsymbol{\nu}$, and $\boldsymbol{\alpha}$. Before estimating models, the scale for the latent variables (i.e., $\boldsymbol{\eta}$) needs to be defined, for example by fixing the first factor loading of each latent variable to unity.

Discrete data with the PML estimation method

To deal with discrete data, the PML estimation method assumes an underlying normally distributed continuous latent response variable. A variable y_{ig} for individual i on item g with C_{g} response scales stems from an underlying continuous variable y^{*}_{ig} with a normal distribution $N\left(y^{*}_{\mathrm{ig}} \mid 0, \sigma^{2}_{\mathrm{g}}\right)$ and $\tau_{\mathrm{g,c}}$ values that refer to thresholds

$$y_{\mathrm{ig}} = c_{\mathrm{g}} \Leftrightarrow \tau_{\mathrm{g,c-1}} < y^{*}_{\mathrm{ig}} < \tau_{\mathrm{g,c}} \tag{5}$$

for categories $c_{\mathrm{g}} = 1, 2, \ldots, C_{\mathrm{g}}$, with $\tau_{\mathrm{g,0}} = -\infty$ and $\tau_{\mathrm{g,C}} = +\infty$. Instead of calculating a full likelihood, PML estimation breaks down the complex likelihood. The log-likelihood of the PML estimation method for an individual is then calculated as the sum of $p^{\star} = p(p-1)/2$ components, each component being the bivariate log-likelihood of two variables (i.e., g and h):

$$\begin{aligned}\log l_i &= \sum_{g=1}^{p-1} \sum_{h=g+1}^{p} \left[\log f\left(y_{ig}, y_{ih}; \boldsymbol{\theta}\right)\right] \\ &= \sum_{g<h} \left[\log f\left(y_{ig}, y_{ih}; \boldsymbol{\theta}\right)\right].\end{aligned} \tag{6}$$

The total log-likelihood of the data is the sum of all individual contributions in a random sample $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_I\}$ of size I and equals

$$\log L(\boldsymbol{\theta}; \mathbf{Y}) = \sum_{i=1}^{I} \log l_i. \tag{7}$$

The exact form of $f\left(y_{ig}, y_{ih}; \boldsymbol{\theta}\right)$ in Equation 7 for discrete indicators g and h equals

$$\log f\left(y_{ig}, y_{ih}; \boldsymbol{\theta}\right) = \sum_{a=1}^{C_g} \sum_{b=1}^{C_h} I\left(y_{ig} = a, y_{ih} = b\right) \log \omega\left(y_{ig} = a, y_{ih} = b; \boldsymbol{\theta}\right), \tag{8}$$

with

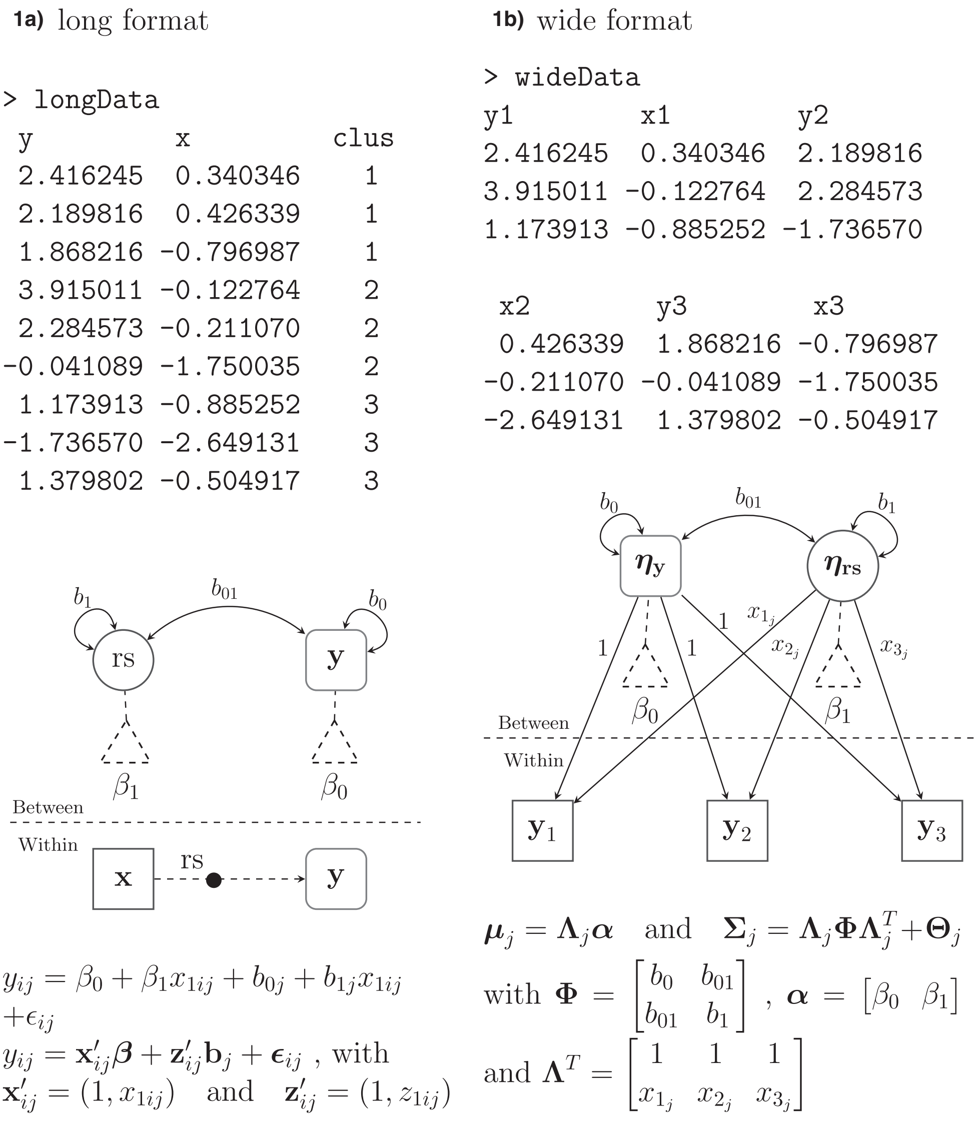
$$\begin{aligned}\omega\left(y_{ig} = a, y_{ih} = b; \boldsymbol{\theta}\right) &= \int_{\tau_{g,a-1}}^{\tau_{g,a}} \int_{\tau_{h,b-1}}^{\tau_{h,b}} f\left(y^*_{ig}, y^*_{ih}; \boldsymbol{\theta}\right) dy^*_{ig} dy^*_{ih}, \\ &= \Phi\left(\tau_{g,a}, \tau_{h,b}; \rho_{gh}\right) - \Phi\left(\tau_{g,a-1}, \tau_{h,b}; \rho_{gh}\right) \\ &\quad - \Phi\left(\tau_{g,a}, \tau_{h,b-1}; \rho_{gh}\right) + \Phi\left(\tau_{g,a-1}, \tau_{h,b-1}; \rho_{gh}\right),\end{aligned} \tag{9}$$

where ρ_{gh} is the model implied correlation between y^*_{ig} and y^*_{ih}, and $\Phi(\tau_1, \tau_2; \rho)$ is the bivariate cumulative normal distribution with correlation ρ evaluated at the point (τ_1, τ_2). Before estimating the model with the PML method, the metric for y^* needs to be determined. Two popular ways are to fix the total variance (so-called delta parameterization: $\boldsymbol{\Theta} = \boldsymbol{\Delta}^{-2} - \mathrm{diag}(\boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^* = \boldsymbol{\Lambda}\ (\boldsymbol{I} - \boldsymbol{B})^{-1} \boldsymbol{\Psi} (\boldsymbol{I} - \boldsymbol{B})^{-1T}\ \boldsymbol{\Lambda}^T$ and $\boldsymbol{\Delta}$ the scaling factors) or to fix the residual variance (so-called theta parameterization: $\boldsymbol{\Delta}^{-2} = \mathrm{diag}(\boldsymbol{\Sigma}^*) + \boldsymbol{\Theta}$). Other ways of scaling are also possible (see Lee et al., 1990, 1992).

Research has shown that the PML estimation method provides accurate and efficient results in the SEM context (see Jöreskog & Moustaki, 2001; Katsikatsou et al., 2012) as well as in the broader framework of composite maximum likelihood estimators (see Lindsay, 1988; Varin, 2008). In theory, Equation 6 is general and can deal with any type of data (continuous or discrete, and combinations thereof). Barendse and Rosseel (2020) estimated SEM with the PML estimation method for a mixture of binary and continuous data, by estimating Pearson, tetrachoric, polychoric, and polyserial correlations. Standard errors and missing-data procedures for the PML estimation method have been developed by Katsikatsou et al. (2012) and Katsikatsou and Moustaki (2017).

Multilevel SEM in the WF approach

Before investigating the PML estimation method for multilevel data in the WF approach, we will first explain the WF framework for continuous data. With continuous multilevel data the LF and WF approaches obtain identical results (see Mehta & Neale, 2005). A random intercept, random slope regression model in the LF approach can be estimated in the WF approach as a single-level confirmatory factor analysis in the SEM framework (see Bryk & Raudenbush, 1987; Chou et al., 1998; MacCallum et al., 1997; McArdle & Epstein, 1987; Mehta & Neale, 2005; Meredith & Tisak, 1990). The random slope allows the relationship between the covariate and the dependent variable ($\boldsymbol{y}$) to be different for each cluster. The mean of the random slope represents the fixed effect of the covariate and the variance of the random slope reflects the variability. Figure 1a and 1b show the data layout, a graphical representation of the model, and the formula of a random intercept and random slope multilevel linear regression model for covariate $\boldsymbol{x}$ in the LF approach and in the WF approach, respectively. In Figure 1b for the WF approach there is no single covariance matrix for the entire sample as with conventional SEM, but a cluster-dependent covariance matrix (Mehta & Neale, 2005), where $\boldsymbol{\Lambda}_{\mathrm{j}}$ is now cluster-specific. To estimate this model in the WF approach, we have to use casewise calculations to compute a likelihood with cluster-specific vectors, not unlike the full-information maximum likelihood approach that was introduced in SEM for handling missing data (Arbuckle, 1996). Consequently, a multilevel random intercept and random slope model can only be estimated with software that allows for casewise estimation to deal with a cluster-specific model implied covariance and mean structure.



