A Bayesian Approach for Estimating Multilevel Latent Contextual Models

Steffen Zitzmann,1 Oliver Lüdtke,1,2 Alexander Robitzsch,1,2 and Herbert W. Marsh3,4

1Leibniz Institute for Science and Mathematics Education
2Centre for International Student Assessment
3Institute for Positive Psychology and Education, Australian Catholic University
4King Saud University, Saudi Arabia

In many applications of multilevel modeling, group-level (L2) variables for assessing group-level effects are generated by aggregating variables from a lower level (L1). However, the observed group mean might not be a reliable measure of the unobserved true group mean. In this article, we propose a Bayesian approach for estimating a multilevel latent contextual model that corrects for measurement error and sampling error (i.e., sampling only a small number of L1 units from a L2 unit) when estimating group-level effects of aggregated L1 variables. Two simulation studies were conducted to compare the Bayesian approach with the maximum likelihood approach implemented in Mplus. The Bayesian approach showed fewer estimation problems (e.g., inadmissible solutions) and more accurate estimates of the group-level effect than the maximum likelihood approach under problematic conditions (i.e., small number of groups, predictor variable with a small intraclass correlation). An application from educational psychology is used to illustrate the different estimation approaches.

Keywords: Bayesian estimation, contextual analysis, multilevel latent contextual model, multilevel modeling, structural equation modeling

Multilevel modeling (MLM) is one of the most prominent modeling approaches in the social sciences. One major advantage of MLM is that it can be used to simultaneously relate outcome variables to predictors at the individual and group levels (e.g., Goldstein, 2011; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). However, many constructs in the social sciences cannot be measured without error, and it is well known that this measurement error leads to biased and less efficient estimates of population parameters. In the last two decades, substantial progress has been made in integrating MLM with structural equation modeling (SEM) by employing multiple indicators to correct for the unreliability that occurs in the assessment of constructs (e.g., Lee & Poon, 1998; McDonald, 1993; Mehta & Neale, 2005; B. O. Muthén & Asparouhov, 2011; Skrondal & Rabe-Hesketh, 2004).

In many applications of MLM, group-level (L2) variables for assessing group-level effects are generated by aggregating variables from a lower level (L1). For example, in educational psychology, features of the learning environment (e.g., classroom climate, teacher behavior; see Marsh et al., 2012) are measured by aggregating individual student ratings at the class level and then, in the next step, using these aggregated ratings to predict learning outcomes. However, it has been pointed out in the methodological literature that the observed group mean might not provide a reliable measure of the unobserved true group mean (e.g., Asparouhov & Muthén, 2007; Croon & van Veldhoven, 2007; Grilli & Ramichini, 2011; Lüdtke et al., 2008; Pokropek, 2015; Preacher, Zyphur, & Zhang, 2010; Shin & Raudenbush, 2010).

Lüdtke, Marsh, Robitzsch, and Trautwein (2011; see also Marsh et al., 2009; Rabe-Hesketh, Skrondal, & Pickles, 2004) argued that two main sources of error can be distinguished when assessing group-level effects by aggregating individual data from a lower level: unreliability that is due to measurement error at L1 and L2 and unreliability that is
due to sampling only a small number of L1 units (e.g., persons) within each L2 unit. To correct for these two types of error, they discussed a multilevel latent contextual model for estimating group-level effects of aggregated L1 variables. This model has been described as a doubly latent model for estimating contextual effects because it takes into account measurement error at L1 and L2 (based on multiple L1 indicators) as well as L2 sampling error (latent aggregation from L1 to L2). The multilevel latent contextual model is currently implemented in Mplus (L. K. Muthén & Muthén, 2012) and is based on maximum likelihood (ML) estimation (see also the GLLAMM module in Stata; Rabe-Hesketh et al., 2004).

However, extensive simulation research has identified two major obstacles to applying the multilevel latent contextual model in research practice. First, serious estimation problems such as nonconvergence or inadmissible estimates (e.g., negative variances) have been observed for the doubly latent approach with small intraclass correlations and small sample sizes at L1 and L2 (Li & Beretvas, 2013; Lüdtke et al., 2011). Second, under certain conditions (e.g., a small number of L2 groups, a small intraclass correlation), the multilevel latent contextual model, although approximately unbiased, was found to be much more variable (i.e., to have a larger empirical sampling variance) than the biased traditional MLM approach, which uses the observed group mean as a predictor variable. This has resulted in the counterintuitive finding that—under certain conditions—the multilevel latent contextual model provides less accurate estimates of the group-level effect than the manifest approach in terms of the root mean squared error (RMSE), which combines bias and variability. This bias–accuracy trade-off constitutes a major challenge for the general applicability of the multilevel latent contextual model (Li & Beretvas, 2013; see also Ledgerwood & Shrout, 2011).

In this article, we introduce a Bayesian approach for estimating the multilevel latent contextual model, and we evaluate it with simulation studies. We believe that a Bayesian approach has the potential to mitigate the limitations of the multilevel latent contextual model. First, several scholars have argued that Bayesian methods offer a promising approach for estimating multilevel models with latent variables, particularly when the focus is on estimating variance components, and the number of groups is small (see Depaoli & Clifton, 2015; Hox, van de Schoot, & Matthijssse, 2012; see also Hamaker & Klugkist, 2011). Using appropriate prior distributions for the parameters, the Bayesian approach guarantees that the model estimates will be within the admissible range (Can, van de Schoot, & Hox, 2015). Second, Bayesian methods allow for the stabilization of parameter estimates by specifying weakly informative prior distributions for critical parameters (Baldwin & Fellingham, 2013; Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2015). In a recent study, Zitzmann, Lüdtke, and Robitzsch (2015) showed that a Bayesian approach with a weakly informative prior distribution for the group-level variance of the predictor produced more accurate estimates of group-level effects than an ML approach (see also Depaoli & Clifton, 2015). However, the authors did not take into account the role of measurement error and considered only the sampling error.

This article is organized as follows. We begin by briefly describing the multilevel latent contextual model. We then present two studies. In Study 1, where there is one latent predictor variable, we discuss how the multilevel latent contextual model can be estimated by applying a Bayesian approach and present a simulation study that compares the proposed Bayesian approach with the ML approach. In Study 2, we add a manifest group-level predictor to the multilevel latent contextual model. We show both analytically and with a simulation how an unreliable assessment of the latent predictor at the group level results in biased estimates of the effect of the manifest group-level predictor. We again compare the Bayesian approach with the ML approach. Finally, we use an empirical example from educational psychology to demonstrate how the Bayesian approach can be specified with the WinBUGS software (Spiegelhalter, Thomas, Best, & Lunn, 2003) and in Mplus.

**MULTILEVEL LATENT CONTEXTUAL MODEL**

In this section, we briefly present the main features of the multilevel latent contextual model, which is essentially a multilevel SEM. Different frameworks for conceptualizing multilevel SEMs have been proposed (see Rabe-Hesketh, Skrondal, & Zheng, 2012, for a discussion). In the following, we use the within–between framework, which decomposes observed indicator variables into within-group components (denoted by $W$) and between-group components (denoted by $B$). Two standard structural equation models are then defined: one within-group model and one between-group model. More specifically, the multivariate response vector $y_{ij}$ for person $i$ in group $j$ (e.g., student $i$ in class $j$) is decomposed as follows:

$$ y_{ij} = y_{ij}^B + y_{ij}^W $$

where $y_{ij}^B$ is a vector of random effects between groups, and $y_{ij}^W$ is a vector of residuals within groups.

In the next step, measurement models that relate the within- and between-group components of the multiple indicators to the latent constructs at the individual and group levels are specified:

$$ \text{Level 1: } y_{ij}^W = \Lambda^w \eta_{ij}^w + \epsilon_{ij}^w $$

$$ \text{Level 2: } y_{ij}^B = v + \Lambda^g \eta_{ij}^g + \epsilon_{ij}^g $$

(2)
where \( v \) is a vector of intercepts; and \( A^W \) and \( A^B \) are factor loading matrices linking the within-parts \( y_{ij}^W \) and between-parts \( y_{ij}^B \) to the latent constructs \( \eta_{ij}^W \) and \( \eta_{ij}^B \) at L1 and L2. The latent constructs \( \eta_{ij}^W \) and \( \eta_{ij}^B \) are assumed to be multivariate normally distributed, and the \( \epsilon_{ij}^W \) and \( \epsilon_{ij}^B \) terms denote within- and between-group errors that have multivariate normal distributions around zero with covariance matrices \( \Theta^W \) and \( \Theta^B \) at L1 and L2. The model in Equation 2 has been extensively discussed in the literature on multilevel confirmatory factor analysis (e.g., Geldhof, Preacher, & Zyphur, 2014; Mehta & Neale, 2005; B. O. Muthén, 1994).

The structural models that relate the latent constructs at L1 and L2 to one another are given as follows:

\[
\begin{align*}
\text{Level 1} &: \quad \eta_{ij}^W = B^W \eta_{ij}^B + \Gamma^W x_{ij} + \zeta_{ij}^W \\
\text{Level 2} &: \quad \eta_{ij}^B = \alpha + B^B \eta_{ij}^B + \Gamma^B x_{ij} + \zeta_{ij}^B
\end{align*}
\]  

(3)

where \( \alpha \) is a vector of regression intercepts and \( B^W \) and \( B^B \) are coefficient matrices for the effects of the latent constructs \( (\eta_{ij}^W \text{ and } \eta_{ij}^B) \) on each other at L1 and L2. The influence of the exogenous covariates \( x_{ij} \) at L1 and L2 are given by the coefficient matrices \( \Gamma^W \) and \( \Gamma^B \). Note that exogenous covariates are manifest variables that are assumed to be measured without error. The \( \zeta_{ij}^W \) and \( \zeta_{ij}^B \) terms denote structural residuals that are multivariate normally distributed with means of zero and covariance matrices \( \Psi^W \) and \( \Psi^B \) at L1 and L2. It is important to emphasize that in the multilevel latent contextual model, the effects of the latent constructs at the group level given by \( B^B \) are corrected for both measurement error and the unreliability that results from sampling only a finite number of persons from a potentially infinite population (e.g., a small number of students from a school). Both kinds of error are discussed in more detail in the next section.

**STUDY 1: ONE LATENT PREDICTOR VARIABLE**

In this section, we begin with the simple case of a multilevel latent contextual model with one latent predictor variable that is measured at the individual level. We further assume that the latent construct is measured by \( k = 1, \ldots, p \) multiple indicators (e.g., the items or subscales of a test). According to Equation 2, the indicator \( y_{kij} \) is decomposed as follows:

\[
y_{kij} = \gamma_{kij} + \eta_{kij} = v_k + \lambda_k^{W,ij} + \epsilon_{kij}^W + \lambda_k^{B,ij} \eta_{ij}^W + \epsilon_{kij}^B
\]  

(4)

where \( v_k \) is the indicator-specific mean, \( \lambda_k^{W,ij} \) is the within-group factor loading, \( \lambda_k^{B,ij} \) is the between-group factor loading, and \( \epsilon_{kij}^W \) and \( \epsilon_{kij}^B \) are the indicator-specific residuals at L1 and L2, which are assumed to be uncorrelated across indicators. As the factor loadings are allowed to vary across indicators, Equation 4 represents a congeneric measurement model for the unobserved true scores \( \eta_{ij}^W \) and \( \eta_{ij}^B \) at L1 and L2. To identify the metric of both latent variables \( \eta_{ij}^W \) and \( \eta_{ij}^B \), the loadings of the first indicator at L1 (\( \lambda_1^W = 1 \)) and at L2 (\( \lambda_1^B = 1 \)) can be fixed to 1.

