# The role of sample cluster means in multilevel models: a view on endogeneity and measurement error issues

Leonardo Grilli and Carla Rampichini
grilli@ds.unifi.it rampichini@ds.unifi.it
Department of Statistics 'Giuseppe Parenti', University of Florence

# The role of sample cluster means in multilevel models: a view on endogeneity and measurement error issues

**Abstract**

The paper explores some issues related to endogeneity in multilevel models, focusing on the case where the random effects are correlated with a level 1 covariate in a linear random intercept model. We consider two basic specifications, without and with the sample cluster mean. It is generally acknowledged that the omission of the cluster mean may cause omitted-variable bias. However, it is often neglected that the inclusion of the sample cluster mean in place of the population cluster mean entails a measurement error that yields biased estimators for both the slopes and the variance components. In particular, the contextual effect is attenuated, while the level 2 variance is inflated. We derive explicit formulae for measurement error biases that allow us to implement simple post-estimation corrections based on the reliability of the covariate. In the first part of the paper, the issue is tackled in a standard framework where the population cluster mean is treated as a latent variable. Later we consider a different framework arising when sampling from clusters of finite size, where the latent variable methods may have a poor performance, and we show how to effectively modify the measurement error correction. The theoretical analysis is supplemented with a simulation study and a discussion of the implications for effectiveness evaluation.

**Keywords**: cluster mean, contextual effect, effectiveness evaluation, random effects, reliability.

# 1 Introduction

Regression analysis with data from observational studies is often threatened by endogeneity, namely a lack of independence of the model errors from the covariates, which yields biased estimators of the model parameters. Two major sources of endogeneity, which will be considered in the paper, are covariate omission and covariate measurement error.

Multilevel random effects models have at least one error term at each hierarchical level, so the endogeneity can concern errors at any level. Our contribution considers two-level random intercept models and focuses on the *level 2 endogeneity* arising when the level 2 errors (random effects) are correlated with level 1 covariates. This issue is well known in the setting of panel data (Hausman and Taylor, 1981), but the topic has recently received attention also in a more general perspective: see Skrondal and Rabe-Hesketh (2004), Fielding (2004), Ebbes *et al*. (2004), Kim and Frees (2007) and Snijders and Berkhof (2008).

Let us consider a random intercept model with a level 1 covariate $X_{ij}$,

$$Y_{ij} = \gamma_0 + \gamma_1 X_{ij} + v_j + e_{ij}$$

where $i = 1, 2, \ldots, n_j$ is the elementary (level 1) index and $j = 1, 2, \ldots, J$ is the cluster (level 2) index. For example, in a panel setting the elementary units are waves and the clusters are individuals, while in a cross-sectional framework the elementary units are individuals and the clusters are entities such as institutions or geographical areas. Moreover, $X_{ij}$ is a level 1 covariate with slope $\gamma_1$, $v_j$ are level 2 errors (random effects) and $e_{ij}$ are level 1 errors.

Level 2 endogeneity is characterized by $E(v_j \mid X_{ij}) \neq 0$, with the consequence that the standard estimators of $\gamma_1$ are biased. Note that $Cov(v_j, X_{ij}) \neq 0$ is a sufficient, though not necessary, condition for level 2 endogeneity.

If $E(v_j \mid X_{ij})$ is assumed to be a linear function of the cluster mean $\overline{X}_j$, a straightforward remedy to endogeneity is to add $\overline{X}_j$ to the model equation (Mundlak, 1978). From another point of view (Neuhaus and Kalbfleish, 1998; Snijders and Berkhof, 2008), the inclusion of $\overline{X}_j$ as a further regressor is just a way to disentangle the between-cluster and within-cluster effects, whose difference is known as the *contextual effect*, a key concept in social sciences and education (Raudenbush and Willms, 1995). However, it is usually not recognized that in most cases $\overline{X}_j$ is a *sample* cluster mean used to measure a *population* cluster mean: as a consequence, the model including $\overline{X}_j$ is affected by measurement error and thus the contextual effect is attenuated, while the level 2 variance is inflated. In the paper we deal with the measurement error issue, studying the biases and proposing simple post-estimation corrections based on the reliability of the covariate. In order to get simple expressions, we focus on the balanced case, i.e. clusters of equal size $n$, though we also consider the extension to unbalanced hierarchies. The properties of the corrected estimators are evaluated through a simulation study.

Our approach is complementary to Croon and van Veldhoven (2007) and Lüdtke *et al.* (2008), who deal with the attenuation of the contextual effect in a structural equation perspective, while Shin and Raudenbush (2010) tackle the issue in the framework of multivariate multilevel models with missing data. Our analysis is peculiar in many respects: (*i*) we interpret the measurement error of the sample cluster mean in terms of level 2 endogeneity, namely a correlation between covariates and random effects; this allows us to establish connections with the case where the cluster mean is omitted, in order to outline a comprehensive framework for analyzing the issues related to the cluster mean; (*ii*) compared to the existing literature, which focuses on the contextual coefficient, we explicitly investigate also the bias of the level 2 variance, showing that its pattern is only partially related to the bias of the contextual coefficient; (*iii*) we exploit the reliability of the covariate to develop simple post-estimation corrections for both the contextual coefficient and the level 2 variance; (*iv*) we propose an effective adjustment for the case where the values of the covariate are sampled from clusters of finite size: this is a relevant case that is not properly handled by the current methods relying on latent variable modelling; (*v*) we discuss how the issues related to the cluster mean affect the effectiveness evaluation of educational institutions.

In our treatment the covariate $X_{ij}$ is assumed to be measured without error, so the measurement error only affects the sample cluster mean $\overline{X}_j$ just because it is a measure of a population cluster mean. The case of multilevel models where a covariate itself is measured with error is treated for example by Woodhouse *et al.* (1996), Hutchison (2004), Ferrão and Goldstein (2009) and Lüdtke *et al.* (2009).