1a) long format

```
> longData
  y          x          clus
  2.416245  0.340346    1
  2.189816  0.426339    1
  1.868216 -0.796987    1
  3.915011 -0.122764    2
  2.284573 -0.211070    2
 -0.041089 -1.750035    2
  1.173913 -0.885252    3
 -1.736570 -2.649131    3
  1.379802 -0.504917    3
```

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + b_{0j} + b_{1j} x_{1ij} + \epsilon_{ij}$$

$$y_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} + \mathbf{z}'_{ij}\mathbf{b}_j + \boldsymbol{\epsilon}_{ij} \text{ , with } \mathbf{x}'_{ij} = (1, x_{1ij}) \quad \text{and} \quad \mathbf{z}'_{ij} = (1, z_{1ij})$$

1b) wide format

```
> wideData
y1         x1         y2
2.416245   0.340346   2.189816
3.915011  -0.122764   2.284573
1.173913  -0.885252  -1.736570

 x2         y3         x3
 0.426339   1.868216  -0.796987
-0.211070  -0.041089  -1.750035
-2.649131   1.379802  -0.504917
```

$$\boldsymbol{\mu}_j = \boldsymbol{\Lambda}_j\boldsymbol{\alpha} \quad \text{and} \quad \boldsymbol{\Sigma}_j = \boldsymbol{\Lambda}_j\boldsymbol{\Phi}\boldsymbol{\Lambda}_j^T + \boldsymbol{\Theta}_j$$

$$\text{with } \boldsymbol{\Phi} = \begin{bmatrix} b_0 & b_{01} \\ b_{01} & b_1 \end{bmatrix}, \ \boldsymbol{\alpha} = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}$$

$$\text{and } \boldsymbol{\Lambda}^T = \begin{bmatrix} 1 & 1 & 1 \\ x_{1_j} & x_{2_j} & x_{3_j} \end{bmatrix}$$

FIGURE 1 A random intercept and slope model with the corresponding data, figure and formula in both (a) the LF approach and (b) WF approach. *Note*: Triangles indicate for the mean or intercepts. In (1a), $\mathbf{x}'_{\mathrm{ij}}$ equals the fixed effects matrix and $\mathbf{z}'_{\mathrm{ij}}$ indicates the random effects matrix. In (1b), the rounded boxes represent the within- and between- components of the original variable $\boldsymbol{y}$ for three units per cluster. $\boldsymbol{\eta}_{\mathbf{y}}$ corresponds to the unit-specific version of $\boldsymbol{y}$ at the between-level, which equals a factor with loadings fixed at unity. $\boldsymbol{\eta}_{\mathrm{rs}}$ corresponds to a random slope at the between-level with indicators fitted to the values of x_{j}. Residual variances (not shown here) contain equality restrictions to ensure only one residual variance is estimated.

A multilevel regression model can be extended to a multilevel SEM. Multilevel random intercept SEM in the LF approach is described by different authors (see Bryk & Raudenbush, 1987; McDonald, 1993; McDonald & Goldstein, 1989; Muthén, 1989, 1990; Schmidt, 1969). Mehta and Neale (2005), Curran (2003), and Bauer (2003) were among the first authors to describe random intercept SEM in the WF approach. Barendse and Rosseel (2020) described a general SEM WF framework for both continuous and discrete data and proposed steps to perform random intercept models. However, a complication arises in the case of a covariate that exists at the within-level and the between-level (see Lüdtke et al., 2008). In a more general sense, this also holds for every regression coefficient that exists at the within-level and the between-level. To obtain the correct between-level effect, one has to disentangle the between effect by estimating the between-level effect and subtract this from the within-level effect using a definition variable (see Lüdtke et al., 2008). To avoid using definition variables to calculate the between effect, we adjusted the estimation of models in the WF approach in a way that is more intuitive than the parameterization presented in Barendse and Rosseel (2020). In this new approach we disentangle the between-and within-levels by creating latent variables, which are then used to build models at the within-level and the between-level.

METHODS

We first reformulate the WF approach of Barendse and Rosseel (2020) to allow for a more intuitive model parameterization in the presence of a covariate or a regression coefficient at both the within and the between-level (see Lüdtke et al., 2008). Then we explain how to apply the PML estimation method with discrete data in the WF approach.

Reformulation of the WF approach

Based on the steps introduced by Barendse and Rosseel (2020), we formulate adjusted steps that first disentangle the between- and within-level of the model and then explicitly build models at both levels.

1. Rearrange the data in such a way that each row corresponds to a single cluster.
2. Disentangle each continuous endogenous variable into a between part (e.g., $\boldsymbol{\eta}_{\mathbf{y}_b}$) and a within part (e.g., $\boldsymbol{\eta}_{\mathbf{y}_w}$) by introducing new variables with factor loadings fixed to unity. For the between variables, representing the random intercepts of the model, one has to construct a latent variable where the indicators correspond to the unit-specific observations of that variable.
3. Construct a model with the newly introduced within-level variables involving the variables that belong to a single unit in a cluster and repeat this model as many times as the maximum cluster size.
4. Put equality constraints on all parameters across units in a cluster in the within part of the model. For example, in a one-factor model, equality constraints are necessary on the factor loadings, factor variances, and error variances. If variables are both at the within and the between-level, the intercepts at the within-level should be fixed to zero.
5. Construct a model at the between-level with the newly constructed between-level latent variables.

The solid lines in Figures 2 and 3 show a random intercept mediation model and a random intercept factor model, respectively. Appendix A shows the lavaan syntax for a multilevel mediation model (i.e., the model shown in Figure 2) and Appendix C shows the lavaan syntax for a multilevel factor model (i.e., the model shown in Figure 3).

This multilevel random intercept SEM can be extended with a random slope. Rockwood (2020) gives a detailed description of multilevel models in the LF approach with random slopes (e.g., Figure 6 in this paper shows a one-factor model in the LF approach with a random slope). In the WF approach

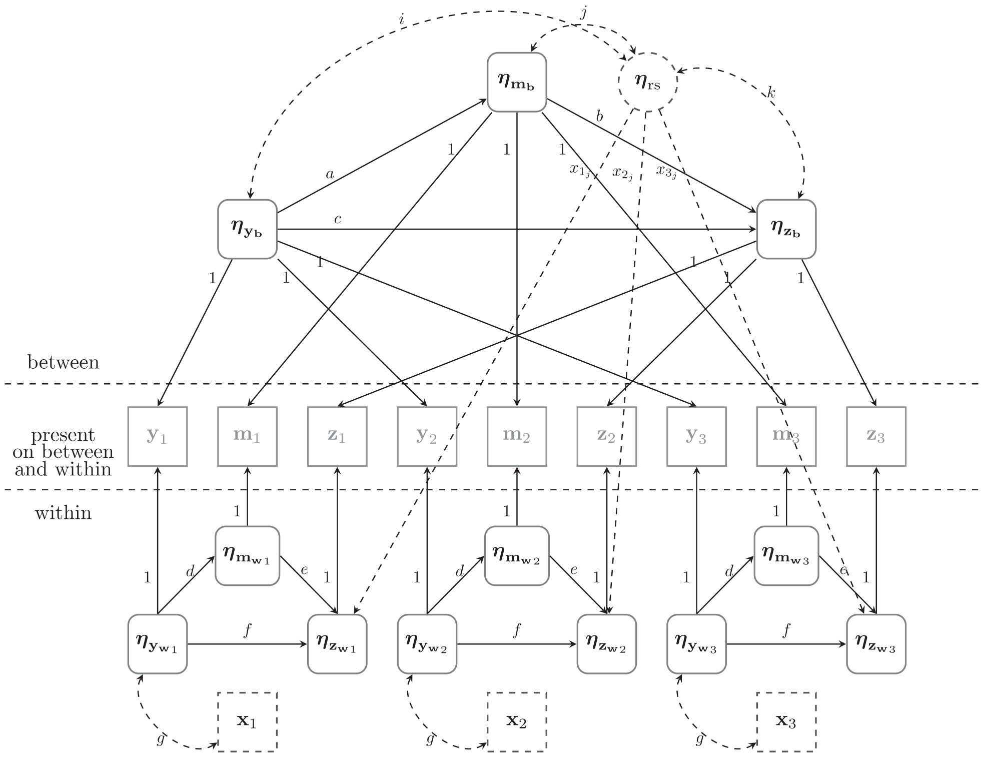


FIGURE 2 A multilevel mediation model in the WF approach with three units in each cluster. *Note*: Mediation model three variables (***y***,***m***, and ***z***} with three units per cluster). The rounded boxes represent the within and between components of the original variables for three units per cluster. Labels *a*, *b*, *c* reflect the between-level regressions and *d*, *e*, *f* reflect the within-level regressions. The dashed lines indicate all parameters related to the random slope $\boldsymbol{\eta}_{\mathrm{rs}}$ with the dependent variable as indicators and factor loadings that are fixed to the values of the covariate x_j. Residual variances (not shown here) also contain equality restrictions at the within-level. Appendix A shows the corresponding lavaan syntax for the random intercept model and Appendix B shows the additional lavaan syntax to add a random slope. Appendix E shows the alternative lavaan syntax with discrete variables.

we can also extend multilevel random intercept models with one or more random slope(s), by adding an additional step:

6. For each regression coefficient at the within-level that we allow to vary across clusters, we construct a latent variable ($\boldsymbol{\eta}_{\mathrm{rs}}$) at the between-level with the dependent variable as indicators and factor loadings that are fixed to the values of the covariate. The latent variable $\boldsymbol{\eta}_{\mathrm{rs}}$ represents the random slope in the model (see the dashed lines in Figures 2 and 3, and Appendices B and D for additional syntax to include a random slope).

Multilevel discrete data with the PML estimation method in the WF approach

With multilevel data, the pairwise likelihood is obtained as the product of bivariate likelihoods for within-cluster pairs of units (see Renard et al., 2004). Equations 6–9 are applicable for multilevel data by substituting individual *i* for unit *j*, where each separate row represents a single cluster instead of a single unit (see Figure 1). If the PML estimation method is used with multilevel discrete data, only the within-level needs to be adjusted. Instead of estimating error variances at the within-level, we estimate thresholds with equality constraints. Renard et al. (2004), Bellio and Varin (2005), Tibaldi et al. (2007), and Cho and Rabe-Hesketh (2011) showed how PML estimation can be used for