In multilevel research, the intraclass correlation (ICC) provides important information about the multilevel structure of the data. For the latent predictor variable, the intraclass correlation is defined by:

\[
\text{ICC}_\eta = \frac{\psi^B}{\psi^B + \psi^W}
\]  

(5)

where \( \psi^B \) is the variance of the latent between-group portion \( \eta_{ij}^B \) and \( \psi^W \) is the variance of the latent within-group portion \( \eta_{ij}^W \). Thus, the ICC\(_\eta \) indicates the proportion of the total variance of the latent predictor variable that can be attributed to between-group differences.

In the following, we are interested in predicting an observed dependent variable \( y_{dep,ij} \) that is measured without measurement error and can be decomposed into its within and between portions as follows:

\[
y_{dep,ij} = y_{dep,ij}^W + y_{dep,ij}^B = 1 \cdot \eta_{ij}^W + 1 \cdot \eta_{ij}^B + \xi_{ij}
\]  

The structural models at L1 and L2 are then given as:

\[
\begin{align*}
\text{Level 1} &: \quad \eta_{ij}^W = B^W \eta_{ij}^B + \zeta_{ij}^W \\
\text{Level 2} &: \quad \eta_{ij}^B = \alpha + B^B \eta_{ij}^B + \zeta_{ij}^B
\end{align*}
\]  

(6)

where \( \beta^W \) is the within-group (L1) regression coefficient, which describes the relationship between \( \eta_{ij}^W \) and \( \eta_{ij}^W \) within groups, and \( \beta^B \) is the between-group (L2) regression coefficient, which describes the relationship between the group components \( \eta_{ij}^B \) and \( \eta_{ij}^B \).

The structural part of the multilevel latent contextual model can be written as follows:

\[
y_{dep,ij} = \eta_{ij}^W + \eta_{ij}^B = \alpha + \beta^W \eta_{ij}^W + \beta^B \eta_{ij}^B + \xi_{ij}^W + \xi_{ij}^B.
\]  

(7)

It can now be seen that the dependent variable is decomposed into the parts that can be predicted by the within- and between-group components of the latent predictor and the corresponding residual terms at L1 and L2. The model is illustrated in Figure 1.

The congeneric measurement model for the latent predictor can be further simplified by assuming parallel indicators; that is, the factor loadings are equal at L1 (\( \lambda_k^W = \lambda_k^W \))

---

1 It is not crucial to assume that the dependent variable is measured without error because unreliability in the dependent variable does not result in biased estimates of unstandardized regression coefficients (e.g., Buonaccorsi, 2010).
and at L2 ($\lambda^{B}_{1} = \lambda^{B}_{2}$), and the measurement error variances are also equal at L1, $\text{Var}(\epsilon^{W}_{ij}) = \text{Var}(\epsilon^{B}_{ki})$, and at L2, $\text{Var}(\epsilon^{B}_{ij}) = \text{Var}(\epsilon^{B}_{kj})$. In research practice, the model with parallel indicators at L1 and L2 in Equation 7 is often approximated by a traditional multilevel model in which the latent between-group component $\eta^{B}_{ij}$ is assessed by using the observed group mean, which is calculated by averaging across items and persons within a group: $\bar{y}_{ij} = \frac{1}{pm} \sum_{k=1}^{n} \sum_{l=1}^{p} y_{klj}$, where $n_j$ denotes the number of persons within a group $j$. In the traditional multilevel model, this observed group mean of the total score is treated like a manifest variable rather than a latent variable. Using the group-mean-centered predictor variable at L1, the traditional approach would be specified as follows (see Raudenbush & Bryk, 2002):

$$y_{\text{dep.ij}} = \alpha^{W}_{DM} + \beta^{W}_{DM}(\bar{y}_{\text{ij}} - \bar{y}_{\text{*.}}) + \beta^{B}_{DM} \bar{y}_{\text{*.}} + \epsilon^{W}_{ij} + \epsilon^{B}_{ij}. \quad (8)$$

As can now be seen, the traditional approach in Equation 8 approximates the multilevel latent contextual model from Equation 7. The latent variables $\eta^{W}_{ij}$ and $\eta^{B}_{ij}$ are substituted by the manifest variables $\bar{y}_{\text{ij}}$ and $\bar{y}_{\text{ij}} - \bar{y}_{\text{*.}}$ at L1 and L2, and $\beta^{W}_{DM}$ and $\beta^{B}_{DM}$ are the corresponding within-group and between-group coefficients. It is apparent that the traditional approach does not take into account the error that is present in the manifest scores. As both types of error are ignored—measurement error and sampling error—the approach could be labeled a doubly manifest (DM) approach for estimating group-level effects (Lüdtke et al., 2011). It can be shown that $\beta^{W}_{DM}$ is a biased estimator of the within-group regression coefficient $\beta^{W}$ that underestimates the absolute size of the true relationship within groups (e.g., Cohen, Cohen, West, & Aiken, 2003; Kenny, 1979; see Equation 14 in Lüdtke et al., 2011). The bias of $\beta^{B}_{DM}$ is a direct function of the reliability of the manifest score at L1, which depends on the number of items when indicators are parallel. However, the bias of the estimator $\beta^{B}_{DM}$ of the between-group coefficient $\beta^{B}$ depends on the reliability of the observed group mean, which is also influenced by the L2 sampling error. Assuming parallel indicators at L1 and L2, the reliability of the observed group mean (Kane, Gillmore, & Crooks, 1976; see also Schweig, 2014) is given by:

$$\text{Rel}(\bar{y}_{\text{*.}}) = \frac{\psi^{B}_{ij}}{\psi^{B}_{ij} + \frac{1}{n} \psi^{W}_{ij} + \frac{1}{p} \beta^{B} + \frac{1}{pm} \theta^{W}} \quad (9)$$

where $\theta^{B}$ is the variance of the between-group errors $\epsilon^{B}_{ij}$, and $\theta^{W}$ is the variance of the within-group errors $\epsilon^{W}_{ij}$. As Equation 9 shows, the reliability is a function of the sampling error (i.e., $\frac{1}{n} \psi^{W}_{ij}$), the measurement error at L2 (i.e., $\frac{1}{p} \beta^{B}$), and the measurement error at L1 (i.e., $\frac{1}{pm} \theta^{W}$).

Assuming equal group sizes (i.e., omitting the index $j$ in $n_j$), it can be shown analytically (Lüdtke et al., 2011; see also Shin & Raudenbush, 2010) that the bias of the estimator of the between-group coefficient in the traditional multilevel model in Equation 8 is given by:

$$\text{Bias}(\beta^{B}_{DM}) = \frac{1 - \text{ICC}_{n}}{\text{ICC}_{n}} \cdot \text{Rel}(\bar{y}_{\text{*.}}) \cdot \beta^{W}$$

$$- \left\{1 - \text{Rel}(\bar{y}_{\text{*.}})\right\} \cdot \beta^{B}. \quad (10)$$

As can be seen, the bias depends on the reliability of the manifest group means as well as the size of the within-group ($\beta^{W}$) and between-group ($\beta^{B}$) coefficients in the population. It is also evident that the bias of the estimator of the between-group coefficient is a function of sampling error ($n$ and $\text{ICC}_{n}$) and measurement error (number of indicators $p$), and that the bias decreases with an increasing number of groups and a larger ICC. However, even when there is no sampling error (when the ICC$(n)$ is 1), $\beta^{B}_{DM}$ is still a biased estimator because measurement error attenuates the effect of the manifest group mean.

**Specification of Prior Distributions**

Bayesian approaches have recently received considerable attention in the literature on MLM (e.g., Gelman & Hill, 2007; Hamaker & Klugkist, 2011; Zhang, Hamagami, Wang, Nesselroade, & Grimm, 2009), SEM (e.g., Arminger & Muthén, 1998; Kaplan & Depaoli, 2012; Lee, 2007; B. O. Muthén & Asparouhov, 2012; Song & Lee, 2009), and multilevel SEM (e.g., Ansari & Jedidi, 2000; Depaoli & Clifton, 2015). It is a crucial feature of Bayesian estimation that it uses the prior as a source of information in addition to the data. The prior is
multiplied by the likelihood function. One refers to the result as the posterior distribution, which forms the basis for the construction of Bayesian point and interval estimators. This idea can be formally written as the well-known Bayes’s theorem: posterior ∝ likelihood • prior. In practice, the posterior is often approximated by using Markov chain Monte Carlo (MCMC) algorithms that iteratively sample from conditional distributions, constituting an MCMC chain. Bayesian point and interval estimates can be derived when the chain has converged and approximates the posterior distribution sufficiently well (for an introduction, see Christensen, Johnson, Branscum, & Hanson, 2011; Jackman, 2009; Lynch, 2007).

The prior distributions that we used to estimate the multilevel latent contextual model were as follows. We specified uninformative or weakly informative priors for the parameters of the measurement model in Equation 4 and the structural model in Equation 6, respectively. Given mean-centered indicator variables, a normal distribution with zero mean and a small precision of .001 (or equivalently a large variance of 1,000) was considered uninformative for the intercepts and the loadings (except for the first one) in the measurement model of the latent predictor as well as for the intercept and coefficients in the structural model: \( \psi_k \sim N(0, 1000) \) for all \( k \), \( \lambda_k^W \sim N(0, 1000) \), \( \lambda_k^B \sim N(0, 1000) \) for \( k > 1 \), \( \alpha \sim N(0, 1000) \), \( \beta_k^W \sim N(0, 1000) \), \( \beta_k^B \sim N(0, 1000) \).

Although it is often less critical to select priors for regression coefficients, there is an ongoing discussion about which prior distribution should be specified for variance components (e.g., Browne & Draper, 2006; Chung, Rabe-Hesketh, Dorie, Gelman, & Liu, 2013; Gelman, 2006; Natarajan & Kass, 2000). However, the inverse gamma (IG) distribution is most commonly used in research practice (Browne & Draper, 2006). With a shape/scale parameter of .001, it is approximately uniform for a large range of possible values and is thus meant to be uninformative (but see Gelman, 2006, for the argument that this prior can be quite informative when variance components are very small). Thus, an approximately uninformative IG distribution with a shape/scale parameter equal to .001 was specified for the within-group variances in the multilevel latent contextual model: \( \psi_k^W \sim IG(.001, .001) \), \( \theta_k^W \sim IG(.001, .001) \) for all \( k \), \( \text{Var} \left( \psi_k^B \right) \sim IG(.001, .001) \). Furthermore, the Bayesian approach provides the opportunity to stabilize parameter estimates by specifying weakly informative prior distributions (e.g., Baldwin & Fellingham, 2013). One application of an informative prior is related to the observation that, under particular conditions, estimates of the between-group variances can be close to zero (Chung et al., 2013). To penalize for near-zero estimates, we set the shape/scale parameter to .1 (see Depaoli & Clifton, 2015): \( \psi_k^B \sim IG(.1, .1) \), \( \theta_k^B \sim IG(.1, .1) \) for all \( k \), \( \text{Var} \left( \psi_k^B \right) \sim IG(.1, .1) \). According to this prior, the 2.5% quantile of the distribution was 0.094 and the 97.5% quantile was 1.082 • 10\(^{15} \). As can be seen, large estimates of the between-group variances were still allowed, so the prior could be considered weakly informative.

