Psychological Methods
2011, Vol. 16, No. 4, 444–467

1082-989X/11/$12.00 DOI: 10.1037/a0024376

A 2 × 2 Taxonomy of Multilevel Latent Contextual Models: Accuracy–Bias Trade-Offs in Full and Partial Error Correction Models

Oliver Lüdtke
Humboldt University and University of Tuebingen

Herbert W. Marsh
University of Oxford

Alexander Robitzsch
Federal Institute for Education Research, Innovation, and Development of the Austrian School System, Salzburg, Austria

Ulrich Trautwein
University of Tuebingen

In multilevel modeling, group-level variables (L2) for assessing contextual effects are frequently generated by aggregating variables from a lower level (L1). A major problem of contextual analyses in the social sciences is that there is no error-free measurement of constructs. In the present article, 2 types of error occurring in multilevel data when estimating contextual effects are distinguished: unreliability that is due to measurement error and unreliability that is due to sampling error. The fact that studies may or may not correct for these 2 types of error can be translated into a 2 × 2 taxonomy of multilevel latent contextual models comprising 4 approaches: an uncorrected approach, partial correction approaches correcting for either measurement or sampling error (but not both), and a full correction approach that adjusts for both sources of error. It is shown mathematically and with simulated data that the uncorrected and partial correction approaches can result in substantially biased estimates of contextual effects, depending on the number of L1 individuals per group, the number of groups, the intraclass correlation, the number of indicators, and the size of the factor loadings. However, the simulation study also shows that partial correction approaches can outperform full correction approaches when the data provide only limited information in terms of the L2 construct (i.e., small number of groups, low intraclass correlation). A real-data application from educational psychology is used to illustrate the different approaches.

Keywords: multilevel modeling, measurement error, sampling error, latent variables, structural equation modeling

Supplemental materials: http://dx.doi.org/10.1037/a0024376.supp

Multilevel modeling (MLM) is one of the central research methods employed by applied researchers in the social sciences. In contrast to single-level regression analysis, MLM allows relationships among variables located at different levels to be explored simultaneously (Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). In the most typical application of MLM, outcome variables are related to several predictor variables at the individual level (e.g., students, employees) and at the group level (e.g., schools, work groups, neighborhoods). However, a major problem in the social sciences is that there is no error-free measurement of constructs, and it is well known that this measurement error can lead to biased and less efficient estimates of population parameters. Hence, there is widespread application of structural equation modeling (SEM), in which multiple indicators are used to correct for unreliability in the assessment of constructs. In the past decade, substantial progress has been made in integrating SEM and MLM techniques within a single analytic framework that can be implemented in applied research (Bovaird, 2007; L. K. Muthén & Muthén, 2007; Skrondal & Rabe-Hesketh, 2004; see also Goldstein & McDonald, 1988; McDonald, 1993).

In several disciplines of psychology, group-level variables (L2) for assessing contextual effects are frequently generated by aggregating variables from a lower level (L1). In industrial and organizational psychology, for instance, service climate is measured by aggregating individual customer perceptions of the quality of service (Masterson, 2001; Schneider, White, & Paul, 1998). In small group research, group characteristics, such as group efficacy, are assessed by reference to the individual ratings of group members (Moritz & Watson, 1998). Clinical psychologists use individual group members' perceptions of the group's therapeutic environment to assess group climate (Johnson, Burlingame, Olsen, Davies, & Gleave, 2006). In educational psychology, student-level ratings are aggregated at the class level to obtain general informa-

This article was published Online First July 25, 2011.

Oliver Lüdtke, Institute of Psychology, Humboldt University, Berlin, Germany, and Department of Education, University of Tuebingen, Tuebingen, Germany; Herbert W. Marsh, Department of Education, University of Oxford, Oxford, England; Alexander Robitzsch, Federal Institute for Education Research, Innovation, and Development of the Austrian School System, Salzburg, Austria; Ulrich Trautwein, Department of Education, University of Tuebingen.

We would like to thank Tihomir Asparouhov, Bengt Muthén, John Fletcher, and Benjamin Nagengast for comments on earlier versions of this article and Susannah Goss for editorial assistance.

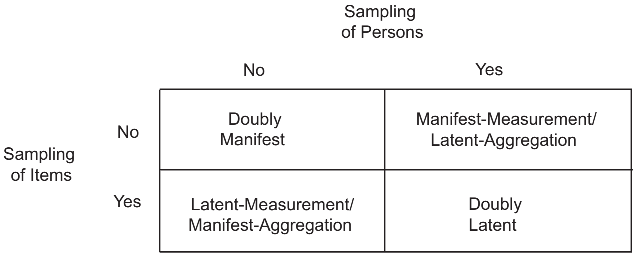
Correspondence concerning this article should be addressed to Oliver Lüdtke, Humboldt University, Unter den Linden 6, 10099 Berlin, Germany. E-mail: oliver.luedtke@hu-berlin.de

tion about the learning environment (Lüdtke, Robitzsch, Trautwein, & Kunter, 2009; Ryan, Gheen, & Midgley, 1998).

In the construction of these group-level variables, two main sources of error in multilevel data can be distinguished. First, unreliability can be due to measurement error in assessing constructs at either the individual or group level. Second, unreliability can be due to sampling only a finite sample from a potentially infinite population. Lüdtke et al. (2008) discussed the role of sampling error for assessing contextual effects in MLM and introduced a multilevel latent covariate (MLC) approach that takes into account unreliability at L2 that is caused by sampling error when estimating group effects in MLM. However, the MLC approach is limited to manifest scale scores and thus does not take the effect of measurement error into account.

The present article explores several approaches to correcting for error in multilevel data. Specifically, we distinguish between *partial correction* approaches that correct for only one of the two sources of error (i.e., sampling or measurement error) and a *full correction* approach that adjusts for both sources of error. This results in a 2 × 2 taxonomy of latent covariate models (see Figure 1) that correct for error in terms of (a) the sampling of items (i.e., that use multiple indicators to correct for measurement error, a traditional focus of confirmatory factor analysis and SEM studies) and/or (b) the sampling of persons (i.e., that use latent variables to correct for sampling error in the aggregation of L1 constructs to form L2 constructs). Traditionally, contextual analysis models were often doubly manifest (manifest variable, aggregation of observed indicators; see top left cell of Figure 1), correcting for neither measurement error nor sampling error.

The article is organized as follows. We start by briefly introducing the basic contextual analysis model, arguing that the estimation of group effects can be distorted by two types of error in multilevel data: unreliability that is due to measurement error and unreliability that is due to sampling error. We show mathematically how unreliable assessment of constructs at L2 results in biased estimates of group effects for the doubly manifest approach. The different approaches to correcting unreliable assessment of group constructs are then introduced, and the results of two simulation studies examining their statistical properties are reported. An empirical example from educational psychology is used to illustrate the different approaches and to demonstrate how they can be specified in the software Mplus. Finally, we offer suggestions for applied researchers and propose directions for further research.



		Sampling of Persons	
		No	Yes
Sampling of Items	No	Doubly Manifest	Manifest-Measurement/ Latent-Aggregation
	Yes	Latent-Measurement/ Manifest-Aggregation	Doubly Latent

Figure 1. A 2 × 2 taxonomy of multilevel latent contextual models designed to correct for measurement error (through sampling of items) and sampling error (through sampling of persons).

Assessing Group Effects in Multilevel Modeling

We assume a two-level structure with persons nested within groups and an individual-level variable *X* (e.g., socioeconomic status) predicting the dependent variable *Y* (e.g., reading achievement). Applying the MLM notation as it is used by Raudenbush and Bryk (2002), we have the following relation at the first level:

$$\text{Level 1:}\quad Y_{ij} = \beta_{0j} + \beta_{1j}(X_{ij} - \bar{X}_{\bullet j}) + r_{ij}, \tag{1}$$

where Y_{ij} is the outcome for person *i* in group *j* predicted by the intercept β_{0j} of group *j* and the regression slope β_{1j} in group *j*. The predictor variable X_{ij} is centered at the respective group mean $\bar{X}_{\bullet j}$. This group mean centering of the individual-level predictor yields an intercept equal to an expected value of Y_{ij} for an individual whose value on X_{ij} is equal to the person's group mean. At L2, the L1 intercepts β_{0j} and slopes β_{1j} are dependent variables:

$$\text{Level 2:}\quad \beta_{0j} = \gamma_{00} + \gamma_{01}\bar{X}_{\bullet j} + u_{0j}$$

$$\beta_{1j} = \gamma_{10}, \tag{2}$$

where γ_{00} and γ_{10} are the L1 intercepts and γ_{01} is the slope relating $\bar{X}_{\bullet j}$ to the intercepts from the L1 equation (Equation 1). As can be seen, only the L1 intercepts have an L2 residual u_{0j}. MLMs that allow only the intercepts to deviate from their predicted value are also called *random intercept models* (e.g., Raudenbush & Bryk, 2002). In these models, group effects are allowed to modify only the mean level of the outcome for the group; the distribution of effects among persons within groups (e.g., slopes β_{1j}) is left unchanged. Inserting the L2 equations into the L1 equation gives

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\bullet j}) + \gamma_{01}\bar{X}_{\bullet j} + u_{0j} + r_{ij}. \tag{3}$$

It is now easy to see that γ_{10} is the within-group (L1) regression coefficient describing the relationship between *Y* and *X* within groups and that γ_{01} is the between-group (L2) regression coefficient that indicates the relationship between group means $\bar{Y}_{\bullet j}$ and $\bar{X}_{\bullet j}$ (Cronbach, 1976). A contextual effect is present if γ_{01} is different from γ_{10}, meaning that the relationship at the aggregated level is stronger or weaker than the relationship at the individual level (e.g., an effect of school-average socioeconomic status on reading achievement after controlling for individual students' socioeconomic status).[1]

Let us now assume that a contextual model holds in the population (i.e., the group sample mean becomes the group population mean) and that the within-group and between-group relationships are described by the within-group regression coefficient β_w and the

[1] Instead of using group mean centering of the predictor variables—where the group mean of the L1 predictor is subtracted from each case—researchers often center the predictor at its grand mean. In grand mean centering, the grand mean of the L1 predictor is subtracted from each L1 case. However, it can be shown that in the case of the random intercept model, the group mean model and the grand-mean-centered model are mathematically equivalent (see Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995). Because our analysis is limited to random intercept models, centering of predictor variables is not a critical issue in this article. In the following, our investigation of the analysis of group effects in MLM focuses on the group-mean-centered case.

between-group regression coefficient β_b (see Snijders & Bosker, 1999, p. 29). Usually, these coefficients are estimated by drawing a sample of L2 groups from the population. In the next stage, a finite sample of L1 individuals is obtained for each sampled L2 group. However, it is well known that if X_{ij} and $\bar{X}_{\bullet j}$ are measured with error, γ_{01} and γ_{10} can result in biased and less efficient estimates of the true population parameters β_w and β_b. Fortunately, progress has been made in recent decades in conceptualizing error in multilevel data (Goldstein, Kounali, & Robinson, 2008; Kamata, Bauer, & Miyazaki, 2008; Marsh et al., 2009; Rabe-Hesketh, Skrondal, & Pickles, 2004; but see also Kane, Gillmore, & Crooks, 1976). In the following section, we build on this work and present a framework for assessing error in multilevel data.

Error in Multilevel Data

The starting point is the basic premise of classical test theory that an observed score X is the sum of the underlying true score U_x plus error of measurement R_x: $X = U_x + R_x$ (e.g., Lord & Novick, 1968). By extending this definition to the multilevel case, an observed score X_{ij} of person i in group j can be decomposed into a grand mean μ_x, a true score U_{xij} and an error score R_{xij} at L1, and a true score U_{xj} and an error score R_{xj} at L2:

$$X_{ij} = \mu_x + U_{xj} + U_{xij} + R_{xj} + R_{xij}. \tag{4}$$

Thus, an individual's true score is decomposed into two uncorrelated random variables U_{xj} and U_{xij}, which are distributed with zero means and variances $\mathrm{Var}(U_{xj}) = \tau_x^2$ and $\mathrm{Var}(U_{xij}) = \sigma_x^2$. Similarly, the error score is decomposed into two uncorrelated random variables R_{xj} and R_{xij}, which are distributed with zero means and variances $\mathrm{Var}(R_{xj}) = \tau_{x,e}^2$ and $\mathrm{Var}(R_{xij}) = \sigma_{x,e}^2$. Note that U_{xj} and U_{xij} represent the true and unobserved group-level and individual-level covariates with coefficients β_b and β_w, that we want to estimate using the potentially error-prone observed covariates $\bar{X}_{\bullet j}$ and X_{ij} (see Equation 3).

It will be argued in the following that unreliability at L2 comprises two kinds of error. One is the traditional measurement error R_{xj} that is due to unreliability of the indicator X_{ij} to assess the corresponding group construct. Group-specific influences may distort the measurement of the intended L2 construct. For example, if students rate their teacher's behavior in a certain lesson, situation-specific influences (e.g., the teacher's personal problems impairing the quality of instruction) or specifics of the item content (e.g., students from rural schools reacting differently to certain questions than students from inner-city schools) can result in systematic measurement error in the student ratings within a class. When student ratings are aggregated at the class level, these errors in the individual ratings can distort the assessment of the group-level construct. More formally, this unreliability at L2 can be interpreted as a kind of multidimensionality (any source other than the common factor that is consistent for all students nested within a school) that is present in $U_{xj} + R_{xj}$ at L2 when researchers aim to measure the unobserved group-level covariate U_{xj}.

A second component is error that is due to sampling only a finite sample from a finite or (potentially) infinite population (Brennan, 2001; Lüdtke et al., 2008; Shin & Raudenbush, 2010). If only a small number of L1 individuals are sampled from each L2 group, the observed group average may be a highly unreliable measure of the unobserved true group average U_{xj}. For instance, in educational psychology, where a small proportion of students may be sampled from each participating school, the observed group average is only an approximation of the unobserved "true" group mean. However, in contextual analyses in which L1 scores are aggregated at L2 to assess an L2 construct, it is often not discussed whether it is reasonable to assume that L1 scores are generated by a finite or an infinite sampling process.

Recently, Lüdtke et al. (2008) discussed the role of finite and infinite sampling and distinguished between *formative* and *reflective* L2 constructs (see also Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; Howell, Breivik, & Wilcox, 2007; Marsh et al., 2009; Skrondal & Laake, 2001). The main distinction between the two is that the group is the referent in the aggregation process in reflective L2 constructs (e.g., each member of the group directly rates an L2 construct), whereas the individual is typically the referent in formative L2 constructs (e.g., the L2 aggregation is based on a group average of individual characteristics). For example, if all students in a class rated the homework quality assigned by their teacher (an L2 construct), the aggregate would be a reflective measure, whereas if each student's gender was used to calculate the gender ratio of a class, the resulting aggregate would be a formative measure. The theoretical rationale for aggregations of L1 constructs when assessing a reflective L2 construct is based on classical measurement theory and the domain sampling model. Group characteristics are latent, unobserved constructs that can be inferred on the basis of a finite subset of a potentially infinite number of observers. In this respect, the group members are regarded as exchangeable (in relation to scores reflecting the L2 reflective construct), and the reflective aggregation corresponds to an infinite sampling process. In contrast, aggregation of L1 scores when assessing a formative L2 construct is more problematic from the perspective of determining sampling error. In formative aggregation, the L1 measures are not exchangeable in the sense that individuals within the same group have different L1 true scores and these L1 measures are used to build a group index. For example, if the gender of all students in each of a large number of different classes is known, and the research interest is in the gender ratio of a specific class, it is clear that students are not exchangeable in relation to gender. In that case, it would not be reasonable to assume a potentially infinite population whose gender could be assessed. Hence, formative aggregation corresponds to a finite sampling process, and there should be little or no sampling error for gender ratio as the sampling ratio in each class is close to 100%.

In the present article, we concentrate on the infinite population case that holds for reflective L2 constructs and assumes that a potentially infinite number of L1 units could be sampled from each L2 unit (see also Preacher, Zyphur, & Zhang, 2010). Strategies for dealing with unreliability in formative L2 constructs that are based on the finite population case are given in the General Discussion.

