Psychological Methods
2008, Vol. 13, No. 3, 203–229

1082-989X/08/$12.00 DOI: 10.1037/a0012869

The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies

Oliver Lüdtke
Max Planck Institute for Human Development

Herbert W. Marsh
Oxford University

Alexander Robitzsch
Institute for Educational Progress

Ulrich Trautwein
Max Planck Institute for Human Development

Tihomir Asparouhov
Muthén & Muthén

Bengt Muthén
University of California, Los Angeles

In multilevel modeling (MLM), group-level (L2) characteristics are often measured by aggregating individual-level (L1) characteristics within each group so as to assess contextual effects (e.g., group-average effects of socioeconomic status, achievement, climate). Most previous applications have used a multilevel manifest covariate (MMC) approach, in which the observed (manifest) group mean is assumed to be perfectly reliable. This article demonstrates mathematically and with simulation results that this MMC approach can result in substantially biased estimates of contextual effects and can substantially underestimate the associated standard errors, depending on the number of L1 individuals per group, the number of groups, the intraclass correlation, the sampling ratio (the percentage of cases within each group sampled), and the nature of the data. To address this pervasive problem, the authors introduce a new multilevel latent covariate (MLC) approach that corrects for unreliability at L2 and results in unbiased estimates of L2 constructs under appropriate conditions. However, under some circumstances when the sampling ratio approaches 100%, the MMC approach provides more accurate estimates. Based on 3 simulations and 2 real-data applications, the authors evaluate the MMC and MLC approaches and suggest when researchers should most appropriately use one, the other, or a combination of both approaches.

Keywords: multilevel modeling, contextual analysis, latent variables, structural equation modeling, Mplus

Supplemental materials: http://dx.doi.org/10.1037/a0012869.supp

In the last 2 decades multilevel modeling (MLM) has became one of the central research methods for applied researchers in the social sciences. A major advantage of MLMs over single-level regression analysis lies in the possibility of exploring relationships among variables located at different levels simultaneously (Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). In the typical application of MLM, outcome variables are related to several predictor variables at the individual level (L1; e.g., students, employees) and at the group level (L2; e.g., schools, work groups, neighborhoods).

Different types of group-level variables can be distinguished. The first type can be measured directly (e.g., class size, school budget, neighborhood population). These variables that cannot be broken down to the individual level are often referred to as "*global*" or "*integral*" *variables* (Blakely & Woodward, 2000). The second type is generated by aggregating variables from a lower level. For example, ratings of school climate by individual students may be aggregated at the school level, and the resulting mean can be

Oliver Lüdtke and Ulrich Trautwein, Center for Educational Research, Max Planck Institute for Human Development, Berlin, Germany; Herbert W. Marsh, Department of Education, Oxford University, Oxford, United Kingdom; Alexander Robitzsch, Institute for Educational Progress, Humboldt University, Berlin; Tihomir Asparouhov, Muthén & Muthén, Los Angeles, California; Bengt Muthén, Graduate School of Education & Information Studies, University of California, Los Angeles.

We thank Marcel Croon, Harvey Goldstein, and David Kenny for comments on earlier versions of this article and Susannah Goss for editorial assistance.

Correspondence concerning this article should be addressed to Oliver Lüdtke, Max Planck Institute for Human Development, Center for Educational Research, Lentzeallee 94, 14195 Berlin, Germany. E-mail: luedtke@mpib-berlin.mpg.de

used as an indicator of the school's collective climate. Variables that are obtained through the aggregation of scores at the lower level are known as *contextual* or *analytical variables*. For instance, Anderman (2002), using a large data set with students nested within schools, examined the relations between school belonging and psychological outcomes (e.g., depression, optimism). School belonging was included in the multilevel regression model as both an individual (L1) characteristic and a school (L2) characteristic. School-level belonging was based on the within-school aggregation of individual perceptions of school belonging. In a similar vein, Ryan, Gheen, and Midgley (1998) related student reports of avoidance of help-seeking to student and classroom goals (for other applications, see Harker & Tymms, 2004; Kenny & La Voie, 1985; Lüdtke, Köller, Marsh, & Trautwein, 2005; Miller & Murdock, 2007; Papaioannou, Marsh, & Theodorakis, 2004).

Croon and van Veldhoven (2007) have emphasized the applicability of these issues to many subdisciplines of psychology, including educational, organizational, cross-cultural, personality, and social psychology. Iverson (1991) provided a brief summary of the extensive application of contextual analyses in sociology, dating as far back as Durkheim's study of suicide and including topics as diverse as the racial composition of neighborhoods, village use of contraceptives, local crime statistics, political behavior in election districts, families in the study of socioeconomic status (SES) and schooling, volunteer organizations, churches, workplaces, and social networks. In fact, the issues are central to any area of research in which individuals interact with other individuals in a group setting, leading Iverson to conclude: "This range of areas illustrates how broadly contextual analysis has been used in the study of human behavior" (p. 11).

In the MLM literature, models that include the same variable at both the individual level and the aggregated group level are called *contextual analysis models* (Boyd & Iverson, 1979; Firebaugh, 1978; Raudenbush & Bryk, 2002) or sometimes *compositional models* (e.g., Harker & Tymms, 2004). The central question in contextual analysis is whether the aggregated group characteristic has an effect on the outcome variable after controlling for interindividual differences at the individual level. The effects of the L1 characteristic may or may not be of central importance, depending on the nature of the study and the L1 construct (e.g., Papaioannou et al., 2004).

One problematic aspect of the contextual analysis model is that the observed group average obtained by aggregating individual observations may not be a very reliable measure of the unobserved group average if only a small number of L1 individuals is sampled from each L2 group (O'Brien, 1990; Raudenbush, Rowan, & Kang, 1991). For instance, in educational research, where only a small proportion of students might be sampled from each participating school, the observed group average is only an approximation of the unobserved "true" group mean—a latent variable. When MLMs are used to estimate the contextual analysis model, it is typically assumed that the observed L2 variables based on aggregated L1 variables are measured without error. However, when only a small number of L1 units are sampled from each L2 group, the L2 aggregate measure may be unreliable and result in a biased estimate of the contextual effect.

In the present study we introduce a latent variable approach, implemented in the latent variable modeling software Mplus (Asparouhov & Muthén, 2006; B. O. Muthén, 2002; L. K. Muthén & Muthén, 2007; but see also B. O. Muthén, 1989; Schmidt, 1969), which takes the unreliability of the group mean into account when estimating the contextual effect. Because the group average is treated as a latent variable, we call this approach the *multilevel latent covariate (MLC) model*. In contrast, we label the "traditional" approach, which relies solely on the (manifest) observed group mean, the *multilevel manifest covariate (MMC) model*. The term *manifest* indicates that this approach treats the observed group means as manifest and does not infer from them to an unobserved latent construct that controls for L2 unreliability.

Our article is organized as follows. We start by distinguishing between reflective and formative L2 constructs. We then give a brief description of how the MMC is usually specified in MLMs, outlining the factors that affect the reliability of the group mean and deriving mathematically the bias that results from using the MMC approach to estimate the contextual effect. After introducing the MLC model as it is implemented in Mplus, we summarize the results of simulation studies comparing the statistical properties of the latent and manifest approaches. In addition, we present analyses comparing the Croon and van Veldhoven (2007) two-step approach to our (one-step) MLC approach. We then present two empirical examples using both the latent and the manifest approaches. Finally, on the basis of all of these results, we offer suggestions for the applied researcher and propose directions for further research.

Reflective and Formative L2 Constructs

We argue that the appropriateness of the MLC approach depends in part on the nature of the construct under study. For the present purposes, we propose a distinction between formative and reflective aggregations of L1 constructs (for more general discussion of formative and reflective measurement, see Bollen & Lennox, 1991; Edwards & Bagozzi, 2000; Kline, 2005; also see Howell, Breivik, & Wilcox, 2007). Although our choice of terms is based on a factor analytic rationale, a related distinction is made in the organizational psychology literature (e.g., Bliese, 2000; Bliese, Chan, & Ployhart, 2007; also see Kozlowski & Klein, 2000) between compilation (or configural) models and composition models.

Formative (compilation or configural) aggregations of L1 constructs are considered to be an index of L1 measures within each L2 group (i.e., arrows in the underlying structural equation model go from the L1 indicators to the L2 construct; e.g., Kline, 2005). Formative constructs have the following characteristics: The focus of L1 measures is on an L1 construct, L1 individuals within the same L2 group are likely to have different L1 true scores, and scores for different individuals within the same L2 group are *not* interchangeable. There is no expectation that the individual level and aggregated variables reflect the same construct; thus, corresponding L1 and L2 measures are not assumed to be isomorphic. For formative aggregated L2 constructs, variation among individuals can be thought of as a substantively important group characteristic (i.e., groups are relatively heterogeneous or homogeneous in relation to a specific L1 characteristic). Particularly when the sampling ratio (the percentage of L1 individuals considered within each L2 group) approaches 100%, it is inappropriate to use variation within each L2 group (intraclass correlation [ICC]) to estimate L2 unreliability due to sampling error. Whereas the focus of our research has been on the mean as an aggregate summary used to construct a group (L2) construct, Kozlowski and Klein (2000) emphasize that various indexes of L1 constructs within a group could be used as the L2 aggregated (formative) measure (e.g., minimum, maximum, variation, profile similarity, system dynamics, etc.). For example, let us assume that a researcher wants to evaluate the gender composition of students in each of a large number of different classes and has information for all students within each class. An appropriate L2 aggregate variable (e.g., percentage of girls) can be measured with essentially no measurement or sampling error at either L1 or L2. Students within each class are clearly not interchangeable in relation to gender, and even if a particular class—by chance or design—happens to have a disproportionate number of boys or girls, this feature of the class reflects a true characteristic of that class rather than unreliability due to sampling error. Hence, as emphasized by Kozlowski and Klein (2000), it might be appropriate to consider measures of within-group heterogeneity (diversity) as a potentially useful L2 aggregated (formative) construct. Examples of formative L2 aggregated constructs might include L2 aggregations of L1 characteristics such as race, age, gender, achievement levels, SES, or other background/demographic characteristics of individuals within a group. Making a similar point, Bliese (2000) noted that for pure compilation-based aggregate measures (similar to our formative L2 variables), there is no assumption of within-group agreement and that measures of reliability based on within-group agreement tend to be irrelevant in establishing the construct validity of the L2 measures.

Reflective (or compositional) aggregations of L1 constructs have the following characteristics: The purpose of L1 measures is to provide reflective indicators of an L2 construct, all L1 indicators (typically different individuals within the same group) within each L2 group are designed to measure the same L2 construct, and scores associated with different individuals within the same L2 group are interchangeable. The L2 construct is assumed to "cause" the L1 indicators (i.e., arrows in the underlying structural model go from the latent L2 construct to the L1 indicators). Thus, reflective aggregations are analogous to the typical latent variable approach based on classical measurement theory and the domain sampling model (Kline, 2005; Nunnally & Bernstein, 1994), in which multiple indicators (in this case, multiple persons within each group rather than the multiple items for each construct) are used to infer a latent construct that is corrected for unreliability (based on the number of indicators and the extent of agreement among the multiple indicators) that would otherwise result in biased estimates. Hence, the concept of reflective measurement is consistent with the notion of a generic group-level construct that is measured by individual responses (Cronbach, 1976; Croon & van Veldhoven, 2007). Under these conditions, it is reasonable to use variation within each L2 group (ICC) to estimate L2 sampling error that includes error due to finite sampling and error due to a selection of indicators (i.e., a specific constellation of individuals used to measure a group-level construct). Within-group variation represents lack of agreement among individuals within the same group in relation to an L2 construct rather than a substantively important characteristic of the group. Examples of reflective L2 constructs might include individual ratings of classroom, group, or team climate; individual ratings of the effectiveness of a teacher, coach, or group leader; individual marker ratings of the quality of written compositions, performances, artwork, grant proposals, or journal article submissions. Within the organizational psychology literature, the term referent shift measures (Chan, 1998; Chen, Bliese, & Mathieu, 2005) is used to denote the case in which the referent for a measure shifts from that of the individual (e.g., individual self-efficacy) to that of the group (the efficacy of the group as a whole). Particularly when the referent of the measures is the group as a whole, the resulting aggregated measures might be considered a reflective (or compositional) construct.

The distinction between formative and reflective variables is particularly important in climate research (for further discussion, see Papaioannou et al., 2004). For example, if all individual students within each of a large number of different classes are asked to rate the competitive orientation of their class as a whole, the aggregated L2 construct might be most appropriately represented as an L2 reflective construct. The observed measure is designed to reflect the L2 construct directly and is not intended to reflect a characteristic of the individual student. However, if each individual student is asked to rate his or her own competitive orientation, the

aggregated L2 construct might be more appropriately considered as a formative L2 construct. The observed L1 measure is designed to reflect an L1 construct rather than to be a direct measure of an L2 construct, even if the L2 aggregation of the L1 measures is used to infer an L2 construct. We would expect agreement among different ratings by students within the same class (ICC) to be substantially higher for the L2 reflective construct than for the corresponding L2 formative construct. Whereas lack of agreement among students within the same class on the L2 reflective variable can be used to infer L2 unreliability, lack of agreement on the L2 formative construct reflects within-class heterogeneity in relation to an L1 construct.

Bliese (2000) argued that pure compositional models (like our L2 reflective aggregation measures) require complete isomorphism in which every group member provides exactly the same score so that there is no variation within groups on the relevant L1 construct. Noting that cases of pure isomorphism between L1 and L2 constructs are extremely rare (except, perhaps, in highly artificial situations), he described a "'fuzzy'" composition construct involving both compositional and compilation processes. Similarly, we contend that L2 constructs based on an aggregation of L1 constructs vary along a continuum in which pure reflective and pure formative constructs represent the endpoints. Although we focus on the endpoints of the continuum, we note that most L2 aggregated constructs fall somewhere between the reflective and formative endpoints of this continuum. We also note that we chose the terms formative versus reflective because these terms have better established meanings in the psychometric literature in relation to the underlying structural equation model used to define them. In contrast, there is less consistency in the use of the compilation (or configural) versus composition models in the organizational psychology literature (Bliese, 2000; Kozlowski & Klein, 2000). Indeed, aggregated variables resulting from a formative aggregation process such as SES or stock market indexes are commonly referred to as composite measures, which is the exact opposite of the implicit meaning of compositional models in the organizational psychology literature. Nevertheless, our use of a formative–reflective continuum of L2 aggregated constructs is essentially the same as the pure compositional to pure compilational continuum used by Bliese (2000) in his discussion of fuzzy compositional L2 aggregated variables.