Simulation Study

The aim of the simulation was threefold. First, previous simulations have shown that the multilevel latent contextual model is prone to estimation problems (nonconvergence, inadmissible estimates) when estimation is carried out with an ML approach (Li & Beretvas, 2013; Lüdtke et al., 2011). These estimation problems should not occur with a Bayesian approach (e.g., Hox & Maas, 2001). Second, we believe that one reason for the low accuracy of the multilevel latent contextual model when estimated by an ML approach is a nonnegligible proportion of near-zero estimates of the between-group variance of the predictor variable (see Chung et al., 2013). We assumed that by specifying weakly informative prior distributions for variance parameters at the group level, the estimates of the between-group effect provided by the Bayesian approach would be stabilized and thus more accurate than the estimates provided by the ML approach (Zitzmann et al., 2015). Third, we compared the doubly manifest model that did not correct for error in the predictor variable with two multilevel latent contextual models that differed in their measurement models (parallel vs. congeneric indicators). This allowed us to test how the complexity of the measurement model was related to the quality of the estimates of the between-group effect.

Method

Conditions. The data were generated by the multilevel latent contextual model given in Equations 4 and 6 (see also Figure 1). The latent predictor was measured with three mean-centered and normally distributed indicators, each with a standardized between-group loading of 0.7. The (unstandardized) within- and between-group regression coefficients were set to 0.2 and 0.5, respectively. The manifest outcome variable was unstandardized and normally distributed (i.e., zero mean, unit variance). The unconditional ICC of the outcome variable (not adjusted for the predictor) was 0.3. The (unstandardized) within-group and between-group loadings of the first indicator were both set to 1 to identify the metric of the latent predictor. The indicators could be either parallel (within-group loadings: 1, 1, 1; between-group loadings: 1, 1, 1) or congeneric (within-group loadings: 1, 0.4, 1.6; between-group loadings: 1, 0.4, 1.6). In both cases, the loadings were specified to be invariant across levels (cross-level invariance). The number of groups was set
to either 25 or 100. These numbers were chosen so that the lower number would represent a very small number in multilevel research (see Kreft & De Leeuw, 1998). The group size was either 10 or 30 persons. Small group sizes occur in organizational psychology (e.g., working teams), whereas group sizes are often larger in educational psychology (e.g., school classes). The ICC of the latent predictor was set to either .1 or .3 to simulate ICCs that often occur in the social sciences (see Gulliford, Ukoumunne, & Chinn, 1999). The indicators were assumed to have the same ICC as the latent predictor. For each of the $2 \times 2 \times 2 = 16$ data constellations, 1,000 data sets were generated.

**Analysis models.** Each of the data sets was analyzed with the doubly manifest model and the multilevel latent contextual models with parallel and congeneric indicators. In the model with parallel indicators, all within- and between-group loadings were fixed to 1. In the model with congeneric indicators, only the within- and between-group loadings of the first indicator were fixed to 1 to identify the metric of the latent predictor, whereas the remaining within- and between-group loadings were freely estimated but constrained to be equal across levels. Moreover, in the model with parallel indicators, the within- and between-group error variances were freely estimated but constrained to be equal across the indicators, whereas in the model with congeneric indicators, the error variances were freely estimated without any constraints.

Each model was estimated in Mplus 7.1 by applying ML with robust standard errors (MLR; L. K. Muthén & Muthén, 2012) and in WinBUGS 1.4.3 (Lunn, Thomas, Best, & Spiegelhalter, 2000). Before running the main simulation study, we studied the behavior of the MCMC sampler in a preliminary simulation by inspecting two criteria for each data constellation and each analysis model: (a) the potential scale reduction factor (PSR; Gelman & Rubin, 1992), and (b) the MCMC standard error (MCMCSE) divided by the posterior standard deviation (PSD): $\frac{\text{MCMCSE}}{\text{PSD}}$. More specifically, for each model parameter, we inspected whether the PSR was close to 1 (see B. O. Muthén & Asparouhov, 2012) and whether the $\frac{\text{MCMCSE}}{\text{PSD}}$ ratio was smaller than .05 (see Cho & Bottge, 2015). The PSR is typically used to diagnose whether the chain has converged, and the $\frac{\text{MCMCSE}}{\text{PSD}}$ ratio is used to assess how well the empirical mean of the chain approximates the posterior mean (Hoff, 2009, p. 101). Applying these two criteria suggested that an average chain length of about 80,000 iterations (with a burn-in period of one tenth of the total length) was sufficient to provide a good approximation of the posterior distribution (see Figure E.1 in the supplemental materials for example trace plots). However, the actual chain length varied substantially in accordance with the analysis model and the data constellation. In the worst case ($J = 25, n = 10, \text{ICC}_1 = .1$), the model with congeneric indicators required a total of 180,000 iterations. However, in the most favorable case ($J = 100, n = 30, \text{ICC}_1 = .3$), the doubly manifest model required only 900 iterations. The Bayesian point estimate was defined by the mode of the posterior distribution (MAP; see DeCarlo, Kim, & Johnson, 2011; Zitzmann et al., 2015). The MAP was calculated using kernel density estimates (e.g., Silverman, 1998). The Bayesian confidence interval (BCI) estimate was defined by the 2.5th and the 97.5th percentiles of the posterior distribution (see Box & Tiao, 1992).

**Evaluation criteria.** The performances of the different analysis models and estimation approaches were evaluated on the basis of the following four criteria: frequency of estimation problems, relative bias, relative RMSE, and coverage rate. An ML solution was classified as nonconverged (NC) when the convergence criterion was not met within the maximum number of iterations or when difficulties arose in optimizing the fit function. Furthermore, a solution was classified as inadmissible converged (IAC) when it converged to a solution that included an inadmissible estimate (e.g., a negative variance; see Marsh, Hau, Balla, & Grayson, 1998; see also Wothke, 1993). Solutions that did not fall into these two categories were classified as admissible converged (AC), indicating that no estimation problems occurred. The relative bias of the estimator of the between-group effect $\beta^B$ is given as the difference between the expectation of the estimate $\hat{\beta}^B$ and $\beta^B$ divided by $\beta^B$. A maximum absolute value of 0.10 was considered to constitute an acceptable degree of relative bias (L. K. Muthén & Muthén, 2002). The overall accuracy was assessed by computing the relative RMSE, which is given as the square root of the expectation of the squared deviation of the estimate $\hat{\beta}^B$ from $\beta^B$ divided by $\beta^B$. Finally, the standard errors of $\hat{\beta}^B$ were evaluated in terms of the coverage rate, which is given as the probability that the 95% confidence interval (CI) captures $\beta^B$. A coverage rate between 91% and 98% was considered acceptable (L. K. Muthén & Muthén, 2002).

**Results**

**Estimation problems.** No estimation problems occurred for the Bayesian approach. However, in some data constellations, serious estimation problems were observed when the ML approach was applied. Table 1 presents the AC rate, the NC rate, and the IAC rate for the multilevel latent contextual models with parallel and congeneric indicators. The doubly manifest model estimated with ML did not show any estimation problems and was therefore not included in Table 1. When the data-generating model was the model with parallel indicators, the AC rate was higher for the correctly specified
model with parallel indicators ($M = 95.2\%$, range = $80.9\%$–$100.0\%$) than for the model with congeneric indicators ($M = 89.8\%$, range = $62.0\%$–$100.0\%$) for which loadings were estimated. As can be seen, for both analysis models, the AC rate was lowest when the number of groups was small (i.e., $J = 25$). When the data-generating model assumed congeneric indicators, the percentages of admissible converged solutions were clearly lower for the model with parallel indicators ($M = 66.2\%$, range = $42.3\%$–$93.3\%$) than for the model with congeneric indicators ($M = 90.9\%$, range = $59.7\%$–$99.9\%$). Most of the estimation problems for the “misspecified” model with parallel indicators were due to nonconverged solutions, which increased in particular when the ICC increased.

For the following analyses, the comparison between the ML and Bayesian approaches was based on only those data sets for which the ML approach provided an AC solution. Furthermore, we restricted the comparison of the two estimation approaches to the case where the analysis model matched the data-generating model.

**Comparison of the ML and Bayesian approaches.** Table 2 shows the relative bias, relative RMSE, and coverage rate for the estimators of the between-group effect provided by the ML and Bayesian approaches. As can be seen, when the data were generated by the model with parallel indicators, both estimation approaches provided approximately unbiased estimates when the ICC of the latent predictor was high.

### TABLE 1
Results of Study 1: Rates of Admissible Converged (AC), Nonconverged (NC), and Inadmissible Converged (IAC) Solutions

| No. of Groups | Group Size | Intraclass Correlation | AC | NC | IAC |
|---------------|------------|------------------------|----|----|-----|
|               |            | Data-generating model: PI |    |    |     |
| $J = 25$      | $n = 10$   | ICC$_n = .1$           | 80.9 | 11.0 | 8.1 |
|               |            | ICC$_n = .3$           | 92.5 | 2.5  | 5.0 |
|               | $n = 30$   | ICC$_n = .1$           | 93.7 | 3.4  | 2.9 |
|               |            | ICC$_n = .3$           | 95.9 | 0.6  | 3.5 |
| $J = 100$     | $n = 10$   | ICC$_n = .1$           | 99.0 | 0.6  | 0.4 |
|               |            | ICC$_n = .3$           | 99.9 | 0.0  | 0.1 |
|               | $n = 30$   | ICC$_n = .1$           | 100.0| 0.0  | 0.0 |
|               |            | ICC$_n = .3$           | 100.0| 0.0  | 0.0 |

|                | Data-generating model: CI |    |    |     |
| $J = 25$      | $n = 10$   | ICC$_n = .1$           | 69.6 | 25.4 | 5.0 |
|               |            | ICC$_n = .3$           | 54.2 | 42.6 | 7.2 |
|               | $n = 30$   | ICC$_n = .1$           | 78.6 | 16.7 | 4.7 |
|               |            | ICC$_n = .3$           | 44.8 | 53.9 | 1.3 |
| $J = 100$     | $n = 10$   | ICC$_n = .1$           | 93.3 | 6.1  | 0.6 |
|               |            | ICC$_n = .3$           | 56.6 | 43.4 | 0.0 |
|               | $n = 30$   | ICC$_n = .1$           | 90.5 | 9.4  | 0.1 |
|               |            | ICC$_n = .3$           | 42.3 | 57.7 | 0.0 |

*Note.* PI = multilevel latent contextual model with parallel indicators; CI = multilevel latent contextual model with congeneric indicators; $J =$ number of groups; $n =$ group size; ICC$_n =$ intraclass correlation of the latent predictor.