The rest of the paper is organized as follows. Section 2 describes the data generating

model, where the cluster mean is a latent variable. Section 3 explores the nature of level 2 endogeneity in the model without the cluster mean, while Section 4 deals with the measurement error induced by the use of the sample cluster mean. Section 5 shows how to correct the biases using the reliability of the covariate and reviews the alternative approach based on structural equation models. Section 6 discusses the nature of the cluster mean, showing that the latent variable approach is not appropriate when sampling from clusters of finite size. The subsequent Section 7 derives the adjustments needed in this case. In Section 8 the performances of the estimators are investigated through a simulation study. Section 9 discusses the implications for effectiveness evaluation and Section 10 concludes.

# 2 The *Latent Cluster Mean* model

Let us now define the model of interest, which we assume to have generated the data. The observed variables are a response $Y_{ij}$ and a covariate $X_{ij}$ which both vary within and between clusters indexed by $j$. For the covariate $X_{ij}$ we specify a variance component model

$$X_{ij} = X_j^B + X_{ij}^W \tag{1}$$

with the following assumptions: (X1) $X_j^B$ are iid with mean $\mu_X$ and variance $\tau_X^2 > 0$; (X2) $X_{ij}^W$ are iid with mean 0 and variance $\sigma_X^2 > 0$; (X3) $X_j^B$ and $X_{ij}^W$ are independent.

Assumptions (X1)-(X3) imply the usual variance decomposition $Var(X_{ij}) = \tau_X^2 + \sigma_X^2$. The Intraclass Correlation Coefficient (ICC) is $\rho_X = \tau_X^2/(\tau_X^2 + \sigma_X^2)$.

The assumptions $\tau_X^2 > 0$ and $\sigma_X^2 > 0$ imply that $X_{ij}$ varies both within and between clusters. If the covariate $X_{ij}$ were purely within (i.e. $\tau_X^2 = 0$), level 2 endogeneity would not be an issue; however, purely within covariates are rare in practice.

While $X_{ij}$ is observable, the components $X_j^B$ and $X_{ij}^W$ are unobservable, so in the models they must be replaced with their observable counterparts, i.e. the sample cluster mean $\overline{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$ for $X_j^B$ and the centered covariate $\widetilde{X}_{ij} = X_{ij} - \overline{X}_j$ for $X_{ij}^W$. The consequences of such substitutions will be explored in Sections 4 and 5. For the moment we reason as if $X_j^B$ and $X_{ij}^W$ were observable.

In a regression model of $Y_{ij}$ on $X_{ij}$ the effect of the between component $X_j^B$ is, in general, different from the effect of the within component $X_{ij}^W$. We therefore specify the regression model as a *Latent Cluster Mean* model:

$$Y_{ij} = \alpha + \beta_W X_{ij}^W + \beta_B X_j^B + u_j + e_{ij} \ , \tag{2}$$

where $\beta_W$ is the *within slope* and $\beta_B$ is the *between slope*. In many settings, the between and within slopes are conceptually different and may even have opposite signs, so it is important to distinguish them (Snijders and Bosker, 1999).

In the following we will often use the alternative parametrization

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta X_j^B + u_j + e_{ij} \ , \tag{3}$$

where $\delta = \beta_B - \beta_W$ is the *contextual coefficient* (Raudenbush and Willms, 1995).

Considering an arbitrary cluster $j$ of sample size $n$ and defining $\mathbf{X}_j^W = (X_{1j}^W, \ldots, X_{nj}^W)'$, the assumptions on the model errors are: (Y1) $u_j \mid X_j^B, \mathbf{X}_j^W$ are iid with mean 0 and variance $\tau^2_{Y|X^B X^W}$; (Y2) $e_{ij} \mid X_j^B, \mathbf{X}_j^W$ are iid with mean 0 and variance $\sigma^2_{Y|X^B X^W}$; (Y3) $u_j$ and $e_{ij}$ are independent given $X_j^B$ and $\mathbf{X}_j^W$.

A key part of the assumption on the level 2 errors $u_j$ is $E(u_j \mid X_j^B, \mathbf{X}_j^W) = 0$, which is known as level 2 exogeneity and implies that each level 2 error is uncorrelated with the covariates, i.e. $Cov(u_j, X_{ij}^W) = 0$, $i = 1, \ldots, n$, and $Cov(u_j, X_j^B) = 0$. Level 2 endogeneity arises when the level 2 errors are correlated with the covariates.

In the *Latent Cluster Mean* model (2), under the stated assumptions the residual variance of $Y_{ij}$ decomposes as $\tau^2_{Y|X^B X^W} + \sigma^2_{Y|X^B X^W}$. The ICC is thus $\rho_{Y|X^B X^W} = \tau^2_{Y|X^B X^W} / (\tau^2_{Y|X^B X^W} + \sigma^2_{Y|X^B X^W})$, which equals the residual correlation among the responses of two units belonging to the same cluster.

# 3 Level 2 endogeneity in the *Raw Covariate* model: omitted-variable bias

The *Latent Cluster Mean* model (3) may be wrongly specified by omitting $X_j^B$. In such a case, the term $\delta X_j^B$ is absorbed by the level 2 error and the model reduces to the following *Raw Covariate* model:

$$Y_{ij} = \eta + \beta_W X_{ij} + v_j + e_{ij} \ , \tag{4}$$

where $\eta = (\alpha + \delta \mu_X)$ and $v_j = \delta(X_j^B - \mu_X) + u_j$. However, the estimable slope of $X_{ij}$ is not $\beta_W$ if $X_{ij}$ is correlated with $v_j$. Indeed, $Cov(v_j, X_{ij}) = Cov(v_j, X_j^B) = \delta \tau_X^2$, which is null only if the contextual effect $\delta$ is null. Since $v_j$ depends on $X_{ij}$ only through $X_j^B$, the correlation among $v_j$ and $X_{ij}$ has bounds that depend on the ICC of the covariate. In fact, it can be shown that the squared correlation among the random effects $v_j$ and the covariate $X_{ij}$ is an increasing function of $\delta^2$ and lies in the interval $(0, \rho_X)$.