Contextual Analysis

The Contextual Analysis Model in Multilevel Modeling

In this section, we provide a short description of the contextual analysis model in the traditional multilevel framework. We assume that we have 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:

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

where the variable 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 his or her group's mean. At Level 2, the L1 intercepts β_{0j} and slopes β_{1j} are dependent variables:

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

$$\beta_{1j} = \gamma_{10}, \quad (2)$$

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. 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. Now inserting the L2 equations into the L1 equation we have

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

This notation is referred to as the linear mixed-effect notation (McCulloch & Searle, 2001) and is used, for example, by the mixed module in SPSS and similar procedures in other statistical packages. Equation 3 reveals that the main difference between a single-level regression analysis and an MLM lies in the more complex error structure of the multilevel specification. Furthermore, it is now easy to see that γ_{10} is the within-group regression coefficient describing the relationship between Y and X within groups and that γ_{01} is the between-groups 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 higher than γ_{10}, meaning that the relationship at the aggregated level is stronger than the relationship at the individual level.

Grand-mean centering. Another approach to test for a contextual effect (which is mathematically equivalent under certain conditions; see Raudenbush & Bryk, 2002) is to use a different centering option for the individual-level predictor. 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. Substituting the group-mean $\bar{X}_{\cdot j}$ in Equation 3 by the grand-mean $\bar{X}_{\cdot\cdot}$ gives the following model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \bar{X}_{\cdot\cdot}) + \gamma_{01}\bar{X}_{\cdot j} + u_{0j} + r_{ij}. \quad (4)$$

In contrast to the group-mean centered model, where the predictor variables are orthogonal, the predictors $(X_{ij} - \bar{X}_{\cdot\cdot})$ and $\bar{X}_{\cdot j}$ in this grand-mean centered model are not independent. Thus, γ_{10} is the specific effect of the group mean after controlling for interindividual differences on X. Note that, in the grand-mean centered model, the individual deviations from the grand mean, $(X_{ij} - \bar{X}_{\cdot\cdot})$, also include the person's group deviation from the grand mean. Consequently, a contextual effect is present if γ_{01} is statistically significantly different from zero. 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 Kreft, de Leeuw, & Aiken, 1995). For the fixed effects, the following relation holds for the L2 between-groups regression coefficient: $\gamma_{01}^{\text{grandmean}} = \gamma_{01}^{\text{groupmean}} - \gamma_{10}^{\text{groupmean}}$. The within-group regression coefficient at Level 1 will be the same in both models: $\gamma_{10}^{\text{grandmean}} = \gamma_{10}^{\text{groupmean}}$. Hence, the results for the fixed part of the grand-mean centered model can be obtained from the group-mean centered model by a simple subtraction.[1] Because our analysis is limited to random-intercept models, centering of predictor variables will not be a critical issue in our article. In the remainder of the article, our investigation of the analysis of group effects in MLM focuses on the group-mean centered case.

The reliability of the group mean for reflective aggregations of L1 constructs. One problematic aspect of the contextual analysis model, as described earlier, is that the observed group average $\bar{X}_{\cdot j}$ might be a highly unreliable measure of the unobserved true group average because only a small number of L1 individuals are sampled from each L2 group (O'Brien, 1990). For reflective aggregations of L1 constructs, the reliability of the aggregated L2 construct as a measure of the "true" group mean depends on at least two aspects: the proportion of variance that is located between groups—measured by the ICC—and the number of individuals in the group (Bliese, 2000; Snijders & Bosker, 1999).

In the multilevel literature, the ICC is used to determine the proportion of the total variance that is based upon differences between groups (Raudenbush & Bryk, 2002). The ICC is based on a one-way analysis of variance (ANOVA) with random effects, where the outcome on L1 is the dependent variable and the grouping variable is the independent variable. The ICC is defined as follows:

$$\text{ICC} = \frac{\tau^2}{\tau^2 + \sigma^2}, \quad (5)$$

where τ^2 is the variance between groups and σ^2 is the variance within groups. Thus, the ICC indicates the proportion of total variance that can be attributed to between-groups differences.

For reflective aggregations of L1 constructs, the reliability of the aggregated data $\bar{X}_{\cdot j}$ is estimated by applying the Spearman–Brown formula to the ICC, with n being the number of persons per group (Bliese, 2000; Snijders & Bosker, 1999):

$$\text{L2 Reliability } (\bar{X}_{\cdot j}) = \frac{n \cdot \text{ICC}}{1 + (n - 1) \cdot \text{ICC}}. \quad (6)$$

As can be seen, the reliability of $\bar{X}_{\cdot j}$ (Equation 6) for reflective aggregations of L1 constructs depends on two factors: the proportion of variance that is located between groups (ICC) and the group size (n). In most cases, the mean group size can be entered for n if not all groups are of the same size (see Searle, Casella, & McCulloch, 1992, on how to deal with pronounced differences in group size). For example, assuming that students in 50 classes rate their mathematics instruction, the ICC indicates the reliability of an individual student's rating—sometimes referred to as the single-rater reliability (Jayasinghe, Marsh, & Bond, 2003; Marsh & Ball, 1981). The reliability of the class-mean rating can be estimated by the Spearman–Brown formula, with n being the number of students per class. As is apparent from Equation 6, the reliability of the class-mean rating increases with the number of students (n). In other words, the more students in a class provide ratings, the more reliably the class-mean rating will reflect the true value of the construct being measured. It is worth noting that Equation 6, which determines the reliability of the observed group mean, says nothing about the reliability of the L1 measure. In general, measurement error at Level 1 results in lower reliability of the group means (Raudenbush et al., 1991). However, Equation 6 does not differentiate between L1 variance that is due to measurement error and L1 variance that is due to true differences between L1 individuals. For reflective aggregations of L1 constructs, the assessment of L2 unreliability due to sampling error is—as noted above—analogous in many ways to traditional approaches to reliability based on multiple, interchangeable indicators of each latent construct (i.e., with multiple persons as interchangeable indicators of each latent group mean).

Bias of the between-groups regression coefficient for reflective aggregations of L1 constructs. Let us now assume that a contextual model holds in the population and that the within-group and between-groups relationships are described by the within-group regression coefficient β_{within}

[1] A little more algebra is needed for the conversion of the variance components (see Kreft et al., 1995). Note that the models are no longer equivalent in either the fixed part or the random part when random slopes or nonlinear components are allowed.

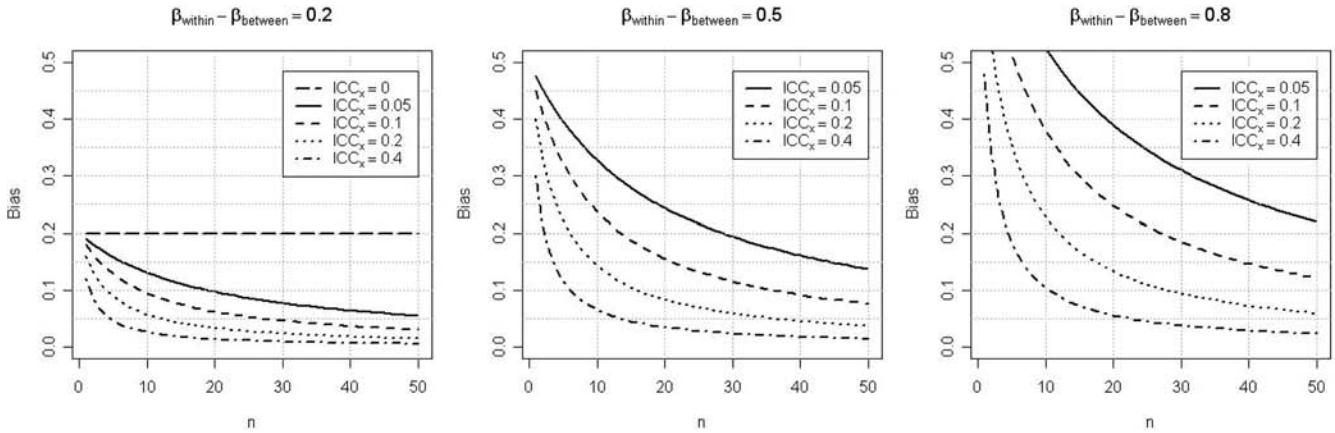

Figure 1. Relationship between the expected bias of the between-groups regression coefficient, the number of Level 1 units within each Level 2 unit (n), and the intraclass correlation coefficient (ICC) of the predictor for three different values of $\beta_{within} - \beta_{between}$.

and the between-groups regression coefficient $\beta_{between}$ (see Snijders & Bosker, 1999, p. 29). We want to estimate these coefficients by sampling a finite sample of L2 groups from the population. In the next stage, a finite sample of L1 individuals is obtained for each sampled L2 group. Bearing in mind the previous formula for the reliability of the group mean, results from the literature on regression analysis suggest that the regression coefficient for the L2 average will be biased. Applying standard results from theory on linear models (Seber, 1977), the expected bias of the within- and between-groups regression coefficients in a contextual analysis model can be determined depending on the reliability of the group mean. Because it is assumed that the individual-level variable is measured without error for reflective aggregations of L1 constructs, the within-group coefficient is an unbiased estimator. In contrast, it can be shown that the between-groups coefficient $\hat{\gamma}_{01}$ is a biased estimator of the between-groups coefficient $\beta_{between}$ (for the derivation, see the Appendix):

$$E(\hat{\gamma}_{01} - \beta_{between}) = (\beta_{within} - \beta_{between}) \cdot \frac{1}{n} \cdot \frac{(1 - ICC)}{ICC + (1 - ICC)/n}. \quad (7)$$

The relationship between the expected bias and the ICC as well as the group size is depicted in Figure 1 for $\beta_{within} - \beta_{between}$ values of .2, .5, and .8. In all three panels, the bias become smaller with larger group sizes n. In other words, when the group mean is more reliable due to a higher n, $\beta_{between}$ can be more precisely approximated by the manifest group mean predictor. The bias also decreases as the ICC increases. As shown in Equation 6, the reliability of the group mean is a direct function of the group size n and the ICC. Indeed, for sufficiently large cluster sizes, the difference between the manifest and latent approaches will be trivially small, even for reflective factors. The bias in the between-groups coefficient has direct consequences for the estimation of the "true" contextual effect for reflective aggregations of L1 constructs. In the group-mean centered model, the contextual effect is calculated as the difference, $\gamma_{01} - \gamma_{10}$, between the between-groups regression coefficient γ_{01} and the within-group coefficient γ_{10}. Assuming perfect measurement of the group mean, this would correspond to a "true" difference of $\beta_{between} - \beta_{within}$. It follows that the bias of the estimated contextual effect can be expressed by:

$$E(\hat{\gamma}_{01} - \hat{\gamma}_{10}) - (\beta_{between} - \beta_{within}) = (\beta_{within} - \beta_{between}) \cdot \frac{1}{n} \cdot \frac{(1 - ICC)}{ICC + (1 - ICC)/n}. \quad (8)$$

This relationship indicates that the contextual effect in the population will be underestimated by the contextual analysis model if $\beta_{within} < \beta_{between}$. However, if $\beta_{within} > \beta_{between}$, the contextual effect will be positively biased.[2] Thus, a low ICC together with small samples of L1 individuals from each L2 group will affect the bias of contextual effect considerably.

In the approach to contextual analysis for reflective aggregations of L1 constructs outlined above, the group-level

[2] This constellation is present in research on the so-called frog-pond effect (Davis, 1966). The most prominent example is the big-fish–little-pond effect: the observation that individual student achievement has a substantially positive effect on academic self-concept, whereas the effect of school- or class-average achievement is consistently negative (Marsh & Hau, 2003). As is apparent from Equation 8, the big-fish–little-pond effect is probably underestimated in absolute terms when the manifest covariate approach is applied.

predictor was formed by aggregating all of the observed measurements in each group (MMC approach). In the next section, we introduce an alternative approach to contextual analysis that infers the latent unobserved group mean from the observed data and that takes into account the unreliability of the group mean (L2 sampling error) when estimating the contextual effect (MLC approach). Historically, nearly all contextual effect studies have used an MMC approach, due in part to technical limitations in most statistical packages that made a latent covariate approach difficult to formulate. With the enhanced flexibility of MLM programs such as Mplus, however, it is now possible to introduce and evaluate an MLC approach (but see Croon & van Veldhoven, 2007, for an alternative implementation). Hence, the purpose of this article is to demonstrate the MLC approach, to evaluate statistical properties with simulated data, to illustrate its application with two actual (real-data) examples, and to critically evaluate its appropriateness from a theoretical and philosophical perspective.

A Multilevel Latent Covariate (MLC) Model

The concept of latent variables was originally introduced in the social and behavioral sciences to represent entities that may be regarded as existing but cannot be measured directly (e.g., Lord & Novick, 1968). For instance, in psychometric research, intelligence is considered a latent variable that cannot be directly observed but can only be inferred from the participants' observed behavior in tests. In these traditional psychometric applications, the values of a latent variable represent participants' scores on a trait or ability. Recently, several methodologists have proposed that the conceptualization of latent variables be broadened to include other circumstances in which unobserved individual values might profitably be included in the model (B. O. Muthén, 2002; Raykov, 2007; Skrondal & Rabe-Hesketh, 2004). In this latent variable framework, latent class membership and missing data are just two examples of latent variables.