### TABLE 2
Results of Study 1: Relative Bias, Relative Root Mean Squared Error (RMSE), and Coverage Rate From Estimating the Between-Group Effect of the Latent Predictor Variable With the Maximum Likelihood Estimation Approach (ML) and the Bayesian Estimation Approach (Bayes)

| No. of Groups | Group Size | Intraclass Correlation | Admissible Converged | Relative Bias ML | Bayes | Relative RMSE ML | Bayes | Coverage Rate ML | Bayes |
|---------------|------------|------------------------|----------------------|-----------------|-------|------------------|-------|------------------|-------|
|               |            | Data-generating model: PI |                      | -0.06 | -0.31 | 2.12 | 1.04 | 94.4 | 97.0 |
| $J = 25$      | $n = 10$   | ICC$_n = .1$           | 80.9 | 2.5  | 0.5  | 0.83 | 0.59 | 93.2 | 98.1 |
|               |            | ICC$_n = .3$           | 92.5 | 88.8 | 17.5 | 1.71 | 1.01 | 94.9 | 96.6 |
|               | $n = 30$   | ICC$_n = .1$           | 93.7 | 78.0 | 17.5 | 0.68 | 0.53 | 93.0 | 97.4 |
|               |            | ICC$_n = .3$           | 95.9 | 78.0 | 17.5 | 1.00 | 1.00 | 99.9 | 99.9 |
| $J = 100$     | $n = 10$   | ICC$_n = .1$           | 99.0 | 98.7 | 0.7  | 98.7 | 0.8  | 99.9 | 99.9 |
|               |            | ICC$_n = .3$           | 99.9 | 62.0 | 1.3  | 62.0 | 1.3  | 99.9 | 99.9 |
|               | $n = 30$   | ICC$_n = .1$           | 100.0| 98.4 | 1.6  | 98.4 | 1.6  | 99.9 | 99.9 |
|               |            | ICC$_n = .3$           | 100.0| 92.6 | 6.1  | 92.6 | 6.1  | 99.9 | 99.9 |

|                | Data-generating model: CI |                      | -0.06 | -0.20 | 2.26 | 1.09 | 96.5 | 97.3 |
| $J = 25$      | $n = 10$   | ICC$_n = .1$           | 59.7 | 10.0 | 0.4  | 0.96 | 0.65 | 92.6 | 98.0 |
|               |            | ICC$_n = .3$           | 88.8 | 88.8 | 10.0 | 92.6 | 98.0 |
|               | $n = 30$   | ICC$_n = .1$           | 87.7 | 87.7 | 10.0 | 92.6 | 98.0 |
|               |            | ICC$_n = .3$           | 93.2 | 93.2 | 10.0 | 92.6 | 98.0 |
| $J = 100$     | $n = 10$   | ICC$_n = .1$           | 98.3 | 98.3 | 10.0 | 92.6 | 98.0 |
|               |            | ICC$_n = .3$           | 99.9 | 99.9 | 10.0 | 92.6 | 98.0 |
|               | $n = 30$   | ICC$_n = .1$           | 100.0| 100.0| 10.0 | 92.6 | 98.0 |
|               |            | ICC$_n = .3$           | 100.0| 100.0| 10.0 | 92.6 | 98.0 |

*Note.* Results are based on only those data sets for which ML provided an admissible converged solution. PI = multilevel latent contextual model with parallel indicators; CI = multilevel latent contextual model with congeneric indicators; $J =$ number of groups; $n =$ group size; ICC$_n =$ intraclass correlation of the latent predictor.
predictor was large (ML: $M = 0.05$, range = 0.04–0.07, Bayes: $M = 0.00$, range = −0.10–0.07). However, in conditions with a small ICC, the Bayesian approach produced a negatively biased estimate, particularly when the number of groups was small. By contrast, the ML approach provided slightly positively biased estimates of the between-group effect in conditions with a large number of groups and a small group size. A very similar picture emerged for the relative bias when the data were generated with the model with congeneric indicators.

Moreover, when the model with parallel indicators was the data-generating model, the estimates obtained with the Bayesian approach ($M = 0.64$, range = 0.32–1.04) were more accurate than the estimates provided by the ML approach ($M = 1.00$, range = 0.31–2.12), particularly when the number of groups was small and the ICC was low. For example, in the condition with a small number of groups, a small group size, and a low ICC, the relative RMSE was 2.12 for the ML approach and only 1.04 for the Bayesian approach. Again, the findings were very similar when the data were generated by the model with congeneric indicators. However, it needs to be pointed out that for both data-generating models, differences in the relative RMSE between the ML and Bayesian approaches vanished in the most favorable condition ($J = 100$, $n = 30$, ICC$_\eta = .3$).

As can also be seen from Table 2, when the model with parallel indicators held in the population, the 95% CI for the ML estimator ($M = 94.7\%$, range = 93.2%–96.3%) and the 95% BCI for the Bayesian approach ($M = 95.7\%$, range = 93.4%–98.1%) produced coverage rates that were close to the nominal 95% rate. However, the Bayesian approach showed a tendency to provide coverage rates that were slightly too high (compared with ML) in conditions with a small number of groups. The results for the coverage rate did not change when the model with congeneric indicators was the data-generating model.

Comparisons of different analysis models within Bayes. We also investigated the differences between the different analysis models for estimating the between-group effect. For these analyses, we only used the Bayesian approach because no estimation problems were produced by the Bayesian approach. Table 3 presents the relative bias, RMSE, and coverage rate for the estimator of the between-group effect for the doubly manifest model, the model with parallel indicators, and the model with congeneric indicators. As can be seen, when the data were generated by the model with parallel indicators, the estimates provided by the doubly manifest model were substantially negatively biased. As expected, the estimated bias was largest when the group size was small and the ICC was low. By contrast, the model with parallel indicators and the model with congeneric indicators provided approximately unbiased estimates of the between-group effect, with the exception that in the worst condition ($J = 25$, $n = 10$, ICC$_\eta = .1$), both models produced slightly negatively biased estimates. However, when the data were generated under the assumption of congeneric indicators, the misspecified model with parallel indicators provided positively as well as negatively biased estimates, particularly in conditions with a large group size and a high ICC. Again, estimates obtained from the doubly manifest model were substantially negatively biased, whereas the correctly specified model with congeneric indicators provided approximately unbiased estimates.

Next, we compared the different analysis models with respect to their relative RMSE. When the data-generating model assumed parallel indicators, the doubly manifest model provided more accurate estimates than the model with parallel indicators and the model with congeneric indicators when the number of groups was small. However, with a large number of groups and a high ICC, the doubly manifest model, which does not take into account error in the predictor variable, was outperformed by the two latent models. For example, in the most favorable condition ($J = 100$, $n = 30$, ICC$_\eta = .3$) the relative RMSE was 0.44 for the doubly manifest model but 0.32 for the model with parallel indicators and 0.31 for the model with congeneric indicators. A very similar pattern of findings was observed when the data were generated under the assumption of congeneric indicators. However, as could be expected, the model with congeneric indicators provided more accurate estimates than the misspecified model with parallel indicators when the number of groups was large.

Finally, the doubly manifest model provided coverage rates that were clearly below 90% (range = 15.3%–87.6%) when the data-generating model assumed parallel indicators. By contrast, the model with parallel indicators and the model with congeneric indicators produced coverage rates that were close to the nominal 95% rate when the data were generated under the assumption of congeneric indicators. However, as could be expected, the model with congeneric indicators provided more accurate estimates than the misspecified model with parallel indicators when the number of groups was large.

In an additional analysis, we also investigated the statistical behavior of the estimates of the within-group effect (see Tables A1 and A2 in the supplemental materials). To summarize, the doubly manifest model provided substantially biased estimates, whereas the model with parallel indicators provided biased estimates only when the data were generated by the model with congeneric indicators. There was no bias for the estimates of the model with congeneric indicators. The doubly manifest estimates were also less accurate in terms of the RMSE than the multilevel latent contextual estimates. Finally, the doubly manifest model provided coverage rates that were too low, the model with congeneric indicators provided acceptable coverage rates, and the model with parallel indicators provided acceptable coverage rates only when the model with parallel indicators was the data-generating model. There were almost no differences between the ML and Bayesian approaches.
TABLE 3
Results of Study 1: Relative Bias, Relative Root Mean Squared Error (RMSE), and Coverage Rate From Estimating the Between-Group Effect of the Latent Predictor Variable With the Bayesian Estimation Approach

| No. of Groups | Group Size | Intraclass Correlation | Relative Bias | Relative RMSE | Coverage Rate |
|---------------|------------|------------------------|---------------|---------------|---------------|
|               |            |                        | DM            | PI            | CI            | DM            | PI            | CI            | DM            | PI            | CI            |
|               |            |                         |               |               |               |               |               |               |               |               |               |
| $J = 25$      | $n = 10$   | ICC$_n = .1$            | $-0.55$       | $-0.18$       | $-0.12$       | $0.73$        | $1.22$        | $1.29$        | $81.5$        | $96.4$        | $96.6$        |
|               |            | ICC$_n = .3$            | $-0.43$       | $-0.05$       | $-0.03$       | $0.53$        | $0.64$        | $0.68$        | $75.0$        | $97.4$        | $97.3$        |
|               | $n = 30$   | ICC$_n = .1$            | $-0.48$       | $-0.07$       | $-0.08$       | $0.71$        | $1.08$        | $1.08$        | $87.6$        | $96.2$        | $96.1$        |
|               |            | ICC$_n = .3$            | $-0.41$       | $0.04$        | $-0.01$       | $0.52$        | $0.60$        | $0.59$        | $80.0$        | $97.1$        | $96.8$        |
| $J = 100$     | $n = 10$   | ICC$_n = .1$            | $-0.56$       | $-0.02$       | $-0.04$       | $0.60$        | $0.72$        | $0.69$        | $28.4$        | $93.6$        | $94.4$        |
|               |            | ICC$_n = .3$            | $-0.44$       | $0.03$        | $0.01$        | $0.46$        | $0.38$        | $0.37$        | $15.3$        | $93.7$        | $93.7$        |
|               | $n = 30$   | ICC$_n = .1$            | $-0.48$       | $0.01$        | $-0.02$       | $0.54$        | $0.56$        | $0.55$        | $52.7$        | $95.3$        | $95.6$        |
|               |            | ICC$_n = .3$            | $-0.41$       | $0.07$        | $0.03$        | $0.44$        | $0.32$        | $0.31$        | $18.4$        | $93.4$        | $96.1$        |

Data-generating model: PI

|               |            |                        |               |               |               |               |               |               |               |               |               |
|               |            |                         |               |               |               |               |               |               |               |               |               |
| $J = 25$      | $n = 10$   | ICC$_n = .1$            | $-0.60$       | $-0.25$       | $-0.15$       | $0.76$        | $1.23$        | $1.28$        | $77.1$        | $97.8$        | $96.5$        |
|               |            | ICC$_n = .3$            | $-0.46$       | $0.00$        | $0.03$        | $0.55$        | $0.67$        | $0.71$        | $70.5$        | $99.2$        | $97.2$        |
|               | $n = 30$   | ICC$_n = .1$            | $-0.52$       | $-0.04$       | $-0.07$       | $0.74$        | $1.16$        | $1.15$        | $85.8$        | $96.9$        | $96.0$        |
|               |            | ICC$_n = .3$            | $-0.45$       | $0.12$        | $0.04$        | $0.53$        | $0.61$        | $0.58$        | $74.9$        | $98.6$        | $97.0$        |
| $J = 100$     | $n = 10$   | ICC$_n = .1$            | $-0.59$       | $0.04$        | $-0.01$       | $0.62$        | $0.75$        | $0.70$        | $20.3$        | $94.8$        | $95.8$        |
|               |            | ICC$_n = .3$            | $-0.47$       | $0.18$        | $0.05$        | $0.50$        | $0.46$        | $0.38$        | $9.4$         | $93.1$        | $93.6$        |
|               | $n = 30$   | ICC$_n = .1$            | $-0.53$       | $0.05$        | $-0.04$       | $0.59$        | $0.64$        | $0.57$        | $38.3$        | $94.1$        | $94.8$        |
|               |            | ICC$_n = .3$            | $-0.46$       | $0.22$        | $0.03$        | $0.48$        | $0.45$        | $0.31$        | $9.5$         | $88.1$        | $95.6$        |

Data-generating model: CI

Note: DM = doubly manifest model; PI = multilevel latent contextual model with parallel indicators; CI = multilevel latent contextual model with congeneric indicators; $J = $ number of groups; $n = $ group size; ICC$_n = $ intraclass correlation of the latent predictor.