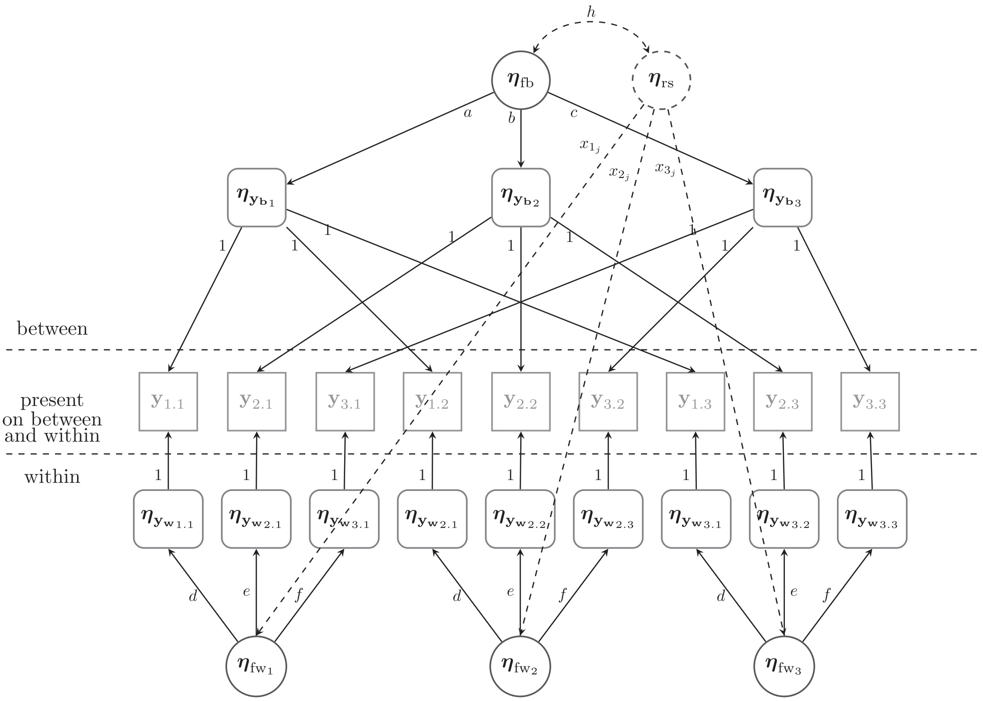


FIGURE 3 A multilevel factor model in the WF approach with three items and three units within each cluster. *Note*: The rounded boxes represent the within- and between- components of the original variables $\mathbf{y}_{1.1}$ to $\mathbf{y}_{3.3}$ for three units per cluster. $\boldsymbol{\eta}_{\mathrm{y_{b_1}}}$, $\boldsymbol{\eta}_{\mathrm{y_{b_2}}}$, and $\boldsymbol{\eta}_{\mathrm{y_{b_3}}}$ correspond to the unit-specific observation of that variable at the between-level, where labels *a*, *b*, *c* reflect the between-level factor loadings. For identification purposes, one factor loading or the variance of $\boldsymbol{\eta}_{\mathrm{fb}}$ must be fixed to unity. The structures below $\mathbf{y}_{1.1}$ to $\mathbf{y}_{3.3}$ account for the within-level latent variables (i.e., $\boldsymbol{\eta}_{\mathrm{fw_1}}$, $\boldsymbol{\eta}_{\mathrm{fw_2}}$, and $\boldsymbol{\eta}_{\mathrm{fw_3}}$) with parameter labels *d*, *e*, *f* for the factor loadings. For identification purposes, one factor loading of each within factor ($\boldsymbol{\eta}_{\mathrm{fw}}$) or the variance of $\boldsymbol{\eta}_{\mathrm{fw}}$ must be fixed to unity. The dashed lines indicate a random slope $\boldsymbol{\eta}_{\mathrm{rs}}$ with the dependent variable as indicators and factor loadings that are fixed to the values of the covariate x_{j}. The parameter *h* is an indicator of the correlation between the between-level factor and the random slope. Identical parameter labels indicate equality constraints in the model. Residual variances (not shown here) also contain equality restrictions. Appendix C shows the corresponding lavaan syntax for the random intercept model, and Appendix D shows the additional lavaan syntax to add a random slope. Appendix F shows the alternative lavaan syntax with discrete variables.

generalized (random intercept, random slope) multilevel regression models with a binary outcome. For fitting multilevel SEM with discrete data, we need to adjust our formulated step 2:

2a. For each discrete endogenous variable at the within-level, construct a latent variable (i.e., $\boldsymbol{y*}$), where the indicators correspond to the unit-specific observation of that variable. The thresholds are estimated with equality constraints on thresholds across units within a cluster (see step 4).

Figure 4 shows a generalized multilevel regression model with both a random intercept and a random slope and four response options. This model is similar to the one presented in Figure 1b, but now contains four response options instead of continuous data. Appendix E shows the lavaan syntax for a multilevel mediation model with discrete data and Appendix F shows the lavaan syntax for a multilevel factor model with discrete data. The syntax in Appendix B or D can be added to include a random slope.

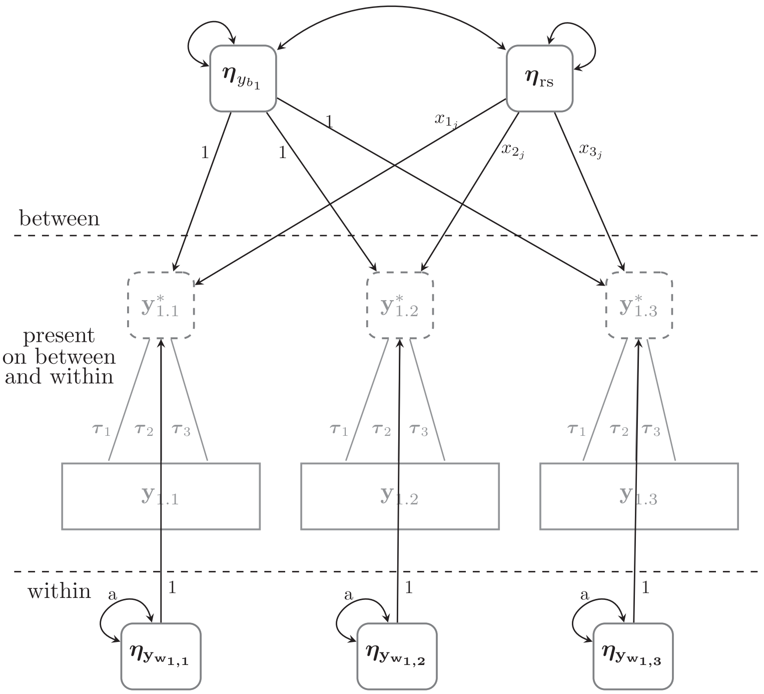

$\eta_{y_{b_1}}$
η_{rs}
1
1
1
x_{1_j}
x_{2_j}
x_{3_j}
between
$\mathbf{y}^*_{1.1}$
$\mathbf{y}^*_{1.2}$
$\mathbf{y}^*_{1.3}$
present on between and within
τ_1 τ_2 τ_3
τ_1 τ_2 τ_3
τ_1 τ_2 τ_3
$\mathbf{y}_{1.1}$
$\mathbf{y}_{1.2}$
$\mathbf{y}_{1.3}$
within
a
1
a
1
a
1
$\eta_{\mathbf{y}_{\mathbf{w}_{1,1}}}$
$\eta_{\mathbf{y}_{\mathbf{w}_{1,2}}}$
$\eta_{\mathbf{y}_{\mathbf{w}_{1,3}}}$

FIGURE 4 Random intercept and random slope model with three units in each cluster. *Note*: With four response options, three thresholds are estimated per variable. ($\boldsymbol{\eta}_{\mathbf{y}_{\mathrm{b}_1}}$) refers to a random intercept, ($\boldsymbol{\eta}_{\mathbf{y}_{\mathrm{w}_1}}$) refers to the within representation of the variables, $\boldsymbol{\tau}$ refers to the thresholds, and $\mathbf{y}*$ refers to the underlying latent response variables. Identical parameter labels indicate equality constraints. Figure 4 is similar to Figure 1 with the new parameterization and discrete data.

SIMULATION STUDY

To investigate computationally intensive multilevel SEM models with discrete data in the WF approach using the PML estimation method, we ran a simulation study with a multilevel mediation model and two multilevel factor models (i.e., one with one random slope and one with two random slopes). All the models have at least six latent variables (i.e., random slopes or factors) in the model. Within each type of model, we vary response scales (two-point, four-point) and the number of clusters (250, 500, 1000, 2000). The cluster size is always fixed at three. In a fully crossed design, these factors yield $3 \times 2 \times 4 = 24$ different conditions. The performance of the PML method is evaluated by calculating the relative bias. The relative bias is calculated for the estimated parameters as % bias = $(\hat{\theta}-\theta)/\theta \times 100$ and for the standard errors as % bias = $(SE\text{-}SD)/SD \times 100$, where SE is the mean of the estimated standard errors across replications and the SD refers to the standard deviation of the parameter estimates across replications.

Data generation

Figures 5 and 6 show the data-generation model in the LF approach for a mediation model and a factor model, respectively. The mediation model is generated with three random slopes and the factor model is generated with one or two random slopes. The dashed lines represent the random slopes (i.e., $\boldsymbol{\eta}_{\mathrm{rs}_1}$, $\boldsymbol{\eta}_{\mathrm{rs}_2}$, $\boldsymbol{\eta}_{\mathrm{rs}_3}$). The solid lines of models in the LF approach are identical to those in the WF approach presented in Figures 2 and 3. In all models we generated more variance at the within-level than at the between-level to mimic real data examples (see Snijders & Bosker, 1999). For example, using the parameter values of the basic factor model from Figure 6, the intraclass correlation for all indicators equals

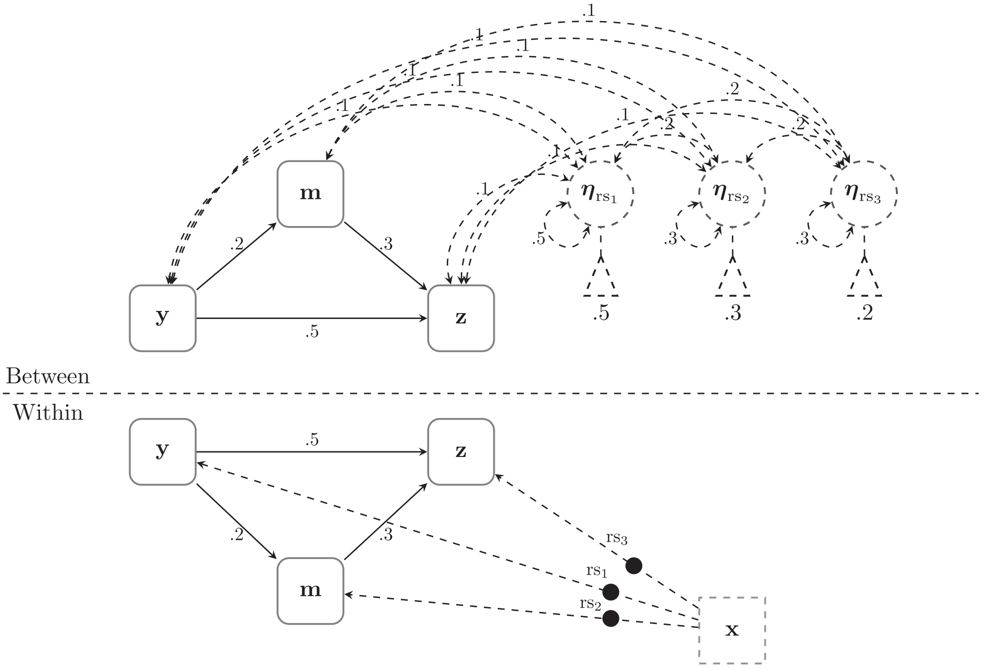


FIGURE 5 A data generation model for a multilevel mediation model in the LF approach. *Note*: The solid lines represent the general mediation model. The rounded boxes represent the within- and between components of the original variables for three units per cluster. Variables $\boldsymbol{y}$, $\boldsymbol{m}$, and $\boldsymbol{z}$ are variables in the mediation model that appear at both the within and the between-level. The dashed lines indicate all parameters related to the random slopes ($\boldsymbol{\eta}_{\mathrm{rs}_1}$, $\boldsymbol{\eta}_{\mathrm{rs}_2}$, $\boldsymbol{\eta}_{\mathrm{rs}_3}$) of covariate $\boldsymbol{x}$.