Assessing Sampling Error

Sampling error is conceptualized as the error that is due to observing only a finite sample of the whole population. The sampling error is inversely related to the precision of a statistical estimate that measures how close estimates from different samples are to one another (see Lohr, 1999). Furthermore, sampling error affects the estimation of the cluster-level true score U_{xj} because the average of the unit-level true scores has some variance, which, as in simple random samples, decreases with increasing sample size. Suppose that there are n_j units observed in a cluster. The precision of the cluster mean $\bar{X}_{\bullet j}$ for the cluster-level true score is then

$$\frac{\mathrm{Var}(U_{xj})}{\mathrm{Var}(\bar{X}_{\bullet j})} = \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2 + \sigma_x^2/n_j + \sigma_{x,e}^2/n_j}. \tag{5}$$

Note that σ_x^2/n_j is the established expression for the sampling variance of observed group averages (Lohr, 1999). If we assume that there is no measurement error at L1 and L2 ($\sigma_{x,e}^2 = \tau_{x,e}^2 = 0$), the precision of a group-average score (see also Raudenbush & Bryk, 2002) is given by

$$\frac{\tau_x^2}{\tau_x^2 + (\sigma_x^2/n_j)}. \tag{6}$$

In the literature on reliability of multilevel data (Bliese, 2000; LeBreton & Senter, 2008), this measure is sometimes called the ICC(2) and is used to determine the reliability of aggregated individual-level data ($\bar{X}_{\bullet j}$) in terms of sampling only a finite number of L1 units from each L2 unit. Thus, it can be interpreted as the reliability of the group mean in relation to sampling error. In most cases, the reliability of a collection of group means can be estimated by using the mean group size for n_j if not all groups are of the same size (see Searle, Casella, & McCulloch, 1992, on how to deal with pronounced differences in group size). However, in research practice, the individual data X_{ij} can rarely be measured with completely reliable indicators.

Measurement Error at L1 and L2

Measurement error attenuates the relationship between the observed score and the true score. In the classical test theory model, an individual's measured score is decomposed into a true score plus error of measurement. To avoid confounding discussion of measurement error and sampling error, we assume in the following that there is no sampling error (e.g., that all L1 units were sampled from each L2 unit). Using the decomposition of the observed score X_{ij} in Equation 4, we can then define the reliability as the percentage proportion of variance of the measure that is due to true score variance:

$$\mathrm{Rel}(X_{ij}) = \frac{\tau_x^2 + \sigma_x^2}{\tau_x^2 + \sigma_x^2 + \tau_{x,e}^2 + \sigma_{x,e}^2}. \tag{7}$$

A straightforward extension of this formula to two-level data would be to define the reliability separately for L1 and L2. The reliability of X_{ij} at L1 is then given by the ratio of true score L1 variance to the total variance at L1:

$$\mathrm{Rel}_{\mathrm{L1}}(X_{ij}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{x,e}^2}. \tag{8}$$

Accordingly, the reliability of X_{ij} at L2 reflects the possibility that not all the variation in the group means is due to differences in the common factor and is given by the following formula:

$$\mathrm{Rel}_{\mathrm{L2}}(X_{ij}) = \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2}. \tag{9}$$

Note that these two reliabilities are mainly of theoretical interest. The measurement error variance is not usually known in applications. Sometimes sensitivity analyses are conducted, in which reasonable guesses of the amount of measurement error variance are used to estimate how unreliability affects the results of the analyses (Goldstein et al., 2008). However, the most common approach is based on the assumption that independent replications (multiple indicators) exist. These replications are treated as measurement indicators for assessing the latent construct (e.g., items or subscales of a test). The multiple indicators are then used to assess the degree of measurement error variance.

Let us assume that a construct is measured by multiple indicators. By extending the classical measurement model, we can decompose the single indicator X_{kij} as follows (Kamata et al., 2008; B. O. Muthén, 1991):

$$X_{kij} = \mu_{xk} + \lambda_{k,W}U_{xij} + R_{xkij} + \lambda_{k,B}U_{xj} + R_{xkj};\ k = 1, \ldots, K, \tag{10}$$

where μ_{xk} are the indicator-specific means, $\lambda_{k,W}$ are the within-factor loadings, $\lambda_{k,B}$ are the between-factor loadings, and R_{xkij} and R_{xkj} are the indicator-specific error scores at L1 and L2, which are assumed to be uncorrelated across indicators. Again, U_{xij} and U_{xj} are the unobserved true scores at L1 and L2. In classical test theory, Equation 10 represents a congeneric measurement model, because the factor loadings are allowed to vary across indicators. The measures are said to be parallel if the factor loadings are equal at L1 ($\lambda_{k,W} = \lambda_{k',W}$) and L2 ($\lambda_{k,B} = \lambda_{k',B}$) and the measurement error variances are also equal at L1, $\mathrm{Var}(R_{xkij}) = \mathrm{Var}(R_{xk'ij}) = \sigma_{x,e}^2$, and L2, $\mathrm{Var}(R_{xkj}) = \mathrm{Var}(R_{xk'j}) = \tau_{x,e}^2$. Based on the assumption of uncorrelated indicator-specific errors, the observed score variance can be separated into true score variance and error variance (Traub, 1994).

For ease of interpretation, we assume in the following that all factor loadings are set to 1 and that X is mean centered. Equation 10 can then be written as follows:

$$X_{kij} = \underbrace{U_{xj} + R_{xkj}}_{\text{between}} + \underbrace{U_{xij} + R_{xkij}}_{\text{within}};\ k = 1, \ldots, K. \tag{11}$$

If it is now assumed that the true score U_{xij} is measured by $k = 1, \ldots, K$ parallel items X_{kij}. The observed score is obtained by averaging across the K parallel items

$$\bar{X}_{\bullet ij} = \sum_{k=1}^{K} X_{kij}/K.$$

The reliability of the observed score that results from aggregating across items is then given by the following formula (see Equation 8):

$$\mathrm{Rel}_{\mathrm{L1}}(\bar{X}_{\bullet ij}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{x,e}^2/K}. \tag{12}$$

Because the items are assumed to be parallel, these measurement error variances are equal. As is known for parallel measures, the reliability of the observed score is a direct function of the number of items. Increasing the number of items reduces the error variance in the denominator of Equation 12 and thus increases the reliability of the observed score. Correspondingly, the reliability at L2 is given by the following formula (see Equation 9):

$$\mathrm{Rel}_{\mathrm{L2}}(\bar{X}_{\bullet ij}) = \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2/K}. \tag{13}$$

Note that this reliability is based only on measurement error and does not take into account the unreliability that is due to sampling error. In the next section, we show how unreliable assessment of L2 constructs results in biased estimates of between-group relations.

Biased Estimation of Between-Group Relations Under Multilevel Error

We now assume that a contextual model holds in the population and that the within-group and between-group relationships are described by the within-group regression coefficient β_w and the between-group regression coefficient β_b. We want to estimate these coefficients by drawing a sample of L2 groups from the population. In the next stage, a finite sample of L1 individuals is obtained for each sampled L2 group. It is further assumed that both X_{ij} and $\bar{X}_{\bullet j}$ are measured with error and that we are interested in estimating the relationship for the true variables U_{xij} and U_{xj}. To make the derivations of expected bias of the within- and between-group regression coefficients more manageable, we restrict our attention to the case in which the true score U_{xij} is measured by $k = 1, \ldots, K$ parallel items X_{kij} and the observed score is obtained by averaging across the K parallel items.

Based on the reliabilities at L1 and L2, it is possible to derive the bias for within-group and between-group regression coefficients (see Appendix A). As the observed score at L1 measures the true score with reliability $\mathrm{Rel}_{\mathrm{L1}}(X_{ij})$, it is apparent that the within-group coefficient $\hat{\gamma}_{10}$ is a biased estimator of the within-group coefficient β_w:

$$E(\hat{\gamma}_{10} - \beta_w) = -\beta_w \cdot (1 - \mathrm{Rel}_{\mathrm{L1}}). \tag{14}$$

This is the classical attenuation formula that states that the association between measured variables is equal to the true correlation multiplied by some number that is less than or equal to 1 (e.g., Kenny, 1979). Hence, in the case of a single predictor variable that is measured with error, $\hat{\gamma}_{10}$ will underestimate the true relationship.

To derive the bias of the between-group coefficient, we have to take both types of error into account: measurement error due to the unreliability of the items and sampling error. It can be shown that the between-group coefficient $\hat{\gamma}_{01}$ is a biased estimator of the between-group coefficient β_b:

$$E(\hat{\gamma}_{01} - \beta_b) = -\beta_b \cdot \left(1 - \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{1}{1 + r \cdot b/n}\right) + \beta_w \cdot \frac{b}{n} \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{1}{1 + r \cdot b/n}, \tag{15}$$

where $r = \mathrm{Rel}_{\mathrm{L2}}/\ \mathrm{Rel}_{\mathrm{L1}}$ is defined as the ratio of the reliabilities at L1 and L2, $b = \sigma_x^2/\tau_x^2$ is the ratio of the variance within groups to the variance between groups, and n is the number of L1 units within each L2 unit. As can be seen, the bias depends primarily on the reliability at L2 ($\mathrm{Rel}_{\mathrm{L2}}$), the proportion of variance in X that is located between groups (b), and the group size n.

For demonstration purposes, we simplify the formula for the bias of the between-group coefficient by assuming that $\mathrm{Rel}_{\mathrm{L1}} = \mathrm{Rel}_{\mathrm{L2}} = \mathrm{Rel}$. The following relation then holds for the bias of the between-group regression coefficient:

$$E(\hat{\gamma}_{01} - \beta_b) = -\beta_b \cdot \left(1 - \mathrm{Rel} \cdot \frac{1}{1 + b/n}\right) + \beta_w \cdot \frac{b}{n} \cdot \mathrm{Rel} \cdot \frac{1}{1 + b/n}. \tag{16}$$

It can now be seen that with $n \to \infty$, the bias of the between-group regression coefficient depends only on the reliability of measurement of X. The relationship between the expected absolute bias and the intraclass correlation (ICC; .05, .10, .20, and .40), which indicates the proportion of variance located between groups in the predictor as well as the group size, and the reliability (.6, .75, .90, and 1.0) is depicted in Figure 2 for a β_b value of 0.7 and a β_w value of 0.2. As is to be expected, the absolute bias decreases with increasing reliability of the manifest score. However, even with perfect reliability (Rel = 1), the between-group regression coefficient $\hat{\gamma}_{01}$ is a biased estimator of the between-group coefficient β_{b} because only a finite sample of L1 units ($n < 50$) is sampled from each L2 unit (Lüdtke et al., 2008). In all four panels, the bias becomes smaller with larger group sizes n. In other words, when the group mean is more reliable due to a higher n, β_b can be more precisely approximated by the manifest group mean predictor. In addition, the bias decreases as the ICC of the predictor increases.

In the traditional approach to assessing group effects in MLM, the group-level predictor was formed by aggregating all the observed measurements in each group. Given the relations in Equation 15, parameter estimates obtained from this approach are expected to be dramatically biased in certain data constellations (e.g., if the ICC is low or only a small number of L1 units are sampled from each L2 unit). In the next section, we discuss alternative approaches to assessing group effects that take into account the unreliability of the group mean potentially caused by sampling error and measurement error.

Correcting for Unreliability: A 2 × 2 Taxonomy of Multilevel Latent Contextual Models

In this section, an example from educational psychology (the effect of homework quality on students' motivation) is used to illustrate different approaches to correcting for unreliability when estimating group effects. In this example, each student is regarded as an independent observer of the quality of the homework assigned by his or her teacher (Trautwein, Lüdtke, Schnyder, & Niggli, 2006; see also Lüdtke et al., 2009). In this case, the referent is the teacher or the homework assigned to the class, and responses are aggregated across all students within a class to provide an indicator of homework quality. At the *individual level*, student ratings represent the individual student's perception of homework quality. Scores aggregated to the *classroom level* reflect shared perceptions of homework quality in which idiosyncrasies associ-

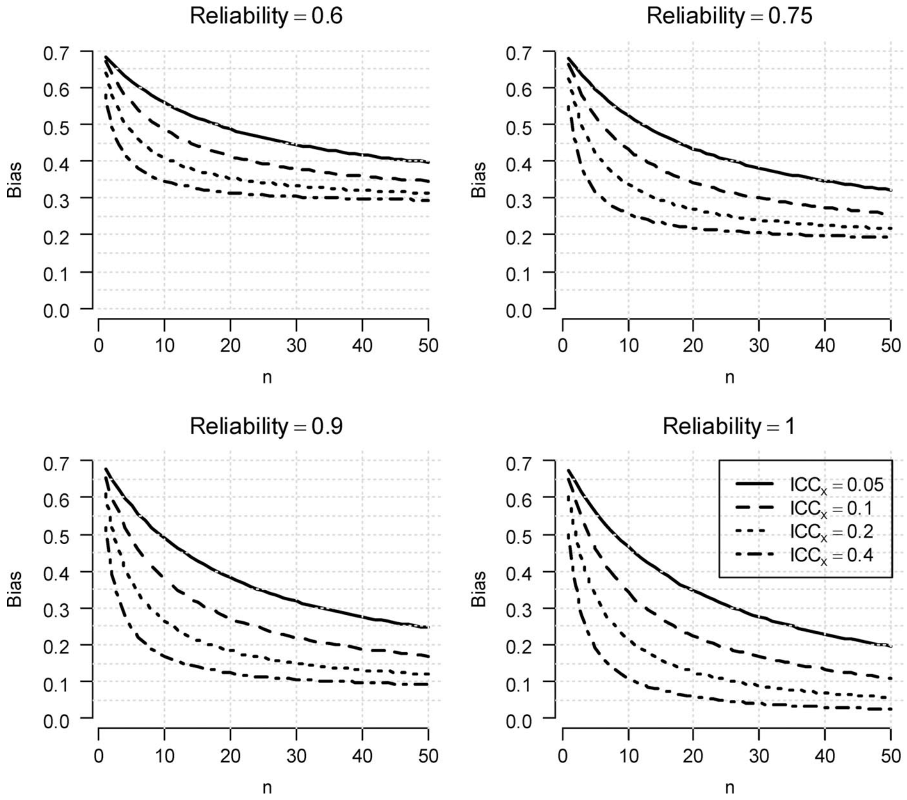


Figure 2. Relationship between the expected absolute bias of the between-group regression coefficient, the reliability of the predictor variable, the number of L1 units within each L2 unit (*n*), and the intraclass correlation (ICC).

ated with individual students' responses tend to cancel each other out (Lüdtke, Trautwein, Kunter, & Baumert, 2006; Miller & Murdock, 2007; Papaioannou, Marsh, & Theodorakis, 2004). In line with the rationale for reflective L2 constructs, all students rate the same construct (i.e., quality of the homework assigned by their teacher) and are treated as indicators of the corresponding L2 construct. Thus, L1 student responses are used to construct an L2 reflective construct that captures a certain characteristic of the classroom learning environment—in this case, homework quality (see Lüdtke et al., 2008).