The flexibility of this modeling framework is expressed in the definition provided by Skrondal and Rabe-Hesketh (2004, p. 1): "We simply define a latent variable as a random variable whose realizations are hidden from us." As a consequence of this generality, the latent variable framework is able to integrate MLM and structural equation modeling (SEM; see Raykov, 2007, for an application to longitudinal analysis) and is currently implemented, for example, in the Mplus (L. K. Muthén & Muthén, 2007) and GLLAMM (Skrondal & Rabe-Hesketh, 2004) software. In the present study, we used the MLC approach to consider the group effect as an unobserved latent variable that has to be inferred from the observed data. More specifically, 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, 2006). As is typical within SEM, the estimate of the group-level coefficient is then corrected for the unreliable measurement of the latent group mean by the observed group mean. In the present study, our MLC approach was implemented using a maximum likelihood procedure in Mplus, which provides estimates that are consistent and asymptotically efficient within a very flexible approach to estimating latent variable models.[3]

The basis for the latent covariate approach is that each variable is decomposed into unobserved components, which are considered latent variables (Asparouhov & Muthén, 2006; B. O. Muthén, 1989; Schmidt, 1969; Snijders & Bosker, 1999; see also Rabe-Hesketh, Skrondal, & Pickles, 2004). The dependent variable Y and the independent variable X can be decomposed as follows:

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

$$Y_{ij} = \mu_y + U_{yj} + R_{yij}, \tag{9}$$

where μ_x is the total mean of X, U_{xj} are group-specific deviations, and R_{xij} are individual deviations. The same decomposition holds for Y. Note that X_{ij} and Y_{ij} are observed variables, whereas U_{xj}, U_{yj}, R_{xij}, and R_{yij} are unobserved. We are interested in estimating the relationship between these unobserved variables at the individual and the group level:

$$R_{yij} = \beta_{\text{within}} R_{xij} + \varepsilon_{ij}$$

$$U_{yj} = \beta_{\text{between}} U_{xj} + \delta_j. \tag{10}$$

The two equations can be combined into one by substituting Equation 10 into Equation 9. The dependent variable Y is then predicted by the individual and group-specific deviations:

$$Y_{ij} = \mu_y + U_{yj} + R_{yij} = \mu_y + \beta_{\text{between}} U_{xj} + \beta_{\text{within}} R_{xij} + \delta_j + \varepsilon_{ij}. \tag{11}$$

It is now easy to see that Equation 11 is approximated by the group-mean centered model expressed by Equation 3,

[3] The version of Mplus (Version 4.2) used in the present investigation is based on an accelerated EM algorithm for analysis of maximum likelihood estimation of a two-level structural equation model with missing data (Asparouhov & Muthén, 2003). This model incorporates random coefficients and integrates the modeling frameworks of hierarchical linear models and two-level structural equation models. It provides robust estimates of the asymptotic covariance of the maximum likelihood estimates and the chi-square test. Note that the models considered here can be fitted with the approach described by Lee and Poon (1998) that handles only random-intercept models but that Mplus takes a more general approach with random slopes (see also supplemental materials available online for a more detailed description).

which is based on observed variables. The latent unobserved group deviation U_{xj} corresponds to the observed group means $\bar{X}_{.j}$ and the individual deviation R_{xij} to $(X_{ij} - \bar{X}_j)$.

It is worth noting that the latent covariate approach to contextual analysis can also be implemented in traditional multilevel programs such as HLM or MLwiN by using a stepwise procedure (Goldstein, 1987; Hox, 2002). In the first step, a within- and between-groups covariance matrix is estimated using a multivariate MLM. Hox (2002) demonstrated how a multivariate model can be estimated using MLM software designed to estimate univariate models (for a recent application of a multivariate MLM, see Bauer, Preacher, & Gil, 2006). In the second step, the within- and between-groups coefficients are estimated based on these covariance matrices.[4] Of course, the multivariate approach is much more limited than the implementation of the latent covariate approach in Mplus in that it can only be applied easily to random-intercept models.

Recently, Croon and van Veldhoven (2007) proposed a two-stage latent variable approach. The unobserved group mean for each L2 unit is calculated using weights obtained from applying basic ANOVA formulas. These adjusted group means form the basis for an ordinary least squares (OLS) regression analysis at the group level. Croon and van Veldhoven showed analytically and by means of simulation studies that an OLS regression analysis based on the observed group means results in biased estimates, whereas the results based on the adjusted group means are unbiased. However, in contrast to our full information maximum likelihood (FIML) MLC approach, their two-stage procedure is only a limited information approach. The model parameters of the two-stage procedure are thus likely to be less efficient than those of the FIML SEM approach (Wooldridge, 2002). As part of the present investigation, we conducted a simulation study to evaluate the differences between these two implementations of the MLC approach.

To summarize, researchers using aggregated individual data to assess the effects of group characteristics are often confronted with the problem that the observed group average score is a rather unreliable measure of the unobserved group mean. For reflective aggregations of L1 constructs, the unreliability of the group mean can lead to biased estimation of contextual effects, particularly when the number of observations per group is small and when the ICC of the corresponding individual observations is low. Our new MLC approach regards the unobserved group mean as a latent variable, consistent with the reflective aggregation of L1 constructs. In the following section, we present a simulation study comparing the statistical properties of the new MLC approach with the traditional MMC approach that assumes the observed group mean to be perfectly reliable.

Study 1: Simulation Study Comparing the Multilevel Latent Covariate (MLC) and Multilevel Manifest Covariate (MMC) Approaches

The simulation study was designed to generate data that resemble the data structures typically found for reflective aggregations of L1 constructs in psychological and educational research. The purpose of the simulation was to explore the statistical behavior of the MMC and MLC approaches under a variety of conditions approximating those encountered in actual practice.

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 3. Each generated data set was analyzed using the MMC and the MLC approach. The conditions manipulated were the number of L2 groups (50, 100, 200, and 500), the number of observations per L2 group (5, 10, 15, and 30), and the ICC of the predictor variable (.05, .10, .20, and .30). In the following, we explain why these particular levels were selected.

Number of L2 groups. The numbers of L2 groups were set to $K = 50$, 100, 200, or 500. A sample of 50 groups is common in educational and organizational research (e.g., Maas & Hox, 2005), although many MLM studies are conducted with fewer than 50 L2 groups. At the same time, a growing number of large-scale assessment studies, including educational assessments, such as the Early Childhood Longitudinal Study (ECLS) and the National Education Longitudinal Study (NELS), have sampled up to 1,000 schools. Hence, we included conditions with 200 and 500 groups. Covering a broad range of L2 groups enables us to study asymptotic behavior in the latent variable approach. More specifically, we anticipate that the variability of the estimator is likely to be sensitive to the number of L2 groups.

Number of observations per L2 group. We then manipulated the number of observations per L2 group to $n = 5$, 10, 15, and 30. A group size of 5 is normal in small group research, where contextual models are also applied (see Kenny, Mannetti, Pierro, Livi, & Kashy, 2002). Group sizes of 20 and 30 reflect the numbers that typically occur in

[4] We ran a small simulation study to compare the results of a contextual analysis model using the latent covariate approach in Mplus with the two-step approach based on the output of traditional MLM software. The results were the same.

educational research assessing class or school characteristics.[5]

ICC of predictor variable. The ICC of the predictor variable (i.e., the amount of variance located between groups) was set at ICC = .05, .10, .20, or .30. Intraclass correlations rarely show values greater than .30 in educational and organizational research (Bliese, 2000; James, 1982).[6]

For each of $4 \times 4 \times 4 = 64$ conditions, 1,500 simulated data sets were generated. The regression coefficients are specified as follows: 0 for the intercept, .2 for β_{within}, and .7 for $\beta_{between}$. Because the contextual effect $\beta_{context}$ equals $\beta_{between} - \beta_{within}$, these values imply a contextual effect of .5. The ICC for the dependent variable is .2. Because the amount of variance explained at Level 2 depends on the ICC of the predictor variable, the following R^2 values at Level 2 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 Level 1 ranged from .04 for ICC = .05 to .05 for ICC = .30.[7] For every cell, the 1,500 repetitions were simulated and analyzed with Mplus using FIML (L. K. Muthén & Muthén, 2007).

In our simulation study, we focused on three aspects of the estimator for the contextual effect in reflective aggregations of L1 constructs: the bias of the parameter estimate, the variability of the estimator, and the accuracy of the standard error. The relative bias indicates the accuracy of the estimator for the contextual effect. Let $\hat{\beta}_{context}$ be the estimator of the population parameter $\beta_{context}$, then the relative percentage bias is given by $100 \times [(\bar{\hat{\beta}}_{context} - \beta_{context})/\beta_{context}]$. To assess the variability of the estimator, we computed the root-mean-square error (RMSE) by taking the square root of the mean square difference of the estimate and the true parameter. The accuracy of the standard error of the contextual effect is analyzed by determining the observed coverage of the 95% confidence interval (CI). Coverage was given a value of 0 if the true value was included in the confidence interval and a value of 1 if the true value was outside the confidence interval.

To determine which of the study's conditions contributed to the relative bias, the RMSE, the coverage, and the empirical standard deviation of the contextual effect, we conducted ANOVAs using the relative bias, RMSE, coverage, and empirical standard deviation of the estimator as the dependent variables and each manipulated condition (method, number of L2 groups, number of L1 individuals within each L2 group, ICC of predictor variable) as a factor. The ANOVAs were conducted at the cell level, with each cell average being treated as one observation (so that the four-way interaction could not be separated from the error). To describe the practical significance of the conditions, we calculated the η^2 effect size for all main effects and for the two- and three-way interactions.

Results and Discussion

No problems were encountered in estimating the coefficients of the MLC and the MMC model; the estimation procedure converged in all 96,000 simulation data sets.

Relative percentage bias. Table 1A shows the relative bias in the parameter estimates for all four conditions. To determine the relative bias, the cell mean for each of the 1,500 repetitions was calculated, subtracted from the population values, and then divided by the population value. For the MLC approach, the relative percentage bias ranged in magnitude from −14.4 to 19.6 ($M = 0.6$, $SD = 4.7$). As expected from the mathematical derivation in Equation 8, the relative bias for the MMC approach was larger, with values ranging from −79.3 to −7.0 ($M = -36.9$, $SD = 20.8$). In contrast to the MLC approach, the MMC approach underestimated the contextual effect, especially for conditions with low ICCs. Consequently, the differences between the MLC approach and the MMC approach were particularly pronounced for low ICCs and small numbers of L1 individuals within each L2 group.

The source of the relative bias was further investigated by conducting a four-way factorial ANOVA with relative bias as the dependent variable (see Table 2). The largest effect was the main effect of method ($\eta^2 = .61$). Almost two-thirds of the variance in the relative bias across the conditions was explained by the difference between the MLC and the MMC approach. The number of L1 individuals within

[5] To evaluate the robustness of the MLC approach in the case of unbalanced group sizes, we conducted an additional, restricted simulation for a subset of the balanced group conditions ($n = 15$; $K = 50$, 100, and 200; ICC = .05, .10, .20, and .30). In the unbalanced condition, the group sizes were uniformly distributed between $n = 10$ and $n = 20$ (cf. a constant $n = 15$ in the balanced condition). Four-way ANOVAs (ICC; number of L2 units, unbalanced vs. balanced design; MLC vs. MMC approach) were conducted. The variance explained by balanced versus unbalanced design and its interactions with other independent variables was, as expected, less than 0.1% for bias, RMSE, and coverage.

[6] We also tried to include a condition with a very low ICC (.01) in our simulation study. However, Mplus showed serious convergence problems under this condition.

[7] Additional simulation studies, which will not be reported here, showed that the magnitude of the R^2 value at Level 1 only marginally affects the results of the simulation for the contextual effect. Typically, larger R^2 values lead to smaller standard errors of the parameter estimates, which are reflected in less variable estimates of the regression coefficients. However, the sample sizes at Level 1 of the conditions of our multilevel simulation study are large and the corresponding standard error of the L1 regression coefficient is therefore of small magnitude.

Table 1
Study 1: Fitting the Multilevel Manifest Covariate Model and the Multilevel Latent Covariate Model as a Function of the ICC of the Predictor Variable, the Number of Level 1 Units Within Each Level 2 Unit, and the Number of Level 2 Units

	No. (n) of Level 1 units within each Level 2 unit							
	$n = 5$		$n = 10$		$n = 15$		$n = 30$	
No. (K) of Level 2 units	Latent	Manifest	Latent	Manifest	Latent	Manifest	Latent	Manifest
A: Relative percentage bias of contextual effect								
$K = 50$								
ICC = .05	−14.4	−79.3	17.7	−64.2	19.6	−54.9	4.9	−38.6
ICC = .10	0.4	−64.9	6.5	−48.6	0.6	−39.7	4.8	−20.9
ICC = .20	−5.0	−46.5	2.8	−28.6	2.7	−20.6	−1.0	−13.3
ICC = .30	−6.0	−31.7	0.7	−19.0	−1.0	−14.8	0.2	−7.4
$K = 100$								
ICC = .05	−2.1	−78.4	10.3	−64.7	8.2	−54.9	2.3	−38.4
ICC = .10	−1.0	−64.5	−0.4	−48.3	−1.8	−39.5	1.1	−22.8
ICC = .20	−0.3	−44.3	2.3	−27.9	1.9	−20.0	0.5	−11.7
ICC = .30	−3.3	−32.2	0.1	−19.3	1.8	−12.4	−0.7	−8.1
$K = 200$								
ICC = .05	0.2	−78.9	−2.2	−67.0	3.7	−55.9	−0.3	−39.5
ICC = .10	−1.9	−64.3	0.6	−47.8	2.0	−36.9	1.6	−22.1
ICC = .20	−2.2	−45.3	−1.2	−29.4	−0.1	−21.3	−0.5	−12.3
ICC = .30	−4.1	−32.4	0.2	−18.9	0.2	−13.5	0.2	−7.2
$K = 500$								
ICC = .05	−6.2	−79.2	0.5	−65.3	0.9	−55.6	0.0	−39.1
ICC = .10	−3.3	−64.6	0.6	−47.2	0.5	−37.3	0.8	−22.5
ICC = .20	−0.6	−44.4	−0.4	−28.8	0.3	−20.9	0.5	−11.4
ICC = .30	−2.3	−31.7	0.0	−18.9	−0.4	−13.9	0.3	−7.0
B: Root-mean-square error of contextual effect								
$K = 50$								
ICC = .05	1.18	0.44	0.81	0.38	0.70	0.34	0.46	0.31
ICC = .10	0.64	0.36	0.43	0.29	0.31	0.26	0.24	0.21
ICC = .20	0.29	0.27	0.20	0.19	0.18	0.17	0.13	0.13
ICC = .30	0.19	0.21	0.13	0.14	0.11	0.12	0.09	0.09
$K = 100$								
ICC = .05	0.74	0.41	0.51	0.35	0.41	0.31	0.27	0.25
ICC = .10	0.42	0.34	0.25	0.27	0.22	0.23	0.15	0.16
ICC = .20	0.19	0.24	0.13	0.17	0.11	0.13	0.09	0.10
ICC = .30	0.13	0.18	0.09	0.12	0.08	0.09	0.06	0.07
$K = 200$								
ICC = .05	0.54	0.40	0.29	0.35	0.29	0.30	0.19	0.23
ICC = .10	0.22	0.33	0.17	0.25	0.14	0.20	0.12	0.14
ICC = .20	0.14	0.24	0.09	0.16	0.08	0.12	0.07	0.09
ICC = .30	0.09	0.17	0.06	0.11	0.06	0.08	0.04	0.05
$K = 500$								
ICC = .05	0.34	0.40	0.19	0.33	0.15	0.28	0.12	0.21
ICC = .10	0.14	0.33	0.11	0.24	0.09	0.19	0.07	0.12
ICC = .20	0.08	0.23	0.06	0.15	0.05	0.11	0.04	0.07
ICC = .30	0.06	0.16	0.04	0.10	0.03	0.07	0.03	0.04
C: Percentage coverage rate for contextual effect								
$K = 50$								
ICC = .05	93.5	40.1	92.6	61.6	94.1	72.7	93.4	83.9
ICC = .10	94.8	48.9	92.8	68.5	90.5	76.1	92.5	88.1
ICC = .20	94.1	62.4	92.6	78.7	90.4	83.3	91.6	88.7
ICC = .30	94.2	74.8	93.7	83.0	93.7	85.6	93.1	91.7