$(1^2 \times .5 + .2)/((1^2 \times .5 + .2) + (1^2 \times 1 + 1)) = .260$. The multilevel models are generated with three variables with three units in each cluster to limit the estimation time in the simulation study. Continuous data were drawn from a multivariate normal distribution. Discrete data with two-point response scales are obtained by discretizing the continuous data such that the population proportions equal approximately .50 per category and four-point response scales are obtained by discretiszing the data such that the population proportions equal approximately .16, .34, .34, and .16. We have chosen the mean of the random slope (see the triangles in Figures 5 and 6) to be equal to the variance of the random slope. The residual variances equal an identity matrix to resemble the theta parameterization. For each condition 500 data sets are generated.

Estimation

The R package lavaan (version 0.6–12: Rosseel, 2012) is used for the calculations. Casewise estimation is necessary to estimate multilevel models with random slopes. To perform casewise calculations in this study, we wrote custom scripts to estimate the likelihood for each cluster separately.[1] Based on the theory of Katsikatsou et al. (2012), we used a closed-form solution to calculate the PML standard errors casewise (i.e., cluster-specific). In the estimated factor models, we fixed the first factor loading to unity to set the metric of the factor.

[1] All R scripts (e.g., data-generation scripts and scripts to analyse the data) are available at https://osf.io/346vj/.

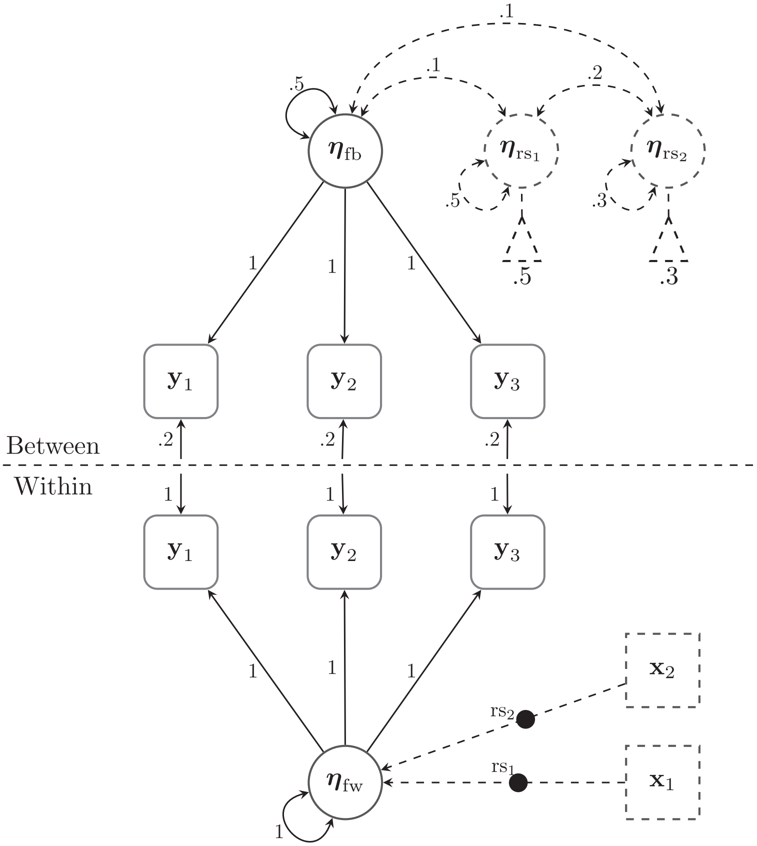


FIGURE 6 A data generation model for a multilevel factor model in the LF approach. *Note*: The solid lines represent the general model and the dashed lines indicate the random slopes. The rounded boxes represent the within- and between-components of the original variables for three units per cluster. $\boldsymbol{\eta}_{\text{fw}}$ denotes the variance of the within-level variance and $\boldsymbol{\eta}_{\text{fb}}$ denotes the variance of the between-level variance. The dashed lines indicate all parameters related to the random slopes ($\boldsymbol{\eta}_{\text{rs}_1}$ and $\boldsymbol{\eta}_{\text{rs}_2}$) of the covariates *x*.

Results

After applying each of the 12000 data sets, we found that all models converged. Inspection of the results shows that especially models with 250 clusters and two-point scales show unreasonable parameter estimates and standard errors. We removed all results that included at least one parameter or standard error at more than four standard deviations from the mean[2]. This resulted in 3.90% deletions for the mediation model with three random slopes, 2.78% deletions for the factor model with one random slope, and 4.38% deletions for the factor model with two random slopes. Below we describe the relative bias (expressed as percentages) of the estimated mediation model and the factor models for the within- and between-level separately. Notice that the coefficients displayed in the plots refer to Equations 3 and 4.

[2] All raw results with and without outliers are available at https://osf.io/346vj/.

Mediation model

Within-level

Bias of the parameter estimates

Figure 7 shows the relative bias of the three regression coefficients in the mediation model at the within-level. The relative bias is lower across all conditions with a larger number of clusters and a four-point response scale. In particular, the condition with 250 clusters with a two-point response scale influenced the overall results. Raw results show that all the parameter estimates across all conditions with a four-point scale have less than 1.3% bias and that all conditions with a two-point scale and 1000 or 2000 clusters have less than 1.5% bias.

Bias standard errors

To study the efficiency of the estimates, we calculated the relative bias of the standard errors. The relative bias of the standard errors associated with the within-level regression coefficients is shown in Figure 8. High percentages of relative bias were found in conditions with 250 clusters and a two-point response scale (about 70% to 275%). To put this in context, an absolute bias between the standard errors and the standard deviations of the parameter estimates that equals .09 resulted in about 100% bias. Conditions with a four-point response scale with at least 500 clusters and conditions with a two-point response scale with at least 1000 clusters were more efficient, with less than 6% bias and less than 8% bias, respectively.

Between-level

Bias of the parameter estimates

The relative bias of the regression coefficients across all conditions in the mediation model is shown in Figure 9. Almost all conditions had very low percentages of relative bias. We only observed a little more

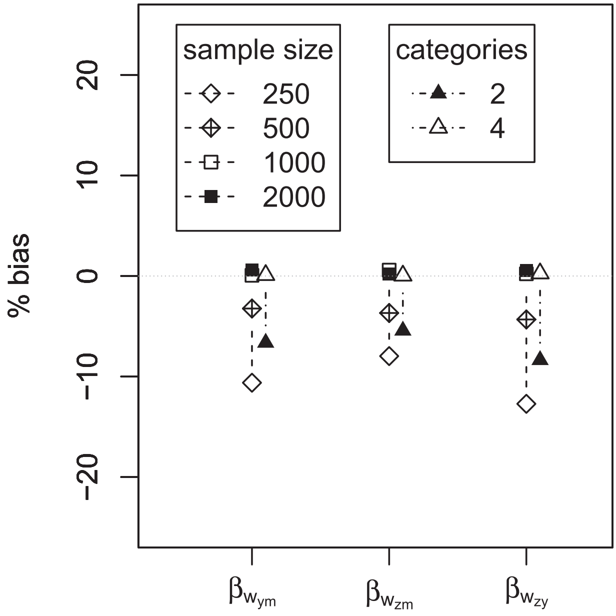


FIGURE 7 Relative bias of the three regression coefficients of the mediation model at the within-level

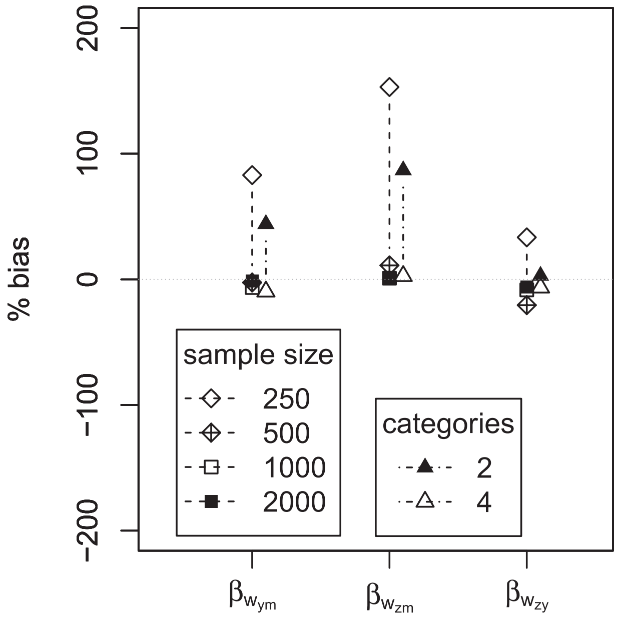


FIGURE 8 Relative bias of the standard error related to the three regression coefficients of the mediation model at the within-level

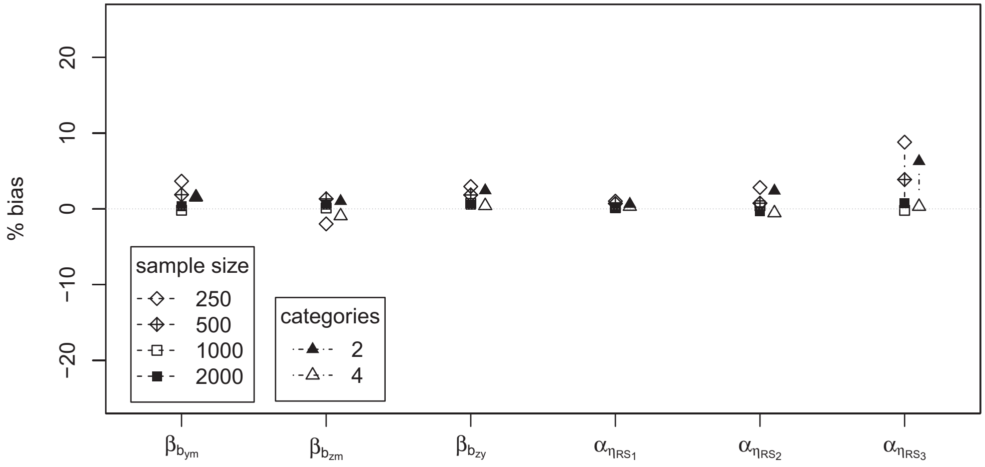


FIGURE 9 Relative bias of the regression coefficients of the mediation model at the between-level

bias in the mean of $\boldsymbol{\eta}_{rs_3}$. Figure 10 shows the relative bias of the parameters related to the variances and covariances in the mediation model. For the sake of efficiency, we only show the variance of $\boldsymbol{z}$ at the between-level (i.e., $\boldsymbol{\eta}_z$). Results show that the random slopes have more bias than the regression coefficients. Figure 10 shows low percentages of bias across all conditions in the estimated covariances and higher percentages of bias in the variances. Overall, we observe similar patterns at the between-level to those we observed at the within-level, and more accuracy across all conditions with more clusters and a four-point response scale.