Therefore, when $\delta \neq 0$ the *Raw Covariate* model (4) is affected by level 2 endogeneity, which can be seen as a consequence of omitting the population cluster mean $X_j^B$ from the *Latent Cluster Mean* model (3). Alternatively, such endogeneity can be viewed as stemming from a wrong equality assumption on the between and within slopes in model (2): in such a case, the estimable slope of the *Raw Covariate* model is a weighted average of $\beta_B$ and $\beta_W$ (Snijders and Bosker, 1999, sec. 3.6).

The level 2 endogeneity of the *Raw Covariate* model also implies that the estimable variances are different from the theoretical variances, which are $\sigma^2_{Y|X} = Var(e_{ij}) = \sigma^2_{Y|X^B X^W}$ and

$$\tau^2_{Y|X} = Var(v_j) = \delta^2 \tau_X^2 + \tau^2_{Y|X^B X^W}. \tag{5}$$

In fact, when $\delta \neq 0$, the bias in the estimate of $\beta_W$ also affects the estimates of the variances. Denoting with $\psi$ the bias of the slope, which cannot be expressed in closed form, it turns out that the estimable variance at level 1 is

$$\psi^2 \sigma_X^2 + \sigma^2_{Y|X^B X^W}, \tag{6}$$

and the estimable variance at level 2 is

$$(\delta - \psi)^2 \tau_X^2 + \tau^2_{Y|X^B X^W}. \tag{7}$$

Both variances depend on the bias of the slope $\psi$ and exceed the corresponding population variances $\sigma^2_{Y|X^B X^W}$ and $\tau^2_{Y|X^B X^W}$.

The endogeneity issue is less important when the clusters are large, since the estimate of the slope of $X_{ij}$ in the *Raw Covariate* model tends to $\beta_W$ as $n \to \infty$ (Raudenbush and Willms, 1995). Thus, in designs with large clusters the *Raw Covariate* model gives an approximately unbiased estimate of $\beta_W$ regardless of the contextual effect. At the same time, as the bias of the slope tends to zero, the estimable level 2 variance tends to $\tau^2_{Y|X}$, which is larger than $\tau^2_{Y|X^B X^W}$.

A popular solution to the endogeneity problem is the *fixed effects* approach, i.e. replacing the random effects with cluster-specific intercepts (Wooldridge, 2002). This approach allows us to unbiasedly estimate the within slope, but it precludes estimating the between slope and the contextual coefficient. Moreover, the level 2 variance is not a model parameter and it can only be estimated quite inefficiently as the variance of the estimated fixed effects.

# 4 Level 2 endogeneity in the *Sample Cluster Mean* model: measurement error bias

The level 2 endogeneity of the *Raw Covariate* model can be avoided by allowing between and within effects to be different, as in the *Latent Cluster Mean* model (2). However, this model cannot be fitted since $X_j^B$ and $X_{ij}^W$ are unobservable. Their sample counterparts are the sample cluster mean $\overline{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$ and the centered covariate $\widetilde{X}_{ij} = X_{ij} - \overline{X}_j$, respectively. In other words, the unobservable split $X_{ij} = X_j^B + X_{ij}^W$ is replaced with the observable split $X_{ij} = \overline{X}_j + \widetilde{X}_{ij}$. Note that $\overline{X}_j = X_j^B + \overline{X}_j^W$ and $\widetilde{X}_{ij} = X_{ij}^W - \overline{X}_j^W$, where $\overline{X}_j^W$ is the sample cluster mean of the latent within components $X_{ij}^W$. Thus both $X_j^B$ and $X_{ij}^W$ are measured with error: this is an instance of *classical error model* (Carroll *et al.*, 2006) with the peculiarity that the measurement errors of the two covariates have the same absolute value but opposite signs.

The sample cluster mean has variance

$$Var(\overline{X}_j) = Var(X_j^B) + Var(\overline{X}_j^W) = \tau_X^2 + \sigma_X^2/n \tag{8}$$

and reliability

$$\lambda_X = \frac{Var(X_j^B)}{Var(\overline{X}_j)} = \frac{\tau_X^2}{\tau_X^2 + \sigma_X^2/n} = \left(1 + \frac{1}{(\tau_X^2/\sigma_X^2)n}\right)^{-1}. \tag{9}$$

The reliability of the sample cluster mean as a measure of the cluster component $X_j^B$ lies in the interval $(0, 1)$ and it is an increasing function of the product of the variance ratio $\tau_X^2/\sigma_X^2$ by the cluster size $n$. For example, a reliability $2/3$ is obtained with $n = 2$ and $\tau_X^2 = \sigma_X^2$ (a panel data configuration) or with $n = 10$ and $\tau_X^2 = 0.20\sigma_X^2$ (a cross-section configuration).

The estimable version of the *Latent Cluster Mean* model (2) with $\widetilde{X}_{ij}$ in place of $X_{ij}^W$ and $\overline{X}_j$ in place of $X_j^B$ will be called *Sample Cluster Mean* model:

$$Y_{ij} = \alpha + \beta_W \widetilde{X}_{ij} + \beta_B \overline{X}_j + z_j + e_{ij} \tag{10}$$

or, alternatively,

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta \overline{X}_j + z_j + e_{ij} \ , \tag{11}$$

where $z_j = u_j - \delta \overline{X}_j^W$, with $E(z_j) = 0$ and $Var(z_j) = \delta^2 \sigma_X^2 / n + \tau^2_{Y|X^B X^W}$. We note that, due to the bias on $\delta$, the previous expression is different from the estimable level 2 variance, which is given later in formula (15).

The within slope $\beta_W$ can be unbiasedly estimated since $\widetilde{X}_{ij}$ is a purely within covariate and thus orthogonal to both $z_j$ and $\overline{X}_j$. On the other hand, the between slope $\beta_B$ and the contextual coefficient $\delta$ are estimated with bias: in fact, $Cov(z_j, \overline{X}_j) = -\delta \sigma_X^2 / n$, so if $\delta \neq 0$ then $\overline{X}_j$ is endogenous. The relevance of level 2 endogeneity can be summarized by the squared correlation among the random effects $z_j$ and the sample cluster mean $\overline{X}_j$, which is an increasing function of $\delta^2$ and lies in the interval $(0, 1 - \lambda_X)$.