Table 1 (*continued*)

No. (*K*) of Level 2 units	No. (*n*) of Level 1 units within each Level 2 unit							
	n = 5		*n* = 10		*n* = 15		*n* = 30	
	Latent	Manifest	Latent	Manifest	Latent	Manifest	Latent	Manifest
	C: Percentage coverage rate for contextual effect (*continued*)							
K = 100								
ICC = .05	95.3	13.4	96.1	33.3	93.1	51.2	94.1	76.0
ICC = .10	94.7	23.0	92.6	45.3	92.0	59.9	94.1	83.8
ICC = .20	95.2	38.9	93.4	66.4	92.9	75.5	95.6	88.3
ICC = .30	94.1	54.4	92.3	71.6	93.8	83.8	94.9	89.5
K = 200								
ICC = .05	95.2	0.9	96.4	6.6	94.6	24.6	92.5	59.2
ICC = .10	95.5	1.2	94.6	18.6	94.0	37.5	93.4	73.0
ICC = .20	94.0	10.2	94.6	35.1	94.0	55.9	91.9	78.4
ICC = .30	93.7	24.9	97.0	52.6	94.6	67.3	95.8	84.7
K = 500								
ICC = .05	93.0	0.0	94.9	0.1	95.5	0.9	94.1	21.5
ICC = .10	94.7	0.0	93.5	0.5	94.1	5.0	94.4	43.0
ICC = .20	95.0	0.3	94.4	4.7	95.2	20.6	94.7	64.7
ICC = .30	94.1	3.1	94.3	15.9	94.4	32.4	95.1	69.5

Note. ICC = intraclass correlation of predictor variable; latent = multilevel latent covariate model; manifest = multilevel manifest covariate model.

each L2 group ($\eta^2 = .09$) and the ICC ($\eta^2 = .09$) had smaller but still substantial effects on the relative bias. When the ICC and the number of L1 individuals within each L2 group were high, the average magnitude of relative bias was low. The number of L2 groups had no effect on the magnitude of the relative bias. Furthermore, two two-way interactions were found to have an effect on the relative bias. First, the number of L1 individuals within each L2 group had a stronger effect on the magnitude of the relative bias for the MMC approach than for the MLC approach. Second, the effect of the ICC on the relative bias was more pronounced in the MMC approach than in the MLC approach.

Table 2
Study 1: Eta-Squared Values for Analysis of Variance Effects of the Simulation Conditions on Bias, RMSE, Coverage, and Variability

Variable	Bias	RMSE	Coverage	Variability
Main effects				
Method	0.61	0.00	0.53	0.14
K	0.00	0.14	0.11	0.18
n	0.09	0.15	0.08	0.05
ICC	0.09	0.44	0.03	0.24
2-way interactions				
Method × K	0.00	0.06	0.12	0.04
Method × *n*	0.06	0.00	0.08	0.05
Method × ICC	0.13	0.04	0.03	0.13
K × *n*	0.00	0.01	0.01	0.01
K × ICC	0.00	0.05	0.00	0.06
n × ICC	0.00	0.03	0.00	0.02
3-way interactions				
Method × *K* × *n*	0.00	0.01	0.01	0.01
Method × *K* × ICC	0.00	0.04	0.00	0.03
Method × *n* × ICC	0.01	0.02	0.00	0.03
K × *n* × ICC	0.00	0.00	0.01	0.00
Error	0.00	0.01	0.01	0.01

Note. RMSE = root-mean-square error; method = multilevel latent covariate model versus multilevel manifest covariate model; *K* = number of Level 2 units; *n* = number of Level 1 units within each Level 2 unit; ICC = intraclass correlation of predictor variable.

RMSE. Next we assessed the variability of the MLC and the MMC estimator (see Table 1B). The root-mean-square error (RMSE) was computed for every cell by taking the square root of the mean square difference of the estimate and the true parameter. It is interesting that in many conditions, the RMSE was higher for the MLC approach than for the MMC approach. The RMSE ranged in magnitude from 0.03 to 1.18 ($M = 0.22$, $SD = 0.22$) for the MLC approach, and from 0.04 to 0.44 ($M = 0.21$, $SD = 0.10$) for the MMC approach. The difference in the RMSE was particularly pronounced in the conditions with low ICCs and small numbers of L2 groups. For instance, in the condition with ICC = .05, $n = 5$, and $K = 50$, the RMSE was 1.18 for the MLC approach and 0.44 for the MMC approach. As the number of L2 units (*K*) increased, however, the MLC approach began to outperform the MMC approach; the RMSE for the MLC model approached zero, whereas that for the MLC model approached the value of the bias estimate given in Equation 8. ANOVAs with RMSE as the dependent variable revealed that the largest effect was found for the ICC. More than one-third of the variance in the RMSE was explained by the main effect of ICC. In addition, the sample sizes at Level 1 (*n*) and Level 2 (*K*) had a substantial impact

on the RMSE. Despite the large differences in certain conditions, there was no main effect for method. However, inspection of the two-way interactions revealed that the MLC approach performed better than the MMC approach in terms of RMSE when the number of L2 units was large ($\eta^2 = .06$) and the ICC was high ($\eta^2 = .04$). Moreover, a significant three-way interaction was found between method, number of L2 units, and ICC ($\eta^2 = .04$). This interaction indicates that the ICC had a higher influence on the difference between the two methods when the number of L2 units was low.

The RMSE assesses variability of the estimates as well as bias in relation to the known population value. To separate these two aspects, we calculated the empirical standard deviation across the 1,500 replications within each cell. As expected given the results for the RMSE, the estimates of the MLC approach were more variable than those of the MMC approach. The empirical standard deviation ranged in magnitude from 0.03 to 1.18 ($M = 0.22$, $SD = 0.22$) for the MLC approach and from 0.02 to 0.24 ($M = 0.09$, $SD = 0.05$) for the MMC approach. ANOVAs with the empirical standard deviation as the dependent variable showed that a substantial amount of variance was explained by the factor method ($\eta^2 = .14$). Similarly, the difference between the MLC and MMC approach was more pronounced when the ICC was low ($\eta^2 = .13$). In other respects, the results were nearly identical to those reported for the RMSE. From a statistical perspective, the smaller variability in the estimates of the MMC approach might be expected, because the group mean of the covariate is treated as observed. In contrast, the group mean is unobserved in the MLC approach, which naturally results in greater variability of the estimates.

Coverage. The accuracy of the standard errors for the MLC approach was evaluated in terms of the coverage rate, which was assessed using the 95% CIs (see Table 1C). Coverage rates for the MLC approach were better than for the MMC approach, reflecting the established finding that standard errors are underestimated if a predictor contains unreliability (e.g., Carroll, Ruppert, & Stefanski, 1995). The coverage rate ranged from 93.4 to 97.0 ($M = 94.0$, $SD = 1.3$) for the MLC approach, near the nominal coverage rate of 95%. In contrast, the coverage rate for the MMC approach ranged from 0 to 91.7 ($M = 47.7$, $SD = 31.1$). ANOVAs indicated that more than half of the variance in coverage was due to method (manifest vs. latent approaches). Evaluation of the two-way interactions (see Table 2) revealed that differences between the latent and manifest approaches were more pronounced when the number of L1 individuals within each L2 group was low ($\eta^2 = .08$) and the number of L2 groups was large ($\eta^2 = .12$). The negative effect of the number of L2 groups on the coverage rate for the MMC approach was due to the smaller CIs in conditions where the number of L2 groups was high. Because the bias was generally independent of the number L2 groups, the narrower CIs in conditions with high numbers of L2 groups for the biased MMC approach increased the probability that the CIs would not cover the true value. In addition, the coverage rates were affected by the main effects of the number of L1 individuals within each L2 group ($\eta^2 = .08$), the ICC of the predictor variable ($\eta^2 = .03$), and the number of L2 groups ($\eta^2 = .11$).

Summary

Overall, the results from the simulation study confirmed the findings from the mathematical derivation showing that the MMC approach is biased for reflective aggregations of L1 constructs, whereas the bias for the MLC approach is negligible. Furthermore, the MLC approach performed well in terms of the coverage rate, which was near the nominal rate of .95. However, results for the RMSE showed that in certain data constellations (e.g., small numbers of L2 groups, small ICCs, and small group sizes) the MMC approach was less variable than the MLC approach. The differences between the two approaches in terms of variability were most pronounced with small group sizes ($n < 10$), small ICCs (ICC $< .10$), and only a modest number of L2 groups ($K = 50, 100$). When the number of L2 units increased, the RMSE for the MLC approach converged to zero, in contrast to that for the MMC approach, which remained positive. From asymptotic theory, it follows that the FIML latent variable estimation approach yields consistent estimates of the contextual effect. However, the results of the simulation study suggest that large numbers of L2 groups (e.g., $K = 500$) are sometimes needed for these asymptotic properties to hold. Furthermore, when interpreting the findings of the simulation study, it is important to keep in mind that the MLC as opposed to the MMC approach directly corresponded to the model used to generate data in our simulation study. Hence, from a purely statistical point of view, the increased variability of the MLC approach in certain data constellations (e.g., small numbers of L2 groups, small ICCs, and small sample sizes) seems to be a limitation to the applicability of that approach. However, as will be argued in a later section, the choice between the approaches also depends strongly on the nature of the group construct under investigation and on the underlying aggregation process.

Study 2: A Two-Stage Implementation of a Multilevel Latent Covariate (MLC) Approach

As mentioned above, Croon and van Veldhoven (2007) recently proposed a two-stage latent variable approach. The unobserved group means for the covariate are calculated using weights obtained from applying basic ANOVA formulas. These adjusted group means form the basis for an

OLS regression analysis at the group level. Because their two-stage procedure is only a limited information approach, Croon and van Veldhoven suggested that it may be less efficient than a full information latent variable approach such as that implemented in Mplus:

> The stepwise estimation method proposed in this article is a limited information approach that does not directly maximize the complete likelihood function for the data under the model considered. Although the full information maximum likelihood approach would lead to the asymptotically most efficient estimates of the model parameters, the limited information approach is probably not much less efficient. A more systematic comparison of both approaches is needed here. [. . .] A disadvantage of the maximum likelihood approach is that it requires rather complex optimization procedures that are not yet incorporated into any readily available software package. (p. 55)

In response to this suggestion, we conducted an additional simulation study to compare the full information MLC approach (using the readily available Mplus package) to the two-stage approach proposed by Croon and van Veldhoven.[8] The model used to generate the data was the same as in Study 1. However, because we aimed to demonstrate—consistent with suggestions by Croon and van Veldhoven—that an FIML approach such as the MLC approach would outperform their two-stage approach, we conducted only a partial replication of the full simulation design. Hence, we see the present simulation as an empirical demonstration that the full information MLC should be preferred. The conditions manipulated were the number of L2 groups (50, 200), the number of observations per L2 group (10, 30), and the ICC of the predictor variable (.1, .2, .3). Again, the magnitude of the contextual effect was set to 0.5. In our analysis of this simulation study, we focus on the bias and the RMSE of the estimator for the contextual effect.

Results and Discussion

Bias. Table 3A shows the relative percentage bias in the parameter estimates for all four conditions. Overall, similar to the MLC approach, the stepwise approach was almost unbiased. The relative percentage bias for the two-stage approach ranged in magnitude from −0.9 to 12.6 (M = 1.98, SD = 3.7), whereas that for the MLC approach ranged from −0.8 to 6.1 (M = 1.30, SD = 2.1). As anticipated by Croon and van Veldhoven (2007), the two-stage approach showed a larger bias than our one-step FIML approach in the condition where the ICC is low and the sample sizes at Level 1 and Level 2 are small.

RMSE. As shown in Table 3B, the results for the RMSE were almost identical. A substantial difference between the two implementations of the MLC approach was present in one condition only (number of L2 units = 50, number of L1 units within each L2 unit = 10, ICC = .1), where the RMSE was .57 for the two-stage approach and .41 for the FIML approach.