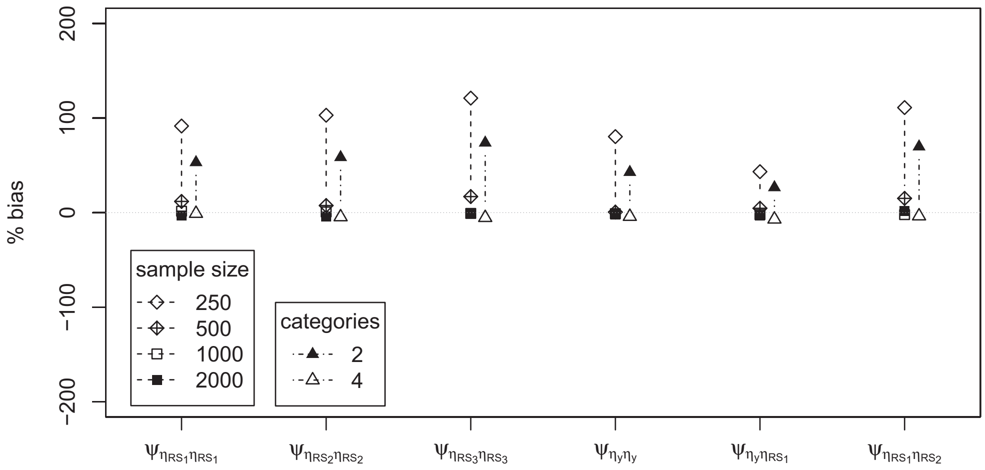


FIGURE 10 Relative bias of the (co)variances of the mediation model at the between-level

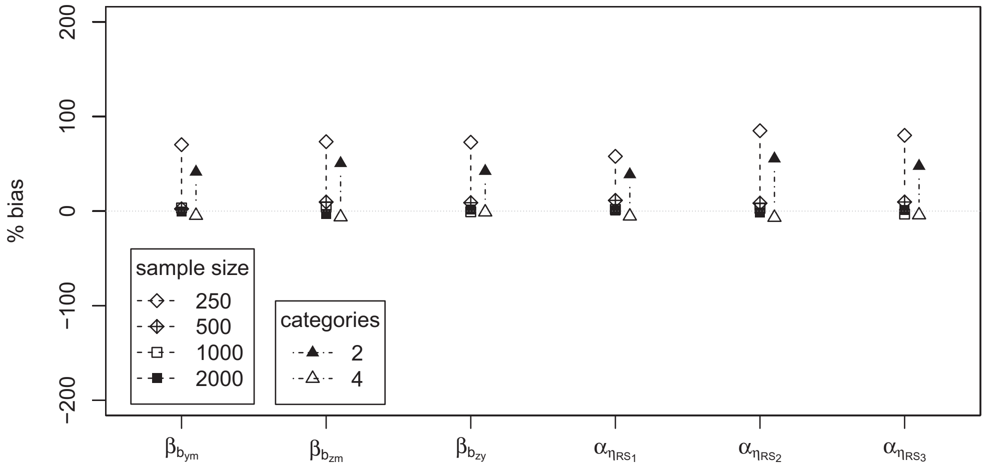


FIGURE 11 Relative bias of the standard error related to the regression coefficients of the mediation model at the between-level

Bias standard errors

Figures 11 and 12 show the relative bias of the standard errors for all parameters at the between-level. Once again, high percentages of bias were found in conditions with fewer clusters and two-point response scales. The bias related to the variances is higher than the bias related to the regression coefficients. We identified similar patterns of high and low percentages of relative bias to those we found at the within-level. In particular, a sample size of 250 with a two-point response scale showed very high percentages of relative bias across all parameters (about 65% to 230%). All other conditions show reasonable percentages of relative bias.

Factor models with random slopes

As the result tendencies of the factor model with one random slope are very similar to the results with two random slopes, we will only show the results of the factor model with two random slopes.

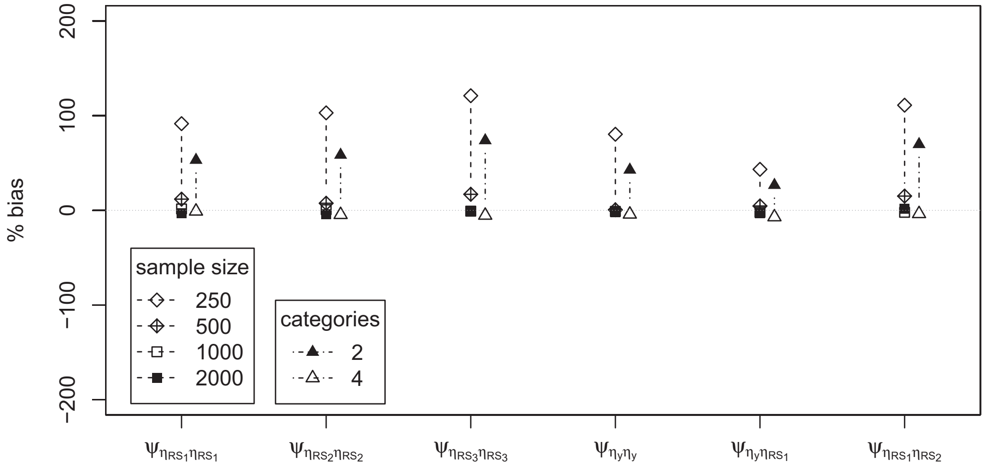


FIGURE 12 Relative bias of the standard error related to the (co)variances of the mediation model at the between-level

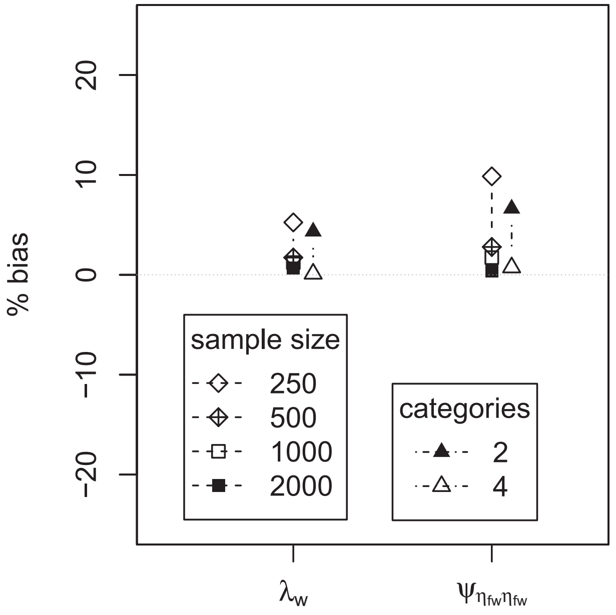


FIGURE 13 Relative bias of one of the factor loadings and the variance of the factor at the within-level

In general, the bias in the factor model with one random slope is a little lower across both parameter estimates and standard errors. The results of the factor model with one random slope are shown in the online supplementary materials.

Within-level

Bias of the parameter estimates

Figure 13 shows the relative bias of one of the factor loadings (we chose the third factor loading) and the variance of the within-level factor in a factor model with two random slopes. The factor variance ($\boldsymbol{\eta}_{\mathrm{fw}}$) shows higher percentages of bias than the factor loading (λ_{w}). Overall, the percentages of relative bias were quite low. Raw results show that the condition with 250 clusters with a two-point response scale showed a maximum of 17% relative bias. All other conditions have a maximum of 5.5% relative bias.

Bias standard errors

The relative bias of the within-level parameter estimates is shown in Figure 14. Notice that we used a different scale on the *y*-axis as the relative percentages of bias were much lower in the factor models than those observed in the mediation model. The relative bias found across all conditions with a four-point response scale was <5%. In conditions with a two-point response scale a maximum of 15% relative bias was observed.

Between-level

Bias of the parameter estimates

The relative bias of the parameter estimates is shown in Figures 15 and 16. The variances and covariances show higher percentages of bias than the other parameters. In general, we observe a similar trend with more bias in conditions with fewer clusters and fewer response scales. Except in conditions with a cluster size of 250 and a two-point response scale, we observe a maximum of 12% bias.

Bias standard errors

Figures 17 and 18 show the relative bias of the standard errors related to the parameter estimates at the between-level. The factor loading and the variance of the between-level show a pattern in which more relative bias is related to a smaller number of clusters. The bias of the standard errors related to the other parameter estimates is lower. Taking into account that the percentages of bias in the standard errors increases quite quickly, the percentages of relative bias were quite low. In addition, we observe here that calculating bias across all conditions slightly increases the bias of the standard errors. For example, the raw results with 1000 and 2000 clusters show no more than 10% relative bias.

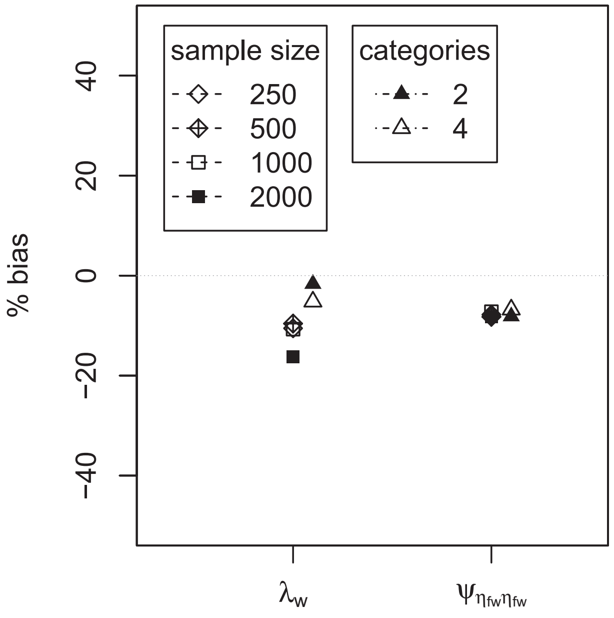


FIGURE 14 Relative bias of the standard errors related to one of the factor loadings and the variance of the factor model at the within-level