In general, the correlation among $z_j$ and $\overline{X}_j$ induced by the measurement error of $\overline{X}_j$ yields biased estimates of $\alpha$, $\beta_B$, $\delta$ and $\tau^2_{Y|X^B X^W}$, while $\beta_W$ and $\sigma^2_{Y|X^B X^W}$ are unbiasedly estimated.

Let us first derive the bias on $\beta_B$. The estimable between slope is denoted with $\beta_{B,m}$, where the subscript $m$ stands for measurement error due to the sample cluster mean. To see how $\beta_{B,m}$ is related to $\beta_B$ and $\beta_W$, let us consider the model obtained by taking the sample cluster means of the variables in the *Sample Cluster Mean* model (10), namely $\overline{Y}_j = \alpha + \beta_B \overline{X}_j + z_j + \overline{e}_j$. In the balanced case, by the least squares criterion $\beta_{B,m} = Cov(\overline{X}_j, \overline{Y}_j) / Var(\overline{X}_j)$, which implies

$$\beta_{B,m} = \lambda_X \beta_B + (1 - \lambda_X) \beta_W = \beta_B - (1 - \lambda_X) \delta \ . \tag{12}$$

Therefore, the within slope $\beta_B$ is overestimated if $\delta < 0$ and underestimated if $\delta > 0$. In both cases the bias is a decreasing function of the reliability and vanishes when $\lambda_X = 1$.

The estimable contextual coefficient $\delta_m$ is

$$\delta_m = \beta_{B,m} - \beta_W = \lambda_X (\beta_B - \beta_W) = \lambda_X \delta \ , \tag{13}$$

so the population contextual coefficient $\delta$ is attenuated by the reliability of the covariate, with relative bias $-(1 - \lambda_X)$.

The measurement error also affects the intercept of the *Sample Cluster Mean* model: indeed, from (10) and (12) it follows that the estimable intercept is

$$\alpha_m = \alpha + (1 - \lambda_X) \delta \mu_X \ . \tag{14}$$

The measurement error caused by the sample cluster mean also affects the estimation of the level 2 variance. In fact, the estimable slope of $\overline{X}_j$ is $\delta_m$ rather than $\delta$, so the actual

level 2 error in model (11) is $(\delta - \delta_m)\overline{X}_j + z_j$. The estimable level 2 variance is thus

$$\begin{aligned}
\tau^2_{Y|X^B X^W,m} &= Var[(\delta - \delta_m)\overline{X}_j + z_j] \\
&= Var[(1-\lambda_X)\delta(X^B_j + \overline{X}^W_j) - \delta\overline{X}^W_j + u_j] \\
&= Var[(1-\lambda_X)\delta X^B_j - \lambda_X\delta\overline{X}^W_j + u_j] \\
&= (1-\lambda_X)^2\delta^2\tau^2_X + \lambda^2_X\delta^2\frac{\sigma^2_X}{n} + \tau^2_{Y|X^B X^W} \\
&= (1-\lambda_X)\delta^2\tau^2_X + \tau^2_{Y|X^B X^W} \quad . \qquad (15)
\end{aligned}$$

Therefore, the *Sample Cluster Mean* model entails an overestimation of the population level 2 variance $\tau^2_{Y|X^B X^W}$. On the contrary, the level 1 variance $\sigma^2_{Y|X^B X^W}$ is unbiasedly estimated, so the ICC is overestimated.

The two cases of endogeneity discussed so far are summarized in Table 1.

Table 1: Two types of endogeneity arising when the *Latent Cluster Mean* model is wrongly specified: omitted variable (*Raw Covariate* model) and measurement error (*Sample Cluster Mean* model).

| | *Raw Covariate model* | *Sample Cluster Mean model* |
|---|---|---|
| Model equation | $Y_{ij} = \eta + \beta_W X_{ij} + v_j + e_{ij}$ | $Y_{ij} = \alpha + \beta_W X_{ij} + \delta\overline{X}_j + z_j + e_{ij}$ |
| Regressor omission | yes (if $\delta \neq 0$) | no |
| Measurement error | no | yes (if $\lambda_X < 1$) |
| Level 2 error covariance | $Cov(v_j, X_{ij}) = \delta\tau^2_X$ | $Cov(z_j, \overline{X}_j) = -\delta\frac{\sigma^2_X}{n}$ |
| Estimable $\beta_W$ | $\beta_W + \psi$ | $\beta_W$ |
| Estimable $\delta$ | - | $\lambda_X\delta$ |
| Estimable level 1 variance | $\psi^2\sigma^2_X + \sigma^2_{Y|X^B X^W}$ | $\sigma^2_{Y|X^B X^W}$ |
| Estimable level 2 variance | $(\delta-\psi)^2\tau^2_X + \tau^2_{Y|X^B X^W}$ | $(1-\lambda_X)\tau^2_X\delta^2 + \tau^2_{Y|X^B X^W}$ |

It is instructive to compare the *Raw Covariate* model (4), which has a single covariate $X_{ij}$, with the *Sample Cluster Mean* model (11), which has covariates $X_{ij}$ and $\overline{X}_j$. Both models are affected by level 2 endogeneity when $\delta \neq 0$. However, in the *Raw Covariate* model the endogeneity arises from the omission of the relevant covariate $X^B_j$, while in the *Sample Cluster Mean* model the endogeneity is due to the measurement error caused by using $\overline{X}_j$ instead of $X^B_j$. In the *Sample Cluster Mean* model the problem is less serious since the slope of $X_{ij}$ is not affected and a simple correction is available for the slope of $\overline{X}_j$. Note that in the *Sample Cluster Mean* model the covariance between the random effects and the sample cluster mean depends not only on the model parameters, but also on the design through the cluster size $n$.