Table 3
Study 2: Fitting Two Alternative Implementations of a Multilevel Latent Covariate Approach as a Function of the ICC of the Predictor Variable, the Number of Level 1 Units Within Each Level 2 Unit, and the Number of Level 2 Units

No. (K) of Level 2 units	No. (n) of Level 1 units within each Level 2 unit			
	n = 10		n = 30	
	Two-stage	FIML	Two-stage	FIML
A: Relative percentage bias of contextual effect				
K = 50				
ICC = .10	12.6	6.1	−0.9	−0.8
ICC = .20	4.8	4.9	0.5	0.6
ICC = .30	2.5	1.7	0.0	0.1
K = 200				
ICC = .10	1.9	1.0	1.3	1.2
ICC = .20	0.7	0.4	−0.2	−0.2
ICC = .30	0.2	0.1	0.5	0.5
B: Root-mean-square error of contextual effect				
K = 50				
ICC = .10	0.57	0.41	0.24	0.24
ICC = .20	0.19	0.19	0.14	0.14
ICC = .30	0.13	0.12	0.09	0.09
K = 200				
ICC = .10	0.17	0.17	0.11	0.11
ICC = .20	0.09	0.09	0.07	0.07
ICC = .30	0.07	0.07	0.04	0.04

Note. Two-stage = two-stage multilevel latent covariate approach; FIML = full information maximum likelihood multilevel latent covariate approach; ICC = intraclass correlation of predictor variable.

Summary

To summarize, comparison of two alternative implementations of the MLC approach showed that the two approaches yielded very similar results, except under the condition with a small sample size at both levels and a low ICC. In this condition, the MLC approach outperformed the two-stage approach in terms of both bias and RMSE. This is not surprising, given that the MLC approach is based on FIML, which uses all information inherent in the raw data. In contrast, the two-stage approach is a limited information approach that relies on a stepwise procedure. Hence, it can be concluded that the problems of the MLC approach with small sample sizes at L2, as outlined in the previous simulation study, are even more serious for the two-stage ap-

[8] The present investigation does not address the case in which a dependent variable is directly measured at Level 2 (see Croon & van Veldhoven, 2007). However, it is possible to include a directly observed L2 dependent variable in the Mplus program. In Example 9.4 of the Mplus manual (L. K. Muthén & Muthén, 2007, p. 236), the variable z indicates a mediator that is directly measured at L2. On the basis of this variable, the example can be modified to run the model used in Croon and van Veldhoven (2007).

proach. In conclusion, given appropriate statistical software, there appears to be no reason to choose the two-stage approach proposed by Croon and van Veldhoven (2007) over the one-step FIML MLC approach presented here, although the two approaches will yield nearly identical results for many situations.

Study 3: The Role of the Sampling Ratio in the Multilevel Latent Covariate (MLC) Approach

Sampling ratio is a critical issue that has not received sufficient attention in the development of the MLC approach (but see Goldstein, 2003), which implicitly assumes that L1 cases are sampled from an infinite sample of L1 cases within each L2 group. This is a reasonable assumption for a reflective aggregation process in which a generic group-level construct is assumed to be measured by the corresponding constructs at the individual level. Croon and van Veldhoven (2007, p. 55) thus regarded the group mean of the L2 construct as a latent variable and treated the corresponding individual scores at Level 1 as "reflective indicators for that variable." However, this reflective measurement assumption is not appropriate for formative L2 constructs. For example, in the earlier discussion of gender composition, the percentage of girls in a class can be measured with essentially no measurement error at Level 1 or sampling error at Level 2—consistent with the assumption underlying the MMC approach. In this example, the MLC approach would not be appropriate. In general, the MMC approach seems better suited than the MLC approach whenever the sampling ratio approaches 1.0 for a formative L2 construct.

What happens in a formative aggregation process in which the sampling ratio does not approach 1.0? The true value of the L2 aggregated variable is unknown because the entire cluster has not been sampled. For example, let us assume that researchers want to assess school-average SES by sampling 5 students from a school of 1,000 students, a sampling ratio of .005. Using the MMC approach, the average SES of this sample would not provide an error-free estimate of the school-average SES. In scenarios with such a low sampling ratio, the MLC approach might be used to correct the estimator of the contextual effect for L2 unreliability due to sampling error.

To address these concerns, we conducted a simulation study to further investigate the suitability of the MLC approach for a formative aggregation process in which the true L2 group average is not known. In contrast to the previous simulations, we assumed that the number of L1 units within each L2 group was some finite number (e.g., 100). Although the number of L1 units was fixed, the L2 units were randomly sampled from a population. Hence, we utilized a two-step procedure to generate populations with finite L1 sample sizes and a fixed number of randomly drawn L2 units. More specifically, in the first step, a certain number of clusters were drawn (e.g., $K = 500$ L2 units with $n = 100$ L1 units within each L2 unit) to establish a population model with finite sample size within each L2 unit. In the next step, a sample was drawn from this finite population according to a particular sampling ratio (e.g., 20%). This two-step procedure was replicated 1,000 times for each condition, and we analyzed the resulting data sets using both the MLC and the MMC approaches. The following conditions were manipulated: the number of L2 groups ($K = 100$, 500), the number of L1 observations per L2 group in the finite population ($n = 25$, 100, 500), the ICC of the predictor variable (ICC = .10, .30), and the sampling ratio (SR; the percentage of L1 observations considered within each L2 group: SR = 20%, 50%, 80%, 100%). Thus, for example, L2 group averages were based on 5 cases per class when the number of students within the class (n) was 25 and the sampling ratio was 20%.

Results and Discussion

Bias. Table 4A shows the relative percentage bias in the parameter estimates for all four conditions. Overall, there was a tendency for the MLC approach to be positively biased—to overestimate the true contextual effect ($M = 9.8$, $SD = 10.2$, range = 0.2 to 39.1)—and for the MMC approach to be negatively biased ($M = -7.8$, $SD = 12.0$, range = −52.0 to 0.2). The difference between the MLC approach and the MMC approach was particularly marked when the finite sample size (i.e., number of L1 units within each L2 unit) and the ICC were small. This finding was confirmed by two significant interactions between method and L1 sample size ($\eta^2 = .23$) and method and ICC ($\eta^2 = .09$) in a five-way ANOVA.

In the worst combinations (ICC = .10 and SR = .2 in Table 4A), the bias was extremely positive (33.0% and 29.8%) in the MLC approach and extremely negative (−51.4% and −52.0%) in the MMC approach. Whenever the MLC approach led to a substantial positive bias, the MMC approach led to a substantial negative bias. However, the pattern of differences was not symmetrical. In particular, the size of the negative bias in the MMC approach declined sharply as the sampling ratio increased (and disappeared for SR = 1.0), whereas the positive bias for the MLC approach did not vary systematically with SR.

It is not surprising that the MMC approach is unbiased when the sampling ratio is 1.0, given that all cases are sampled from the finite population. However, the MLC approach is most positively biased under these conditions, because it assumes that the samples were drawn from an infinite population. When the sampling ratio is low (.2), the negative bias of the MMC approach is larger than the positive bias of the MLC approach. In these conditions, the MMC approach is negatively biased because it does not

correct for the unreliability of the aggregated L2 variable, whereas the MLC approach is positively biased because it overcorrects the contextual effect based on biased estimates of unreliability of the aggregated L2 variable. With increasing magnitude of the number of L1 units in the finite population, the bias of both the latent and the manifest approach is reduced. However, except for the lowest sampling ratio, the absolute value of bias based on the MMC approach was systematically smaller than that of the MLC approach.

To further study the bias of both approaches, a condition with a very low sampling ratio (SR = .05) was included for $n = 500$. As expected, the MLC approach was unbiased for a low ICC ($K = 100$: 2.4%; $K = 500$: 2.2%) as well as for a high ICC ($K = 100$: 0.0%; $K = 500$: 0.6%). In contrast, the MMC approach was seriously biased for such a low sampling fraction, with the bias being more pronounced for a low ICC ($K = 100$: −25.7%; $K = 500$: −25.7%) than for a high ICC ($K = 100$: −8.4%; $K = 500$: −8.3%).

RMSE. As shown in Table 4B, the RMSE for both methods was of a similar magnitude. The RMSE for the MLC approach ranged in magnitude from 0.03 to 0.46 ($M = 0.11$, $SD = 0.09$). For the MMC approach, the RMSE ranged from 0.03 to 0.29 ($M = 0.10$, $SD = 0.06$). In general, the RMSE was high when the number of L1 units within each L2 unit, the number of L2 units, and the ICC were low. The MLC approach showed a higher RMSE than the MMC approach in some conditions (e.g., when the number of L2 units was 100, the ICC was .10, and the number of L1 units within each L2 unit was 25). Furthermore, in line with the results from the previous simulation studies, the RMSE for the MLC approach was affected by the number of L2 units.

In the additional condition that we ran with a very low sampling ratio (SR = .05), when n was substantial ($n = 500$), only slight differences were found between the two approaches. As expected, the RMSE for the MLC approach was larger for a modest number of L2 units (ICC = .1: 0.18; ICC = .3: 0.08) than for a larger number of L2 units (ICC = .10: 0.08; ICC = .30: 0.04). The RMSE for the MMC approach was almost identical for a modest number of L2 units (ICC = .10: 0.18; ICC = .30: 0.08) but slightly higher for a large number of groups (ICC = .10: 0.14; ICC = .30: 0.06).

Coverage. As in Study 1, the accuracy of the standard errors was evaluated in terms of the coverage rate, which was assessed using the 95% CIs. As shown in Table 4C, the coverage rates were generally better for the MLC approach than for the MMC approach, ranging from 32.8 to 96.0 ($M = 87.8$, $SD = 13.1$) for the MLC approach and from 0.2 to 95.6 ($M = 83.5$, $SD = 21.6$) for the MMC approach.

In addition, we looked at the coverage for a condition with a very low sampling ratio (SR = .05) and $n = 500$. As expected, the MLC approach showed coverage rates near the nominal coverage rate of 95% for a modest number of L2 units (ICC = .10: 94.3%; ICC = .30: 94.5%) and for a large number of L2 units (ICC = .10: 94.0%; ICC = .30: 94.3%). In contrast, the CIs of the MMC approach were not accurate. The probability that the CIs do not cover the true value was higher for the conditions with a high number of L2 units (ICC = .10: 39.0%; ICC = .30: 73.5%) than for the conditions with a modest number of L2 units (ICC = .10: 81.0%; ICC = .30: 88.6%).

Summary

Overall, this simulation study showed that, given a formative aggregation process, the results for the manifest and latent approach depend on the size of the finite population that is assumed to generate the observed data. When the sample size and sampling ratio are both small, both approaches perform poorly—albeit in counterbalancing directions. Particularly when the sample size is low ($n = 25$) and/or the sampling ratio is high, the MLC approach suffers from the fact that it assumes an infinite population for each L2 unit, whereas estimates based on the MMC approach show little or no bias.[9] However, when the sampling ratio is low and the sample size is high, the MLC approach appears to behave more favorably than the MMC approach. For instance, in the conditions with a large number of L2 units ($K = 500$), a low sampling ratio (20%), and a large number of L1 units ($n = 100$), the MLC approach outperformed the MMC approach in terms of bias as well as RMSE. When the number of L1 units is further increased (e.g., $n = 500$) and the sampling ratio is low (e.g., SR = .05), the finite population sampling model is almost equivalent to the infinite population sampling model. Hence, the results would be nearly identical to the findings reported in the simulation study above, in which an infinite population was assumed (see Tables 1 and 2).

Studies 4 and 5: Two Applications of Manifest and Latent Variable Approaches With Actual Data

We next present two examples illustrating the difference between the latent and the manifest approaches to contex-

[9] Additional, unreported simulations showed that the bias for the MLC approach becomes even more extreme when the sampling ratio approaches 1.0 and the number of cases is very small. In the most extreme situation, with $n = 10$, ICC = .10, $K = 100$, and SR = 1.0, the bias for the MLC approach was very positive (91.6%), whereas the MMC approach was almost unbiased (−0.8%). The RMSE was also considerably higher for the MLC approach (0.53) than for the MMC approach (0.12). When the number of L2 units was increased to $K = 500$ and the ICC was set to 0.30, the MLC approach performed better (bias: 23.3%; RMSE: 0.13), but it was still outperformed by the manifest approach (bias: −0.3%; RMSE: 0.04).

Table 4
Study 3: Fitting the Multilevel Manifest Covariate Model and the Multilevel Latent Covariate Model as a Function of the ICC of the Predictor Variable, the Number of Level 1 Units Within Each Level 2 Unit, the Number of Level 2 Units, and the Sampling Ratio

	No. (n) of Level 1 units within each Level 2 unit					
	$n = 25$		$n = 100$		$n = 500$	
No. (K) of Level 2 units	Latent	Manifest	Latent	Manifest	Latent	Manifest
A: Relative percentage bias of contextual effect						
$K = 100$						
ICC = .10						
SR = .2	33.0	−51.4	11.8	−24.2	0.8	−8.3
SR = .5	39.1	−22.3	9.9	−7.7	2.0	−2.5
SR = .8	37.3	−6.9	8.9	−2.9	1.5	−1.3
SR = 1	37.4	−0.6	8.0	−1.7	1.5	−1.0
ICC = .30						
SR = .2	9.3	−25.8	2.4	−8.7	0.2	−2.3
SR = .5	9.8	−8.4	2.5	−2.3	1.0	−0.1
SR = .8	9.2	−2.6	2.9	−0.2	0.8	−0.2
SR = 1	9.9	0.2	2.6	0.1	0.5	−0.2
$K = 500$						
ICC = .10						
SR = .2	29.8	−52.0	9.3	−25.4	1.7	−8.2
SR = .5	35.5	−23.2	8.9	−9.2	1.3	−4.2
SR = .8	36.0	−7.3	9.3	−3.8	1.6	−2.6
SR = 1	36.5	−1.1	9.4	−1.3	1.4	−2.1
ICC = .30						
SR = .2	9.3	−25.4	2.5	−8.5	0.6	−2.2
SR = .5	9.4	−8.8	2.4	−2.6	0.6	−0.8
SR = .8	9.5	−2.3	2.5	−0.7	0.2	−0.9
SR = 1	9.4	−0.3	2.5	−0.3	0.4	−0.8
B: Root-mean-square error of contextual effect						
$K = 100$						
ICC = .10						
SR = .2	0.46	0.29	0.21	0.18	0.15	0.14
SR = .5	0.31	0.17	0.16	0.14	0.14	0.14
SR = .8	0.27	0.13	0.15	0.13	0.14	0.14
SR = 1	0.26	0.13	0.15	0.14	0.13	0.14
ICC = .30						
SR = .2	0.16	0.16	0.09	0.09	0.07	0.07
SR = .5	0.11	0.09	0.07	0.07	0.07	0.07
SR = .8	0.10	0.08	0.07	0.07	0.07	0.07
SR = 1	0.09	0.08	0.07	0.07	0.07	0.07
$K = 500$						
ICC = .10						
SR = .2	0.21	0.27	0.10	0.14	0.07	0.08
SR = .5	0.20	0.13	0.08	0.08	0.06	0.08
SR = .8	0.20	0.08	0.08	0.08	0.06	0.08
SR = 1	0.20	0.07	0.08	0.07	0.06	0.08
ICC = .30						
SR = .2	0.08	0.13	0.04	0.06	0.03	0.04
SR = .5	0.06	0.06	0.03	0.04	0.03	0.04
SR = .8	0.06	0.04	0.03	0.03	0.03	0.03
SR = 1	0.06	0.04	0.03	0.03	0.03	0.04