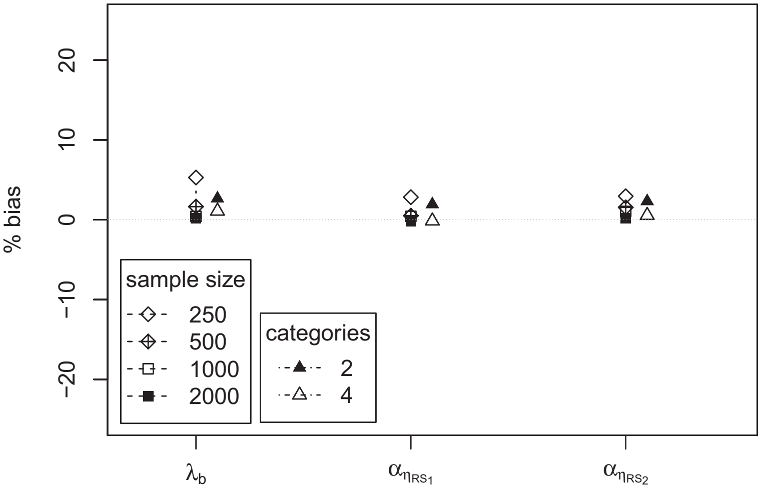


FIGURE 15 Relative bias related to the factor loading and the regression coefficients of the covariate of the factor model at the between-level

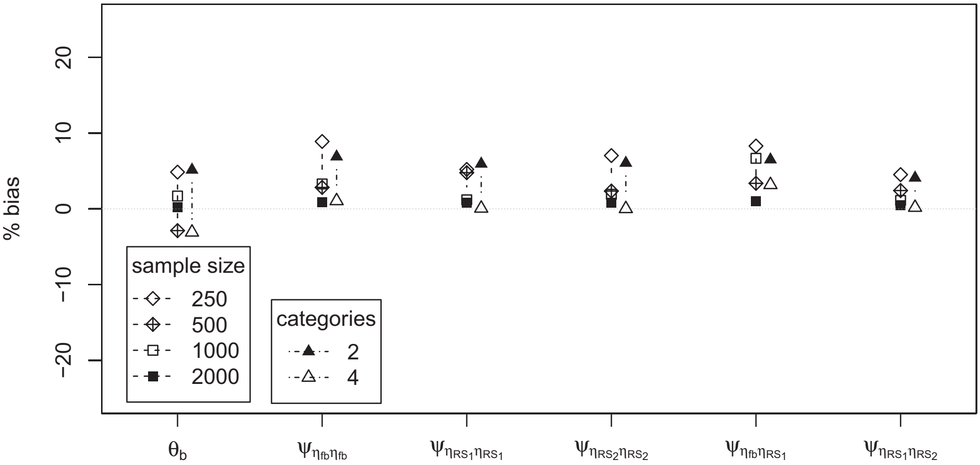


FIGURE 16 Relative bias related to the (co)variances of the factor model at the between-level

DISCUSSION

In this study we combined the challenges of multilevel- and discrete-data SEM using the PML estimation method. In pursuing this, we first reformulated the general framework of multilevel SEM in the WF approach of Barendse and Rosseel (2020) into a parameterization that is more intuitive and avoids using definition variables in case covariates or regression coefficients exist at both the within- and between-level. In this approach, the between- and within-levels are disentangled via latent variables that are then used to specify models at each level separately. Then we extended the reformulated WF approach for multilevel models with random slopes. In this study, we specifically focused on estimating structural equation models with discrete data and many latent variables (six or more, including random slopes) as these are computationally too intensive to estimate with the multilevel marginal maximum likelihood estimation method. The least squares estimation methods cannot be used either because they do not allow for random slopes due to relying on summary statistics. The PML estimation method, on

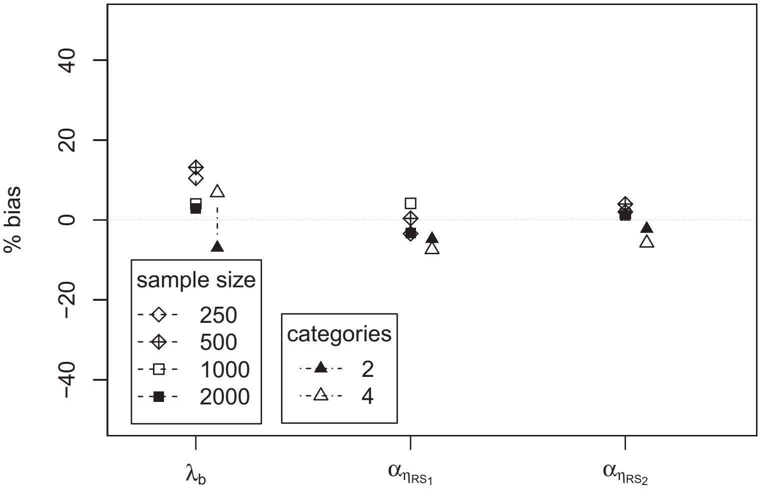


FIGURE 17 Relative bias of the standard error related to the factor loading and the regression coefficients of the covariate of the factor model at the between-level

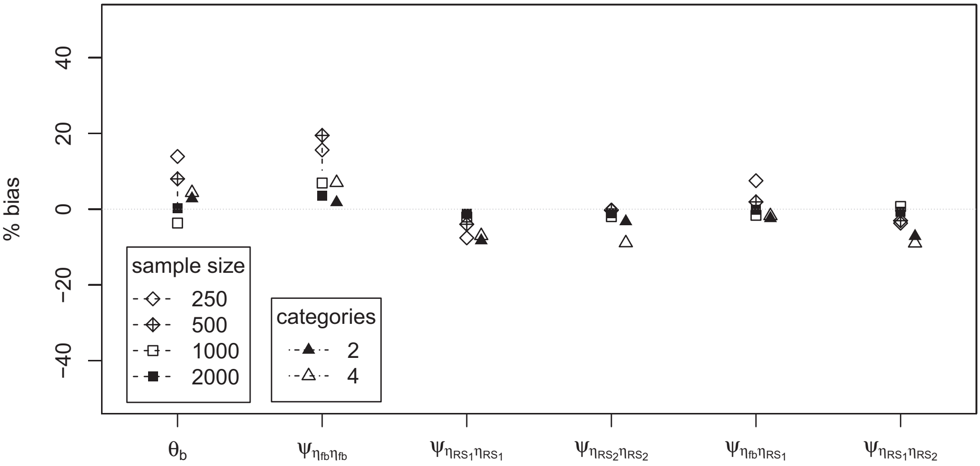


FIGURE 18 Relative bias of the standard error related to the (co)variances of the factor model at the between-level

the other hand, is well suited for these kinds of models, as PML allows for random slopes and calculates the products of bivariate likelihoods that are computationally easy to handle.

We conducted a small simulation study to examine whether the PML estimation method can deal with complex multilevel data including random slopes. Results across all conditions show that the PML estimation method can deal with three-variable mediation models with three random slopes and three-variable factor models with one or two random slopes. The mediation model showed higher percentages of relative bias than the factor models. This may have to do with the number of random slopes. The simulation study also showed that a factor model with one random slope can easily be extended to a model with two random slopes, with only slightly less accurate and efficient results. In general, a larger number of clusters and four-point response scales leads to low percentages of relative bias. The condition with 250 clusters and a two-point response scale showed higher percentages of relative bias. More specifically, the most prominent bias of the standard errors is observed in the mediation model. Overall, we observed more bias at the within-level than at the between-level. Additional research to estimate the same multilevel model with scripts that do not make use of equality constraints at the within-level revealed that the equality constraints did not cause the additional bias. The bias at the within-level may be improved by increasing the number of units in each cluster. Generally speaking, we conclude that the PML estimation method is quite accurate and efficient when there are enough data to estimate the parameters. Only the condition with 250 clusters and a two-point response scale lacked sufficient accuracy and efficiency.

Comparing our results to those obtained in Barendse and Rosseel (2020), we conclude that models with random slopes need a larger number of cluster to be accurate and efficient. The random slopes bring an extra level of variability to the model. More research is needed to investigate whether this can be improved by increasing the number of units in the cluster or using techniques such as bounded estimation (De Jonckere & Rosseel, 2022). In this study we used a different parameterization than in the study of Barendse and Rosseel (2020). This parameterization does not influence the results in most of the models. However, in models with regression coefficients that exist at both the within- and between-level, the newly described parameterization is highly recommended as it causes fewer difficulties in model estimation (i.e., it avoids calculation the between-level with the within-level using definition variables). Compared to the previous approach, we estimate more latent variables in the newly described approach. However, the PML estimation is not affected by this as it can handle many latent variables.

Adding random slopes complicates the estimation of a multilevel structural equation model, but it is still a sophisticated method to deal with within-level covariates. We recommend that each within-level covariate is first investigated via a random slope. Ignoring a random slope creates a misspecification in the model. This misspecification can result in incorrect standard errors and incorrect fit statistics. In principle, there are different ways of incorporating covariates that can be combined when fitting multilevel models. In case a random slope does not contain enough variance, one can continue with a fixed covariate. If one is not interested in the effect of the covariate, it is also possible to treat the variable as an exogenous covariate and regress out the covariates first and perform all other calculations on the residual correlations (Katsikatsou, 2017). When the covariate is a dichotomy, it is also possible to investigate the effect of the covariate in a multigroup analysis. Depending on the model, one can chose and combine different ways to incorporate covariates.

Even though PML estimation can estimate complex multilevel models in the WF approach, there are a number of limitations. As the PML with multilevel data is obtained as the product of bivariate likelihoods for within-cluster pairs of units and variables, the PML in the WF approach is slower with large cluster sizes. Further research should investigate whether we can improve the estimation by, for example, deleting the bivariate pairs that do not contribute much to the likelihood or applying a two-step approach where the thresholds are fixed in the second step to reduce the number of estimated parameters.

Notwithstanding the difficulties of the PML estimation method in the WF approach, the PML estimation method seems promising and suited to fit complex multilevel structural equation models in the WF approach. The suggested two-level models with PML estimation methods can in theory also be extended to more than two levels. In addition, PML is computationally efficient enough to deal with many latent variables at both the within- and between-level. That does not mean that estimation is fast. Indeed, it currently takes a couple of hours to estimate a model. Another advantage of the PML estimation method with multilevel data is that the number of columns (variables × units) can be larger than the number of rows (number of clusters). In this simulation study we used three units in each cluster to limit the estimation time, but models with larger cluster sizes can be estimated. Finally, the PML estimation method can deal with all types of data – discrete, continuous, and combinations thereof.

In this paper we only focused on frequentist estimation methods and did not take into account the multilevel Bayesian estimation method (e.g., Fox, 2010), which is a full-information method that is able to estimate random slopes. For further research, it would be interesting to compare the PML estimation method to the multilevel Bayesian estimation method, which can also estimate complex multilevel models.

CONFLICT OF INTEREST

All authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data has been shared (following the link https://osf.io/346vj/).

REFERENCES

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. Marcoulides & R. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Erlbaum.

Asparouhov, T., & Muthén, B. (2007). Computationally efficient estimation of multilevel high-dimensional latent variable models. *Proceedings of the 2007 JSM meeting in Salt Lake City, Utah, Section on Statistics in Epidemiology*, pp. 2531–2535.