We also compared the WinBUGS software with the Bayesian estimation procedure that is implemented in the popular SEM software Mplus. To this end, we conducted an additional simulation in both programs in which we examined the relative bias, relative RMSE, and coverage rate for the estimator of the between-group effect in the models with parallel and congeneric indicators. The results for WinBUGS and Mplus were almost identical (see Table A3 in the supplemental materials).

Discussion

The main findings from the simulation can be summarized as follows. First, in some of the conditions, serious estimation problems were observed when the ML approach was used to estimate the multilevel latent contextual model. Estimation problems were due to nonconverged solutions or inadmissible estimates. As expected, these estimation problems could be avoided by using a Bayesian estimation approach. Thus, from the perspective of an applied researcher, adopting a Bayesian approach seems to be particularly promising when the ML approach does not converge or converges to an inadmissible solution (see Depaoli & Clifton, 2015). Second, the results for the relative RMSE, which combines bias and variability, suggested that the Bayesian estimator of the between-group effect can be more accurate than the ML estimator, particularly when sample sizes are small and the ICC is low. However, these differences between the Bayesian estimator and the ML estimator vanished when the number of groups was large. Third, consistent with previous research, the estimates of the between-group effect provided by the doubly manifest model were more accurate than the estimates produced by the multilevel latent contextual model (bias-accuracy trade-off; Lüdtke et al., 2011; see also Ledgerwood & Shrout, 2011). There is a turning point when the approximately unbiased estimator of the latent model becomes more accurate than the estimator of the manifest model. Further analyses of the simulation data indicated that the Bayesian approach required less data (i.e., a smaller number of groups and a smaller group size).
to reach this turning point than the ML approach. For the model with congeneric indicators, Figure 2 shows that differences in the RMSE between the manifest and latent models depended on the estimation approach. For example, in the condition with a large number of groups (\(J = 100\)) and a small ICC (\(\text{ICC}_\eta = .1\)), the Bayesian estimator of the between-group effect had almost the same RMSE as the biased estimator of the doubly manifest model, whereas the ML estimator was clearly less accurate. To conclude, these findings suggest that the Bayesian approach offers a promising alternative to an ML approach. It reduces estimation problems in the multilevel latent contextual model and can also increase the accuracy in the estimation of group-level effects (see Depaoli & Clifton, 2015).

One limitation of the simulation was that it was restricted to only one predictor variable that was measured at L1. In the next section, we discuss the multilevel latent contextual model for the case of two predictor variables in which one predictor is directly measured at the group level.

**STUDY 2: ADDING A MANIFEST GROUP-LEVEL PREDICTOR VARIABLE**

MLM is often used to assess how variables directly measured at the group level (e.g., class size, treatment condition) affect outcomes at the individual level. In cluster or group randomized trials in educational research, classrooms or schools are randomly assigned to different treatment conditions (Aydin, Leite, & Algina, in press; Murray, 1998). MLM is then applied to estimate the effect of the intervention on individual student outcomes. In observational studies, researchers rely on the naturally occurring variation between groups to estimate the effect of directly measured group-level variables. For example, educational psychologists are interested in assessing the effect of school type (e.g., academic track schools vs. nonacademic track schools) on cognitive and noncognitive student outcomes. In this section, we investigate the multilevel latent contextual model with a directly measured group-level predictor and one latent predictor measured at L1.

**Multilevel Latent Contextual Model With a Manifest Group-Level Predictor Variable**

The multilevel latent contextual model with a manifest group-level variable is an extension of the model presented in Study 1. In addition to the latent predictor \(\eta_{1ij}\), which is measured by multiple indicators \(y_{kij} (k = 1, \ldots, p)\) and splits into a between-group part \(\eta_{B1j}\) and a within-group part \(\eta_{W1ij}\), the model assumes a manifest group-level predictor

![Figure 2](image-url)
\( y_{\text{dep},ij} = \eta_{ij}^y + \eta_{ij}^B = \alpha + \beta^B \eta_{ij}^W + \beta^B \eta_{ij}^B + \gamma x_j + \zeta_j + \zeta_{ij} \) (11)

where the regression coefficient \( \gamma \) indicates the effect of \( x_j \) on the dependent variable. The measurement model for the latent predictor \( \eta_{ij} \) is the same as in Equation 4 when an additional manifest group-level predictor is added. In the traditional MLM approach for estimating the effect of group-level predictors, the observed group mean \( \bar{y}_{ij} \) would be used instead of \( \eta_{ij} \) and the group-mean-centered individual scores \( \hat{y}_{ij} - \bar{y}_{ij} \) would be used instead of \( \eta_{ij}^W \). The traditional multilevel model would then be specified as follows:

\[
y_{\text{dep},ij} = \alpha_{DM} + \beta^B_{DM}(\bar{y}_{ij} - \bar{y}_{*}) + \beta^B_{DM}\bar{y}_{*} + \gamma_{DM}x_j + \zeta_j + \zeta_{ij}.
\] (12)

It can now be shown that \( \beta^B_{DM} \), \( \hat{\beta}^B_{DM} \), and \( \hat{\gamma}_{DM} \) are biased estimators of the corresponding regression coefficients \( \beta^W \), \( \beta^B \), and \( \gamma \) in the multilevel latent contextual model presented in Equation 11 (for the analytical derivation, see Appendix). For the between-group effect of the predictor that is measured at L1, the bias (assuming that \( \eta_{ij} \) and \( x_j \) are standardized) is given by:

\[
\text{Bias}(\hat{\beta}^B_{DM}) = \frac{1}{n} \left[ 1 - \frac{1 - \text{ICC}_\eta}{\text{ICC}_\eta} \cdot \frac{\text{Rel}(\hat{y}_{*})}{1 - \rho^2_{\eta} \cdot \text{Rel}(\hat{y}_{*})} \right] \cdot \beta^B
\] (13)

where \( \rho_{\eta} \) indicates the correlation between the between-group component of the latent predictor and the manifest group-level predictor. As can be seen, when the two predictors are uncorrelated at the group level \( (\rho_{\eta} = 0) \), the expression for the bias of the between-group coefficient simplifies to the case with only one predictor variable measured at L1 (see Equation 10). However, with correlated predictors \( (\rho_{\eta} \neq 0) \), the bias becomes a function of the reliability of the observed group mean, the ICC of the latent predictor, and the correlation of the two predictors at the group level. Note that with correlated predictors, the bias of the estimator \( \hat{\beta}^B_{DM} \) will be even stronger than in the case with uncorrelated predictors.

It is interesting to note that the doubly manifest model can also provide a biased estimator of the effect of the observed group-level predictor that is assumed to be measured without error. It can be shown that \( \hat{\gamma}_{DM} \) is a biased estimator of \( \gamma \):

\[
\text{Bias}(\hat{\gamma}_{DM}) = -\rho_{\eta} \cdot \sqrt{\text{ICC}_\eta} \cdot \text{Bias}(\hat{\beta}^B_{DM}).
\] (14)

This relationship indicates that the bias depends on the correlation between the predictors at the group level as well as on the ICC of the latent predictor and the bias of the estimator of the between-group effect. As can be seen, when there are uncorrelated predictors at the group level \( (\rho_{\eta} = 0) \), \( \hat{\gamma}_{DM} \) is an unbiased estimator of \( \gamma \). However, when there are positively correlated predictors at the group level \( (\rho_{\eta} > 0) \), \( \hat{\gamma}_{DM} \) will be positively biased when a positive-between-group effect exists \( (\beta^B > 0) \). For the case where \( \beta^W = 0.2, \beta^B = 0.5, \gamma = .2, \) and \( \rho_{\eta} = .3 \), Figure 3 shows the expected bias as a function of the ICC of the latent predictor, the group size, and the number of indicators (again assuming that the standardized loadings are .7 at L2). As can be seen, the bias grows larger as the ICC increases. However, increasing the group size or the number of indicators has a negative effect on the bias. Thus, in the general condition in which the predictor variables are

FIGURE 3 Theoretically expected bias from using the doubly manifest model to estimate the effect of the manifest group-level predictor variable (\( \gamma \)). \( n = \) group size; \( \text{ICC}_\eta = \) intraclass correlation of the latent predictor; \( p = \) number of indicators of the latent predictor.
correlated, all regression coefficients in the doubly manifest model in Equation 12 will be biased if the predictor measured at L1 is affected by sampling and measurement error.

**Specification of Prior Distributions**

As in the previous section, we suggest a Bayesian approach for estimating the multilevel latent contextual model with an additional manifest group-level predictor. In this model, we assume that the between-group component of the latent predictor and the manifest group-level predictor are multivariate normally distributed. For these additional model parameters, appropriate prior distributions have to be selected in the Bayesian approach. More specifically, we specified an uninformative normal distribution for the mean of the manifest group-level predictor: \( \mu_{\eta_i} \sim N(0, 1000) \); and a somewhat informative IG distribution for its variance: \( \sigma_{\eta_i}^2 \sim IG(.1, .1) \). Moreover, as suggested by Lunn, Jackson, Best, Thomas, and Spiegelhalter (2013; see also Liu, Zhang, & Grimm, 2016), we used an uninformative uniform distribution ranging from \(-1\) to \(1\) as a prior distribution for the correlation between \( \eta_{ij} \) and \( x_i \): \( \rho_{\eta x} \sim U(-1, 1) \). The prior distributions for all other parameters were specified as in Study 1.

**Simulation Study**

The simulation had two goals. First, the simulation was expected to confirm the results of the analytical derivation that showed that the doubly manifest model can provide substantially biased estimates of the effect of a manifest group-level predictor when the error in a predictor variable measured at L1 is not taken into account. Second, we compared the ML approach with the Bayesian approach and assumed that the advantages of the Bayesian approach for estimating the multilevel latent contextual model would also transfer to the situation in which an additional manifest group-level variable was added to the model.