(*Table continues*)

Table 4 (*continued*)

No. (*K*) of Level 2 units	No. (*n*) of Level 1 units within each Level 2 unit					
	n = 25		*n* = 100		*n* = 500	
	Latent	Manifest	Latent	Manifest	Latent	Manifest
C: Percentage coverage rate for contextual effect						
K = 100						
ICC = .10						
SR = .2	92.8	36.0	92.1	80.2	92.9	90.3
SR = .5	84.7	82.3	92.8	91.9	94.1	93.7
SR = .8	81.3	91.6	93.5	93.8	93.6	92.8
SR = 1	78.6	92.7	92.7	93.1	94.4	93.3
ICC = .30						
SR = .2	93.7	61.1	93.5	87.3	93.5	92.8
SR = .5	91.2	87.6	93.9	92.6	92.8	92.3
SR = .8	89.6	91.1	92.8	93.0	93.3	92.8
SR = 1	89.1	92.5	94.3	94.0	93.4	92.9
K = 500						
ICC = .10						
SR = .2	86.5	0.2	90.3	38.7	94.7	88.5
SR = .5	54.5	42.9	89.0	86.1	94.0	92.0
SR = .8	39.1	88.7	88.3	92.3	95.4	93.2
SR = 1	32.8	92.5	88.5	92.7	95.1	93.8
ICC = .30						
SR = .2	88.5	11.5	93.2	73.9	94.9	93.8
SR = .5	80.3	74.3	93.1	92.8	94.8	94.9
SR = .8	74.8	92.3	92.6	94.0	96.0	95.6
SR = 1	74.1	93.1	93.1	94.6	94.1	93.7

Note. Latent = multilevel latent covariate model; manifest = multilevel manifest covariate model; ICC = intraclass correlation of predictor variable; SR = sampling ratio.

tual analysis. The first example utilizes students' ratings of their teachers' behavior, a reflective aggregation of L1 constructs in which the referent is an L2 construct. The central question is whether the individual and shared perceptions of a specific teaching behavior are related to students' achievement outcomes. Because the contextual variable is based on different students' perceptions of a specific teacher behavior—an L2 referent—it seems reasonable to assume that students within each class are interchangeable in relation to this L2 reflective construct.

The second example is a classic illustration of contextual analysis, namely the question of whether the school composition in terms of SES affects students' reading literacy (Raudenbush & Bryk, 2002). Again, L1 scores (individual student SES) are used to assess the L2 construct (school-average SES). In this case, however, the aggregation of L1 constructs is formative; the referent is the (L1) individual student and the aggregated L2 construct is an index of L1 measures that may be very heterogeneous. Because SES can be measured with a reasonably high level of reliability at L1 and the number of students within each L2 group is substantial, the reliability of the L2 aggregate (school-average SES) may be sufficient. Furthermore, within-school variability in SES is a potentially interesting characteristic of the school (i.e., heterogeneity of SES).

We selected these two examples to illustrate that, from a theoretical perspective, the appropriateness and the reasons for applying the MLC approach may depend on the nature of the specific construct under study.

Study 4: Teacher Behavior—Contextual Analysis of a Reflective L2 Construct

In educational research, it is widely posited that individual students' learning outcomes are affected by teacher behaviors. Empirical studies draw on different data sources to elucidate aspects of the learning environment. One simple and efficient research strategy is to ask students to rate several specific teacher behaviors. In this approach, each student is regarded as an independent observer of the teacher, the referent is the teacher, and responses are aggregated across all students within a class to provide an indicator of teacher behavior. At the individual level, student ratings represent the individual student's perception of the teacher behavior. Scores aggregated to the classroom level reflect shared perceptions of teacher behavior in which idiosyncrasies associated with the responses of individual students tend to cancel each other out (Lüdtke, Trautwein, Kunter, & Baumert, 2006; Miller & Murdock, 2007; Papaioannou et al., 2004). Several studies—many using MLM—

have provided empirical support for the predictive validity of these individual and shared perceptions of features of the learning environment with respect to student outcomes (Kunter, Baumert, & Köller, 2007; Lüdtke et al., 2005; Urdan, Midgley, & Anderman, 1998).

Background to the Application

In this first example, we examine students' perceptions of a specific teaching behavior. Students were asked to rate how easily distracted their mathematics teacher was (teacher distractibility) on three items (sample item: "Our mathematics teacher is easily distracted if something attracts his/her attention"). The scale was developed on the basis of Kounin (1970) and covers 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). Consistent with the rationale for the reflective aggregation of L1 constructs to form an L2 construct that is the primary focus of study, 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 reflects a specific teacher characteristic, namely, distractible teaching style (Cronbach, 1976; Miller & Murdock, 2007).

We used the German sample of lower secondary students who participated in the Third International Mathematics and Science Study (TIMSS; 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 for the student ratings was .08, 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. On the basis of Equation 8, the MMC approach might be expected to underestimate the strength of the relationship between perceived distractible teaching style and mathematics achievement at the class level.

Both the MLC and the MMC approaches were specified in Mplus 4.2 (for Mplus codes, see the supplemental material available online). Students' perceptions of their teachers' distractibility and mathematics achievement scores were standardized across the entire sample (z-score with $M = 0$, $SD = 1$) at the individual level. For the MMC approach, the standardized distractibility was aggregated but not restandardized at the class level (thus, class-level effects are measured in terms of student-level standard deviations).

Results and Discussion

The parameter estimates for both approaches were nearly identical except for the L2 (between-groups) regression coefficient $\hat{\gamma}_{01}$ and the L2 residual variance (see Table 5). As expected on the basis of both the mathematical derivation of the bias and the simulation study, the regression coefficient for teacher distractibility at the class level was larger in the MLC approach than in the MMC approach, but it also exhibited a larger standard error. Classes with teachers who were perceived as showing a high level of distractibility in lessons had lower levels of achievement than did classes with teachers who were perceived to be less distractible. At the student level, there was no effect of the individual students' perception of their teachers' teaching style on individual achievement. Given that variables were standardized at the individual level, the large regression coefficient obtained at the class level seems unusual. The reason for these large values at the group level is that the standard deviation of the aggregated group-level predictor is often smaller than 1. When the regression coefficient for the MLC approach is interpreted in relation to the class-level standard deviation of teacher distractibility, it decreases to .45. However, there are currently no agreed-upon standards

Table 5
Study 4: Empirical Analysis Results of the Effects of Students' Perception of Their Teachers' Distractibility on Mathematics Achievement

Variable	Latent			Manifest		
	Coefficient	*SE*	Variance component	Coefficient	*SE*	Variance component
Fixed effect						
$\hat{\gamma}_{00}$ intercept	−0.06	0.07		−0.06	0.07	
$\hat{\gamma}_{01}$ distractible teaching (average)	−1.23	0.31		−0.83	0.19	
$\hat{\gamma}_{10}$ distractible teaching (student)	−0.03	0.02		−0.03	0.02	
Random effect						
$Var(u_{0j})$			0.42			0.46
$Var(r_{ij})$			0.48			0.48

Note. N for Level 1 = 2,133; N for Level 2 = 108. Average cluster size = 19.75. All parameter estimates except the intercept and $\hat{\gamma}_{10}$ are statistically significantly different from zero ($p < .001$).

for how to calculate standardized regression coefficients in MLM. Standardization strategies in MLM remain a topic of research (e.g., Raudenbush & Bryk, 2002).

Students' ratings of their teachers' behavior seem to be a good example of an L2 reflective construct where the rationale of the MLC approach is appropriate. In this example, the main purpose of the L2 measurements is to assess an L2 group-level construct—the behavior of a particular teacher as perceived by his or her students. In his seminal paper on multilevel issues in educational research, Cronbach (1976) was very clear about the role of students' perceptions in assessing aspects of the learning environment. In a discussion of the Learning Environment Inventory (LEI), he argued that

> The purpose of the LEI is to identify differences among classrooms. For it, then, studies of scale homogeneity or scale intercorrelation should be carried out with the classroom group as unit of analysis. Studying individuals as perceivers within the classrooms could be interesting, but is a problem quite separate from the measurement of environments. (p. 918)

From this point of view, it is reasonable to correct for factors that impinge the measurement of that class-level construct. In the MLC approach, the restriction to small samples of students within classes and disagreement among students are taken into account when estimating the effect of aggregated student ratings on achievement.

Study 5: School-Average SES—A Contextual Effect Analysis of a Formative L2 Construct

Educational researchers believe that a student's performance in school is affected by the characteristics of his or her fellow students (Marsh, Kong, & Hau, 2000; Willms, 1985). For example, several researchers have posited that aggregated school SES or mean ability affects individual student outcomes (e.g., student achievement or academic self-concept), even after controlling for the individual effects of these L1 constructs—a contextual effect. Raudenbush and Bryk (2002, p. 139) define such a contextual effect to exist "when the aggregate of a person-level characteristic, $\bar{X}_{\cdot j}$, is related to the outcome, Y_{ij}, even after controlling for the effect of the individual characteristic, X_{ij}."

Background to the Application

In the present example, individual students' SES is used to assess the effect of school-average SES on reading literacy after controlling for individual SES, drawing on data from the German sample (Baumert et al., 2002) of the Programme for International Student Achievement (PISA) 2000 study (Organisation for Economic Co-operation and Development, 2001). The analyses are based on 4,460 students in 189 schools, giving a mean of 23.6 students per school. Note that the PISA study sampled 15-year-old students from schools rather than classes. Hence, in contrast to Study 4, where the L2 groups were classes, the present example focused on the school at Level 2 (meaning that the sampling ratio is much lower). The ICC for SES was .22, indicating that a substantial amount of the variance in students' SES was located at the school level.

Results and Discussion

The results for both manifest and latent models are reported in Table 6. SES and reading scores were standardized (z-score with $M = 0$, $SD = 1$) at the individual level. For the MMC approach, the standardized SES was aggregated but not restandardized at the school level. As expected, the differences between the two approaches in the parameters based on student-level data were negligible. The effect of students' SES $\hat{\gamma}_{10}$ on reading achievement, the L1 residual, and the intercept $\hat{\gamma}_{00}$ were almost the same. In contrast, estimates at the school level differed across the two approaches. As expected on the basis of Equation 8, the effect of school-average SES was higher in the MLC approach, which corrects for unreliability of the school-average SES scores. Because we group-centered the L1 predictor vari-

Table 6
Study 5: Empirical Analysis Results of the Effects of Socioeconomic Status (SES) on Reading Achievement

	Latent			Manifest		
Variable	Coefficient	*SE*	Variance component	Coefficient	*SE*	Variance component
Fixed effect						
$\hat{\gamma}_{00}$ intercept	−0.01	0.03		−0.02	0.03	
$\hat{\gamma}_{01}$ SES (average)	1.52	0.06		1.29	0.06	
$\hat{\gamma}_{10}$ SES (student)	0.10	0.01		0.10	0.01	
Random effect						
$Var(u_{0j})$			0.08			0.15
$Var(r_{ij})$			0.43			0.43

Note. N for Level 1 = 4,460; N for Level 2 = 189. Average cluster size = 23.6. All parameter estimates except the intercept are statistically significantly different from zero ($p < .001$).

able, the compositional effect was determined by subtracting the within-school regression coefficient from the between-schools regression coefficient: $\hat{\gamma}_{01} - \hat{\gamma}_{10}$ (e.g., Raudenbush & Bryk, 2002). We obtained a compositional effect of 1.42 for the MLC approach and 1.19 for the MMC approach. The contextual effect can be interpreted as the difference expected in reading literacy between two students with the same individual SES who attend schools differing by one unit in mean SES. One unit in mean SES corresponds to one standard deviation at the individual level metric because the aggregated individual SES score were not restandardized. Similarly, when the regression coefficients at the school level are interpreted in relation to the school-level standard deviation of SES, they decrease in absolute size. For instance, the regression coefficient for the MLC decreases to 0.71 when interpreted in relation to the school-level standard deviation. Again, this demonstrates that the magnitude of the parameter estimates is very sensitive to the different standardization strategies.

The central question now is whether we can justifiably correct for unreliability based on the MLC approach in the present example. In contrast to Study 4, where students' responses serve as reflective indicators of the L2 construct (each student provides a single fallible estimate of the teacher's distractibility), in Study 5 student responses reflect the L1 construct SES. Variability in individual student levels of SES in a given school clearly reflects systematic true score variation in a well-defined L1 construct in which the referent is the individual student (consistent with the formative approach to aggregation). Hence, it does not seem appropriate to assume a reflective aggregation process for school-average SES. However, another reason for choosing a MLC approach might be a low sampling ratio. On average, 23.6 students were sampled from each school in the present example. Assuming an average school size of 500, only 5% of the pupils from each school were sampled. Given such a low sampling ratio in connection with a potentially very large number of L1 units within each L2 unit, application of the MLC approach may be justified. On the basis of these assumptions, one can even go further and use the results of the simulation study to infer the direction of the bias. For instance, given these assumptions, in the present example a researcher can infer that the true value for the contextual effect of average SES is expected to be closer to the estimate of the MLC approach than to the estimate of the MMC approach (see Table 4A). However, as this example shows, the choice of the analysis model (MLC or MMC approach) is very sensitive to the assumptions made about the underlying population sampling model. For instance, if classes (with *n*s of approximately 25) and not schools were chosen as L2 units and the sampling ratio approached 1.0, we would have more confidence in the MMC approach. For a formative process in which the sampling ratio is small or moderate, our simulation results might provide preliminary evidence about the relative size (and direction) of biases under the manifest and latent approaches. Resolving this problem is clearly beyond the scope of this study. It is, however, important that applied researchers are aware of the problem, which does not seem to have been clearly demonstrated in previous research.