For ease of interpretation, we assume in the following that in the population model all factor loadings are set to 1 and that *X* is mean centered (see Equation 11). The latent scores U_{xij} and U_{xj} are then related to the outcome variable Y_{ij} in the structural model and the relation between observed indicators and latent and error scores is given by the measurement model:

$$Y_{ij} = \mu_y + \beta_w U_{xij} + \beta_b U_{xj} + \delta_j + \varepsilon_{ij}$$

$$X_{kij} = \underbrace{U_{xj} + R_{xkj}}_{\text{between}} + \underbrace{U_{xij} + R_{xkij}}_{\text{within}}; k = 1, \ldots, K, \qquad (17)$$

where β_w is the within-group regression coefficient describing the relationship within L1 units, β_b is the between-group regression coefficient indicating the relationship between the L2 group means, and ε_{ij} and δ_j are normally distributed (with an expected value of 0) and uncorrelated residuals at L1 and L2. We now present four models that differ in how they assess the unobserved scores U_{xij} and U_{xj}.[2]

Ignoring Error in Multilevel Data: A Doubly Manifest Approach

We start with the most basic model, in which both measurement and sampling error are assumed to be zero when constructing the independent variable at the individual and group level (i.e., individual and aggregated homework quality). At L1, the covariate is calculated by

$$\bar{X}_{\bullet ij} = \frac{1}{K}\sum_{k=1}^{K} X_{kij} = U_{xij} + U_{xj} + \bar{R}_{x\bullet ij} + \bar{R}_{x\bullet j},$$

averaging across the *K* items for each individual *i* in group *j*. Correspondingly, at L2, the covariate is calculated by summing across the *K* items and the n_j persons in each group *j*:

[2] It is assumed that the dependent variable is measured without error. However, unreliability of the dependent variable does not result in bias of unstandardized regression coefficients.

$$\bar{X}_{\bullet\bullet j} = \frac{1}{K \cdot n_j} \sum_{i=1}^{n_j} \sum_{k=1}^{K} X_{kij} = U_{xj} + \bar{U}_{x\bullet j} + \bar{R}_{x\bullet\bullet j} + \bar{R}_{x\bullet j}.$$

Hence, in the structural equation (see Equation 17), U_{xij} and U_{xj} are substituted by $(\bar{X}_{\bullet ij} - \bar{X}_{\bullet\bullet j}) = U_{xij} + \bar{R}_{x\bullet ij} - \bar{U}_{x\bullet j} - \bar{R}_{x\bullet\bullet j}$ and $\bar{X}_{\bullet\bullet j}$. In contrast to the unobserved group mean U_{xj}, the observed group mean $\bar{X}_{\bullet\bullet j}$ is affected by the sampling error component $\bar{U}_{x\bullet j}$,[3] the measurement error component at L2 $\bar{R}_{x\bullet j}$ (e.g., group-specific factors, such as a particularly difficult test in the last lesson, that affect ratings on specific items), and the measurement error component at L1 $\bar{R}_{x\bullet\bullet j}$ (e.g., individual-specific factors, such as mood, that affect ratings on specific items).

Thus, the traditional multilevel model (see Equation 3) based on manifest scores fails to take both types of unreliability into account. Because the two sources of error are not considered when estimating group effects, we label this approach the *doubly manifest* approach to estimating group effects in MLM. Figure 3 presents path diagrams for the four approaches to estimating group effects in MLM for the homework quality example (see Mehta & Neale, 2005, for a different graphical presentation of MLM with recticular action model notation; see also Curran & Bauer, 2007). Let us assume that three items were used to assess students' evaluation of the homework quality. As shown in Figure 3, the doubly manifest approach uses observed scores for homework quality at the within and between level (indicated by squares).[4]

Correcting for Sampling Error: A Manifest-Measurement/Latent-Aggregation Approach

One problematic aspect of the manifest contextual analysis model is that the observed average $\bar{X}_{\bullet\bullet j}$ may be a highly unreliable measure of the unobserved group average because only small numbers of L1 students are sampled from each L2 school (O'Brien, 1990). Lüdtke et al. (2008; see also Croon & van Veldhoven, 2007; Goldstein et al., 2008; Preacher et al., 2010; Shin & Raudenbush, 2010) introduced an MLC approach that takes sampling error into account when group effects are estimated. In this approach, the unobserved group mean is regarded as a latent variable that is measured with a certain amount of precision by the group mean of the observed data (Asparouhov & Muthén, 2007). The basis for the MLC is that the observed scores of the independent variable are decomposed into unobserved components, which are considered as latent variables (Asparouhov & Muthén, 2007; B. O. Muthén, 1989; Schmidt, 1969; Snijders & Bosker, 1999). More specifically, the observed scores $\bar{X}_{\bullet ij}$ that are obtained by averaging across the K items are modeled as $\bar{X}_{\bullet ij} = \widehat{U}_{xj} + \widehat{U}_{xij}$, whereas the actual decomposition is given as follows:

$$\bar{X}_{\bullet ij} = \underbrace{U_{xj} + \bar{R}_{x\bullet j}}_{\text{between}} + \underbrace{U_{xij} + \bar{R}_{x\bullet ij}}_{\text{within}}. \tag{18}$$

Note that $\bar{X}_{\bullet ij}$ is an observed variable, whereas U_{xj} and U_{xij} are unobserved latent variables. As can be seen, $\widehat{U}_{xj}$ corrects for error that is due to sampling only a finite sample, but the resulting true score estimates at L1 and L2 are still affected by measurement error (e.g., variability that is not due to the common factor). In the structural model, the dependent variable Y_{ij} is then predicted by the individual and group-specific deviations:

$$Y_{ij} = \mu_y + \beta_w \widehat{U}_{xij} + \beta_b \widehat{U}_{xj} + \delta_j + \varepsilon_{ij}. \tag{19}$$

It is clear that in the MLC, measurement error at L1 and L2 distorts the estimation of the corresponding regression coefficients at β_w and β_b. In addition, in Equation 19, the individual deviations $\widehat{U}_{xj}$ are given by the observed deviations, as is also shown in Figure 3, in which homework quality is a latent variable (represented as a circle) at L2 but a manifest variable (represented as a square) at L1.

Two critical issues are connected to the MLC approach. First, Lüdtke et al. (2008) presented simulation studies showing that the MLC approach provided approximately unbiased estimates. However, they also showed that the estimates produced by the MLC are substantially more variable in certain data constellations (e.g., small number of L2 groups, small ICCs, and small number of L1 units within each L2 unit) than the estimates obtained with the traditional doubly manifest approach. Second, the MLC does not take measurement error into account. We therefore refer to the MLC as the *manifest-measurement/latent-aggregation* approach—"manifest measurement" because it starts with scale scores or single indicators and "latent aggregation" because it corrects for unreliability due to sampling error.

Correcting for Measurement Error: A Latent-Measurement/Manifest-Aggregation Approach

Let us assume that the independent variable is measured by multiple indicators ($k = 1, \ldots, K$). However, the cluster means of the indicators $\bar{X}_{k\bullet j}$ that are the result of manifest aggregations of the observed indicators to the group level are treated as indicators of the common factor in the L2 measurement model. Thus, the measurement model at L1 is

$$X_{kij} - \bar{X}_{k\bullet j} = \lambda_{k,W} U_{xij} + R_{xkij};\ k = 1, \ldots, K. \tag{20}$$

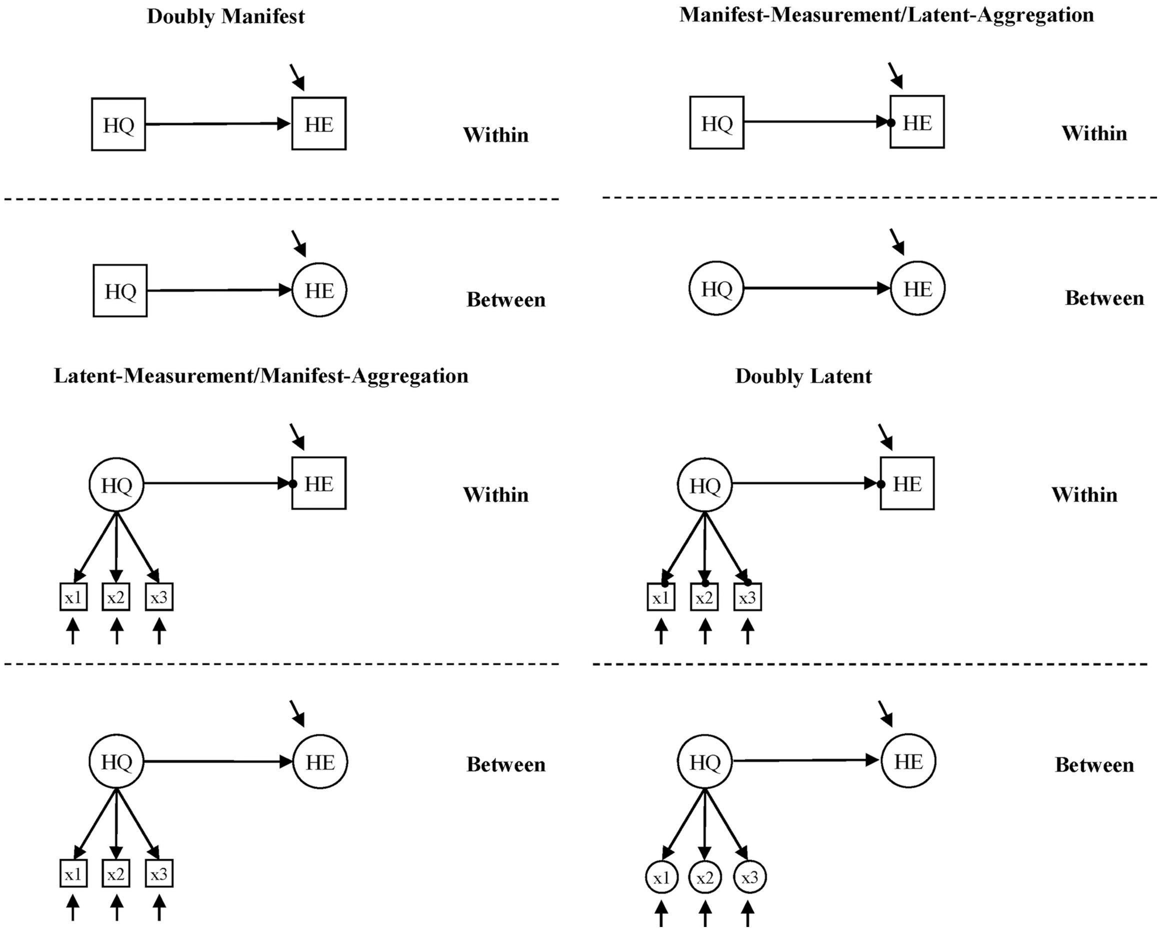
The assumed measurement model at L2 is

$$\bar{X}_{k\bullet j} = \lambda_{k,B} U_{xj} + R_{xkj};\ k = 1, \ldots, K. \tag{21}$$

Now assuming again, for ease of interpretation, that the factor loadings are 1 and that the observed indicators are mean centered, we can average over students within groups for all items. At L1, the application of multiple indicators allows an error-free assess-

[3] Note that the term $\sum_i \bar{U}_{xij} = \bar{U}_{x\bullet j}$ is only zero when all units are sampled from an L2 unit j.

[4] It needs to be added that the motivation for using the doubly manifest approach is not always to estimate the parameters of the model given in Equation 17. For example, when researchers are interested in assessing the causal effect of an L1 covariate, a group-mean-centered covariate provides a consistent estimate of the causal effect of that covariate if the model is correctly specified, except that there are omitted group-level confounders (see Angrist & Pischke, 2009, for applications of these models in economics).



Doubly Manifest

Manifest-Measurement/Latent-Aggregation

Latent-Measurement/Manifest-Aggregation

Doubly Latent

Figure 3. Schematic path diagrams for four alternative multilevel models used to estimate group effects. Circles indicate latent variables; squares indicate observed (manifest) variables. Within and between levels are separated by a dashed line. A dot at the end of the within (L1) regression (either from dependent on independent variables or from observed variables on latent factors) indicates that the corresponding intercept is random at the between level, representing the latent aggregation process. Thus, both single and multiple indicators at the between level (L2) are represented by squares if they are manifest (based on class-average values formed outside the model) and circles if they are latent (formed through the latent aggregation process as part of the model estimation process). All between-level random intercepts are represented as latent variables. HQ = homework quality ($x1$–$x3$ represent the multiple indicators of HQ); HE = homework effort; within = student level (L1); between = class level (L2).

ment of the unobserved U_{xij}, whereas at L2 the following relation holds for the observed indicator:

$$\bar{X}_{k\bullet j} = U_{xj} + \bar{U}_{x\bullet j} + R_{xkj} + \bar{R}_{xk\bullet j}. \quad (22)$$

In a one-factor model, the R variables are identified as measurement error. However, sampling error is not taken into account in that approach, and the unobserved group mean U_{xj} is approximated by the term $\breve{U}_{xj}$, which comprises not only the unobserved group mean U_{xj} but also the sampling error $\bar{U}_{x\bullet j}$. The resulting structural model is

$$Y_{ij} = \mu_y + \beta_w U_{xij} + \beta_b \breve{U}_{xj} + \delta_j + \varepsilon_{ij}. \quad (23)$$

For example, using this approach, the L1 homework quality factor would be based on responses to the three L1 indicators, and the L2 (class-average) homework quality factor would be based on the responses to simple (manifest) class-average values of the corresponding L1 indicators (see the path diagram in Figure 3 in which L2 indicators of homework quality are represented as squares, representing manifest variables, rather than as circles, representing latent variables). Hence, L2 homework quality is latent in the sense that it is based on multiple indicators. However, it is manifest in relation to aggregation from L1 to L2 in the sense that it does not correct for sampling error due to sampling of persons. We thus call this approach the *latent-measurement/manifest-aggregation* approach, reflecting that it is latent in terms of the measurement model ("latent-measurement") but manifest in terms of not taking the sampling error into account ("manifest-aggregation"). Given that the doubly manifest approach showed a smaller mean-squared error than the manifest-measurement/latent-aggregation approach in certain conditions of a previous simulation study (Lüdtke et al., 2008), we expect the latent-measurement/latent-aggregation approach to be a promising approach for correcting error in multilevel data when the data provide only limited information in terms of the L2 construct (i.e., small number of groups, low ICC).

Correcting for Measurement Error and Sampling Error: A Doubly Latent Approach

A more comprehensive multilevel model in terms of correcting for unreliability would be a model that we call *doubly latent* in the sense that it takes into account measurement error at L1 and L2 (based on multiple L1 indicators) and L2 sampling error (latent aggregation from L1 to L2, based on variation in individual perceptions of homework quality within classes). Thus, in the doubly latent approach, X_{kij} is decomposed as in Equation 10. As shown in Figure 3, in the doubly latent approach, the indicators of homework quality at L2 and the factor at L1 are considered as latent variables. The latent aggregation of the indicators in the doubly latent approach is represented by a dot (indicating that the intercepts are random at the between level) for each indicator in the path diagram. This type of model was also previously described by Rabe-Hesketh et al. (2004) as a special case of their generalized linear latent and mixed model framework (p. 181).

The four models for estimating group effects can be classified along the two sampling processes introduced above (see 2 × 2 taxonomy in Figure 1). In the doubly latent approach, both sampling of items and sampling of persons are taken into account. Hence, this approach provides the most comprehensive correction in terms of error variance when it comes to estimating group effects. The manifest-measurement/latent-aggregation and latent-measurement/manifest-aggregation approaches make only partial corrections because only a single source of error is considered: the unreliability that is due to sampling items or the unreliability that is due to sampling persons. The doubly manifest approach provides an uncorrected estimator of the group effect because neither sampling of items nor sampling of persons is taken into account. Following Carroll, Ruppert, Stefanski, and Crainiceanu (2006, p. 62), the manifest-measurement/latent-aggregation and latent-measurement/manifest-aggregation approaches can also be seen as "compromise estimators" because they allow for only partial correction of the biased group effect. However, compromise estimators sometimes outperform both uncorrected (doubly manifest) and corrected estimators (doubly latent) in terms of mean-squared error. Especially when sample sizes are small or the data provide less information (e.g., low ICC of the predictor variable), partial correction estimators can show better mean-squared-error performance than unbiased complete correction estimators (Carroll et al., 2006). In the next section, we evaluate the statistical properties of the four approaches to correcting for unreliability when estimating group effects with simulated data.

Study 1: Comparing Different Approaches to Correcting for Unreliability

The simulation study was designed to generate data that resemble the data structures typically found when assessing group effects in psychological and educational research.