Barendse, M., & Rosseel, Y. (2020). Multilevel modeling in the 'wide format'approach with discrete data: A solution for small cluster sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(5), 1–26.

Bauer, D. J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics*, *28*(2), 135–167.

Bellio, R., & Varin, C. (2005). A pairwise likelihood approach to generalized linear models with crossed random effects. *Statistical Modelling*, *5*(3), 217–227.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an em algorithm. *Psychometrika*, *46*(4), 443–459.

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

Browne, M. W. (1984). Asymptotically distribution–free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62–83.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, *101*(1), 147–158.

Cho, S.-J., & Rabe-Hesketh, S. (2011). Alternating imputation posterior estimation of models with crossed random effects. *Computational Statistics & Data Analysis*, *55*(1), 12–25.

Chou, C.-P., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(3), 247–266.

Curran, P. J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research*, *38*(4), 529–569.

De Jonckere, J., & Rosseel, Y. (2022). Using bounded estimation to avoid nonconvergence in small sample structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *29*(3), 412–427.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. Springer Science & Business Media.

Guo, B., Perron, B. E., & Gillespie, D. F. (2008). A systematic review of structural equation modelling in social work research. *British Journal of Social Work*, *39*(8), 1556–1574.

Hedeker, D., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research*, *36*(3), 347–387.

Katsikatsou, M. (2017). *The pairwise likelihood method for structural equation modelling with ordinal variables and data with missing values using the r package lavaan*. Retrieved February 06, 2021, from. http:users.ugent.bey~rosseellavaanpmlPL_Tutorial.pdf.

Katsikatsou, M., & Moustaki, I. (2017). Pairwise likelihood estimation for confirmatory factor analysis models with ordinal variables and data that are missing at random. *Submitted to Journal of Royal Statistical Society Series C*, *75*, 23–45

Katsikatsou, M., Moustaki, I., Yang-Wallentin, F., & Jöreskog, K. G. (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics & Data Analysis*, *56*(12), 4243–4258.

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. Guilford publications.

Koomen, H., Verschueren, K., & Pianta, R. C. (2007). *Leerling Leerkracht Relatie Vragenlijst-Handleiding g. [Student-Teacher Relationship Scale: Manual]*. Bohn Stafleu van Loghum.

Lau, Y., Htun, T. P., Lim, P. I., Ho-Lim, S., & Klainin-Yobas, P. (2015). Maternal, infant characteristics, breastfeeding techniques, and initiation: Structural equation modeling approaches. *PLoS One*, *10*(11), e0142861.

Lee, S., Poon, W., & Bentler, P. M. (1990). Full maximum likelihood analysis of structural equation models with polytomous variables. *Statistics & Probability Letters*, *9*(1), 91–97.

Lee, S., Poon, W., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, *57*(1), 89–105.

Lindsay, B. G. (1988). Composite likelihood methods. *Contemporary Mathematics*, *80*(1), 221–239.

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.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*(1), 201–226.

MacCallum, R. C., Kim, C., Malarkey, W. B., & Kiecolt-Glaser, J. K. (1997). Studying multivariate change using multilevel models and latent curve models. *Multivariate Behavioral Research*, *32*(3), 215–253.

Mahlke, J., Schultze, M., Koch, T., Eid, M., Eckert, R., & Brodbeck, F. C. (2016). A multilevel cfa–mtmm approach for multisource feedback instruments: Presentation and application of a new statistical model. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(1), 91–110.

McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, *58*, 110–133.

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), 259–283.

Meredith, W., & Tisak, J. (1990). Latent curve analysis. *Psychometrika*, *55*(1), 107–122.

Moorman, S. M. (2016). Dyadic perspectives on marital quality and loneliness in later life. *Journal of Social and Personal Relationships*, *33*(5), 600–618.

Muthén, B., Du Toit, S. H., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Psychometrika*, *75*, 1–45.

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

Muthén, B. O. (1990). Mean and covariance structure analysis of hierarchical data. In *Department of Statistics (UCLA statistics series)* (Vol. *62*, pp. 1–64). University of California.

NLSAH. (2005). National longitudinal study of adolescent health. Retrieved October 30, 2006, from. http://www.cpc.unc.edu/addhealth

Renard, D., Molenberghs, G., & Geys, H. (2004). A pairwise likelihood approach to estimation in multilevel probit models. *Computational Statistics & Data Analysis*, *44*(4), 649–667.

Rockwood, N. J. (2020). Maximum likelihood estimation of multilevel structural equation models with random slopes for latent covariates. *Psychometrika*, *85*(2), 275–300.

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

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

Snijders, T., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Tibaldi, F. S., Verbeke, G., Molenberghs, G., Renard, D., Noortgate, W., & Boeck, P. (2007). Conditional mixed models with crossed random effects. *British Journal of Mathematical and Statistical Psychology*, *60*(2), 351–365.

Varin, C. (2008). On composite marginal likelihoods. *Advances in Statistical Analysis*, *92*(1), 1–28.

Varin, C., Reid, N. M., & Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica*, *21*(1), 5–42.

Xiong, B., Skitmore, M., & Xia, B. (2015). A critical review of structural equation modeling applications in construction research. *Automation in Construction*, *49*, 59–70.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Barendse, M. T., & Rosseel, Y. (2023). Multilevel SEM with random slopes in discrete data using the pairwise maximum likelihood. *British Journal of Mathematical and Statistical Psychology*, *76*, 327–352. https://doi.org/10.1111/bmsp.12294

APPENDIX A | LAVAAN SYNTAX FOR A MULTILEVEL MEDIATION MODEL IN THE WF APPROACH

```
# Syntax corresponding to Figure 2 with three observations in each cluster
model <- '

    ### Step 2: Disentangle each continuous endogenous variable into a
    ### between-part and a within-part
    # create a separate within-level
    wy.1 =~1*y.1; wy.2 =~1*y.2; wy.3 =~1*y.3 # $\eta_{y_w}$ in Figure 2
    wm.1 =~1*m.1; wm.2 =~1*m.2; wm.3 =~1*m.3 # $\eta_{m_w}$ in Figure 2
    wz.1 =~1*z.1; wz.2 =~1*z.2; wz.3 =~1*z.3 # $\eta_{z_w}$ in Figure 2
    # create a separate between-level
    by =~1*y.1 + 1*y.2 + 1*y.3 # $\eta_{y_b}$ in Figure 2
    bm =~1*m.1 + 1*m.2 + 1*m.3 # $\eta_{m_b}$ in Figure 2
    bz =~1*z.1 + 1*z.2 + 1*z.3 # $\eta_{z_b}$ in Figure 2
    # zero residual variances
    y.1 ~~0*y.1; y.2 ~~0*y.2; y.3 ~~0*y.3
    m.1 ~~0*m.1; m.2 ~~0*m.2; m.3 ~~0*m.3
    z.1 ~~0*z.1; z.2 ~~0*z.2; z.3 ~~0*z.3

    ### Steps 3 and 4: Construct a model with all the newly introduced
    ### within-level variables with equality constraints
    # variances in the model
    wy.1 ~~ yvar*wy.1; wy.2 ~~ yvar*wy.2; wy.3 ~~ yvar*wy.3 # Var($\eta_{y_w}$) in Figure 2
    wm.1 ~~ mvar*wm.1; wm.2 ~~ mvar*wm.2; wm.3 ~~ mvar*wm.3 # Var($\eta_{m_w}$) in Figure 2
    wz.1 ~~ zvar*wz.1; wz.2 ~~ zvar*wz.2; wz.3 ~~ zvar*wz.3 # Var($\eta_{z_w}$) in Figure 2
    # within regressions
    wz.1 ~ zy.w*wy.1 + zm.w*wm.1 # $e$ and $f$ in Figure 2
    wz.2 ~ zy.w*wy.2 + zm.w*wm.2 # $e$ and $f$ in Figure 2
    wz.3 ~ zy.w*wy.3 + zm.w*wm.3 # $e$ and $f$ in Figure 2
    wm.1 ~ my.w*wy.1 # $d$ in Figure 2
    wm.2 ~ my.w*wy.2 # $d$ in Figure 2
    wm.3 ~ my.w*wy.3 # $d$ in Figure 2

    ### Step 5: Construct a model with all the newly introduced
    ### between-level variables
    # regressions at between-level
    bz ~ zy.b*by + zm.b*bm # $b$ and $c$ in Figure 2
    bm ~ my.b*by # $a$ in Figure 2
    # variances and means at between-level
    by ~~ by; bm ~~ bm; bz ~~ bz # Var($\eta_{y_b}$), Var($\eta_{m_b}$), and Var($\eta_{z_b}$) in Figure 2
    by ~1; bm ~1; bz ~1# E($\eta_{y_b}$), E($\eta_{m_b}$), and E($\eta_{z_b}$) in Figure 2
'
# fitting the model with ML
fit <- lavaan(model, data = wideData)
summary(fit)
```

APPENDIX B | ADDITIONAL LAVAAN SYNTAX TO ADD A COVARIATE WITH RANDOM SLOPES IN THE MEDIATION MODEL

```
    ### Step 6: Add a random slope for variable z
    # variance of the covariate $x$ at the within-level
    x.1 ~~ xvar*x.1; x.2 ~~ xvar*x.2; x3 ~~ xvar*x.3 # Var($x$) in Figure 2
    # covariance $y$ and $x$ (exogenous!)
    wy.1 ~~ covyx*x.1; wy.2 ~~ covyx*x.2; wy.3 ~~ covyx*x.3 # $g$ in Figure 2
    # mean $x$
    x.1 ~mx*1; x.2 ~ mx*1; x.3 ~ mx*1 # E($x$)

    # covariate at the between-level
    # create a random slope for $z$ with covariate $x$
    RS =~999*wz.1 + 999*wz.2 + 999*wz.3 # 999 indicate the unit specific values of $x$;
      $\eta_{rs}$ with unit specific $x$ values in Figure 2
    RS ~~ RS # Var($\eta_{rs}$) in Figure 2
    RS ~1 # E($\eta_{rs}$) in Figure 2
    RS ~~ by + bm + bz # $i$, $j$, and $k$ in Figure 2
```