Discussion

Multilevel analyses are frequently used to estimate the effects of group-level (L2) constructs in the social sciences. When using aggregated individual data to assess an L2 construct within the MMC approach, however, the observed group mean might not be a reliable measure of the unobserved, latent group characteristic. We compared two approaches to the analysis of contextual models: a new MLC approach that corrects for the unreliable assessment of the latent group mean when estimating MLMs and the traditional MMC approach, which relies on manifest group means that are assumed to be perfectly reliable.

Statistical Considerations

By means of mathematical derivation, we showed that the MMC approach results in biased estimates of contextual effects for reflective aggregations of L1 constructs, particularly when the ICC and L1 sample sizes within groups are small. This result was confirmed by a simulation study, which also showed that the MLC approach is generally unbiased. Although the contextual effects estimated within the MLC approach were larger, they were also substantially more variable in certain data constellations (e.g., small number of L2 groups, small ICCs, and small *n*) than those obtained using the traditional MMC approach. Indeed, due to this trade-off, the results of Study 4 suggested that the likelihood of obtaining statistically significant results was similar for both approaches. Although this article clearly does not provide sufficient evidence to suggest that this result will generalize more broadly, it is a relevant consideration for further research. More generally, because the contextual effect estimates are so variable within the MLC approach, results based on a given sample may deviate substantially from the true population parameter—as can be demonstrated by a simple inspection of the standard error.

What are the consequences of these mixed findings for the statistical properties of the estimator of the contextual effect in cases of reflective aggregation? The MMC approach is used almost exclusively in research practice. Our results suggest that for the single predictor case at least the sizes of contextual effects published are likely to be conservative. Although the new MLC approach is unbiased, the large sampling variability in certain data constellations (e.g.,

small number of L2 groups, small ICCs, and small *n*) suggests that it should be applied only very cautiously in such cases. Although asymptotically the MLC approach provides the most efficient and consistent estimator of the contextual effect and is thus asymptotically superior to the MMC approach, the results of our simulation study suggest that large numbers of L2 groups (*K*) may be needed for these asymptotic properties to hold. Particularly for studies where the number of L2 groups and the number of L1 cases within each group are modest, the latent variable approach can be recommended only when the ICC is very large.

We also compared the MLC approach to the two-stage latent variable approach for L2 reflective constructs proposed by Croon and van Veldhoven (2007). Consistent with their speculations, our simulation study showed that the two approaches yielded very similar results, except under the condition with small sample sizes at both L1 and L2 and a low ICC. For this data constellation, the FIML MLC approach outperformed the two-stage approach. Because the two-stage approach is only a limited information approach, the FIML implementation should be generally preferred. Although it was not a focus of our study, we also note that the potentially cumbersome two-stage approach to estimating standard errors requires further consideration. When discussing critical issues for future research in multilevel latent variable modeling, Croon and van Veldhoven (2007, p. 55) emphasized that "efforts should be made to develop a reliable numerical and generally applicable procedure that yields the full information maximum likelihood estimates of the model parameters." Mplus provides such a flexible latent variable model that integrates several different analysis models within a unified MLM framework (L. K. Muthén & Muthén, 2007). Hence, we recommend that the one-step approach demonstrated here be used instead of the two-stage approach.

Another even simpler approach (related to the Croon & van Veldhoven, 2007, two-stage approach) would be to estimate the reliability of the group mean via Equation 6 and then to correct the estimated regression coefficient for the unreliability of the group mean (see Grilli & Rampichini, 2007). To this end, the well-known correction for attenuation formula can be used to adjust the between-groups regression coefficient by multiplying it by $\frac{1}{\text{Rel}(\bar{X}_j)} = \frac{\tau^2 + \sigma^2/n}{\tau^2}$. This disattenuation approach thus consists of two steps. First, the MMC model is specified to estimate the between-groups regression coefficient. Second, the estimated between-groups regression coefficient is corrected for unreliability by the attenuation formula. In other words, the MMC approach is used to implement an MLC approach. A major drawback of that approach is that it is well defined only for balanced group sizes. However, even if the group sizes vary moderately, it may be acceptable to use the mean group size (see Snijders & Bosker, 1999, for a discussion of different adjustment formulas). In addition, the standard error of the regression coefficient needs to be adjusted when applying the disattenuation approach (e.g., using a bootstrapping procedure; see Carpenter, Goldstein, & Rasbash, 2003). Further research should evaluate how this disattenuation approach is related to the FIML implementation of the MLC approach as well as to the MMC approach for balanced and unbalanced group sizes. Although it may be premature to recommend that this disattenuation approach be used routinely, it does provide the applied researcher with an initial indication of the size of the bias that might be expected within the MMC approach when the MLC approach is appropriate. However, further investigation of the approach is required, particularly for small sample sizes.

Another important application of the MLC approach is when the true value of an L2 formative construct is unknown because the entire cluster has not been sampled. In this case, the MLC approach is used to control for a low sampling ratio and the limited reliability of the L2 formative construct. In Study 3, we tested the MLC approach's suitability for adjusting for the effect of a small sampling ratio from a finite population. The results showed that, when the finite sample size of the L2 units is at least moderate (e.g., 100) and the sampling ratio is low (e.g., 20%; 20 cases from a finite population of 100), the MLC approach outperformed the MMC approach in terms of both bias and RMSE. Critically, however, the bias associated with the two approaches was in opposite directions. Importantly, for formative aggregations, a sampling ratio tending to 0 corresponds with the assumptions of the MLC approach, whereas a sampling ratio of 1 corresponds with the assumptions of the MMC approach.

The relative sizes of the counterbalancing biases associated with the two approaches varied systematically with sample size and sampling ratio. When the number of L1 cases within each L2 group is sufficiently large, the manifest and latent approaches give similar results (because L2 unreliability is negligible). However, when the sampling ratio and sample size are both small, and the two approaches are thus likely to give very different results, the only conclusion that the applied researcher can make with confidence is that the true value is on average somewhere between the results of the MMC and the MLC approaches. Hence, the most reasonable recommendation is to use both approaches to determine whether their results point to substantively different conclusions. If so, conclusions must be made with caution.

What are the consequences of these findings for applying the MLC approach to group-level constructs based on formative aggregation? In research practice, it is often difficult to determine the true cluster sizes because ad hoc samples are frequently drawn without a sampling scheme. In Study 5, for example, there was an average of 23.6 students per school. If we assume that the average school size is at least

500 students, then the sampling ratio is less than .05. However, the sample was limited to students who were 15 years of age, and the average number of 15-year-old students per school might be just 100. We are not necessarily arguing for this alternative interpretation, but we use this example to illustrate why the computation of sampling ratio might not be straightforward. Furthermore, the true cluster sizes are likely to vary, perhaps substantially, across different L2 groups. Thus, further research is needed to study the effect of varying finite sample sizes of the L2 clusters on the estimator of the contextual effect based on L2 formative constructs. Although our simulation results do not provide a sufficient basis for making detailed recommendations, we would like to offer the following guidelines (subject to further research).

For formative L2 constructs (as in Study 5), the MMC approach can comfortably be used when the sampling ratio approaches 1.0. Even when the sampling ratio is moderate (at least .5), the MMC approach seems to provide relatively unbiased estimates as long as the sample size and ICC are large. The MLC approach should be considered instead of the MMC approach when the sampling ratio is very small and the numbers of L2 groups and L1 cases in each L2 group are large. In all other cases, the applied researcher should apply both the latent and manifest approaches and compare the results of each. If the results of the two approaches differ substantively, the applied researcher should be cautious in drawing any conclusions. Because these recommendations are not entirely satisfactory in providing clear-cut advice, this is an area in which more research is needed. We also note that there may be applications where the true population mean for each L2 group is known (e.g., based on information external to the study such as gender ratio of a school), although only a sample of participants is considered. In this special case, it would be appropriate to use the known true population mean for each group (instead of the sample mean based on data actually collected) and to interpret the sampling ratio in relation to this L2 construct as being 100%. Hence, for formative L2 constructs in which the true (population) sample mean is known for each L2 group, we recommend that the MMC is always appropriate and should be used instead of the MLC approach described here.

For reflective L2 constructs, the researcher should always use the MLC approach in preference to the MMC approach—even when the apparent sampling ratio is very large. The rationale for this recommendation, as for the domain sampling rationale in the classical approach to measurement, is that there is a potentially infinite number of L1 indicators that could be sampled. There are, however, important qualifications to this recommendation. When the number of groups, the number of cases within each group, and the ICC are all modest, the contextual effect estimates for the MLC approach are—because of their larger variability—less accurate than those of the MMC approach. However, we cannot recommend the MMC approach because the parameter estimates are likely to be very negatively biased and the apparently small standard errors are likely to substantially underestimate sampling variability (at least in terms of making population inferences). Rather, we recommend that the results of contextual effects studies for reflective L2 constructs based on small *ns* at L1 and, in particular, small numbers of L2 groups be interpreted very cautiously unless the ICCs are substantial.

Theoretical Considerations

In multilevel studies, group-level constructs are often constructed by aggregating individual data at the group level. The theoretical rationale for the aggregation process may differ. In this article, we distinguish two quite different aggregation processes—reflective and formative aggregation—that represent opposite ends of a continuum.

At the reflective aggregation end, the aggregation process assumes an isomorphic relationship between the individual-level data and the group-level construct. In other words, a generic group-level construct is assumed to be measured by the corresponding constructs at the individual level. A typical research paradigm was presented in Study 4, in which L1 students' responses were treated as observers of their L2 teacher's behavior and the referent was the L2 teacher. Ideally, each student would assign the same rating, such that the responses of students in the same class would be interchangeable. Because the L1 perceptions of each student were designed to measure the same L2 construct, variation within each class can be regarded as L2 unreliability (Cronbach, 1976; Van Mierlo, Vermunt, & Rutte, 2008). This situation is analogous to typical assumptions in test construction, in which differences between multiple items are assumed to reflect measurement error that varies as a function of the number of items and the size of correlations among items. Further examples of group-level constructs that rely on this aggregation process are multiple L1 assessors' evaluations of the L2 quality of grant proposals (Jayasinghe et al., 2003), L1 students' evaluations of L2 teaching effectiveness (Marsh, 1987, 2007), and multiple interchangeable L1 markers assessing the quality of student essays. In each of these situations in which the referent is a group-level construct, the reflective aggregation of L1 constructs to form L2 constructs and the application of the MLC approach seem reasonable.

At the formative aggregation end of the continuum, the aggregation process assumes that the group-level variable is merely an index of a well-defined L1 construct that is aggregated to the L2 group level. In other words, the aggregation is not based on multiple interchangeable observations of a single entity but on different characteristics associated with discrete (noninterchangeable) individuals. In fact, under appropriate

circumstances (high levels of reliability at L1 and large numbers of L1 cases within each L2 group or a sampling ratio that approaches 1.0), it is reasonable to argue that nearly all of the within-group variability reflects true score differences among different individuals within each group. As such, L2 unreliability becomes trivial in size and, perhaps, ignorable (consistent with the MMC approach). Particularly when the main aim of a construct is to reflect individual differences at L1 (e.g., academic achievement, SES, individual demographic characteristics such as race, age, and gender) and the referent is an L1 construct, the level of interrater agreement associated with the aggregated L2 construct may be of no consequence to construct validity at L1 (Bliese, 2000). This situation was demonstrated in Study 5, in which school-average SES was determined by aggregating individual-level (L1) student SES.

Limitations of the Manifest and Latent Approaches: Directions for Further Research

As formulated in this article, both the manifest and the latent covariate approaches begin with manifest scale scores at L1, largely ignoring the potential to estimate and control for measurement error at L1—even when L1 measures are based on multiple indicators. Under appropriate circumstances, the integration of multiple L1 indicators into the analyses would allow researchers using either of the two approaches to differentiate between L1 measurement error and L2 unreliability due to sampling error associated with the aggregation process in moving from L1 to L2 aggregations. This would be particularly valuable for the MMC approach, offering researchers a way to estimate L2 unreliability in L2 formative measures (see Kline, 2005, for discussion of how to take account of measurement error in observed exogenous variables). Furthermore, low reliability of individual measures can give the appearance of substantial estimates of group-level contextual effects. However, these estimates are biased because the aggregate measure is more reliable and "mops up" variance that would be explained at the individual level with more reliable measures—the so-called phantom effect or contextual fallacy (Harker & Tymms, 2004; see also Lüdtke, Robitzsch, & Köller, 2002). Historically, limitations in statistical software posed intractable problems in integrating confirmatory factor analysis/SEM and MLM models (e.g., Mehta & Neale, 2005; B. O. Muthén, 1991). However, even with conventional MLM programs that do not explicitly incorporate multiple indicators of each L1 construct, it is possible to incorporate information about L1 measurement error (e.g., Goldstein, 2003; Raudenbush & Bryk, 2002). More sophisticated statistical packages are increasingly providing applied researchers with added flexibility to incorporate multiple indicators at L1 while addressing the multilevel structure of their data.

In the present article, we looked at models in which only the intercepts were allowed to vary (random-intercept models). However, in research practice models that allow slopes to vary between L2 units are of great interest (random-slope models). For instance, following our application in Study 5, the relationship between SES and reading achievement could be different in high and low SES schools (see also Example 9.10 in L. K. Muthén & Muthén, 2007). Further research is needed to investigate the behavior of the MLC and MMC approaches when identifying these cross-level interactions.

Implicit in our presentation of the MLC approach for L2 reflective measures is the assumption that L1 individuals can be regarded as indicators and that these indicators are drawn from a population of infinite potential indicators to represent each L2 group, analogous to assumptions made in the domain sampling approach to classical measurement theory. From this perspective, sampling ratio is not a critical concern for reflective L2 measures, although concerns about having adequate numbers of L1 cases in each L2 group are still relevant. Although the correction for unreliability in L2 aggregates in the MLC approach is generally appropriate under the assumption that it is at least hypothetically possible to have an infinitely large sample size, this assumption may be questionable in situations in which it is not feasible to have a large number of L1 cases for each L2 group.