Conditions

The population model used to generate the data was a random intercept model with one explanatory variable at the individual level and one explanatory variable at the group level, as specified in Equation 17. The explanatory variable was measured by a varying number of indicators, and the dependent variable was assessed by a single indicator. Thus, the data-generating model assumed that the doubly latent model holds in the population. Each data set generated was analyzed with all four approaches (doubly manifest, manifest-measurement/latent-aggregation, latent-measurement/manifest-aggregation, and doubly latent). The conditions manipulated were the number of L2 groups (50, 100, and 200), the number of observations per L2 group (five, 15, and 30), the latent ICC of the predictor variable (.05, .10, .20, and .30), the standardized factor loading (0.6 and 0.8), and the number of indicators (three and seven) used to assess the independent variable. Figure 4 shows the data-generating model and population parameters used in the condition in which the ICC of the predictor variable is .10, the standardized factor loadings are 0.6, and the number of indicators is three. Note that Figure 4 reports unstandardized population parameters (see Appendix B for details of the derivation of the parameters of the measurement model in the data-generating model). In the following, we explain why these particular conditions were selected.

Number of L2 groups. The number of L2 groups was set to J = 50, 100, or 200. A sample of 50 groups is common in educational and organizational research (e.g., Maas & Hox, 2005), although many MLM studies involve fewer than 50 L2 groups. At the same time, growing numbers of large-scale assessment studies,

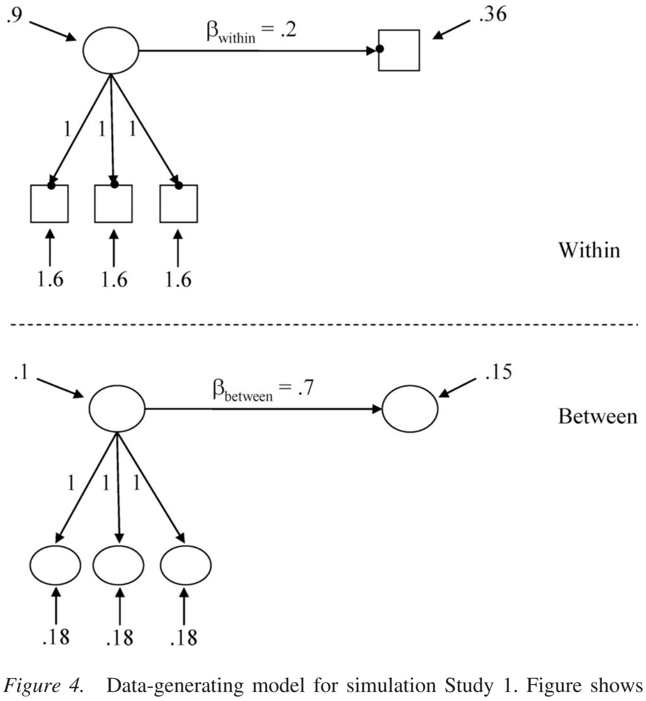


Figure 4. Data-generating model for simulation Study 1. Figure shows the population parameters for the following conditions: intraclass correlation = .10, number of indicators = 3, and standardized factor loadings = 0.6.

including educational assessments such as the Early Childhood Longitudinal Study and the National Education Longitudinal Study, are being conducted. Hence, we included conditions with 100 and 200 groups. Covering a broad range of L2 groups enables us to study asymptotic behavior in the latent variable approach.

Number of observations per L2 group. We then manipulated the number of observations per L2 group to $n = 5$, 15, and 30. A group size of five is normal in small group research, where contextual models are also applied (see Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). Group sizes of 15 and 30 reflect the numbers that typically occur in educational psychology assessing class or school characteristics.

ICC of predictor variable. The ICC of the predictor variable (i.e., the amount of variance located between groups) at the latent level was set at .05, .10, .20, or .30. The error variances were specified to have the same ICC. Hence, the observed scores have the same ICC as the latent scores. ICCs rarely show values greater than .30 in educational and organizational research (Bliese, 2000; James, 1982; Lüdtke et al., 2006).

Standardized factor loadings. The standardized factor loading can be interpreted as the reliability of a single indicator. A standardized factor loading of 0.6 means that 36% of the observed variance in that indicator is explained by differences in the true score. The standardized factor loadings were set to $\lambda = 0.6$ and $\lambda = 0.8$. In addition, we assumed that the unstandardized factor loadings were invariant across L1 and L2 in the data-generating model (see Figure 4).

Number of indicators. The number of indicators used to measure the predictor variable was varied between three and seven. Thus, for the conditions in which the standardized factor loadings were set to 0.6, the reliability was .63 when the number of items was set to three and .80 when the number of items was set to seven. For the conditions in which the standardized factor loadings were set to 0.8, the reliability was .84 when the number of items was set to three and .93 when the number of items was set to seven.

For each of the $3 \times 3 \times 4 \times 2 \times 2 = 144$ conditions, 1,000 simulated data sets were generated. The regression coefficients were specified as follows: 0 for the intercept, 0.2 for β_w, and 0.7 for β_b. Because the contextual effect β_c equals $\beta_b - \beta_w$, these values imply a contextual effect of 0.5. The ICC for the dependent variable was .2. Because the amount of variance explained at L2 depends on the ICC of the predictor variable, the following R^2 values at L2 were obtained for the different simulation conditions: .12 for ICC = .05, .25 for ICC = .10, .49 for ICC = .20, and .74 for ICC = .30. The corresponding R^2 values at L1 ranged from .04 for ICC = .05 to .05 for ICC = .30. For every cell, the 1,000 repetitions were simulated and analyzed with Mplus 4.1 via maximum-likelihood estimation with robust standard errors (denoted MLR in Mplus; L. K. Muthén & Muthén, 2007). Because invariant loadings were assumed in the data-generating model, factor loadings for the latent-measurement/manifest-aggregation and doubly latent approaches were freely estimated but specified to be invariant across L1 and L2. Item residual variances were freely estimated at L1 and L2. The first loading of the first indicator at L1 and L2 was fixed to 1 in order to identify the metric of the latent variable. In total, 576,000 analyses were conducted. To estimate parameters in MLMs, Mplus uses a general approach based on an accelerated expectation–maximization algorithm that provides maximum likelihood estimates for two-level structural equation models with missing data (B. O. Muthén & Asparouhov, 2008).

In our simulation study, we primarily focused on two aspects of the estimator for the between-group effect: the relative percentage bias of the parameter estimate and the root-mean-square error (RMSE). Relative bias was estimated by comparing the mean parameter estimate from each design cell with the corresponding population parameter. Let $\hat{\beta}_b$ be the estimator of the population parameter β_b; the relative percentage bias is then given by $100 \times [(\hat{\beta}_b - \beta_b)/\beta_b]$. The overall accuracy of the parameter estimates was assessed with the RMSE. The RMSE was computed by taking the square root of the mean square difference of the estimate and the true parameter. When a parameter estimate is unbiased, the RMSE quantifies the sampling variance (i.e., variability) of that parameter. For biased parameter estimates, the RMSE combines bias and variability into an overall measure of accuracy. In addition, we evaluated the performance of the standard errors by considering the standard error ratio (*SE* ratio), which is estimated by dividing the average standard error in each cell by the empirical standard deviation of the estimator. An *SE* ratio larger than 1.0 means that the standard errors on average overestimate the sampling variability of the estimator, whereas an *SE* ratio smaller than 1.0 indicates that the standard errors show a tendency to underestimate the sampling variability of the estimator.

Results and Discussion

Convergence rate. Given the complexity of the model, we also evaluated potential problems of estimation by examining the rate of model nonconvergence. The outcome of each model was categorized as nonconverged if the output obtained from Mplus indicated that there were problems with the estimation. Inspection of the error messages of nonconverged solutions indicated that in almost all cases, the computational problems were caused by a nonpositive definite estimated between covariance matrix. Most convergence problems occurred for the doubly latent approach, particularly when the number of groups was small and the ICC was low. For instance, in the condition with $n = 5$, $J = 50$, $\lambda = 0.6$, and seven items, only 13% of the solutions converged for an ICC of .05 and 57% for an ICC of .10. However, when the number of L1 units within each L2 unit was increased ($n = 15$), 89% of the solutions converged for an ICC of .05 and 100% for an ICC of .10. Only parameters of converged solutions were considered in the analysis of the simulation.[5]

Bias. As expected from the mathematical derivation in Equation 16, the estimated relative percentage bias for the doubly manifest approach was large, with values ranging from −73.7 to

[5] To check the sensitivity of the simulation results, we conducted additional analyses using Mplus 5.2. Interestingly, the percentage of converged solutions increased dramatically for the doubly latent approach. Further analyses of the Mplus outputs revealed that a large percentage of models classified as nonconverged due to a nonpositive definite covariance matrix by Version 4.1 were classified as "terminated normally" in Version 5.2. However, closer inspection showed that the converged solutions in Version 5.2 still suffered from a nonpositive covariance matrix (e.g., negative residual variances). Including the cases "converged" in Version 5.2 for the doubly latent approach increased the RMSE of the parameters, reflecting that extreme parameter estimates usually exhibit larger variability.

−12.1 (M = −40.8, SD = 14.8) across the conditions. The estimated relative percentage bias for the manifest-measurement/latent-aggregation approach, which corrects for L2 sampling error, ranged from −46.0 to −0.2 (M = −18.9, SD = 11.4); that for the latent-measurement/manifest-aggregation approach, from −57.8 to 2.7 (M = −24.6, SD = 15.6). The estimator with the smallest estimated relative percentage bias was provided by the doubly latent approach, with values ranging from −16.6 to 116.1 (M = 7.7, SD = 14.3). Tables 1–12 in the supplemental materials provide detailed information on the relative percentage bias, RMSE, and SE ratio of all four approaches.

We also investigated the source of the relative percentage bias by conducting a five-way factorial analysis of variance (ANOVA) with estimated relative percentage bias as the dependent variable and the five simulation conditions as independent variables (see Table 1). The largest effect was the main effect of method (η^2 = .49). Almost half the variance in the estimated relative percentage bias across the conditions was explained by the difference between the four approaches. The other main effects were all of a considerably smaller magnitude (less than 2.5% of the variance was explained). In addition, three two-way interactions were found to have an effect on the relative percentage bias. Additional analyses of variance conducted separately for each method showed that both the number of L1 individuals within each L2 group and the ICC had a strong effect on the magnitude of the relative percentage bias in the latent-measurement/manifest-aggregation approach but no effect in the manifest-measurement/latent-aggregation approach. Furthermore, the number of items and the size of the loadings had a strong effect on the magnitude of the estimated relative percentage bias in the manifest-measurement/latent-aggregation and doubly manifest approaches but almost no effect in the latent-measurement/manifest-aggregation or doubly latent approach. No three- or four-way interactions explained more than 1% of the variance in the estimated relative percentage bias.

Table 1
Percentage Variance Explained for Variance Effects of the Simulation Conditions on Bias, Root-Mean-Square Error, and Standard Error Ratio

Variable	Bias	RMSE	*SE* ratio
Main effects			
Method	48.7	3.7	4.5
J	0.1	3.7	0.0
n	1.8	10.8	6.7
ICC	2.4	21.4	3.9
K	0.1	0.4	0.2
λ	1.8	1.7	0.1
Two-way interactions			
Method × n	2.5	4.9	5.1
Method × ICC	2.7	8.9	3.5
Method × λ	3.2	1.2	0.0
Method × K	1.1	0.1	0.3
Method × J	0.1	2.2	0.1
ICC × n	0.0	4.6	9.8
ICC × J	0.0	1.4	1.0
J × n	0.0	0.3	2.3
Three-way interactions			
Method × ICC × n	0.9	3.7	8.5
J × ICC × n	0.2	0.3	2.2

Note. Two-way or higher interactions that explain less than 1% of the variance for each outcome are not reported. RMSE = root-mean-square error; method = doubly manifest, manifest-measurement/latent-aggregation, latent-measurement/manifest-aggregation, doubly latent; J = number of L2 units; n = number of L1 units within each L2 unit; ICC = intraclass correlation of predictor variable; K = number of items; λ = loadings.

The main findings for estimated relative percentage bias are also depicted in Figures 5A–5D. As shown in Figure 5A for small groups (n = 5), the doubly manifest, manifest-measurement/latent-aggregation, and latent-measurement/manifest-aggregation approaches are negatively biased, whereas the doubly latent approach is positively biased in extreme conditions (small ICC and low reliability). In this context, we need to consider that the distribution of the parameter estimates of the converged solutions is left skewed. Thus, taking the median of the parameter distribution for these extreme conditions produced a value that is closer to the true parameter. As shown in Figures 5B and 5D, the manifest-measurement/latent-aggregation and doubly latent approaches are both approximately unbiased when reliability is high (λ = 0.8): Both approaches correct for sampling error. On the other hand, as shown in Figures 5C and 5D, the latent-measurement/manifest-aggregation approach is approximately unbiased when the ICC is high and the group size is large. The manifest-measurement/latent-aggregation approach is still biased under these conditions because it does not correct for measurement error.[6]

RMSE. Next, we assessed the overall accuracy of the parameter estimates of the four approaches by estimating the RMSE. In line with previous simulation results, the estimated RMSE ranged from 0.10 to 0.53 (M = 0.30, SD = 0.11) for the doubly manifest approach and from 0.07 to 1.07 (M = 0.29, SD = 0.20) for the manifest-measurement/latent-aggregation approach. The estimated RMSE for the latent-measurement/manifest-aggregation approach ranged from 0.06 to 0.47 (M = 0.23, SD = 0.11). The estimated RMSE for the doubly latent approach was substantially larger, with values ranging from 0.05 to 3.87 (M = 0.54, SD = 0.68).

ANOVAs with the estimated RMSE as the dependent variable revealed that the largest effect was found for the ICC. More than one fifth of the variance in the RMSE was explained by the main effect of the ICC. Furthermore, the sample sizes at L1 (n) had a substantial impact, explaining about one tenth of the variance in the RMSE. In total, six significant two-way interactions were found. The most pronounced—between method and ICC—indicated that the ICC had a particularly strong effect on the RMSE in

[6] In additional analyses, we also investigated the estimated relative percentage bias of the within-group coefficient. As was to be expected for the latent-measurement/manifest-aggregation and doubly latent approaches, estimated bias was approximately zero, ranging from −1.7 to 3.7 (M = 0.2, SD = 0.8) and from −1.3 to 5.9 (M = 0.5, SD = 1.2), respectively. However, the manifest-measurement/latent-aggregation and doubly manifest approaches underestimated the size of the within-group effect, ranging from −37.9 to −4.2 (M = −20.1, SD = 10.9) and from −38.3 to −5.4 (M = −20.3, SD = 10.8), respectively. Five-way ANOVAs (ICC, number of L2 units, number of L1 units within each L2 unit, number of items, and size of loadings) with estimated relative percentage bias as the dependent variable showed that for both the manifest-measurement/latent-aggregation approach and the doubly manifest approach, the independent variables size of loadings (62%) and number of items (34%) explained most of the variance.