The *Raw Covariate* model and the *Sample Cluster Mean* model can be fitted via likelihood methods such as FIML (Full Information Maximum Likelihood) and REML (REstricted Maximum Likelihood). FIML and REML are two versions of the Generalized Least Squares estimator for the fixed effects that differ in the estimation of the variance components (Skrondal and Rabe-Hesketh, 2004): FIML is efficient, but it underestimates the level 2 variance, so our setting we prefer to use the unbiased, even if less efficient, REML.

# 5 Correcting the measurement error biases of the *Sample Cluster Mean* model

In order to overcome the measurement error problem due to the use of the sample cluster mean, two main routes are possible: (*i*) fit the *Sample Cluster Mean* model and then correct the estimates using the reliability; and (*ii*) directly fit the *Latent Cluster Mean* model, which is a structural equation model.

## 5.1 Correction of measurement error biases via the reliability

The expressions of the biases derived in Section 4 can be exploited to correct the biased estimates yielded by the *Sample Cluster Mean* model. The key quantity for these corrections is the reliability of the sample cluster mean $\lambda_X$, that can be estimated by plugging estimates of $\sigma^2_X$ and $\tau^2_X$ into equation (9). Unbiased estimates of $\sigma^2_X$ and $\tau^2_X$ can be obtained by fitting a variance component model for $X_{ij}$, or using the so-called ANOVA formulae based on the observed between and within sum of squares (Snijders and Bosker, 1999).

In most applications the parameter of main interest is the contextual coefficient $\delta$, which can be unbiasedly estimated with a simple correction derived from formula (13):

$$\widehat{\delta}_c = \frac{\widehat{\delta}_m}{\widehat{\lambda}_X} \ , \tag{16}$$

where the subscript $c$ means *corrected* and the estimate of $\delta_m$ is obtained from the *Sample Cluster Mean* model.

The expectation and sampling variance of $\widehat{\delta}_c$ can be approximated via the first-order Taylor approximation for the ratio of two random variables (Casella and Berger, 2001):

$$E\left(\widehat{\delta}_c\right) = E\left(\frac{\widehat{\delta}_m}{\widehat{\lambda}_X}\right) \simeq \frac{\delta_m}{\lambda_X} = \delta \tag{17}$$

and

$$Var\left(\widehat{\delta}_c\right) = Var\left(\frac{\widehat{\delta}_m}{\widehat{\lambda}_X}\right) \simeq \left(\frac{\delta_m}{\lambda_X}\right)^2 \left[\frac{Var(\widehat{\delta}_m)}{\delta^2_m} + \frac{Var(\widehat{\lambda}_X)}{\lambda^2_X}\right] \ , \tag{18}$$

where the formula for the variance is obtained using $Cov(\widehat{\delta}_m, \widehat{\lambda}_X) = 0$. The sampling variance (18) can be estimated by plugging in the point estimates of $\delta_m$ and $\lambda_X$ and their estimated sampling variances (the sampling variance of $\widehat{\lambda}_X$ can be computed via the delta method).

Even if the corrected estimator $\widehat{\delta}_c$ is approximately unbiased, it follows from (18) that the sampling variance of $\widehat{\delta}_c$ is higher than the sampling variance of the standard estimator $\widehat{\delta}_m$. Thus, the convenience of the correction should be evaluated in terms of mean squared errors (MSE). Noting that in large samples $E(\widehat{\delta}_m - \delta) \simeq -(1 - \lambda_X)\delta$ and $E(\widehat{\delta}_c - \delta) \simeq 0$, the two estimators have approximately the following mean squared errors:

$$MSE(\widehat{\delta}_m) \simeq Var(\widehat{\delta}_m) + (1 - \lambda_X)^2\delta^2$$
$$MSE(\widehat{\delta}_c) \simeq \tfrac{1}{\lambda^2_X} Var(\widehat{\delta}_m) + \tfrac{\delta^2}{\lambda^2_X} Var(\widehat{\lambda}_X).$$

The comparison between the two MSE is not trivial. Anyway, for given values of $\lambda_X$ and $\delta$, the correction is convenient when $Var(\widehat{\lambda}_X)$ is small, i.e. when the number of clusters $J$ is large. The pattern of the MSE with respect to $\delta$ and $J$ is explored via the simulations described in Section 8.2.

Another important quantity in the applications is the level 2 variance $\tau^2_{Y|X^B X^W}$. An approximately unbiased estimator can be derived from expression (15):

$$\widehat{\tau}^2_{Y|X^B X^W,c} = \widehat{\tau}^2_{Y|X^B X^W,m} - (1 - \widehat{\lambda}_X)\widehat{\tau}^2_X \widehat{\delta}_c \ . \tag{19}$$

In principle, it is possible to derive a Taylor approximation of the sampling variance of $\widehat{\tau}^2_{Y|X^B X^W,c}$ but this is not relevant as Wald tests for variance components are not appropriate. The usual test for the nullity of a variance component is a LRT with a halved *p*-value (Snijders and Bosker, 1999), but it is not simple to define an analogous test based on the corrected variance (19). A proper test can be obtained with the structural equation approach presented in Section 5.2.

In unbalanced designs, the value of the reliability $\lambda_X$ changes with the cluster size, so there is no more a unique value of $\lambda_X$. There are two main ways to obtain a pooled value of $\lambda_X$ to be used for correcting the measurement bias: (*i*) compute the reliability using the average cluster size $\lambda_X(\overline{n})$, where $n$ in formula (9) is replaced with the average cluster size $\overline{n} = J^{-1}\sum_{j=1}^{J} n_j$; (*ii*) compute the reliability $\lambda_{X(j)}$ for each cluster and then take the average reliability $\overline{\lambda}_X = J^{-1}\sum_{j=1}^{J}\lambda_{X(j)}$.

In balanced designs, $\overline{\lambda}_X = \lambda_X(\overline{n})$, while in unbalanced designs $\overline{\lambda}_X < \lambda_X(\overline{n})$, and the difference increases with the degree of unbalancedness. The simulations reported in Section 8.3 show that $\overline{\lambda}_X$ is closer to the actual attenuation factor and yields a satisfactory correction in most cases.