**Method**

The data-generating model was a multilevel latent contextual model with parallel indicators and a manifest group-level predictor (see Equation 11). The manifest group-level predictor was standardized at the group level, and its regression coefficient was set to 0.2. The within-group and between-group loadings of the latent predictor were all set to 1 and were not varied further in the simulation. The number of groups was set to either 50 or 200 with a group size of either 5 or 20 persons each. The ICC of the latent predictor was \( 0.05 \) or \( 0.2 \). The very low value (0.05) was chosen to include a condition that is expected to be particularly problematic for the ML approach. The correlation between the between-group component of the latent predictor (\( \eta_{ij} \)) and the manifest group-level predictor (\( x_i \)) was set to zero (uncorrelated) or 0.3 (correlated; see Croon & Van Veldhoven, 2007). For each of the \( 2 \times 2 \times 2 = 16 \) conditions, 1,000 data sets were generated. Each data set was analyzed with the doubly manifest model (see Equation 12) and the multilevel latent contextual model with parallel indicators. Again, the PSR and the MCMCSE ratio were used to assess the convergence behavior of the MCMC algorithm in the different conditions.\(^5\)

**Results**

**Estimation problems.** First, we evaluated potential estimation problems for the ML approach (see also Table B.1 in the supplemental materials). Again, the doubly manifest model did not show any estimation problems. However, for the multilevel latent contextual model with parallel indicators, estimation problems were observed under some of the conditions. For example, for the most problematic condition (\( J = 50, n = 5, ICC_{\eta} = .05, \rho_{\eta x} = .3 \)), only \( 69.6\% \) of the solutions were classified as admissible converged. The percentage of solutions with estimation problems decreased as the ICC of the latent predictor increased and as the group size and number of groups increased. In the most favorable condition (\( J = 200, n = 20, ICC_{\eta} = .20 \)), no estimation problems were identified for the latent model. The occurrence of estimation problems was not affected by whether correlated or uncorrelated predictors were specified in the data-generating model.

**Comparison of the ML and Bayesian approaches.** Table 4 shows the relative bias, RMSE, and coverage rate for the estimator of the effect of the manifest group-level predictor provided by the ML and Bayesian approaches. When uncorrelated predictors were specified in the data-generating model, both the ML approach and the Bayesian approach provided approximately unbiased estimates. This was also true when the predictors were correlated, with the exception of the condition with a small number of groups, a small group size, and a low ICC, in which both approaches were slightly positively biased (ML = 0.10, Bayes = 0.12). Next, an inspection of the relative RMSE showed that the Bayesian approach (\( M = 0.35, range = 0.19–0.53 \)) provided more accurate estimates of the effect of the manifest group-level predictor than the ML approach did (\( M = 0.47, range = 0.20–1.05 \)). Especially in conditions with a small number of groups and a low ICC, the Bayesian approach clearly outperformed the ML approach. In addition, the difference between the ML approach and the Bayesian approach in

\(^5\) The chain length that was used to approximate the posterior distributions varied substantially. In the worst case scenario (\( J = 50, n = 5, ICC_{\eta} = .05 \)), the model with parallel indicators required a chain length of 90,000 iterations. However, in the most favorable case (\( J = 200, n = 20, ICC_{\eta} = .3 \)), the doubly manifest model required only 900 iterations.
TABLE 4
Results of Study 2: Relative Bias, Relative Root Mean Squared Error (RMSE), and Coverage Rate From Estimating the Effect of the Manifest Group-Level Predictor Variable With the Maximum Likelihood Estimation Approach (ML) and the Bayesian Estimation Approach (Bayes)

| No. of Groups | Group Size | Intraclass Correlation | Admissible Converged | Relative Bias ML | Relative Bias Bayes | Relative RMSE ML | Relative RMSE Bayes | Coverage Rate ML | Coverage Rate Bayes |
|---------------|------------|------------------------|----------------------|------------------|---------------------|------------------|---------------------|------------------|---------------------|
| \( J = 50 \)  | \( n = 5 \) | ICC\(_n\) = .05         | 70.5                 | 0.07             | 0.04                | 0.85             | 0.51                | 96.2             | 99.6               |
|               |            | ICC\(_n\) = .20        | 84.4                 | 0.01             | 0.00                | 0.56             | 0.47                | 94.9             | 98.5               |
| \( n = 20 \)  |            | ICC\(_n\) = .05        | 92.2                 | 0.00             | 0.00                | 0.50             | 0.42                | 94.7             | 97.8               |
| \( n = 20 \)  |            | ICC\(_n\) = .20        | 99.0                 | −0.01            | −0.01               | 0.40             | 0.39                | 94.3             | 96.9               |
| \( J = 200 \) | \( n = 5 \) | ICC\(_n\) = .05        | 87.8                 | −0.02            | −0.01               | 0.35             | 0.25                | 96.5             | 98.2               |
|               |            | ICC\(_n\) = .20        | 99.2                 | 0.00             | 0.00                | 0.23             | 0.23                | 95.8             | 96.6               |
| \( n = 20 \)  |            | ICC\(_n\) = .05        | 99.9                 | −0.01            | −0.01               | 0.20             | 0.19                | 95.5             | 95.6               |
|               |            | ICC\(_n\) = .20        | 100.0                | 0.01             | 0.01                | 0.20             | 0.20                | 93.8             | 94.5               |

Data-generating model: Correlated predictor variables

| \( J = 50 \)  | \( n = 5 \) | ICC\(_n\) = .05         | 69.6                 | 0.10             | 0.12                | 1.05             | 0.53                | 97.0             | 99.4               |
|               |            | ICC\(_n\) = .20        | 81.3                 | −0.01            | 0.07                | 0.64             | 0.49                | 94.8             | 99.9               |
| \( n = 20 \)  |            | ICC\(_n\) = .05        | 90.4                 | −0.03            | 0.03                | 0.65             | 0.44                | 95.8             | 98.0               |
| \( n = 20 \)  |            | ICC\(_n\) = .20        | 98.4                 | −0.07            | −0.04               | 0.47             | 0.42                | 95.1             | 96.6               |
| \( J = 200 \) | \( n = 5 \) | ICC\(_n\) = .05        | 87.2                 | −0.06            | 0.04                | 0.61             | 0.28                | 97.5             | 97.4               |
|               |            | ICC\(_n\) = .20        | 99.0                 | −0.02            | −0.01               | 0.29             | 0.27                | 96.2             | 96.8               |
| \( n = 20 \)  |            | ICC\(_n\) = .05        | 99.7                 | −0.03            | −0.01               | 0.28             | 0.22                | 95.9             | 95.5               |
|               |            | ICC\(_n\) = .20        | 100.0                | 0.00             | −0.01               | 0.21             | 0.21                | 95.5             | 95.1               |

Note. Results are based on only those data sets for which ML provided an admissible converged solution. \( J = \) number of groups; \( n = \) group size; ICC\(_n\) = intraclass correlation of the latent predictor.

Comparisons of different analysis models within Bayes. In the next step, we compared the doubly manifest model with the model with parallel indicators. As expected by the analytical derivation, the doubly manifest model provided approximately unbiased estimates of the effect of the manifest group-level predictor in the case of uncorrelated predictors (see Table 5). The same was true for the model with parallel indicators. However, with correlated predictors, the doubly manifest model produced positively biased estimates, whereas the estimates provided by the model with parallel indicators were still approximately unbiased, with the exception of a slight positive bias (0.11) in one extreme condition (\( J = 50, n = 5, ICC\(_n\) = .05\)).

With regard to the relative RMSE, the doubly manifest model (\( M = 0.33, \) range = 0.19–0.49) provided slightly more accurate estimates than the model with parallel indicators (\( M = 0.35, \) range = 0.19–0.54), particularly when the number of groups or the group size was small and the ICC was low. However, with increasing information about the aggregated latent predictor (i.e., large number of groups, large group size), the difference between the doubly manifest model and the model with parallel indicators decreased. In one of the most favorable conditions (\( J = 200, n = 20, ICC\(_n\) = .20, \rho_{\theta\chi} = .3\)), the model with parallel indicators even outperformed the doubly manifest model in terms of the relative RMSE. Furthermore, the difference between the models was more pronounced when the predictors were correlated in the population.

Finally, in the case of uncorrelated predictors, both the doubly manifest model (range = 94.1%–96.1%) and the model with parallel indicators (range = 94.5%–99.6%) provided acceptable coverage rates such that the model with parallel indicators had a tendency to provide higher coverage terms of the RMSE was larger when correlated predictors were specified in the data-generating model. However, it is interesting that even in conditions with uncorrelated predictors in the population, the Bayesian approach was more accurate than the ML approach. For example, when the number of groups was small, the group size was small, and the ICC was low, the relative RMSE was 0.85 for the ML approach and only .51 for the Bayesian approach. One possible explanation for these gains in accuracy is that in conditions with small groups and low ICCs, the observed correlations between the two predictors at the group level might deviate substantially from zero in the sample, and the Bayesian approach provides more stable estimates of these associations at the group level. This will in turn also lead to more stable estimates of the effect of the manifest group-level predictor.

Both the ML approach (range = 93.8%–97.5%) and the Bayesian approach (range = 94.5%–99.6%) provided acceptable coverage rates, with the exception that the coverage rate was somewhat too high (99.6%) for the Bayesian approach in the condition with an ICC of .05 and 50 groups with \( n = 5 \). The results for the between-group effect of the latent predictor produced by the ML and Bayesian approaches are in line with the findings presented in the previous simulation (see Tables B2 and B3 in the supplemental materials).
rates than the doubly manifest model in conditions with a small number of groups. The findings for the coverage rates were similar when the predictors were correlated, except that in conditions with a large number of groups \( J = 200 \), the doubly manifest model provided coverage rates that were too low.

**Discussion**

Overall, this simulation confirmed the results of the analytical derivation that showed that the doubly manifest model can produce biased estimates of group-level effects even when the group-level predictor is measured directly at the group level without error. In addition, the simulation was in line with the previous simulation, which showed that the Bayesian approach for estimating a multilevel latent contextual model provides more accurate estimates than the ML approach, particularly when the sample sizes were small and the ICC was low. It is interesting that these gains in accuracy offered by the Bayesian approach were also found for the effect of the manifest group-level predictor when the predictors were assumed to be uncorrelated in the population. The next section illustrates the different estimation approaches with a real data example from educational psychology.

**ILLUSTRATIVE EXAMPLE: EFFECT OF TEACHER BEHAVIOR ON STUDENTS’ ACHIEVEMENT**

Previous research in educational psychology has supported the notion that students’ learning outcomes are affected by teacher behavior. One important specific teacher characteristic is how easily teachers get distracted by their students. In the Third International Mathematics and Science Study (Baumert et al., 1997), students were asked to rate their mathematics teacher’s distractibility on three items (e.g., “Our mathematics teacher is easily distracted if something attracts his/her attention”; see Kounin, 1970). The data set stemmed from the German sample of lower secondary school students and contained observations from 2,131 students distributed across 108 classes with an average class size of 19.73. The ICCs of the student ratings for the three items were 0.03, 0.04, and 0.09. In line with previous research, we expected that student perceptions of their mathematics teacher’s distractibility would be negatively associated with students’ mathematics achievement.

We specified two multilevel latent contextual models in which the latent predictor variable (teacher distractibility) was measured with three indicators at the student level. In the first model, the loading of the first indicator was fixed to 1 at L1 and L2. The other two loadings were freely estimated but constrained to be invariant across levels (cross-level invariance). In the second model, the constraint of invariant loadings across levels was relaxed, and the second and third loadings were freely estimated at L1 and L2. For comparison purposes, we also estimated the doubly manifest model, which ignores error in the predictor variable.