APPENDIX C | LAVAAN SYNTAX FOR A MULTILEVEL FACTOR MODEL IN THE WF APPROACH

```
# Syntax corresponding to Figure 3 with three observations in each cluster
model <- '

    ### Step 2: Disentangle each continuous endogenous variable into a
    ### between-part and a within-part
    # create a separate within-level
    wy1.1 =~1*y1.1; wy2.1 =~1*y2.1; wy3.1 =~1*y3.1 # $\eta_{y_{w_1}}$ in Figure 3
    wy1.2 =~1*y1.2; wy2.2 =~1*y2.2; wy3.2 =~1*y3.2 # $\eta_{y_{w_2}}$ in Figure 3
    wy1.3 =~1*y1.3; wy2.3 =~1*y2.3; wy3.3 =~1*y3.3 # $\eta_{y_{w_3}}$ in Figure 3
    # create a separate between-level
    by1 =~1*y1.1 + 1*y1.2 + 1*y1.3 # $\eta_{y_{b_1}}$ in Figure 3
    by2 =~1*y2.1 + 1*y2.2 + 1*y2.3 # $\eta_{y_{b_2}}$ in Figure 3
    by3 =~1*y3.1 + 1*y3.2 + 1*y3.3 # $\eta_{y_{b_3}}$ in Figure 3

    ### Steps 3 and 4: Construct a model with all the newly introduced
    ### within-level variables with equality constraints
    # variances in the model
    wy1.1 ~~ y1var*wy1.1; wy2.1 ~~ y1var*wy2.1; wy3.1 ~~ y1var*wy3.1 # Var($\eta_{y_{w_1}}$) in Figure 3
    wy1.2 ~~ y2var*wy1.2; wy2.2 ~~ y2var*wy2.2; wy3.2 ~~ y2var*wy3.2 # Var($\eta_{y_{w_2}}$) in Figure 3
    wy1.3 ~~ y3var*wy1.3; wy2.3 ~~ y3var*wy2.3; wy3.3 ~~ y3var*wy3.3 # Var($\eta_{y_{w_3}}$) in Figure 3
    # factor model
    fw1 =~1*wy1.1 + lw2*wy2.1 + lw3*wy3.1 # $\eta_{fw}$ with factor loadings d, e, and f in Figure 3
    fw2 =~1*wy1.2 + lw2*wy2.2 + lw3*wy3.2 # $\eta_{fw}$ with factor loadings d, e, and f in Figure 3
    fw3 =~1*wy1.3 + lw2*wy2.3 + lw3*wy3.3 # $\eta_{fw}$ with factor loadings d, e, and f in Figure 3
    fw1 ~~ NA*fw1 + fwt*fw1 # Var($\eta_{fw}$) in Figure 3
    fw2 ~~ NA*fw2 + fwt*fw2 # Var($\eta_{fw}$) in Figure 3
    fw3 ~~ NA*fw3 + fwt*fw3 # Var($\eta_{fw}$) in Figure 3

    ### Step 5: Construct a model with all the newly introduced
    ### between-level variables
    # factor model
    fb =~1*by1 + by2 + by3 $\eta_{fw}$ with factor loadings a, b, and c in Figure 3
    fb ~~ fb # Var($\eta_{fb}$) in Figure 3
    # intercepts
    by1 ~1; by2 ~1; by3 ~1 # E($\eta_{y_b}$) in Figure 3
    # variances
    by1 ~~ by1; by2 ~~ by2; by3 ~~ by3 # Var($\eta_{y_b}$) in Figure 3
'
# fitting the model with ML
fit <- lavaan(model, data = wideData)
summary(fit)
```

APPENDIX D | ADDITIONAL LAVAAN SYNTAX TO ADD A COVARIATE WITH A RANDOM SLOPE IN THE FACTOR MODEL

```
    # covariate at the between-level
    # create a random slope
    RS =~ (999)*fw1 + (999)*fw2 + (999)*fw3 # 999 indicate the unit-specific values of x;
      $\eta_{rs}$ with unit-specific x values in Figure 3
    # mean of the random slope
    RS ~1# E($\eta_{\mathrm{rs}}$) in Figure 3
    # variance of the random slope
    RS ~~ RS # Var($\eta_{\mathrm{rs}}$) in Figure 3
    # covariance between RS and factor at the between-level
    fb ~~ RS # h in Figure 3
```

APPENDIX E | LAVAAN SYNTAX FOR A MULTILEVEL MEDIATION MODEL WITH DISCRETE DATA IN THE WF APPROACH

```
# syntax corresponding to solid lines in mediation model of Figure 2 with
# three observations in each cluster with discrete data (4-point scales)
# in the theta parameterization using the PML estimation method
model <- '

    ### Step 2: Disentangle each continuous endogenous variable into a
    ### between-part and a within-part
    # take care of the thresholds
    y.1 + y.2 + y.3 | thy1*t1 + 0*t2 + thy3*t3
    m.1 + m.2 + m.3 | thm1*t1 + 0*t2 + thm3*t3
    z.1 + z.2 + z.3 | thy1*t1 + 0*t2 + thy3*t3
    # create star version of variables explicitly
    fy.1 =~1*y.1; fy.2 =~1*y.2; fy.3 =~1*y.3
    fm.1 =~1*m.1; fm.2 =~1*m.2; fm.3 =~1*m.3
    fz.1 =~1*z.1; fz.2 =~1*z.2; fz.3 =~1*z.3
    # zero residual variances
    y.1 ~~0*y.1; y.2 ~~0*y.2; y.3 ~~0*y.3
    m.1 ~~0*m.1; m.2 ~~0*m.2; m.3 ~~0*m.3
    z.1 ~~0*z.1; z.2 ~~0*z.2; z.3 ~~0*z.3
    # create within-level with star version of variables
    wy.1 =~1*fy.1; wy.2 =~1*fy.2; wy.3 =~1*fy.3
    wm.1 =~1*fm.1; wm.2 =~1*fm.2; wm.3 =~1*fm.3
    wz.1 =~1*fz.1; wz.2 =~1*fz.2; wz.3 =~1*fz.3
    # create between-level with star version of variables
    by =~1*fy.1 + 1*fy.2 + 1*fy.3
    bm =~1*fm.1 + 1*fm.2 + 1*fm.3
    bz =~1*fz.1 + 1*fz.2 + 1*fz.3

    ### Steps 3 and 4: Construct a model with all the newly introduced
    ### within-level variables with equality constraints
    # variances in the model, representing theta parametrisation
    wy.1 ~~1*wy.1; wy.2 ~~1*wy.2; wy.3 ~~1*wy.3
    wm.1 ~~1*wm.1; wm.2 ~~1*wm.2; wm.3 ~~1*wm.3
    wz.1 ~~1*wz.1; wz.2 ~~1*wz.2; wz.3 ~~1*wz.3
    # within regressions
    wz.1 ~ zy.w*wy.1 + zm.w*wm.1
    wz.2 ~ zy.w*wy.2 + zm.w*wm.2
    wz.3 ~ zy.w*wy.3 + zm.w*wm.3
    wm.1 ~ my.w*wy.1
    wm.2 ~ my.w*wy.2
    wm.3 ~ my.w*wy.3
    ### Step 5: Construct a model with all the newly introduced
    ### between-level variables
    # regressions at between-level
    bz ~ zy.b*by + zm.b*bm
    bm ~ my.b*by
    # variances and means at between-level
    bz ~~ bz; bm ~~ bm; by ~~ by; bz ~1; bm ~1; by ~1
'
#fit the model with the PML estimation method
Pfit <- lavaan(model, data = wideData, ordered = paste(rep(c("y1","y2",
    "y3"), 3), rep(1:3, each = 3), sep = "."), estimator = "PML",
    parameterization = "theta")
summary(Pfit)
```

APPENDIX F | LAVAAN SYNTAX FOR THE MULTILEVEL ONE-FACTOR MODEL WITH DISCRETE DATA IN THE WF APPROACH

```
# syntax corresponding to solid lines in the factor model of Figure 3 with
# three observations in each cluster with discrete data (4-point scales)
# in the theta parameterization using the PML estimation method
model <- '

    ### Step 2: Disentangle each continuous endogenous variable into a
    ### between-part and a within-part
    # take care of the thresholds
    y1.1 + y2.1 + y3.1 | thy1.1*t1 + 0*t2 + thy3.1*t3
    y1.2 + y2.2 + y3.2 | thy1.2*t1 + 0*t2 + thy3.2*t3
    y1.3 + y2.3 + y3.3 | thy1.3*t1 + 0*t2 + thy3.3*t3
    # create star explicitly
    fy1.1 =~1*y1.1; fy2.1 =~1*y2.1; fy3.1 =~1*y3.1
    fy1.2 =~1*y1.2; fy2.2 =~1*y2.2; fy3.2 =~1*y3.2
    fy1.3 =~1*y1.3; fy2.3 =~1*y2.3; fy3.3 =~1*y3.3
    # zero residual variances
    y1.1 ~~0*y1.1; y2.1 ~~0*y2.1; y3.1 ~~0*y3.1
    y1.2 ~~0*y1.2; y2.2 ~~0*y2.2; y3.2 ~~0*y3.2
    y1.3 ~~0*y1.3; y2.3 ~~0*y2.3; y3.3 ~~0*y3.3
    # create within-level component (with star) of the variables
    wy1.1 =~1*fy1.1; wy2.1 =~1*fy2.1; wy3.1 =~1*fy3.1
    wy1.2 =~1*fy1.2; wy2.2 =~1*fy2.2; wy3.2 =~1*fy3.2
    wy1.3 =~1*fy1.3; wy2.3 =~1*fy2.3; wy3.3 =~1*fy3.3
    # create between-level component (with star) of the variables
    by1 =~1*fy1.1 + 1*fy1.2 + 1*fy1.3
    by2 =~1*fy2.1 + 1*fy2.2 + 1*fy2.3
    by3 =~1*fy3.1 + 1*fy3.2 + 1*fy3.3

    ### Steps 3 and 4: Construct a model with all the newly introduced
    ### within-level variables with that represent theta parametrisation
    # variances in the model
    wy1.1 ~~1*wy1.1; wy2.1 ~~1*wy2.1; wy3.1 ~~1*wy3.1
    wy1.2 ~~1*wy1.2; wy2.2 ~~1*wy2.2; wy3.2 ~~1*wy3.2
    wy1.3 ~~1*wy1.3; wy2.3 ~~1*wy2.3; wy3.3 ~~1*wy3.3
```

```
   # factor model
   fw1 =~1*wy1.1 + lw2*wy2.1 + lw3*wy3.1
   fw2 =~1*wy1.2 + lw2*wy2.2 + lw3*wy3.2
   fw3 =~1*wy1.3 + lw2*wy2.3 + lw3*wy3.3
   fw1 ~~ NA*fw1 + fwt*fw1
   fw2 ~~ NA*fw2 + fwt*fw2
   fw3 ~~ NA*fw3 + fwt*fw3

   ### Step 5: Construct a model with all the newly introduced
   ### between-level variables
   # factor model
   fb =~1*by1 + by2 + by3;
   fb ~~ fb
   # intercepts
   by1 ~1; by2 ~1; by3 ~1
   # variances
   by1 ~~ by1; by2 ~~ by2; by3 ~~ by3
'
#fit the model with the PML estimation method
Pfit <- lavaan(model, data = wideData, ordered = paste(rep(c("y1","y2",
   "y3"), 3), rep(1:3, each = 3), sep = "."), estimator = "PML",
   parameterization = "theta")
summary(Pfit)
```