## 5.2 The structural equation approach

In general, the bias stemming from covariate measurement error can be amended by fitting a structural equation model that includes a measurement model for the covariate. This is true also for the special case of the measurement error of the sample cluster mean investigated by Lüdtke *et al.* (2008); see also Croon and van Veldhoven (2007).

The structural equation approach consists in the simultaneous estimation of the measurement model (1) for the covariate $X_{ij}$ and the regression model (2) for the response $Y_{ij}$. This strategy cannot be easily implemented in standard software. A notable exception is M*plus* (Muthén and Muthén, 2007), which can fit the model via maximum likelihood. Section 8.2 reports some simulation results for the structural equation estimator, in order to make a comparison with the performance of the reliability-corrected estimator of Section 5.1.

The structural equation approach gives standard errors that account for measurement error, so the inferential procedures are correct, e.g. it is straightforward to perform a likelihood ratio test for the level 2 variance. More importantly, this approach can be easily extended to complex models, such as models with several covariates, random slopes and categorical responses.

Lüdtke *et al.* (2008) argue that the structural equation approach is strictly appropriate when the cluster mean is a reflective measure, while it may yield biased results for formative measures depending on the sampling design. In the next Section we discuss the

nature of the cluster mean, reviewing the concepts of reflective and formative constructs. Subsequently, we deal with the case where the structural equation approach is not appropriate, namely when the cluster mean is a formative measure and the population is made of clusters of finite size. We will show that in such a case, the reliability correction is still appropriate as long as the reliability is properly defined and estimated.

# 6 Nature of the cluster mean

The *Latent Cluster Mean* model defined in Section 2 assumes that the population cluster mean is a latent variable measured through the mean of a random sample and thus it is not observable, no matter how large the sample size is. This assumption underlies the post-estimation correction based on the reliability (Section 5.1) and all the latent variable approaches, such as structural equation models (Section 5.2) and the missing data method of Shin and Raudenbush (2010). In applications where the latent cluster mean assumption is not appropriate, the latent variable approaches are not suitable and they may have a poor performance depending on the sampling design. On the other hand, the correction based on the reliability can be easily adjusted as shown in Section 7.

Table 2 summarizes the characteristics of the cluster mean in some relevant cases: in cases *A* and *B* the latent cluster mean assumption is appropriate, while in case *C* the assumption is not appropriate and thus the measurement error correction requires a modification.

Table 2: Characteristics of the cluster mean in some relevant cases.

| *Case* | *Nature of the cluster mean* | *Cluster size in the population* | *Source of within-cluster variance of* $X_{ij}$ | *Variance of* $\overline{X}_j$ |
|---|---|---|---|---|
| *A* | reflective | (irrelevant) | parallel measurement | $\tau_X^2 + \frac{\sigma_X^2}{n}$ |
| *B* | formative | infinite | random sampling | $\tau_X^2 + \frac{\sigma_X^2}{n}$ |
| *C* | formative | finite | random sampling | $\tau_X^2 + \frac{\sigma_X^2}{n}\frac{N-n}{N-1}$ |

In case *A* of Table 2 the population cluster mean is a latent construct and the level 1 units yield parallel measures of such construct. For example, the school climate may be measured by asking each pupil to evaluate it. A construct of this kind, which is measured (but not defined) by level 1 units, is called *reflective* by Lüdtke *et al.* (2008). Another case of parallel measurement arises when the level 1 units are repeated measures in a longitudinal design. When measuring a latent construct the variability in the measures stems from the instrument and does not disappear even if the whole population is observed.

On the other hand, in cases *B* and *C* of Table 2 the construct is *formative*, i.e. it is defined by aggregating the values of the level 1 units, e.g. the school mean of an intake test score. In cases *B* and *C* the variability in the measures arises only from random sampling. In case *B* the size of the clusters in the population is infinite, i.e. the units within a cluster cannot be exhaustively enumerated. For example, the clusters may be different plants yielding a given product or groups of potential users of a certain service. On the contrary, in case $C$ the clusters have finite size, such as the students of a school.

To continue the example of the school climate, a formative approach consists in first measuring the personal feelings of each pupil and then aggregating them using the school mean. This approach falls in case $C$ since the personal feelings refer to the level 1 units and the cluster mean is a way to form a level 2 construct. The issue of composing group-level constructs from individual-level survey data is discussed at length in van Mierlo *et al.* (2009).

# 7 Measurement error when sampling $X_{ij}$ from clusters of finite size

Let us consider in detail the situation where the cluster mean is a formative measure and the level 1 observations are randomly sampled from a population with clusters of finite size. This is case *C* of Table 2, which differs from the other cases because the variance of the cluster mean $\overline{X}_j$ has a different within-cluster component. This in turn affects the reliability of $\overline{X}_j$ and thus the measurement error bias of the contextual coefficient. For simplicity, we consider the balanced case where all the clusters have the same sample size $n$ and, if they are finite, also the same population size $N$. As shown in equation (8), the variance of the sample cluster mean $Var(\overline{X}_j)$ is the sum of two components: the variance of the population cluster mean $Var(X_j^B)$ and the residual variance $Var(\overline{X}_j^W)$ originated within clusters and due to parallel measurement in case *A* and sampling in cases *B* and *C*. This residual variance is the usual sampling variance of the mean $\sigma_X^2/n$ in cases *A* and *B*, since they both imply that $X_{ij}$ follows model (1) with assumptions (X1)-(X3) of Section 2; on the contrary, in case $C$ (clusters of finite size) the variance of $\overline{X}_j$ originated within clusters is the variance of the sample mean under simple random sampling from a finite population

$$\frac{\sigma_X^2}{n}\frac{N-n}{N-1} \simeq \frac{\sigma_X^2}{n}(1-\frac{n}{N}) \ , \tag{20}$$

where $n/N$ is the within-cluster sampling fraction. Thus, $\sigma_X^2/n$ is a good approximation of the actual variance (20) if the within-cluster sampling fraction is low, but it substantially overestimates the actual variance when large portions of the clusters are sampled. In such cases the reliability of the cluster mean should be modified accordingly, yielding the following *adjusted reliability* for sampling $X_{ij}$ from clusters of finite size:

$$\lambda_X^f = \frac{\tau_X^2}{\tau_X^2 + \frac{\sigma_X^2}{n}\frac{N-n}{N-1}} \ . \tag{21}$$

When the population is made of a finite number of clusters of finite size, $\tau_X^2$ and $\sigma_X^2$ are not model parameters, but they are the between and within variances of $X_{ij}$ in the finite population.