The models were estimated with both the ML and Bayesian approaches (see the supplemental materials for the WinBUGS code). For the ML approach, we used Mplus 7.1. For the Bayesian approach, we specified a burn-in period of 50,000 iterations. Each fifth iteration of the subsequent 450,000 iterations was saved, which resulted in 90,000 draws from the posterior distribution. For each model and model parameter, the PSR was 1 and the MCMCSE ratio was well below .05, indicating that WinBUGS approximated the posterior distributions very well (see Hoff, 2009; B. O. Muthén & Asparouhov, 2012; see also Cho & Bottge, 2015). We selected the MAP as the Bayesian point estimate and the PSD as the Bayesian standard error.

Results for the doubly manifest model, the model with invariant loadings, and the model with noninvariant loadings are presented in Table 6. In all three models, there was no effect of distractibility on mathematics achievement at the student level. However, student perceptions of their mathematics teacher’s distractibility had a negative effect at the class level, indicating that students in classes with a more distractible mathematics teacher achieved less in mathematics than did those with a less distractible teacher. As expected from the results of the analytical derivation in Equation 10, the estimate of the class-level
TABLE 6
Results of the Third International Mathematics and Science Study: Maximum-Likelihood (ML) and Bayesian Estimates (Bayes) of the Effects of Perceived Teacher Distractibility on Mathematics Achievement

| Parameter | DM Model | CI Model | CNI Model |
|-----------|----------|----------|-----------|
|           | ML       | Bayes    | ML        | Bayes    | ML        | Bayes    |
|           | Est.     | SE       | Est.      | SE       | Est.      | SE       |
| Loadings  |          |          |           |          |           |          |
| Indicator 1 | 1.000    | —        | 1.000     | —        | 1.000     | —        |
| Indicator 2 | 1.139    | 0.076    | 1.143     | 0.059    | 1.222     | 0.086    | 1.235     | 0.069    |
| Indicator 3 | 0.936    | 0.059    | 0.930     | 0.047    | 0.934     | 0.060    | 0.947     | 0.049    |
| Error Variances |          |          |           |          |           |          |
| Indicator 1 | 0.447    | 0.027    | 0.449     | 0.020    | 0.459     | 0.027    | 0.461     | 0.020    |
| Indicator 2 | 0.273    | 0.024    | 0.268     | 0.021    | 0.249     | 0.027    | 0.247     | 0.023    |
| Indicator 3 | 0.455    | 0.025    | 0.453     | 0.019    | 0.463     | 0.025    | 0.458     | 0.019    |
| Within-Group Coefficient | −0.037 | 0.022    | −0.038    | 0.023    | −0.044    | 0.032    | −0.046    | 0.033    | −0.041    | 0.032    | −0.038    | 0.034    |
| Residual Variance | 0.475    | 0.034    | 0.474     | 0.015    | 0.475     | 0.034    | 0.475     | 0.015    |          |          |          |          |

Note. Fitting the multilevel latent contextual model with congeneric noninvariant indicators with ML had led to an inadmissible solution as indicated by a negative between-group error variance. Therefore, we fixed that variance to zero. DM = doubly manifest model; CI = multilevel latent contextual model with congeneric indicators; CNI = multilevel latent contextual model with congeneric noninvariant indicators; Est. = estimate; SE = standard error.

We presented an application of the Bayesian approach to multilevel SEM, focusing on (a) multilevel contextual models (Study 1), and (b) multilevel models with multiple predictors at L2 (Study 2). In Study 1, we discussed the multilevel latent contextual model, which is often considered the true model in contextual studies, and presented a Bayesian estimation approach. In Study 2, we extended the model to include an additional observed group-level predictor and analytically derived the bias of that coefficient under the traditional MLM approach (i.e., the doubly manifest approach). In two simulation studies, we showed that with a Bayesian approach, estimation of multilevel SEMs could be stabilized (fewer estimation problems), and the accuracy of the estimates of group-level effects could be substantially improved. Finally, we presented an example from educational psychology to show how the Bayesian approach could be specified in WinBUGS and Mplus.

GENERAL DISCUSSION

...effect provided by the doubly manifest model was substantially smaller than those provided by the multilevel latent contextual models. It is interesting to note that fitting the model with noninvariant loadings with ML led to an inadmissible solution as indicated by a negative error variance at L2. To avoid estimation problems, we fixed that variance to zero. As expected, no inadmissible estimates were observed with the Bayesian approach. However, given that the data provided a relatively large amount of information (i.e., relatively large sample sizes), the differences between the Bayesian and ML approaches were small...
However, the following limitations of this study should be mentioned. First, as with most simulation studies, it is difficult to generalize the findings beyond the conditions that we manipulated in our studies. Second, as advised by Zitzmann et al. (2015), we stabilized the group-level effects by adding very little information to the prior distributions for the variances at Level 2. The priors for other model parameters were left uninformative. However, Depaoli and Clifton (2015) showed that the accuracy of the estimate of the between-group effect could be improved even more by incorporating previous knowledge about that parameter. For example, a weakly informative normal distribution could be selected for the between-group effect. Moreover, in Study 2 we specified an uninformative uniform prior ranging from -1 to 1 for the correlation between the between-group component of the latent predictor and the observed group-level predictor. Alternatively, a multivariate inverse Wishart (IW) distribution could have been specified for the covariance matrix (e.g., Gelman & Hill, 2007; B. O. Muthén & Asparouhov, 2012; see Daniels, 1999, for a multivariate alternative to the IW distribution). It would be interesting to see whether the findings would change with an IW prior distribution. Third, we used the Mplus default specification of ML in the comparison with the Bayesian approach. However, estimation problems are expected to be reduced when variances are constrained to be positive, negative variances are fixed to zero, more iterations are run, or better starting values are used. Fourth, only relatively simple multilevel SEMs were discussed. Future research should investigate the potential of a Bayesian approach for stabilizing the estimates of more complex multilevel SEMs with randomly varying slopes and cross-level interactions (see Preacher, Zhang, & Zyphur, 2016). It could be assumed that the Bayesian approach will become even more advantageous for estimating these models. Finally, it should also be noted that the ideas that motivated the use of the Bayesian approach in this investigation can also be found in likelihood-based approaches. For example, penalized likelihood methods (e.g., Chung et al., 2013; DeCarlo et al., 2011; Galindo-Garre, Vermunt, & Bergsma, 2004; McNeish, 2015; Mislevy, 1986) employ weakly informative priors for variance components to avoid boundary estimates (e.g., zero variances; Chung et al., 2013).

To conclude, this article showed that applying the Bayesian approach to multilevel SEM could be a promising alternative to an ML approach, particularly when the data contain only a little information about the group-level variables (e.g., small group sizes, low ICCs). General-purpose Bayesian software (e.g., WinBUGS) as well as software that was designed specifically for latent variable models (e.g., Mplus) can be used to specify the proposed approach.

FUNDING

This research was supported by Grant LU 1636/1-1 from the German Research Foundation (DFG).

SUPPLEMENTAL MATERIAL

Supplemental data for this article can be accessed at www.tandfonline.com/hsem.

REFERENCES

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. Psychometrika, 65, 475–496. doi:10.1007/BF02296339

Arminger, G., & Muthén, B. O. (1998). A Bayesian approach to nonlinear latent variable models using the Gibbs sampler and the Metropolis–Hastings algorithm. Psychometrika, 63, 271–300. doi:10.1007/BF02294856

Asparouhov, T., & Muthén, B. O. (2007). Constructing covariates in multilevel regression (Mplus Web Notes no. 11, version 2). Retrieved from http://www.statmodel.com/download/webnotes/webnotell.pdf

Aydin, B., Leite, W. L., & Algina, J. (in press). The effects of including observed means or latent means as covariates in multilevel models for cluster randomized trials. Educational and Psychological Measurement. doi:10.1177/0013164415618705

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. Psychological Methods, 18, 151–164. doi:10.1037/a0030642

Baumert, J., Lehmann, R. H., Lehrke, M., Schnitz, B., Clausen, M., Hosenfeld, I., . . . Neubrand, J. (1997). TIMSS: Mathematisch-naturwissenschaftlicher Unterricht im internationalen Vergleich: Deskriptive Befunde [TIMSS: Mathematics and science instruction in an international comparison]. Opladen, Germany: Leske & Budrich.

Box, G. E. P., & Tiao, G. C. (1992). Bayesian inference in statistical analysis. New York, NY: Wiley.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis, 1, 513–541. doi:10.1214/06-BA117

Buonaccorsi, J. P. (2010). Measurement error: Models, methods, and applications. New York, NY: CRC.

Can, S., van de Schoot, R., & Hox, J. J. (2015). Collinear latent variables in multilevel confirmatory factor analysis: A comparison of maximum likelihood and Bayesian estimations. Educational and Psychological Measurement, 75, 406–427. doi:10.1177/0013164415457959

Cho, S.-J., & Botte, B. A. (2015). Multilevel multidimensional item response model with a multilevel latent covariate. British Journal of Mathematical and Statistical Psychology, 68, 410–433. doi:10.1111/bmsp.12051

Christensen, R., Johnson, W., Branscum, A., & Hanson, T. E. (2011). Bayesian ideas and data analysis: An introduction for scientists and statisticians. Boca Raton, FL: CRC.

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly informative prior for point estimation of covariance matrices in hierarchical models. Journal of Educational and Behavioral Statistics, 40, 136–157. doi:10.3102/1076998615570945

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. Psychometrika, 78, 685–709. doi:10.1007/s11336-013-9328-2

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences. Mahwah, NJ: Erlbaum.

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. Psychological Methods, 12, 45–57. doi:10.1037/1082-989X.12.1.45
Lee, S.-Y., & Poon, W.-Y. (1998). Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica*, 8, 749–766.

Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling*, 20, 241–264. doi:10.1080/10705511.2013.769391.

Liu, H., Zhang, Z., & Grimm, K. J. (2016). Comparison of inverse Wishart and separation-strategy priors for Bayesian estimation of covariance parameter matrix in growth curve analysis. *Structural Equation Modeling*, 23, 354–367. doi:10.1080/10705511.2015.1057285

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, 16, 444–467. doi:10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. O. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203–229. doi:10.1037/a0012869

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2013). *The BUGS book*. Boca Raton, FL: CRC.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). *WinBUGS*—A Bayesian modelling framework: Concepts, structure, and flexibility. *Statistics and Computing*, 10, 325–337. doi:10.1023/A:1008925926011

Lynch, S. M. (2007). *Introduction to applied Bayesian statistics and estimation for social scientists*. New York, NY: Springer.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33, 181–220. doi:10.1207/S15327906mbr3302_1

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist*, 47, 106–124. doi:10.1080/00220267.2012.670488

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T. Muthén, B. O., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44, 764–802. doi:10.1080/00273171903336656

McDonald, R. R. (1993). A general model for 2-level data with responses missing at random. *Psychometrika*, 58, 575–585. doi:10.1007/BF02294828

McNeish, D. M. (2015). Using lasso for prediction selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, 50, 471–483. doi:10.1080/00273171.2015.1036965

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel factorial score models. In J. J. Hox & J. K. Roberts (Eds.), *Introducing multilevel modeling*: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Introducing multilevel modeling*: Multilevel analysis in a general latent variable framework. New York, NY: Routledge.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177–195. doi:10.1007/BF02293979

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376–398. doi:10.1177/004912419402003006

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel methodology*. Newbury Park, CA: Sage.