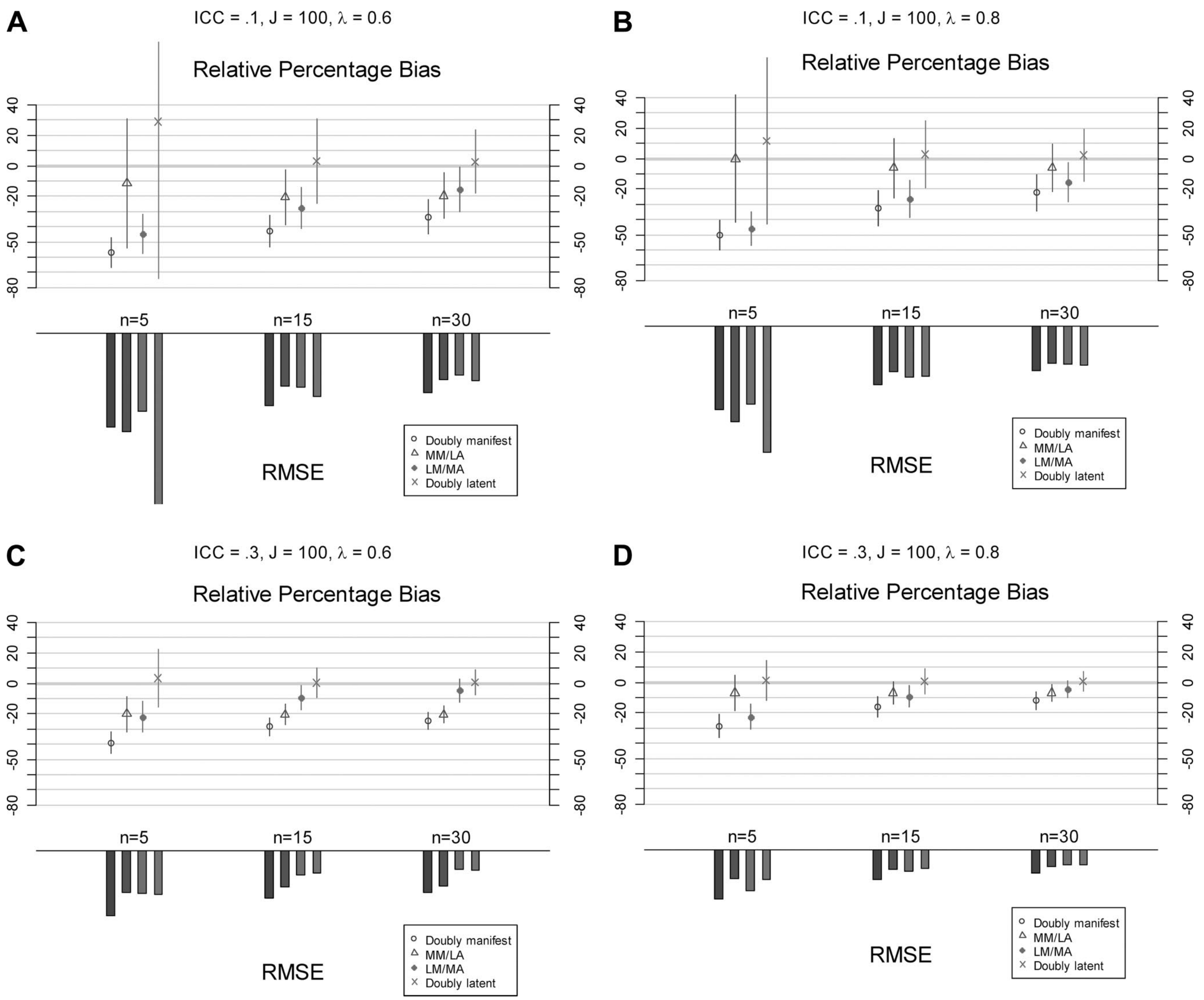


Figure 5. Relative percentage bias and root-mean-square error (RMSE) for selected conditions in the simulation study. (A) Intraclass correlation (ICC) = .1, number of groups (*J*) = 100, loadings (λ) = .6, number of indicators = 7. (B) ICC = .1, *J* = 100, λ = .8, number of indicators = 7. (C) ICC = .3, *J* = 100, λ = .6, number of indicators = 7. (D) ICC = .3, *J* = 100, λ = .8, number of indicators = 7. MM/LA = manifest-measurement/latent-aggregation; LM/MA = latent-measurement/manifest-aggregation.

the doubly latent approach. Moreover, a significant three-way interaction was found. The interaction between method, ICC, and *n* can mainly be attributed to the fact that in the doubly latent approach, the effect of the number of L1 units within each L2 unit on the RMSE is particularly strong when ICC is low. This is also illustrated in Figures 5A–5D, where the RMSE for the doubly latent approach is much higher when *n* is small in the ICC = .1 condition than in the ICC = .3 condition. Particularly, the doubly latent approach and the manifest-measurement/latent-aggregation approach show a large RMSE when little information about the L2 construct is available (small ICC and small number of L1 units within each L2 unit). Increasing the number of L1 units within each L2 unit and/or increasing the ICC has a substantial effect on the RMSE of the estimator in the doubly latent and manifest-measurement/latent-aggregation approaches.

***SE* ratio.** We also studied the behavior of the numerical standard error estimates produced by the different approaches by estimating the *SE* ratio. The standard error estimates of the doubly manifest approach (range: 0.85–1.02; *M* = 0.95, *SD* = 0.03) and the latent-measurement/manifest-aggregation approach (range: 0.80–1.03; *M* = 0.95, *SD* = 0.04) on average showed a tendency to underestimate the sampling variability of the between-group coefficient. In contrast, the manifest-measurement/latent-aggregation approach (range: 0.78–1.57; *M* = 0.96, *SD* = 0.11) and the doubly latent approach (range: 0.69–2.64; *M* = 1.05, *SD* = 0.28) showed a less clear and more variable picture in terms of over- or underestimation. ANOVAs with the *SE* ratio as the dependent variable showed that the ICC and the number of L1 units within each L2 unit play an important role in explaining the differences between the approaches, as indicated by a significant three-way interaction (see Table 1). In particular, the standard error estimates of the doubly latent approach were too large when the number of L1 units within each L2 was small (e.g., $N = 50$) and the ICC was low (e.g., ICC = .05).

Summary

Overall, the results of the simulation study confirmed the findings of the mathematical derivation, showing that the doubly manifest, manifest-measurement/latent-aggregation, and latent-measurement/manifest-aggregation approaches are biased. In terms of the estimated RMSE, the results of the simulation study indicated that—in some conditions—the partial correction of the manifest-measurement/latent-aggregation approach and the latent-measurement/manifest-aggregation approach outperformed the full correction of the doubly latent approach. To provide a more direct comparison of the four approaches, we calculated the best model in terms of RMSE for each cell of the simulation design. Figure 6 shows the percentage by which one of the four approaches outperformed the other three approaches in terms of RMSE. The percentages were calculated separately for each condition by averaging across all other conditions. There was no condition in which the doubly manifest approach emerged to be the best approach. Figure 6A shows that with a small number of groups, the latent-measurement/manifest-aggregation is superior to the other approaches. However, with an increasing number of L2 units, the doubly latent approach shows better performance in terms of the RMSE. Figure 6B indicates that the difference between the latent-measurement/manifest-aggregation approach and the doubly latent approach vanishes with increasing ICC. In Figure 6C, the results for group size confirmed the observation that the doubly latent approach shows better performance with larger L1 sample sizes. Finally, Figure 6D shows that the difference between the manifest-measurement/latent-aggregation approach and the latent-measurement/manifest-aggregation approach diminishes with an increasing number of items.

Study 2: Assessing the Assumption of Invariance of Factor Loadings

A limitation of Study 1 was that the data-generating model assumed for the predictor variable was a very simple measurement model in which loadings were constrained to be invariant (cross-level invariance) and of equal size within L1 and L2. Although it can be questioned that these conditions are met in real data, there are several reasons why the assumption of invariant loadings is typical in research practice (Marsh et al., 2009; Raykov & Penev, 2009; Zyphur, Kaplan, & Christian, 2008). First, invariant cross-level factor loadings equate the metric of the latent factors across levels, making latent factor variances directly comparable

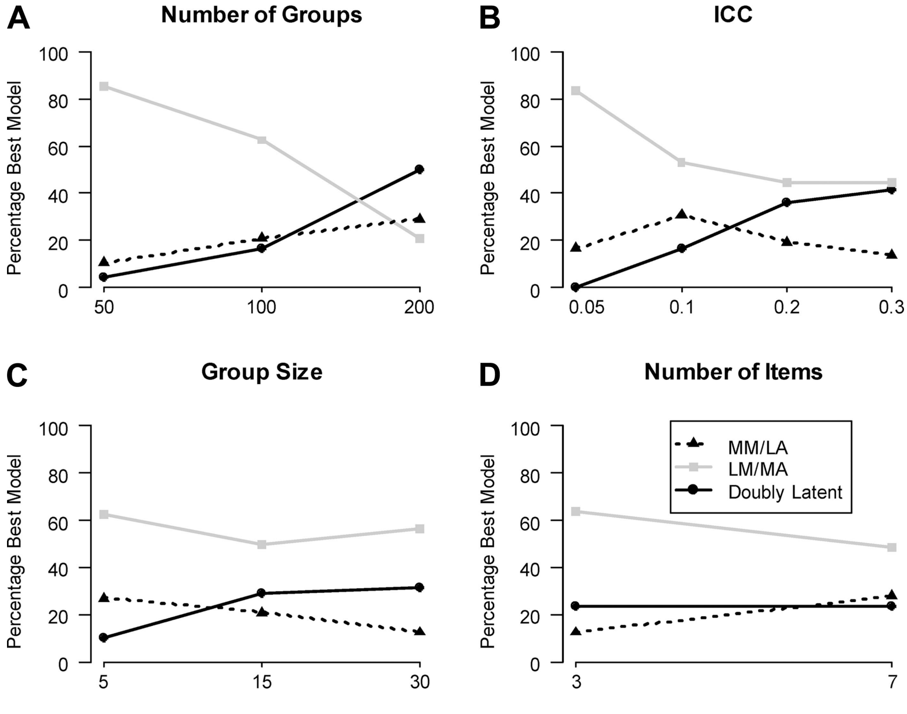


Figure 6. Percentage by which one of the four approaches outperformed the other three approaches in terms of the root-mean-square error of the between-group regression coefficient. Percentages were calculated for number of groups, intraclass correlation (ICC), group size, and number of items separately by averaging across all other conditions. MM/LA = manifest-measurement/latent-aggregation; LM/MA = latent-measurement/manifest-aggregation.

across levels (Mehta & Neale, 2005). More specifically, cross-level invariance establishes a measurement model in which the latent covariate is a simple decomposition of the within (deviation of the latent individual scores from the latent cluster means) and between (deviation of the latent cluster means) components. Second, cross-level invariance constraints facilitate the interpretation of coefficients when contextual effects are assessed, ensuring that there is a common metric at both within and between levels. Third, assuming invariant loadings across levels has the advantage that it reduces the complexity of the measurement model and that information from L1 can be used to estimate the much more unstable factor loadings at L2. Hence, it might be speculated that the trade-offs in accuracy and bias between full and partial correction approaches that were observed in Study 1 might even be more pronounced when the invariance constraints in the measurement model are relaxed. However, assuming invariant factor loadings in the analysis model when in fact the patterns of factor loadings at L1 and L2 differ in the population might result in strongly biased estimates of contextual effects. To address these concerns, we conducted an additional simulation study evaluating the performance of the doubly latent approach and the latent-measurement/manifest-aggregation approach when different patterns of factor loadings at L1 and L2 are specified.

Conditions

The model used to generate the data was the same as in Study 1, except that different patterns of unstandardized factor loadings were assumed at L1 and L2. The conditions manipulated were the number of L2 groups (50 and 200), the number of observations per L2 group (five and 20), and the ICC of the predictor variable (.05, .10, and .20). The latent variable was assumed to be measured by three indicators, and the loading of the first indicator was fixed to 1 at L1 and L2 in order to identify the metric of the latent variable (see Appendix B for details of the derivation of the parameters of the measurement model). The unstandardized loadings of the second and third indicator were held equal within levels but allowed to differ across levels (L1: 0.6, 0.8, 1.0; L2: 1.0, 1.2), resulting in six conditions of factor loading patterns for the three items: one invariant pattern (L1: 1.0, 1.0, 1.0; L2: 1.0, 1.0, 1.0) and five patterns with a “small” (L1: 1.0, 0.8, 0.8; L2: 1.0, 1.0, 1.0; and L1: 1.0, 1.0, 1.0; L2: 1.0, 1.2, 1.2), “medium” (L1: 1.0, 0.6, 0.6; L2: 1.0, 1.0, 1.0; and L1: 1.0, 0.8, 0.8; L2: 1.0, 1.2, 1.2), and “large” degree of noninvariance (L1: 1.0, 0.6, 0.6; L2: 1.0, 1.2, 1.2). For each of the $2 \times 2 \times 3 \times 6 = 72$ conditions, 1,000 simulated data sets were generated. The data sets were analyzed with four analysis models. First, the doubly latent approach was used, with the first loading at L1 and L2 being set to 1 and the second and third loading being freely estimated at L1 and L2 (DL noninvariant). Second, the second and third indicators were held invariant across levels (DL invariant). Third, all three loadings were fixed to 1 at L1 and L2 (DL parallel). Fourth, we used the latent-measurement/manifest-aggregation approach and fixed all loadings to 1 at L1 and L2 (LM parallel).[7] This approach can be regarded as a partial correction approach that strongly reduces the complexity of the estimated model. In all analysis models, item residuals were freely estimated at L1 and L2. In our analysis of this simulation study, we focus on the relative percentage bias and the RMSE of the estimator of the between-group regression coefficient.

Results and Discussion

Convergence rate. Most convergence problems occurred for the doubly latent approach with noninvariant loadings (DL-noninvariant), particularly when the number of groups was small and the ICC was low. The low convergence rate for these conditions was almost independent of the degree of noninvariance in the factor loadings. For instance, in the condition with $n = 5$, $J = 50$, and ICC = .05, only 5% of the solutions converged given a large degree of noninvariance, and only 6% converged for the invariant pattern. Increasing the ICC raised the number of converged solutions; however, even with an ICC of .20, less than 50% of the solutions of the DL-noninvariant model converged for all six patterns of factor loadings. Only when the number of L1 units within each L2 unit was increased ($n = 20$) did almost every solution converge at an ICC of .20. Because of the low convergence rate of the doubly latent approach, we present results only for the conditions with a large number of groups ($J = 200$). Tables 13–18 in the supplemental materials provide details of relative percentage bias and RMSE for all conditions of this simulation.

Bias. Table 2 shows the estimated relative percentage bias in the parameter estimates for different patterns of factor loadings. The DL-noninvariant model was approximately unbiased when the number of L1 units within each L2 unit was large ($n = 20$) but positively biased when the number L1 of units within each L2 unit was small ($n = 5$) and the ICC was small (e.g., ICC = .05). These positively biased estimates were at least in part due to the exclusion of nonconverging solutions, which still occurred even with 200 groups, particularly when the ICC was small. At a medium (L1: 1.0, 1.0, 1.0; L2: 1.0, 1.2, 1.2) or large (L1: 1.0, 0.6, 0.6; L2: 1.0, 1.2, 1.2) degree of noninvariance in factor loadings, the DL-invariant and DL-parallel models provided negatively biased estimates in the case of a large number of L1 units within each L2 unit and a large ICC. This estimated bias was due to the misspecification in the measurement model of the predictor variable for these two approaches. As expected, the parameter estimates of the latent-measurement/manifest-aggregation approach were negatively biased in all conditions. The negative bias of the latent-measurement/manifest-aggregation approach was larger when the number of L1 units within each L2 unit was small.

RMSE. Next, we assessed the RMSE, which combines bias and variability of the parameter estimates (see Table 2). The main finding is that in many conditions, the latent-measurement/manifest-aggregation approach showed a smaller RMSE although the measurement model was misspecified. For instance, when the number of L1 units within each L2 unit was small, the latent-measurement/manifest-aggregation approach outperformed the three doubly latent approaches in terms of estimated RMSE, par-

[7] Additional unreported analyses showed that the latent-measurement/manifest-aggregation approach with parallel loadings (LM parallel) outperformed the other two latent-measurement/manifest-aggregation models with noninvariant and invariant loadings in terms of RMSE. In 44 conditions (61%) of the simulation study, LM parallel was the best model in terms of RMSE. Although these results for the latent-measurement/manifest-aggregation approach clearly depend on the conditions selected in our simulation, we decided to report only the results for the LM-parallel model in order to keep the presentation of Study 2 simple.