In the MLC procedure, ICCs are an index of interrater agreement and one basis for the determination of L2 reliability. With an ICC = 0, which implies a reliability of zero, correction for unreliability becomes problematic. The same would apply to reliability estimates based on multiple items in the classical approach to measurement; this is not an issue that is idiosyncratic to our new approach. Hence, it is not surprising that the largest difference between the MMC approach (that ignores unreliability) and the MLC approach (that corrects for unreliability) is when reliability is very low. Thus, it is important that researchers routinely provide an estimate for the ICC when applying the MLC approach and, if the ICC is close to zero, justify the appropriateness of subsequent analyses and their interpretations. Further research on the most appropriate estimation procedure and interpretation of results based on very small ICCs is warranted.

Conclusion

The simultaneous investigation of individual and group effects is one of the basic features of MLMs. In research practice, the L2 group characteristics are often measured by aggregation from L1 individual measures. Two approaches that differ in their treatment of the aggregated group-level construct were compared. Whereas the MLC approach corrects for the unreliable assessment of the latent group mean when estimating MLMs, the MMC approach relies solely on the observed group mean that is assumed to be measured

with no L2 unreliability. We argue that the appropriateness of either of these approaches depends on the research question and the nature of the L2 construct under study. If a generic group-level construct is assessed through a reflective aggregation of L1 measures to form the L2 construct, then the MLC approach is appropriate and offers many advantages as long as minimal standards are met. However, when the aggregated variable is a formative summary of the observations at the individual level (e.g., school-average SES), the assumptions made by the MLC approach are appropriate only when the sampling ratio is small. In research practice, the distinction between group-level variables based on reflective and formative aggregation is not usually so clear cut, and it is easy to imagine situations in which the theoretical status of the group-level construct is ambiguous and the calculation of the sampling ratio is not straightforward. Hence, it might be useful to analyze the sensitivity of empirical results to both approaches. In conclusion, although the latent covariate approach demonstrated here has wide applicability in relation to a serious limitation of existing research, the appropriateness of its application varies depending on the nature of the data, the number of L2 groups, the number of cases within each group, and the sampling ratio.

References

Anderman, E. M. (2002). School effects on psychological outcomes during adolescence. *Journal of Educational Psychology, 94,* 795–809.

Asparouhov, T., & Muthén, B. (2003). *Full-information maximum-likelihood estimation of general two-level latent variable models with missing data* (Tech. report). Los Angeles: Muthén & Muthén.

Asparouhov, T., & Muthén, B. (2006). *Constructing covariates in multilevel regression.* Mplus Web Notes: No. 11. Retrieved from http://www.statmodel.com/download/webnotes/webnote11.pdf

Bauer, D. J., Preacher, K. J., & Gil, K. M. (2006). Conceptualizing and testing random indirect effects and moderated mediation in multilevel models: New procedures and recommendations. *Psychological Methods, 11,* 142–163.

Baumert, J., Artelt, C., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., et al. (Eds.). (2002). *PISA 2000: Die Länder der Bundesrepublik Deutschland im Vergleich* [PISA 2000: A comparison of the German states]. Opladen, Germany: Leske & Budrich.

Baumert, J., Lehmann, R. H., Lehrke, M., Schmitz, B., Clausen, M., Hosenfeld, I., et al. (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.

Blakely, T. A., & Woodward, A. J. (2000). Ecological effects in multi-level studies. *Journal of Epidemiological Health, 54,* 367–374.

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

Bliese, P. D., Chan, D., & Ployhart, R. E. (2007). Multilevel methods: Future directions in measurement, longitudinal analyses, and nonnormal outcomes. *Organizational Research Methods, 10,* 551–563.

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

Boyd, L., & Iverson, G. (1979). *Contextual analysis: Concepts and statistical techniques.* Belmont, CA: Wadsworth.

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Applied Statistics, 52,* 431–443.

Carroll, R. J., Ruppert, D., & Stefanski, L. A. (1995). *Measurement error in nonlinear models.* London: Chapman & Hall/CRC.

Chan, D. (1998). Functional relations among constructs in the same content domain at different levels of analysis: A typology of composition models. *Journal of Applied Psychology, 83,* 234–246.

Chen, G., Bliese, P. D., & Mathieu, J. E. (2005). Conceptual framework and statistical procedures for delineating and testing multilevel theories of homology. *Organizational Research Methods, 8,* 375–409.

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 variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods, 12,* 45–57.

Davis, J. A. (1966). The campus as a frog pond: An application of the theory of relative deprivation to career decisions of college men. *American Journal of Sociology, 72,* 17–31.

Edwards, J. R., & Bagozzi, R. P. (2000). On the nature and direction of relationships between constructs and measures. *Psychological Methods, 5,* 155–174.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review, 43,* 557–572.

Goldstein, H. (1987). *Multilevel models in educational and social research.* London: Griffin.

Goldstein, H. (2003). *Multilevel statistical models.* London: Arnold.

Grilli, L., & Rampichini, C. (2007). *Endogeneity issues in multilevel linear models.* Unpublished manuscript.

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.

Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement, 15,* 177–199.

Howell, R. D., Breivik, E., & Wilcox, J. B. (2007). Reconsidering formative measurement. *Psychological Methods, 12,* 205–218.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications.* Mahwah, NJ: Erlbaum.

Iverson, G. R. (1991). *Contextual analysis* (Sage University Paper Series on Quantitative Applications in the Social Sciences No. 07–081). Newbury Park, CA: Sage.

James, L. R. (1982). Aggregation bias in estimates of perceptual agreement. *Journal of Applied Psychology, 67,* 219–229.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society Series A: Statistics in Society, 166,* 279–300.

Kenny, D. A., & La Voie, L. (1985). Separating individual and group effects. *Journal of Personality and Social Psychology, 48,* 339–348.

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.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

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

Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). San Francisco: Jossey-Bass.

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.

Kunter, M., Baumert, J., & Köller, O. (2007). Effective classroom management and the development of subject-related interest. *Learning and Instruction, 17,* 494–509.

Lee, S.-Y., & Poon, W.-Y. (1998). Analysis of two-level structural equation models via EM type algorithms. *Statistica Sinica, 8,* 749–766.

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

Lüdtke, O., Köller, O., Marsh, H. W., & Trautwein, U. (2005). Teacher frame of reference and the big-fish–little-pond effect. *Contemporary Educational Psychology, 30,* 263–285.

Lüdtke, O., Robitzsch, A., & Köller, O. (2002). Statistische Artefakte bei Kontexteffekten in der pädagogisch-psychologischen Forschung [Statistical artifacts in educational studies on context effects]. *German Journal of Educational Psychology, 16,* 217–231.

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.

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

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11,* 253–388.

Marsh, H. W. (2007). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–384). New York: Springer.

Marsh, H. W., & Ball, S. (1981). The interjudgmental reliability of reviews for the Journal of Educational Psychology. *Journal of Educational Psychology, 18,* 872–880.

Marsh, H. W., & Hau, K.-T. (2003). Big-fish–little-pond effect on academic self-concept: A cross-cultural (26 country) test of the negative effects of academically selective schools. *American Psychologist, 58,* 364–376.

Marsh, H. W., Kong, C.-K., & Hau, K.-T. (2000). Longitudinal multilevel models of the big-fish–little-pond effect on academic self-concept: Counterbalancing contrast and reflected-glory effects in Hong Kong schools. *Journal of Personality and Social Psychology, 78,* 337–349.

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models.* New York: Wiley.

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

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.

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

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B. O. (2002). Beyond SEM: General latent variable modeling. *Behaviormetrika, 29,* 81–117.

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

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

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

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life. First results from the OECD Programme for International Student Assessment (PISA) 2000.* Paris: Author.

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 and Exercise Psychology, 26,* 90–118.

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

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

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics, 16,* 295–330.

Raykov, T. (2007). Longitudinal analysis with regressions among random effects: A latent variable modeling approach. *Structural Equation Modeling, 14,* 146–169.

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.

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

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

Seber, G. A. F. (1977). *Linear regression analysis.* New York: Wiley.

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: Sage.

Urdan, T., Midgley, C., & Anderman, E. M. (1998). The role of classroom goal structure in students' use of self-handicapping strategies. *American Educational Research Journal, 35,* 101–122.

Van Mierlo, H., Vermunt, J. K., & Rutte, C. G. (2008). Composing group-level constructs from individual-level survey data. *Organizational Research Methods.* Advance online publication. Retrieved June 10, 2008. doi:10.1177/1094428107309322

Willms, D. (1985). The balance thesis: Contextual effects of ability on pupils' examination results in Scotland. *Oxford Review of Education, 11,* 33–41.

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

Appendix

Derivation of Bias for the Multilevel Manifest Covariate (MMC) Approach

In this Appendix, we derive the bias for the MMC approach. Because we are interested in within-group and between-groups relations, the following population model for two variables X and Y will be assumed (see Snijders & Bosker, p. 29):

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

$$Y_{ij} = \mu_y + U_{yj} + R_{yij}.$$

In this model, group (e.g., school) j has specific main effects U_{xj} and U_{yj} for variables X and Y, and the within-group deviations R_{xij} and R_{yij} are associated with individual (e.g., student) i. The population means are denoted μ_x and μ_y, and it is assumed that the Us and the Rs have population means 0. In addition, the Us and the Rs are independent.

The covariance matrix of X and Y at Level 1 and Level 2 can be written as

$$\underset{\text{Level 1 (within)}}{\begin{pmatrix} \sigma_x^2 & \\ \sigma_{xy} & \sigma_y^2 \end{pmatrix}} \quad \underset{\text{Level 2 (between)}}{\begin{pmatrix} \tau_x^2 & \\ \tau_{xy} & \tau_y^2 \end{pmatrix}}.$$

We are interested in estimating the following relationship in the population:

$$Y_{ij} = \mu_y + \beta_{\text{within}} R_{xij} + \beta_{\text{between}} U_{xj} + \delta_j + \varepsilon_{ij}.$$

where μ_y is the grand mean, β_{within} the within-group regression coefficient, β_{between} the between-groups regression coefficient, δ_j a group-specific residual, and ε_{ij} an individual-specific residual. In the group-mean centered case, the following multilevel model would be specified to estimate β_{within} and β_{within}:

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

Under the assumption of equal group sizes n, $\bar{X}_{\cdot j} = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$ is the mean for group j. Furthermore, γ_{00}, γ_{10}, and γ_{01} denote the estimators for μ_y, β_{within}, and β_{between}. The L2 and L1 residuals are given by u_{0j} and r_{ij}. Given the covariance matrix of X and Y at Level 1 and Level 2, the

observed covariance matrix of Y_{ij}, $X_{ij} - \bar{X}_{\cdot j}$, and $\bar{X}_{\cdot j}$ is distributed as follows:

$$Cov\begin{bmatrix} Y_{ij} \\ X_{ij} - \bar{X}_{\cdot j} \\ \bar{X}_{\cdot j} \end{bmatrix} = \begin{pmatrix} \sigma_y^2 + \tau_y^2 & & \\ \sigma_{xy} \cdot (1 - 1/n) & \sigma_x^2(1 - 1/n) & \\ \tau_{xy} + \sigma_{xy}/n & 0 & \tau_x^2 + \sigma_x^2/n \end{pmatrix}.$$

As can be seen, the covariances between Y_{ij}, $X_{ij} - \bar{X}_{\cdot j}$, and $\bar{X}_{\cdot j}$ depend on the common group size as well as on the "true" covariances within and between groups. Employing the OLS principle and bearing in mind that the predictors $X_{ij} - \bar{X}_{\cdot j}$ and $\bar{X}_{\cdot j}$ are uncorrelated, the estimator $\hat{\gamma}_{10}$ for the within-group regression coefficient β_{within} can be obtained as

$$\hat{\gamma}_{10} = \frac{Cov(Y_{ij}, X_{ij} - \bar{X}_{\cdot j})}{Var(X_{ij} - \bar{X}_{\cdot j})} = \frac{\sigma_{xy}(1 - 1/n)}{\sigma_x^2(1 - 1/n)} = \frac{\sigma_{xy}}{\sigma_x^2} = \beta_{\text{within}}.$$

Now let $\text{ICC}_x = \frac{\tau_x^2}{\tau_x^2 + \sigma_x^2}$. denote the ICC for X. The estimator $\hat{\gamma}_{01}$ of the between-groups regression coefficient β_{between} can now be formulated as follows:

$$\begin{aligned}\hat{\gamma}_{01} = \frac{Cov(Y_{ij}, \bar{X}_{\cdot j})}{Var(\bar{X}_{\cdot j})} &= \frac{\tau_{xy} + \sigma_{xy}/n}{\tau_x^2 + \sigma_x^2/n} \\ &= \frac{\tau_{xy}}{\tau_x^2} \cdot \frac{\tau_x^2}{\tau_x^2 + \sigma_x^2/n} + \frac{\sigma_{xy}/n}{\sigma_x^2/n} \cdot \frac{\sigma_x^2/n}{\tau_x^2 + \sigma_x^2/n} \\ &= \beta_{\text{between}} \cdot \left(1 - \frac{1}{n} \cdot \frac{(1 - \text{ICC}_x)}{\text{ICC}_x + (1 - \text{ICC}_x)/n}\right) \\ &\quad + \beta_{\text{within}} \cdot \frac{1}{n} \cdot \frac{(1 - \text{ICC}_x)}{\text{ICC}_x + (1 - \text{ICC}_x)/n}.\end{aligned}$$

Thus, the bias on the between level is now computed as follows:

$$E(\hat{\gamma}_{01} - \beta_{\text{between}}) = (\beta_{\text{within}} - \beta_{\text{between}}) \cdot \frac{1}{n} \cdot \frac{(1 - \text{ICC}_x)}{\text{ICC}_x + (1 - \text{ICC}_x)/n}.$$

As can be seen, the bias depends primarily on the proportion of variance in X that is located between groups (ICC_x) and on the average group size n. If $n \to \infty$, the estimator will be unbiased.

Received August 28, 2007
Revision received May 1, 2008
Accepted May 7, 2008 ■