The adjusted reliability $\lambda_X^f$ is an increasing function of the within-cluster sampling fraction $n/N$ taking values in the interval $(\lambda_X, 1]$: if $n/N \to 0$ then $\lambda_X^f \to \lambda_X$, while if $n = N$ then $\lambda_X^f = 1$. Indeed, when the clusters are fully observed ($n = N$) the variance of $\overline{X}_j$ originated within clusters vanishes, so the measurement error of the sample cluster mean is no more an issue.

In order to estimate the adjusted reliability $\lambda_X^f$, it should be noted that the standard estimators of the cluster variance $\tau_X^2$ are biased when sampling from clusters of finite size. Indeed, the cluster variance is estimated by subtracting the spurious variance due to sampling from the variance of the observed cluster means. This fact is true for ML, REML and ANOVA estimators and it is explicit in the ANOVA formulae

$$\begin{aligned} \widehat{\sigma}_X^2 &= S_{X,w}^2 \\ \widehat{\tau}_X^2 &= S_{X,b}^2 - \frac{\widehat{\sigma}_X^2}{n} \ , \end{aligned} \tag{22}$$

where $S_{X,w}^2$ is the sample-within variance, while $S_{X,b}^2$ is the sample-between variance (Snijders and Bosker, 1999). In the case of clusters of finite size, the estimator of the within variance $\widehat{\sigma}_X^2$ is still unbiased, while the estimator of the between variance $\widehat{\tau}_X^2$ is downward biased, since the spurious variance $\widehat{\sigma}_X^2/n$ is computed under the assumption of random sampling from clusters of infinite size. In order to obtain an unbiased estimator, the spurious variance should be modified as follows:

$$\widehat{\tau}_{X,f}^2 = S_{X,b}^2 - \frac{\widehat{\sigma}_X^2}{n}\frac{N-n}{N-1} \ . \tag{23}$$

Therefore, the adjusted reliability $\lambda_X^f$ should be estimated by plugging in expression (21) the standard ANOVA estimate $\widehat{\sigma}_X^2$ (22) and the modified ANOVA estimate $\widehat{\tau}_{X,f}^2$ (23). When the sampling design is unbalanced, similarly to the solution proposed in Section 5.1, one should compute the adjusted reliability for each cluster and then correct the estimates using the average adjusted reliability.

In summary, when the cluster mean is a formative measure and the level 1 observations are randomly sampled from finite-size clusters it is not appropriate to model the population cluster mean as a latent variable, which is the core assumption of the structural equation approach. In practice, the inappropriateness of the latent cluster mean assumption affects the performance of the estimators, but the severity of the bias is strongly related to the within-cluster sampling fraction. At one extreme, when the within-cluster sampling fraction is one, namely all the units of the clusters are sampled, the measurement error vanishes and the model parameters are unbiasedly estimated using the *Sample Cluster Mean* model. At the other extreme, when the within-cluster sampling fraction is close to zero, the *Sample Cluster Mean* model yields biased estimates; however, the situation is well approximated by the latent cluster mean assumption and thus the structural equation approach offers a satisfactory correction. Intermediate situations are more challenging. In fact, when the within-cluster sampling fraction is far from one but not negligible (say more than 5%), the *Sample Cluster Mean* model yields biased estimates and the structural equation approach has a poor performance (Lüdtke *et al.*, 2008); indeed, the structural equation approach overestimates the contextual coefficient. In the same situations, also the correction based on the standard reliability (9) overestimates the contextual coefficient. However, the use of the reliability adjusted for finite-size clusters (21) solves the problem as shown by the simulation study (see Table 8 later on).

Table 3: Estimates from the *Raw Covariate* and *Sample Cluster Mean* models: Monte Carlo mean for $J = 200$ clusters of size $n = 10$ and varying $\delta$ (data generated by a *Latent Cluster Mean* model with $\tau^2_X = 0.2$, $\sigma^2_X = 1$, $\alpha = 0$, $\beta_W = 1$, $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$).

| $\delta$ | *Raw Covariate* model | | | | *Sample Cluster Mean* model | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $(\beta_B - \beta_W)$ | $\eta$ | $\beta_W$ | $\tau^2_{Y\|X}$ | $\sigma^2_{Y\|X}$ | $\alpha$ | $\beta_W$ | $\delta$ | $\tau^2_{Y\|X^B X^W}$ | $\sigma^2_{Y\|X^B X^W}$ |
| -1.50 | -1.48 | 0.98 | 1.45 | 1.00 | -0.50 | 1.00 | -1.01 | 1.16 | 1.00 |
| -1.00 | -0.99 | 0.98 | 1.19 | 1.00 | -0.35 | 1.00 | -0.66 | 1.06 | 1.00 |
| -0.50 | -0.49 | 0.99 | 1.05 | 1.00 | -0.16 | 1.00 | -0.34 | 1.02 | 1.00 |
| -0.25 | -0.25 | 1.00 | 1.01 | 1.00 | -0.09 | 1.00 | -0.17 | 1.00 | 1.00 |
| 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 1.00 |
| 0.25 | 0.25 | 1.01 | 1.02 | 1.00 | 0.08 | 1.00 | 0.17 | 1.01 | 1.00 |
| 0.50 | 0.50 | 1.01 | 1.05 | 1.00 | 0.17 | 1.00 | 0.34 | 1.02 | 1.00 |
| 1.00 | 0.99 | 1.02 | 1.20 | 1.00 | 0.34 | 1.00 | 0.67 | 1.07 | 1.00 |
| 1.50 | 1.48 | 1.02 | 1.45 | 1.00 | 0.49 | 1.00 | 1.00 | 1.16 | 1.00 |

# 8 Simulation study

We perform a Monte Carlo study in order to assess the bias on the slopes and on the residual variances and to evaluate the finite sample properties of the estimators. The data are generated by the *Latent Cluster Mean* model defined by equations (1) and (3), while the fitted models are the *Raw Covariate* model (4) and the *Sample Cluster Mean* model (11). The estimator is REML in both cases.