Nagin, D. (1999). *Group-based modeling of development*. Cambridge, MA: Harvard University Press.

Nagin, D. S. (1999). Group-based modeling of development. *Cambridge, MA: Harvard University Press.*

Nagin, D. S. (1999). Group-based modeling of development. *Cambridge, MA: Harvard University Press.*

Nagin, D. S. (1999). Group-based modeling of development. *Cambridge, MA: Harvard University Press.*
expected biases associated with adopting the doubly manifest approach were derived elsewhere (e.g., Lüdtke et al., 2011). Therefore, in the following, we consider the case of an additional observed group-level predictor (see also Study 2). If we assume that the indicators $y_{ij}$ of the latent predictor $\eta_1$, the manifest dependent variable $\eta_2$, and the manifest group-level predictor $x$ are centered (i.e., zero mean), then all intercepts become 0. Moreover, if we assume that the $y_{ij}$ are parallel, then the loadings become 1, and the error variances no longer depend on $k$ (i.e., $\text{Var}(\epsilon_{kij}) = \theta^2$, $\text{Var}(\epsilon_{ij}) = \theta^2$). More formally, the measurement model for the latent predictor becomes:

$$y_{ij} = \eta_{1ij} + \eta_{wij} + \epsilon_{ij} + \epsilon_{wij} \quad (A1)$$

and the structural model becomes:

$$y_{\text{dep},ij} = \eta_{1ij}^W + \eta_{1ij}^B + \gamma_{xy} + c^B + c^W \quad (A2)$$

Let the within- and the between-group components $\eta_{1ij}^W$ and $\eta_{1ij}^B$ of the latent predictor $\eta_1$ and the observed group-level predictor $x$ be summarized by the vector $\eta = (\eta_{1ij}^W, \eta_{1ij}^B, x)^T$. Then the vector of the coefficients $\beta = (\beta^W, \beta^B, \gamma)^T$ can be calculated as:

$$\beta = \Sigma_{\eta_1 \eta}^{-1} \Sigma_{\eta_{\text{dep}} \eta_1}$$

(A3)

where:

$$\Sigma_{\eta_1 \eta_1} = \begin{pmatrix} \sigma_{\eta_{1ij}^W} & 0 & 0 \\ 0 & \sigma_{\eta_{1ij}^B} & \sigma_{\eta_{1ij}^B x} \\ 0 & \sigma_{\eta_{1ij}^B x} & \sigma_{x}^2 \end{pmatrix}$$

and

$$\Sigma_{\eta_{\text{dep}} \eta_1} = \begin{pmatrix} \sigma_{\eta_{1ij}^W} & \sigma_{\eta_{1ij}^W \eta_{1ij}^B} & \sigma_{\eta_{1ij}^W x} \\ \sigma_{\eta_{1ij}^W \eta_{1ij}^B} & \sigma_{\eta_{1ij}^B} & \sigma_{\eta_{1ij}^B x} \\ \sigma_{\eta_{1ij}^W x} & \sigma_{\eta_{1ij}^B x} & \sigma_{x}^2 \end{pmatrix}$$

(A4)

$\Sigma_{\eta_1 \eta_1}$ is the covariance matrix of the predictors, and $\Sigma_{\eta_{\text{dep}} \eta_1}$ is the covariance matrix of the outcome variable and the predictors, respectively. In the doubly manifest approach, the group-mean-centered manifest scale score $\bar{y}_{\text{dep},ij} - \bar{y}_{\text{dep}}$ and the observed group mean $\bar{y}_{\text{dep},ij}$ replace the within- and the between-group components $\eta_{1ij}^W$ and $\eta_{1ij}^B$ of the latent predictor $\eta_1$, respectively:

$$y_{\text{dep},ij} = \beta^W_{\text{DM}} \left( \bar{y}_{\text{dep},ij} - \bar{y}_{\text{dep}} \right) + \beta^B_{\text{DM}} \bar{y}_{\text{dep}} + \gamma_{\text{DM}} x_j + c^B + c^W \quad (A5)$$

If we let the vector $y$ be defined as $y = (\bar{y}_{\text{dep},ij}, \bar{y}_{\text{dep}}, \bar{y}_{\text{dep},ij}, x)^T$, and we assume equal group sizes (i.e., omitting the index $j$ in $n_j$), then the covariance matrix $\Sigma_{\text{dep},y}$ is given by:

$$\Sigma_{\text{dep},y} = (I + C) \Sigma_{\text{dep},\eta_1}$$

(A6)

where $I$ is the identity matrix and $C$ is:

---

**APPENDIX**

**DERIVATION OF BIAS FOR DOUBLY MANIFEST APPROACH**

Assuming that the multilevel latent contextual model in Equations 4 and 6 holds in the population, the theoretically
By inserting Equation A12 into Equation A10, we obtain

\[
\beta = \left( -\frac{1}{n} \ 0 \ 0 \right). \tag{A7}
\]

The covariance matrix \( \Sigma_{yy} \) is given by:

\[
\Sigma_{yy} = \begin{pmatrix}
\frac{1}{n} + \frac{1}{p} & 0 & 0 \\
0 & \psi^y + \frac{1}{p} \psi^w + \frac{1}{p} \sigma_\eta^2 & \frac{1}{p} \\
0 & \frac{1}{p} \sigma_\eta^2 & \sigma_\eta^2
\end{pmatrix}.
\]

The manifest regression coefficients \( \beta_{DM} = (\beta^W, \beta^B, \gamma_{DM})^T \) are obtained as an ordinary least squares estimate:

\[
\beta_{DM} = \Sigma_{y'y}^{-1} \Sigma_{y'a} \beta = \Sigma_{y'y}^{-1} (I + C) \Sigma_{y'y} \beta.
\]

Let the estimates of \( \beta_{DM} \) be summarized by the vector \( \hat{\beta}_{DM} = (\hat{\beta}^W, \hat{\beta}^B, \hat{\gamma}_{DM})^T \). Then its bias with respect to \( \beta \) is given as the difference between the expectation of \( \beta_{DM} \) and \( \beta \):

\[
\text{Bias}(\hat{\beta}_{DM}) = E(\hat{\beta}_{DM}) - \beta = \Sigma_{y'y}^{-1} (I + C) \Sigma_{y'y} \beta - \beta = \left\{ \Sigma_{y'y}^{-1} (I + C) \Sigma_{y'y} \beta - I \right\} \beta. \tag{A9}
\]

For ease of notation, let:

\[
s^2 = \frac{n - 1}{n} \psi^W + \frac{n - 1}{pn} \psi^W \quad \text{and} \quad
\]

\[
t^2 = \psi^y + \frac{1}{p} \psi^w + \frac{1}{p} \psi^w + \frac{1}{p} \sigma_\eta^2
\]

be the variances of \( \bar{y}_{*ij} - \bar{y}_{*i} \) and \( \bar{y}_{*ij} \), respectively. Then the inverse of the manifest covariance matrix \( \Sigma_{y'y}^{-1} \) is given by:

\[
\Sigma_{y'y}^{-1} = \begin{pmatrix}
\frac{1}{t^2} & 0 & 0 \\
0 & \frac{\sigma^2_{\psi^W} / t^2}{\sigma^2_{\psi^W} / t^2 - \sigma^2_{\psi^W} / t^2} & 0 \\
0 & 0 & \frac{1}{\sigma^2_{\psi^W} / t^2 - \sigma^2_{\psi^W} / t^2}
\end{pmatrix}.
\]

By inserting Equation A12 into Equation A10, we obtain the following biases:

\[
\text{Bias}(\hat{\beta}^W_{DM}) = \left\{ \frac{n - 1}{n} \psi^W - 1 \right\} \beta^W
\]

\[
\text{Bias}(\hat{\beta}^B_{DM}) = \frac{1}{n} \psi^W \sigma^2_{\psi^W} / \psi^W - \frac{\sigma^2_{\psi^W} / t^2}{\sigma^2_{\psi^W} / t^2 - \sigma^2_{\psi^W} / t^2} \beta^B
\]

\[
\text{Bias}(\hat{\gamma}_{DM}) = -\sigma_{\eta^2} \left\{ \frac{1}{n} \psi^W \sigma^2_{\eta^2} / \psi^W - \frac{1}{\sigma^2_{\psi^W} / t^2 - \sigma^2_{\psi^W} / t^2} \beta^B \right\}. \tag{A13}
\]

To facilitate interpretation, we insert the terms:

\[
\text{Rel}(\bar{y}_{*ij} - \bar{y}_{*i}) = \frac{\psi^W_{ij}}{s^2}, \quad \text{Rel}(\bar{y}_{*i}) = \psi^W_{i} / t^2.
\]

and \( \text{ICC} = \frac{\psi^W_i}{\psi^W_i + \psi^W_{ij}} \). into Equation A13:

\[
\text{Bias}(\hat{\beta}^W_{DM}) = \left\{ \frac{n - 1}{n} \text{Rel}(\bar{y}_{*ij} - \bar{y}_{*i}) - 1 \right\} \beta^W
\]

\[
\text{Bias}(\hat{\beta}^B_{DM}) = \frac{1}{n} \text{ICC} \frac{\sigma^2_{\eta^2} \text{Rel}(\bar{y}_{*i})}{\sigma^2_{\eta^2}} \frac{\psi^W_{ij}}{\psi^W_{i}} \beta^B - \frac{\sigma^2_{\eta^2}}{\sigma^2_{\eta^2} - \frac{\sigma^2_{\eta^2} \psi^W_{ij}}{\psi^W_{i}}} \text{Rel}(\bar{y}_{*i}) \beta^B
\]

\[
\text{Bias}(\hat{\gamma}_{DM}) = -\sigma_{\eta^2} \left\{ \frac{1}{n} \text{ICC} \frac{\psi^W_{ij}}{s^2} - \frac{1}{\sigma^2_{\psi^W} / t^2} \beta^B \right\}. \tag{A15}
\]

Assuming that the latent predictor \( \eta_1 \) and the observed group-level predictor \( x \) are standardized implies:

\[
\rho_{\eta x} = \frac{\sigma_{\eta^2}}{\sqrt{\psi^W_{ij}}} \quad \text{and} \quad \text{ICC} = \psi^W_{i}. \tag{A16}
\]

Finally, substituting Equation A16 into Equation A15 yields:

\[
\text{Bias}(\hat{\beta}^W_{DM}) = \left\{ \frac{n - 1}{n} \text{Rel}(\bar{y}_{*ij} - \bar{y}_{*i}) - 1 \right\} \beta^W
\]

\[
\text{Bias}(\hat{\beta}^B_{DM}) = \frac{1}{n} \text{ICC} \frac{\text{Rel}(\bar{y}_{*i})}{1 - \rho^2_{\eta x}} \beta^B
\]

\[
\text{Bias}(\hat{\gamma}_{DM}) = -\rho_{\eta x} \sqrt{\text{ICC}} \text{Bias}(\hat{\beta}^W_{DM}). \tag{A17}
\]