Table 2

Relative Percentage Bias and Root-Mean-Square Error for the Doubly Latent and Latent-Measurement/Manifest-Aggregation Approaches as a Function of Different Patterns of Loadings (Number of L2 Units = 200)

		n = 5				n = 20			
Pattern of loadings	ICC	DL noninvariant	DL invariant	DL parallel	LM/MA	DL noninvariant	DL invariant	DL parallel	LM/MA
				Relative percentage bias					
L1: 1.0	.05	44.1	38.7	33.4	−55.0	2.7	2.4	2.3	−34.1
L2: 1.0	.10	19.1	13.1	13.2	−45.9	1.2	1.0	1.0	−22.1
	.20	4.3	3.5	3.4	−31.6	0.3	0.2	0.3	−12.1
L1: 1.0	.05	62.3	27.1	16.1	−55.4	3.3	−6.8	−7.3	−36.4
L2: 1.2	.10	15.0	0.7	−0.6	−46.8	1.1	−8.0	−9.0	−26.4
	.20	3.9	−3.8	−6.2	−33.9	0.6	−7.3	−8.9	−18.6
L1: 0.6	.05	36.6	15.7	46.9	−42.4	2.9	−28.6	−8.5	−26.3
L2: 1.2	.10	12.5	−14.7	−1.6	−32.6	0.6	−22.0	−10.0	−18.9
	.20	2.4	−14.9	−9.4	−23.4	0.1	−14.7	−9.6	−13.6
				Root-mean-square error					
L1: 1.0	.05	1.74	1.43	1.16	0.40	0.27	0.26	0.25	0.27
L2: 1.0	.10	0.60	0.40	0.40	0.33	0.15	0.14	0.14	0.18
	.20	0.18	0.16	0.16	0.23	0.08	0.08	0.08	0.11
L1: 1.0	.05	2.16	1.15	0.87	0.40	0.27	0.23	0.23	0.28
L2: 1.2	.10	0.45	0.31	0.30	0.34	0.14	0.13	0.14	0.20
	.20	0.17	0.15	0.14	0.25	0.09	0.09	0.10	0.14
L1: 0.6	.05	0.89	1.71	1.91	0.32	0.23	0.27	0.21	0.23
L2: 1.2	.10	0.38	0.30	0.28	0.25	0.14	0.19	0.14	0.16
	.20	0.15	0.16	0.14	0.18	0.08	0.12	0.09	0.11

Note. For the patterns of loadings, the values of the second and third indicator at L1 and L2 are given. ICC = intraclass correlation; n = number of L1 individuals within each L2 unit; DL noninvariant = doubly latent model with noninvariant factor loadings; DL invariant = doubly latent model with invariant factor loadings; DL parallel = doubly latent model with parallel factor loadings; LM/MA = latent-measurement/manifest-aggregation model with parallel factor loadings.

ticularly when the ICC was low. Note that the differences are reported for a substantial number of L2 units (J = 200); the results for the conditions with a small number of L2 units (J = 50) were even more strongly in favor of the latent-measurement/manifest-aggregation approach. It is also interesting to note that the specification of invariant or parallel loadings in the doubly latent approach (DL invariant and DL parallel) had a positive effect on the estimated RMSE, at least when the misspecification of the measurement model was only moderate. For example, under a medium degree of invariance in factor loadings, the DL-invariant and DL-parallel models provided more accurate estimates than the DL-noninvariant model when the number of L1 units was small. These findings indicate that especially when the ICC is low and the number of L1 units within each L2 unit is small, reducing the complexity of the model by assuming a simpler measurement model can increase the accuracy of the parameter estimates.

Summary

Overall, this simulation showed that a misspecified measurement model that incorrectly assumes invariant or parallel factor loadings can result in biased estimates of the between-group effect in the doubly manifest and latent-measurement/manifest-aggregation approaches. However, the simulation also showed that the specification of invariant or parallel loadings in the analysis model can provide more accurate parameter estimates, even when the model assumed to generate the data has a different pattern of loadings at L1 and L2. Particularly when the data provided less information about the L2 construct (e.g., low ICC, small number of L1 units), the approaches with a simplified measurement model outperformed the doubly latent approach with freely estimated factor loadings in terms of RMSE. Furthermore, the results are in line with Study 1, showing that partial error correction approaches can outperform full correction approaches in certain data constellations. Hence, the assumption of invariant loadings made for interpretational purposes (e.g., same metric across levels) was also justifiable from a statistical perspective: Parameter estimates of the between-group effect in fully estimated measurement models at L1 and L2 showed so much sampling variability that reducing the complexity of the measurement model by placing constraints on the loadings resulted in more accurate estimates with data constellations characteristic of psychological research.

Study 3: Manifest and Latent Variable Approaches With Actual Data: An Application From Educational Psychology

In this real-data application from educational psychology, we examine students' perceptions of a specific teaching behavior (see Lüdtke et al., 2008). Students were asked to rate how easily distracted their mathematics teacher was (teacher distractibility) on three items (e.g., "Our mathematics teacher is easily distracted if something attracts his/her attention"). The scale was developed on the basis of Kounin (1970), and taps teacher behavior that leads to the disruption or discontinuation of learning activities in class. Such behavior makes lessons less efficient and is negatively related to students' learning gains (Gruehn, 2000). From a measurement perspective, all student ratings are supposed to measure the same construct (i.e., the teacher behavior under study), and the referent is the teacher. L1 student responses are thus used to construct an L2 reflective construct that represents a specific teacher characteristic, namely distractible teaching style (Cronbach, 1976; Miller & Murdock, 2007).

The data stemmed from the German sample of lower secondary students who participated in the Third International Mathematics and Science Study (Baumert et al., 1997; Beaton et al., 1996). The data set contains 2,133 students nested within 108 classes (average cluster size = 19.75). The ICC of the student ratings for the three items was .09, .04, and .05, indicating that a moderate proportion of the total variance was located at the class level. The amount of variance located at the student level indicates that there is a considerable lack of agreement among students about the distractibility of their mathematics teacher. The internal consistency (Cronbach's α) of the scale was calculated to be .72.

All four approaches (doubly manifest, manifest-measurement/latent-aggregation, latent-measurement/manifest-aggregation, and doubly latent) were specified in Mplus (see supplemental materials for the Mplus syntax). For the doubly manifest approach, the distractibility scores were aggregated at the class level. For the latent-measurement/manifest-aggregation approach, the three manifest L2 indicators were class averages of the corresponding L1 indicators. The parameter estimates for the four approaches are presented in Table 3.[8] To establish a comparable metric for the L1 and L2 effects of distractibility on mathematics achievement, we set factor loadings for the latent-measurement/manifest-aggregation and doubly latent approaches to be equal across the two levels.

For all four approaches, there was no effect of the individual students' perception of their teachers' teaching style on individual achievement (relations at the student level, L1). Inspection of the L2 effect of distractibility on mathematics achievement revealed that classes with teachers who were perceived as showing a higher level of distractibility at the class level had lower levels of achievement than classes with teachers who were perceived to be less distractible. However, the estimate of the doubly latent approach is almost twice the size of that of the doubly manifest approach. Interestingly, the manifest-measurement/latent-aggregation approach is closer to the doubly latent approach than is the latent-measurement/manifest-aggregation approach. This result indicates that adjusting for sampling error is more relevant in the present example than correcting for measurement error (i.e., the three items allow a reliable assessment of Distractibility at the class level in terms of measurement error).[9]

In addition, we compared the results of the data example with the results of the simulation study for the corresponding data constellation ($n = 15$, ICC = .10, $J = 100$; $\lambda = 0.8$; see Table 3 in the supplemental materials). The order of the parameter estimates in the empirical example closely matches the order of the relative percentage bias in the simulation study: -38.2 (doubly manifest), -14.2 (manifest-measurement/latent-aggregation), -25.7 (latent-measurement/manifest-aggregation), 4.7 (doubly latent). However, when interpreting these results, readers should bear in mind that none of the four approaches addresses the problem of group-level confounders. It is therefore possible that class-level perceived distractibility of the teacher is correlated with other class-level variables (e.g., how often the teacher is distracted by poor student behavior and therefore perceived as distractible) that also affect student achievement.

General Discussion

Traditionally, MLM of group effects has relied on manifest scale scores that do not take the unreliability of measured variables into account. The present study introduced a 2 × 2 taxonomy of multilevel latent contextual models for estimating group effects when a construct is measured with error. Extensive simulations comparing the approaches showed that the doubly latent approach that corrects for measurement error and sampling error results in unbiased estimates of L2 constructs under appropriate conditions. However, when only limited information on the L2 construct is available in the data (e.g., low ICC, small number of groups and number of persons within groups), partial correction approaches (manifest-measurement/latent-aggregation and latent-measurement/manifest-aggregation) can outperform full correction approaches (doubly latent), providing more accurate estimates in terms of RMSE. Hence, in many data constellations characteristic of psychological research, partial error correction approaches may provide more accurate parameter estimates than full error correction approaches.

Discussion and Limitations of the Simulation Study

As is true of any simulation study, the results cannot be generalized beyond the specific conditions implemented. Moreover, a number of limitations of our simulation study need to be mentioned. The first limitation is that the performance of the four approaches was explored only with multivariate normally distributed data. Latent variable models like the full and partial error correction models that use maximum-likelihood procedures to obtain parameter estimates are usually based on assumptions such as multivariate normality of the variables involved. However, the traditional assumption of multivariate normality is not crucial, as several estimation methods are able to circumvent it. For example, the software used in the present study (Mplus; estimator MLR) provides robust standard error estimates based on corrections proposed by Yuan and Bentler (1998). Furthermore, we note that

[8] Note that the results for the doubly manifest and manifest-measurement/latent-aggregation approaches deviate from the results reported in Lüdtke et al. (2008). This deviation is attributable to a difference in the standardization of the predictor variables. In the present study, the scales were not z standardized before the analyses in order to keep them in the same metric as the single indicators that were used for the manifest-measurement/latent-aggregation and doubly latent approaches.

[9] The fit of the two models with multiple indicators was satisfactory: latent-measurement/manifest-aggregation: $\chi^2(6) = 26.56$, comparative fit index = .98, root-mean-square error of approximation = .04; doubly latent: $\chi^2(7) = 24.32$, comparative fit index = .98, root-mean-square error of approximation = .03. The fit of the other two models cannot be evaluated because they are saturated. In the doubly latent model, one item displayed a residual variance that was very close to zero. We fixed that variance to zero in order to avoid estimation problems.

Table 3
Empirical Analysis Results: Effects of Students' Perception of Their Teachers' Distractibility on Mathematics Achievement

Variable	DM		MM/LA		LM/MA		DL	
	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*	Estimate	*SE*
L1 (within)								
Distractibility factor loadings								
Indicator 1					1.000	0.000	1.000	0.000
Indicator 2					1.145	0.075	1.133	0.075
Indicator 3					0.942	0.060	0.939	0.059
Math on distractibility	−0.037	0.022	−0.037	0.022	−0.043	0.032	−0.044	0.032
R^2	.001	.001	.001	.002	.001	.002	.001	.002
L2 (between)								
Distractibility factor loadings								
Indicator 1					1.000	0.000	1.000	0.000
Indicator 2					1.145	0.075	1.133	0.075
Indicator 3					0.942	0.060	0.939	0.059
Math on distractibility	−1.186	0.274	−1.766	0.438	−1.469	0.389	−2.190	0.652
R^2	.167	.069	.232	.104	.180	.084	.255	.116

Note. N = 2,133 (L1), 108 (L2); Average cluster size = 19.75. All parameter estimates except the effect of distractibility on mathematics achievement at L1 are statistically significantly different from zero ($p < .01$). DM = doubly manifest; MM/LA = manifest-measurement/latent-aggregation; LM/MA = latent-measurement/manifest-aggregation; DL = doubly latent.

critical issues of robustness in relation to violations of multivariate normality typically relate to standard errors rather than to parameter estimates per se. Clearly, further research is needed to investigate the behavior of robust and normal theory maximum-likelihood standard error estimates under strong violations of multivariate normality in MLM of group effects (e.g., Hox, Maas, & Brinkhuis, 2010).

Second, we observed serious convergence problems for the doubly latent approach in conditions where little information was available for assessing the L2 construct (i.e., low ICC, small sample size at L1 and L2). The main reason for nonconvergence was a nonpositive definite between covariance matrix. Regularization methods for nonconvergence problems due to inadmissible parameter estimates are discussed in the statistical literature. The basic idea is that a small constant a multiplied by an identity matrix $\mathbf{I}$ is added to the nonpositive covariance matrix $\mathbf{S}$ so that the resulting regularized covariance matrix $\mathbf{S}_a = \mathbf{S} + a\mathbf{I}$ is positive definite (Amemiya, 1985; Yuan & Chan, 2008). Using the regularized between-group covariance matrix $\mathbf{S}_a$ thus allows researchers to obtain parameter estimates even when little information is available for assessing the L2 construct. This approach is similar to using prior information to estimate parameters, as is done in a Bayesian framework (Gelman, Carlin, Stern, & Rubin, 2003). When the between-covariance matrix $\mathbf{S}$ contains little information, reasonable prior information improves the overall inference of the unknown parameter. It would be interesting to conduct simulation studies to investigate how the performance of these regularization methods compares with partial correction approaches, particularly when the sample size at L1 and L2 is small.

Third, it is important to emphasize that the doubly latent model was the "true" model, in that the pattern of free and fixed parameters was the same in the data-generating model used to generate the simulated data and in the approximating model used to fit the data. However, in certain data constellations, the "wrong" partial correction models came closer to the true value of the data-generating model, because biases of the partial correction models were offset by their lower variability when little information was available for assessing the L2 construct. Further simulation research is needed to investigate the behavior of the different approaches when the true model is not the doubly latent model but, for instance, the latent-measurement/manifest-aggregation model. From a statistical point of view, this would require all L1 units from each L2 unit to be sampled and the corresponding L2 construct to be assessed without sampling error. Lüdtke et al. (2008) conducted a simulation study in which the data-generating model was doubly manifest and the analysis models were both the doubly manifest approach and the manifest-measurement/latent-aggregation approach. In this simulation, it was assumed that the number of L1 units within each L2 unit in the data-generating model was some fixed number (e.g., 100). When the sampling ratio approached 100%—particularly when the number of individuals within each group was small—the manifest-measurement/latent-aggregation approach overestimated the unreliability that was due to sampling error, resulting in positively biased estimates of the contextual effect.

Implications for Analysis of Contextual Effects in Psychological Research

Given that the partial error correction approaches sometimes outperformed the full error correction approaches under conditions characteristic of psychological research, what are the implications for the assessment of contextual effects? From a statistical perspective, we believe that a reasonable approach is to juxtapose the different approaches and to use the estimates of the doubly manifest, partial, and full correction approaches as bounds for the true parameter (Marsh et al., 2009; Morgan & Winship, 2007). For instance, in our data example, the estimate of the doubly latent approach was almost twice the size of that of the doubly manifest approach, with the estimates of the manifest-measurement/latent-aggregation and latent-measurement/manifest-aggregation approaches falling in between. Hence, the four approaches provide a range of values indicating how ignoring error, adjusting for sampling error or measurement error, and correcting for both types of error impacts the estimated effect of the teacher's perceived distractibility on student achievement. The two simulation studies provided clear evidence that correcting for different sources of error introduces variability in the parameter estimates. Thus, the partial and full error correction approaches may reduce the bias of parameter estimates but do not necessarily increase the accuracy of the resulting estimates. Thus, balancing the estimates of contextual effects by applying all four error correction approaches is a reasonable strategy for dealing with the observed accuracy–bias trade-offs.[10]

From a more theoretical perspective, the distinction between formative and reflective aggregation of L1 constructs is of central importance for the 2 × 2 taxonomy of MLC models. First, the corrections in the manifest-measurement/latent-aggregation and doubly latent approaches are based on the assumption that an infinite number of L1 units have been sampled from each L2 unit. Thus, the appropriateness of these approaches depends in part on the nature of the construct under study (Lüdtke et al., 2008). For reflective aggregations of L1 constructs, it seems reasonable to assume a potentially infinite number of L1 units within each L2 unit: In this context, an unobserved group construct is assessed by reference to a finite and exchangeable sample of individual observers at L1. However, for formative aggregations of L1 constructs (e.g., gender ratio), it is reasonable to assume an infinite sampling process only under specific circumstances. First, it can be argued that the manifest-measurement/latent-aggregation and doubly latent approaches are appropriate for formative aggregation when the sampling ratio is low (i.e., a small number of L1 units is sampled from each L2 unit) and there is thus considerable unreliability due to sampling error (e.g., Marsh et al., 2009; Preacher et al., 2010). In this case, a finite sampling process may be well approximated by an infinite sampling process. For example, if only five students from a school of 1,000 are sampled, an infinite sampling process could be assumed even if the L2 construct under study was based on formative aggregation. However, in the case of larger sampling ratios, finite sampling corrections need to be applied. The evaluation and development of finite sampling corrections is clearly a topic for future research (e.g., Goldstein et al., 2008; Grilli & Rampichini, 2009). Second, it could be argued that even when all L1 units are sampled from an L2 unit, the observed group mean must still be regarded as only a proxy for an underlying true score. Goldstein et al. (2008) give an example from educational data, in which the average attainment of a school in an achievement test is used as a proxy for that school's long-term intake characteristics (also see Shin & Raudenbush, 2010). In this case, one is not interested in the observed group-average score, but wants to generalize across different conditions that may influence the data generated for a group (e.g., specific time point, specific achievement test). The process of generalizing an observed score across different conditions (i.e., what might have happened if the circumstances had been slightly different) is expressed in the survey sampling literature by the idea of a superpopulation from which the observed data are generated. This idea concedes that the properties of the population are unknown or at least partially unknown, even when all units have been sampled (Särndal, Swensson, & Wretman, 2003).

Directions for Future Research

In the past decade, substantial progress has been made in integrating hierarchical linear modeling or multilevel analysis and SEM techniques within a single analytic framework (e.g., Skrondal & Rabe-Hesketh, 2004). The present analysis examined a relatively simple model in which only intercepts were allowed to vary (random intercept models) and unreliability of a single predictor variable was taken into account. An obvious extension would be multilevel models with cross-level interactions or models that allow slopes to vary between L2 units (e.g., Marsh et al., 2009). It would be important to explore how error in the predictor variable affects the estimation of the variance component of the slope (e.g., LaHuis & Ferguson, 2009) and cross-level interaction between L1 and L2 variables. Using the big-fish-little-pond effect as a substantive example, Marsh et al. (2009) have demonstrated how the doubly latent model can be easily extended to include latent interactions and nonlinear effects. Furthermore, Preacher and colleagues (Preacher, Zhang, & Zyphur, 2011; Preacher et al., 2010) have proposed a general multilevel SEM framework for assessing multilevel mediation that builds on the correction of sampling error used by the manifest-measurement/latent-aggregation and doubly latent approaches when estimating the effect of group-level variables. Further simulation research is needed to investigate how partial and full error correction approaches perform in these more realistic cases of multiple predictor variables. Based on the results of the present study, it might be expected that the partial correction approaches become even more relevant because the chance of unstable between-group covariance matrices increases with the number of variables.

Currently, the four approaches can be estimated by means of the Mplus software and the generalized linear latent and mixed model module for Stata (Rabe-Hesketh et al., 2004). From a statistical point of view, both Mplus and the generalized linear latent and mixed model are based on a frequentist framework. Recently, it has been shown how a Bayesian framework using the WinBUGS (Spiegelhalter, Thomas, Best, Gilks, & Lunn, 2002) software, which is based on Markov chain Monte Carlo methods, can be used to integrate MLM and SEM in a very flexible way (Lee, 2007; Segawa, Emery, & Curry, 2008). A Bayesian approach to MLM has proved to be particularly promising when the number of

[10] Further analysis of the simulation results (Study 1) supports the idea of using the different error correction approaches as bounds for the true parameter. Analysis of the order of the sizes of the parameter estimates for the between-group effect revealed that in almost all replications, the lowest value was provided by the doubly manifest approach and the highest value by the doubly latent approach. The order of the other two approaches depended on whether more sampling error or measurement error was present in the generated data. However, in extreme conditions, when little information on the L2 construct was available in the data (e.g., ICC = .05, $n = 5$, $J = 50$), for a substantial number of data sets (e.g., 30%), the lowest and highest values were not provided by the doubly manifest and doubly latent approaches, respectively. This was particularly the case when the between effect estimated by the doubly manifest approach was close to zero, resulting in unstable corrections of the other approaches.

groups is small or when the model to be estimated is very complicated (Gelman & Hill, 2007; Swaminathan & Rogers, 2008). One special application that demonstrates the flexibility of a Bayesian approach is an integration of a cross-classified multilevel model and a structural equation model that is suited for three-mode data (González, De Boeck, & Tuerlinckx, 2008). In the context of the proposed 2 × 2 taxonomy, this approach would allow researchers to take into account a third component of sampling (e.g., sampling of persons, sampling of items, sampling of situations), thus integrating a general measurement model that takes into account different sources (e.g., facets or modes) of sampling error as given by generalizability theory (Brennan, 2001) with the structural part of a model that aims at estimating relations among constructs. Future research should compare the different methods of combining MLM and SEM.

Conclusion

MLM allows relationships among variables located at group and individual levels to be explored simultaneously. In research practice, group-level variables (L2) for assessing contextual effects are frequently generated by aggregating variables from a lower level (L1). Two types of error occurring in multilevel data when measuring L2 constructs by aggregating from L1 measures can be distinguished: unreliability that is due to measurement error and unreliability that is due to sampling error. This results in a 2 × 2 taxonomy of multilevel latent contextual models that distinguishes between an uncorrected approach (doubly manifest), partial correction approaches (manifest-measurement/latent-aggregation and latent-measurement/manifest-aggregation), and a full correction approach (doubly latent). The appropriateness of each approach depends on the nature of the L2 construct under study, the nature of the research question, and the specific data constellation (number of L2 groups, number of cases within each group, sampling ratio). In the case of a reflective L2 construct and multiple indicators, the full correction of the doubly latent approach should be applied. However, when little information is available for assessing the reflective L2 construct, accuracy–bias trade-offs in parameter estimates can make the partial correction approaches a more appropriate choice. In the case of a formative L2 construct and multiple indicators, the partial correction approach that adjusts only for measurement error would be the natural choice. However, in real-world applications, it is often difficult to decide which of the models of the 2 × 2 taxonomy should be selected (e.g., the nature of L2 construct might be ambiguous in relation to a specific research question, the sampling ratio might be unknown). A reasonable approach is to juxtapose the different error correction approaches and to use the estimates of the different approaches as bounds for the true parameter by showing how ignoring error versus adjusting for sampling error and/or measurement error affects the estimated contextual effect.

References

Amemiya, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite. *American Statistician, 39,* 112–117. doi:10.2307/2682808

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion.* Princeton, NJ: Princeton University Press.

Asparouhov, T., & Muthén, B. (2007). *Constructing covariates in multilevel regression* (Mplus Web Notes No. 11, Version 2). Retrieved from http://www.statmodel.com/download/webnotes/webnote11.pdf

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

Beaton, A. E., Mullis, I. V. S., Martin, M. O., Gonzales, E. J., Kelly, D. L., & Smith, T. A. (1996). *Mathematics achievement in the middle school years: IEA's Third International Mathematics and Science Study.* Chestnut Hill, MA: Boston College.

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundation, extensions, and new directions* (pp. 349–381). San Francisco, CA: Jossey-Bass.

Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin, 110,* 305–314. doi:10.1037/0033-2909.110.2.305

Bovaird, J. A. (2007). Multilevel structural equation models for contextual factors. In T. D. Little, J. A. Bovaird, & N. A. Card (Eds.), *Modeling contextual effects in longitudinal studies* (pp. 149–182). Mahwah, NJ: Erlbaum.

Brennan, R. L. (2001). *Generalizability theory.* New York, NY: Springer-Verlag.

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

Cronbach, L. J. (1976). *Research on classrooms and schools: Formulations of questions, design and analysis.* Stanford, CA: Stanford Evaluation Consortium.

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

Curran, P. J., & Bauer, D. J. (2007). Building path diagrams for multilevel models. *Psychological Methods, 12,* 283–297. doi:10.1037/1082-989X.12.3.283

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174. doi:10.1037/1082-989X.5.2.155

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods, 12,* 121–138. doi:10.1037/1082-989X.12.2.121

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). *Bayesian data analysis.* London, England: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* Cambridge, England: Cambridge University Press.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, England: Arnold.

Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassifications in multilevel models. *Statistical Modelling, 8,* 243–261. doi:10.1177/1471082X0800800302

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53,* 455–467. doi:10.1007/BF02294400

González, J., De Boeck, P., & Tuerlinckx, F. (2008). A double-structure structural equation model for three-mode data. *Psychological Methods, 13,* 337–353. doi:10.1037/a0013269

Grilli, L., & Rampichini, C. (2009). *Measurement error in multilevel models with sample cluster means* (Working Paper 2009/06). Retrieved from Department of Statistics, University of Florence, Italy, website: http://www.ds.unifi.it/ricerca/pubblicazioni/working_papers/2009/wp2009_06.pdf

Gruehn, S. (2000). *Unterricht und schulisches Lernen: Schüler als Quellen der Unterrichtsbeschreibung* [Instruction and learning in school: Students as sources of information]. Münster, Germany: Waxmann.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12,* 205–218. doi:10.1037/1082-989X.12.2.205

Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica, 64,* 157–170. doi:10.1111/j.1467-9574.2009.00445.x

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219–229. doi:10.1037/0021-9010.67.2.219

Johnson, J. E., Burlingame, G. M., Olsen, J. A., Davies, D. R., & Gleave, R. L. (2005). Group climate, cohesion, alliance, and empathy in group psychotherapy: Multilevel structural equation models. *Journal of Counseling Psychology, 3,* 310–321. doi:10.1037/0022-0167.52.3.310

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). Multilevel measurement modeling. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 345–388). Charlotte, NC: Information Age.

Kane, M. T., Gillmore, G. M., & Crooks, T. J. (1976). Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement, 13,* 171–183. doi:10.1111/j.1745-3984.1976.tb00009.x

Kenny, D. A. (1979). *Correlation and causality.* New York, NY: Wiley.

Kenny, D. A., Mannetti, L., Pierro, A., Livi, S., & Kashy, D. A. (2002). The statistical analysis of data from small groups. *Journal of Personality and Social Psychology, 83,* 126–137. doi:10.1037/0022-3514.83.1.126

Kounin, J. S. (1970). *Discipline and group management in classrooms.* New York, NY: Holt, Rinehart & Winston.

Kreft, I. G. G., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21. doi:10.1207/s15327906mbr3001_1

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12,* 418–435. doi:10.1177/1094428107308984

LeBreton, J. M., & Senter, J. L. (2008). Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods, 11,* 815–852. doi:10.1177/1094428106296642

Lee, S.-Y. (2007). *Structural equation modelling: A Bayesian approach.* New York, NY: Wiley. doi:10.1002/9780470024737

Lohr, S. L. (1999). *Sampling: Design and analysis.* Pacific Grove, CA: Duxbury.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

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

Lüdtke, O., Robitzsch, A., Trautwein, U., & Kunter, M. (2009). Assessing the impact of learning environments: How to use student ratings of classroom or school characteristics in multilevel modelling. *Contemporary Educational Psychology, 34,* 120–131. doi:10.1016/j.cedpsych.2008.12.001

Lüdtke, O., Trautwein, U., Kunter, M., & Baumert, J. (2006). Reliability and agreement of student ratings of the classroom environment: A reanalysis of TIMSS data. *Learning Environments Research, 9,* 215–230. doi:10.1007/s10984-006-9014-8

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology, 1,* 86–92. doi:10.1027/1614-2241.1.3.85

Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov. T., Muthén, B., & 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:10.1080/00273170903333665

Masterson, S. S. (2001). A trickle-down model of organizational justice: Relating employees' and customers' perceptions of and reactions to fairness. *Journal of Applied Psychology, 86,* 594–604. doi:10.1037/0021-9010.86.4.594

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

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations modeling. *Psychological Methods, 10,* 259–284. doi:10.1037/1082-989X.10.3.259

Miller, A. D., & Murdock, T. B. (2007). Modeling latent true scores to determine the utility of aggregate student perceptions as classroom indicators in HLM: The case of classroom goal structures. *Contemporary Educational Psychology, 32,* 83–104. doi:10.1016/j.cedpsych.2006.10.006

Morgan, S. L., & Winship, C. (2007). *Counterfactual and causal inferences: Methods and principles for social research.* Cambridge, England: Cambridge University Press.

Moritz, S. E., & Watson, C. B. (1998). Levels of analysis issues in group psychology: Using efficacy as an example of a multilevel model. *Group Dynamics: Theory, Research, and Practice, 2,* 285–298. doi:10.1037/1089-2699.2.4.285

Muthén, B. O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585. doi:10.1007/BF02296397

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354. doi:10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O., & Asparouhov, T. (2008). Growth mixture modeling: Analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbecke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 143–165). Boca Raton, FL: Chapman & Hall/CRC.

Muthén, L. K., & Muthén, B. O. (2007). *Mplus user's guide.* Los Angeles, CA: Muthén & Muthén.

O'Brien, R. M. (1990). Estimating the reliability of aggregate-level variables based on individual-level characteristics. *Sociological Methods and Research, 18,* 473–504. doi:10.1177/0049124190018004004

Papaioannou, A., Marsh, H. W., & Theodorakis, Y. (2004). A multilevel approach to motivational climate in physical education and sport settings: An individual or a group level construct? *Journal of Sport & Exercise Psychology, 26,* 90–118.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantage of multilevel SEM. *Structural Equation Modeling, 18,* 161–182. doi:10.1080/10705511.2011.557329

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods, 15,* 209–233. doi:10.1037/a0020141

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69,* 167–190. doi:10.1007/BF02295939

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raykov, T., & Penev, S. (2009). Estimation of maximal reliability for multiple-component instruments in multilevel designs. *British Journal of Mathematical and Statistical Psychology, 62,* 129–142. doi:10.1348/000711007X255345

Ryan, A. M., Gheen, M. H., & Midgley, C. (1998). Why do some students avoid asking for help? An examination of the interplay among students' academic efficacy, teachers' social–emotional role, and the classroom goal structure. *Journal of Educational Psychology, 90,* 528–535. doi:10.1037/0022-0663.90.3.528

Särndal, C.-E., Swensson, B., & Wretman, J. (2003). *Model assisted survey sampling.* New York, NY: Springer.

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

Schneider, B., White, S. S., & Paul, M. C. (1998). Linking service climate and customer perceptions of service quality: Tests of a causal model. *Journal of Applied Psychology, 83,* 150–163. doi:10.1037/0021-9010.83.2.150

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components.* New York, NY: Wiley.

Segawa, E., Emery, S., & Curry, S. J. (2008). Extended generalized linear latent and mixed model. *Journal of Educational and Behavioral Statistics, 33,* 464–484. doi:10.3102/1076998607307359

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35,* 26–53. doi:10.3102/1076998609345252

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66,* 563–575. doi:10.1007/BF02296196

Skrondal, A., & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal, and structural equation models.* Boca Raton, FL: Chapman & Hall/CRC.

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

Spiegelhalter, D. J., Thomas, A., Best, N. G., Gilks, W. R., & Lunn, D. (2002). *BUGS: Bayesian inference using Gibbs sampling.* Cambridge, England: MRC Biostatistics Unit.

Swaminathan, H., & Rogers, H. J. (2008). Estimation procedures for hierarchical linear models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 469–519). Charlotte, NC: Information Age.

Traub, R. E. (1994). *Reliability for the social sciences: Theory and applications.* Thousand Oaks, CA: Sage.

Trautwein, U., Lüdtke, O., Schnyder, I., & Niggli, A. (2006). Predicting homework effort: Support for a domain-specific, multilevel homework model. *Journal of Educational Psychology, 98,* 438–456. doi:10.1037/0022-0663.98.2.438

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Yuan, K.-H., & Bentler, P. M. (1998). Robust mean and covariance structure analysis. *British Journal of Mathematical and Statistical Psychology, 51,* 63–88.

Yuan, K.-H., & Chan, W. (2008). Structural equation modeling with near singular covariance matrices. *Computational Statistics and Data Analysis, 52,* 4842–4858. doi:10.1016/j.csda.2008.03.030

Zyphur, M. J., Kaplan, S. A., & Christian, M. S. (2008). Assumptions of cross-level measurement and structural invariance in the analysis of multilevel data: Problems and solutions. *Group Dynamics: Theory, Research, and Practice, 12,* 127–140. doi:10.1037/1089-2699.12.2.127

Appendix A

Derivation of Bias for the Doubly Manifest Approach

The following population model will be assumed for two variables X and Y, each measured with multiple indicators (see Snijders & Bosker, 1999, p. 29):

$$X_{kij} = \mu_{xk} + U_{xj} + U_{xij} + R_{xkj} + R_{xkij};\ \ k = 1, \ldots, K$$

$$Y_{lij} = \mu_{yl} + U_{yj} + U_{yij} + R_{ylj} + R_{ylij};\ \ l = 1, \ldots, L,$$

where U_{xj} and U_{yj} are the true scores at L2 and U_{xij} and U_{yij} are the true scores at L1. Furthermore, for each item k, R_{xkj} and R_{ylj} denote measurement error at L2, and R_{xkij} and R_{ykij} denote measurement error at L1. At both levels, L1 and L2, strictly parallel measurements are assumed (i.e., same loading and same error variances within each level).

The covariance matrix of the true scores U_{xij} and U_{yij} at L1 and U_{xj} and U_{yj} at L2 can be written as

$$\begin{array}{cc} \text{L1 (within)} & \text{L2 (between)} \\ \begin{pmatrix} \sigma_x^2 & \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix} & \begin{pmatrix} \tau_{\mathrm{x}}^2 & \\ \tau_{\mathrm{xy}} & \tau_{\mathrm{y}}^2 \end{pmatrix} \end{array}.$$

Because parallel measurements were assumed at both levels, for the error variances of X, it holds that $\mathrm{Var}(R_{xkij}) = \sigma_{x,e}^2$ and $\mathrm{Var}(R_{xkj}) = \tau_{x,e}^2$, for all $k = 1, \ldots, K$. The scale average score $\bar{Y}_{\bullet ij}$ for the dependent variable is calculated by averaging across the L items. Note that the size of the error variances $\mathrm{Var}(R_{ylij}) = \sigma_{y,e}^2$ and $\mathrm{Var}(R_{ylj}) = \tau_{y,e}^2$, for all $l = 1, \ldots, L$, is of no importance for the magnitude of the bias (e.g., Wooldridge, 2002). We are now interested in estimating the following relationship in the population:

(*Appendices continue*)

$$\bar{Y}_{\bullet ij} = \mu_y + \beta_x U_{xij} + \beta_b U_{xj} + \delta_j + \varepsilon_{ij},$$

where μ_y is the grand mean, β_w the within-group regression coefficient, β_b the between-group regression coefficient, δ_j a group-specific residual, and ε_{ij} an individual-specific residual.

The manifest score can be calculated by averaging across the K items at L1,

$$\bar{X}_{\bullet ij} = \frac{1}{K}\sum_{k=1}^{K} X_{kij},$$

and by averaging across items and groups at L2,

$$\bar{X}_{\bullet\bullet j} = \frac{1}{Kn}\sum_{k=1}^{K}\sum_{i=1}^{n} X_{kij}$$

(under the assumption of equal group sizes n). Based on these manifest scores, the following model can then be specified to estimate the relationship in the population:

$$\bar{Y}_{\bullet ij} = \gamma_{00} + \gamma_{10}(\bar{X}_{\bullet ij} - \bar{X}_{\bullet\bullet j}) + \gamma_{01}\bar{X}_{\bullet\bullet j} + u_{0j} + r_{ij},$$

where U_{xij} is approximated by $(\bar{X}_{\bullet ij} - \bar{X}_{\bullet\bullet j})$ and U_{xj} is approximated by $\bar{X}_{\bullet\bullet j}$. Furthermore, γ_{00}, γ_{10}, and γ_{01} denote the estimators for μ_y, β_w, and β_b. The L2 and L1 residuals are given by u_{0j} and r_{ij}. Given the covariance matrix of the true scores at L1 and L2, and given the assumption of parallel measurements, the observed covariance matrix of $\bar{Y}_{\bullet ij}$, $\bar{X}_{ij} - \bar{X}_{\bullet\bullet j}$, and $\bar{X}_{\bullet\bullet j}$ is distributed as follows:

$$\mathrm{Cov}\begin{bmatrix} \bar{Y}_{\bullet ij} \\ \bar{X}_{\bullet ij} - \bar{X}_{\bullet\bullet j} \\ \bar{X}_{\bullet\bullet j} \end{bmatrix} =$$

$$\begin{pmatrix} \sigma_y^2 + \sigma_{y,e}^2/K + \tau_y^2 + \tau_{y,e}^2/K & & \\ \sigma_{xy}\cdot(1 - 1/n) & \sigma_x^2(1 - 1/n) + \sigma_{x,e}^2(1 - 1/n)/K & \\ \tau_{xy} + \sigma_{xy}/n & 0 & \tau_x^2 + \tau_{x,e}^2/K + \sigma_x^2/n + \sigma_{x,e}^2/(nK) \end{pmatrix}.$$

As can be seen, the covariances between $\bar{Y}_{\bullet ij}$, $\bar{X}_{ij} - \bar{X}_{\bullet\bullet j}$, and $\bar{X}_{\bullet\bullet j}$ depend on the common group size, the number of items, the measurement error, and the "true" covariances within and between groups. The reliability of the observed scores at L1 and L2 is now defined as follows (see Equations 12 and 13):

$$\mathrm{Rel}_{\mathrm{L1}}(\bar{X}_{\bullet ij}) = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{x,e}^2/K}, \quad \mathrm{Rel}_{\mathrm{L2}}(\bar{X}_{\bullet ij}) = \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2/K}.$$

Employing the ordinary-least-squares principle and bearing in mind that the predictors $\bar{X}_{\bullet ij} - \bar{X}_{\bullet\bullet j}$ and $\bar{X}_{\bullet\bullet j}$ are uncorrelated, the bias for the within-group coefficient $\hat{\gamma}_{10}$ can be obtained as

$$E(\hat{\gamma}_{10} - \beta_w) = -\beta_w\cdot(1 - \mathrm{Rel}_{\mathrm{L1}}).$$

Now let $\mathrm{ICC}_x = \tau_x^2/(\tau_x^2 + \sigma_x^2)$ denote the ICC for X. To obtain the expected bias of the between-group regression coefficient, we first derive the estimator $\hat{\gamma}_{01}$ of the between-group regression coefficient β_b:

(*Appendices continue*)

$$
\begin{aligned}
\hat{\gamma}_{01} &= \frac{\mathrm{Cov}(\bar{Y}_{\bullet ij}, \bar{X}_{\bullet\bullet j})}{\mathrm{Var}(\bar{X}_{\bullet\bullet j})} \\
&= \frac{\tau_{xy} + \sigma_{xy}/n}{\tau_x^2 + \tau_{x,e}^2/K + \sigma_x^2/n + \sigma_{x,e}^2/(nK)} \\
&= \frac{\tau_{xy}}{\tau_x^2} \cdot \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2/K + \sigma_x^2/n + \sigma_{x,e}^2/(nK)} + \frac{\sigma_{xy}/n}{\sigma_x^2/n} \cdot \frac{\sigma_x^2/n}{\tau_x^2 + \tau_{x,e}^2/K + \sigma_x^2/n + \sigma_{x,e}^2/(nK)} \\
&= \beta_b \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{\tau_x^2 \cdot \mathrm{Rel}_{\mathrm{L1}}}{\tau_x^2 \cdot \mathrm{Rel}_{\mathrm{L1}} + \sigma_x^2 \cdot \mathrm{Rel}_{\mathrm{L2}}/n} + \beta_w \cdot \frac{\sigma_x^2 \cdot \mathrm{Rel}_{\mathrm{L1}} \cdot \mathrm{Rel}_{\mathrm{L2}}/n}{\tau_x^2 \cdot \mathrm{Rel}_{\mathrm{L1}} + \sigma_x^2 \cdot \mathrm{Rel}_{\mathrm{L2}}/n} \\
&= \beta_b \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{\mathrm{Rel}_{\mathrm{L1}} \cdot ICC_x}{\mathrm{Rel}_{\mathrm{L1}} \cdot ICC_x + \mathrm{Rel}_{\mathrm{L2}} \cdot (1 - ICC_x)/n} + \beta_w \cdot \frac{1}{n} \cdot \frac{\mathrm{Rel}_{\mathrm{L1}} \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot (1 - ICC_x)}{\mathrm{Rel}_{\mathrm{L1}} \cdot ICC_x + \mathrm{Rel}_{\mathrm{L2}} \cdot (1 - ICC_x)/n}.
\end{aligned}
$$

In order to further simplify the expression, let $r = \mathrm{Rel}_{\mathrm{L2}}/\ \mathrm{Rel}_{\mathrm{L1}}$ and $b = (1 - \mathrm{ICC}_x)/\mathrm{ICC}_x$. The expectation of the between-group coefficient can then be written as follows:

$$\beta_b \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{1}{1 + r \cdot b/n} + \beta_w \cdot \frac{b}{n} \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{1}{1 + r \cdot b/n}.$$

The expected bias for the between-group regression coefficient can now be calculated as

$$E(\hat{\gamma}_{01} - \beta_b) = -\beta_b \cdot \left(1 - \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{1}{1 + r \cdot b/n}\right) + \beta_w \cdot \frac{b}{n} \cdot \mathrm{Rel}_{\mathrm{L2}} \cdot \frac{1}{1 + r \cdot b/n}.$$

As can be seen, the bias depends primarily on the reliability at L2 ($\mathrm{Rel}_{\mathrm{L2}}$), the proportion of variance in X that is located between groups (b), and the average group size n.

Assuming that $\mathrm{Rel} = \mathrm{Rel}_{\mathrm{L1}} = \mathrm{Rel}_{\mathrm{L2}}$ further simplifies the formula for the expected bias of the between-group regression coefficient:

$$E(\hat{\gamma}_{01} - \beta_b) = -\beta_b \cdot \left(1 - \mathrm{Rel} \cdot \frac{1}{1 + b/n}\right) + \beta_w \cdot \frac{b}{n} \cdot \mathrm{Rel} \cdot \frac{1}{1 + b/n}.$$

It is now evident that if $n \to \infty$, the bias of the between-group regression coefficient depends only on the reliability of measurement of X.

Appendix B

Derivation of Parameters for the Data-Generating Models of the Two Simulation Studies

In the following, details are provided of the derivation of the parameters of the data-generating models in the two simulation studies.

Simulation Study 1

First, let us assume that the predictor variable X is measured by multiple indicators. Under the assumption of mean-centered indicators, a single indicator X_{kij} for person i in group j can be decomposed as follows (Kamata et al., 2008; B. O. Muthén, 1991):

$$X_{kij} = \lambda_{k,W} U_{xij} + R_{xkij} + \lambda_{k,B} U_{xj} + R_{xkj};\ k = 1, \ldots, K,$$

where $\lambda_{k,W}$ are the within-factor loadings, $\lambda_{k,B}$ are the between-factor loadings, R_{xkij} and R_{xkj} are the residuals at L1 and L2, and U_{xij} and U_{xj} are the unobserved true scores at L1 and L2, which are assumed to be normally distributed with zero means and $\mathrm{Var}(U_{xij}) = \sigma_x^2$ and $\mathrm{Var}(U_{xj}) = \tau_x^2$. In Study 1, parallel measures were specified by setting unstandardized factor loadings to 1 at L1 ($\lambda_{k,W} = \lambda_{k',W} = 1$) and L2 ($\lambda_{k,B} = \lambda_{k',B} = 1$) and specifying equal measurement error variances at L1, $\mathrm{Var}(R_{xkij}) = \mathrm{Var}(R_{xk'ij}) = \sigma_{x,e}^2$, and L2, $\mathrm{Var}(R_{xkj}) = \mathrm{Var}(R_{xk'j}) = \tau_{x,e}^2$. After dropping the index k for indicators because parallel measures are assumed, the observed variance of a single indicator can be decomposed as follows:

(*Appendices continue*)

$$\mathrm{Var}(X_{ij}) \;=\; \sigma_x^2 + \sigma_{x,e}^2 + \tau_x^2 + \tau_{x,e}^2.$$

The standardized factor loading for a single indicator at L2 and L1 is now defined as

$$\lambda_{B,st}^2 \;=\; \frac{\tau_x^2}{\tau_x^2 + \tau_{x,e}^2}$$

and

$$\lambda_{W,st}^2 = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_{x,e}^2}.$$

In Study 1, the standardized factor loadings were specified to be equal at L1 and L2 ($\lambda_{B,st} = \lambda_{W,st}$), assuming equal reliability of the indicators at both levels. In addition, it was assumed that the total variance of the predictor variable X equals 1, by specifying $\mathrm{Var}(U_{xj}) = \tau_x^2$ and $\mathrm{Var}(U_{xij}) = 1 - \tau_x^2$.

The item residual variances at L2 and L1 are then given as follows:

$$\tau_{x,e}^2 = \frac{\tau_x^2(1 - \lambda_{B,st}^2)}{\lambda_{B,st}^2}$$

and

$$\sigma_{x,e}^2 = \frac{(1 - \tau_x^2)(1 - \lambda_{W,st}^2)}{\lambda_{W,st}^2}.$$

Thus, for example, the residual variances at L2 and L1 for the data-generating model for the conditions with intraclass coefficient (ICC) = .10 and standardized factor loadings = 0.6 (see Figure 4) are

$$\tau_{x,e}^2 = \frac{.10 \cdot (1 - 0.6^2)}{0.6^2} = 0.18$$

and

$$\sigma_{x,e}^2 = \frac{.90 \cdot (1 - 0.6^2)}{0.6^2} = 1.6.$$

Simulation Study 2

In Study 2, we used the same setup as in Study 1 but relaxed the assumption of parallel measures at L1 ($\lambda_{k,W} \neq \lambda_{k',W}$) and L2 ($\lambda_{k,B} \neq \lambda_{k',B}$) and specified the unstandardized loadings to be noninvariant across the two levels. To fix the metric of the latent variable, we set the first unstandardized factor loading at L1 and L2 to 1 ($\lambda_{1,B} = \lambda_{1,W} = 1$). The standardized factor loadings for a single indicator k at L2 and L1 are then given by

$$\lambda_{k,B,st}^2 = \frac{\lambda_{k,B}^2\tau_x^2}{\lambda_{k,B}^2\tau_x^2 + \mathrm{Var}(R_{xkj})}$$

and

$$\lambda_{k,W,st}^2 = \frac{\lambda_{k,W}^2\sigma_x^2}{\lambda_{k,W}^2\sigma_x^2 + \mathrm{Var}(R_{xkij})}.$$

To arrive at reasonable population parameters, we fixed the standardized factor loading at L2 for each indicator to 0.9 and assumed that the ICC for the latent variable was smaller than the ICC for an observed indicator by factor 0.75 (B. O. Muthén, 1991). Assigning different values to the unstandardized loadings (i.e., L1: 0.6, 0.8, 1.0; L2: 1.0, 1.2) and again specifying the total variance of the latent predictor variable to be 1—that is, $\mathrm{Var}(U_{xj}) = \tau_x^2$ and $\mathrm{Var}(U_{xij})=1-\tau_x^2$—the residual variance for indicator k at L2 is then given by the following relation:

$$\mathrm{Var}(R_{xkj}) = \frac{(1 - \lambda_{k,B,st}^2)\lambda_{k,B}^2\tau_x^2}{\lambda_{k,B,st}^2}.$$

For example, in the condition with unstandardized loadings set to $\lambda_{2,B} = \lambda_{3,B} = 1.2$ and $\lambda_{2,W} = \lambda_{3,W} = 0.8$ and a latent ICC of .2, the residual variance for the second indicator at L2 equals

$$\mathrm{Var}(R_{xkj}) = \frac{(1 - 0.9^2) \cdot 1.2^2 \cdot .2}{0.9^2} = 0.07.$$

Given that the ICC of the latent variable is smaller than the ICC of the manifest indicators by factor 0.75, the residual variance for the second indicator at L1 is $\mathrm{Var}(R_{xkij}) = 0.50$.

Received August 27, 2009
Revision received March 2, 2011
Accepted April 14, 2011 ■