The simulation study comprises several experiments with 1000 independent replications each. The experiments are variations on the following scenario: (*i*) the hierarchical structure is balanced with $J = 200$ clusters of $n = 10$ observations each (2000 observations overall); (*ii*) the values of the covariate $X_{ij}$ are drawn from model (1) as the sum of two independent normal variates with $\mu_X = 1$, $\tau^2_X = 0.2$ and $\sigma^2_X = 1$, implying a reliability $\lambda_X = 2/3$; (*iii*) the values of the response $Y_{ij}$ are drawn from model (3) with $\alpha = 0$, $\beta_W = 1$, $\delta = 1$, normal level 1 and 2 errors with zero means and $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$. In the first part of the simulation study (Tables 3 to 5) the contextual coefficient $\delta$ takes several values in the interval $[-1.5, +1.5]$, while in the second part it is fixed at $\delta = 1$. Note that $\beta_B$ is determined by the relationship $\beta_B = \beta_W + \delta$, for example $\delta = 1$ implies $\beta_B = 2$.

## 8.1 Comparing the *Raw Covariate* and *Sample Cluster Mean* models

Table 3 reports the Monte Carlo means of the REML estimates obtained from the *Raw Covariate* model (4) and *Sample Cluster Mean* model (11). In both models all the parameters are unbiasedly estimated when the contextual coefficient $\delta = \beta_B - \beta_W$ is zero, so in the following we comment only the cases where $\delta \neq 0$.

In the *Raw Covariate* model, $\beta_W$ is estimated with a bias having the same sign as $\delta$ and increasing with the absolute value of $\delta$. Both level 1 and level 2 variances are inflated, according to formulae (6) and (7) of Section 3. Since the bias of $\beta_W$ is small, the bias of the level 1 variance is so tiny that it does not come out in Table 3. The level 2 variance is

Table 4: Estimates from the *Raw Covariate* and *Sample Cluster Mean* models: Monte Carlo mean for $J = 1000$ clusters of size $n = 2$ and varying $\delta$ (data generated by a *Latent Cluster Mean* model with $\tau^2_X = \sigma^2_X = 1$, $\alpha = 0$, $\beta_W = 1$, $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$).

| $\delta$ | *Raw Covariate* model | | | | *Sample Cluster Mean* model | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $(\beta_B - \beta_W)$ | $\eta$ | $\beta_W$ | $\tau^2_{Y\|X}$ | $\sigma^2_{Y\|X}$ | $\alpha$ | $\beta_W$ | $\delta$ | $\tau^2_{Y\|X^B X^W}$ | $\sigma^2_{Y\|X^B X^W}$ |
| -1.50 | -1.12 | 0.62 | 2.26 | 1.14 | -0.50 | 1.00 | -1.00 | 1.75 | 1.00 |
| -1.00 | -0.70 | 0.70 | 1.49 | 1.09 | -0.33 | 1.00 | -0.67 | 1.33 | 1.00 |
| -0.50 | -0.34 | 0.84 | 1.12 | 1.03 | -0.17 | 1.00 | -0.33 | 1.08 | 1.00 |
| -0.25 | -0.17 | 0.92 | 1.03 | 1.01 | -0.08 | 1.00 | -0.17 | 1.02 | 1.00 |
| 0.00 | 0.00 | 1.00 | 1.00 | 1.00 | 0.00 | 1.00 | 0.00 | 1.00 | 1.00 |
| 0.25 | 0.17 | 1.08 | 1.03 | 1.01 | 0.08 | 1.00 | 0.17 | 1.02 | 1.00 |
| 0.50 | 0.34 | 1.16 | 1.12 | 1.03 | 0.16 | 1.00 | 0.33 | 1.09 | 1.00 |
| 1.00 | 0.70 | 1.30 | 1.50 | 1.09 | 0.33 | 1.00 | 0.67 | 1.34 | 1.00 |
| 1.50 | 1.13 | 1.37 | 2.28 | 1.14 | 0.50 | 1.00 | 1.00 | 1.76 | 1.00 |

inflated to a greater extent, so the ICC is overestimated.

As discussed in Section 5, the *Sample Cluster Mean* model yields an unbiased estimator of the within slope $\beta_W$ and a biased estimator of the contextual coefficient $\delta$: according to formula (13), the estimate of $\delta$ is attenuated by the reliability $\lambda_X = 2/3$. The level 2 variance is inflated and depends on $\delta$ as shown by formula (15). On the contrary, the level 1 variance is immune from bias, so the ICC is overestimated.

In the simulations reported in Table 3 the reliability of the covariate is $\lambda_X = 2/3$ and thus the attenuation of $\delta$ is about $2/3$. It is worth to note that any configuration $(n, \tau^2_X, \sigma^2_X)$ with the same value of $\lambda_X$ yields the same attenuation of $\delta$, but the pattern is different for the inflation of the level 2 variance, which also depends on the cluster variance of the covariate $\tau^2_X$ as shown by expression (15). To highlight this point, we replicate the simulation for an alternative design with the same reliability $\lambda_X = 2/3$ but a smaller cluster size $n = 2$ and a larger cluster variance of the covariate $\tau^2_X = 1$ (the number of clusters is set to $J = 1000$ in order to maintain the total sample size $nJ = 2000$). This design may arise in a panel study with two waves or a cross-section study with two units per cluster, e.g. a study on eyes or twins. The simulation results reported in Table 4 confirm the theoretical findings for the *Sample Cluster Mean* model.

Moreover, the entries of Table 4 allows us to point out some interesting properties of the *Raw Covariate* model. Firstly, the small cluster size $n = 2$ entails a substantial bias on $\beta_W$ (recall from Section 3 that the magnitude of the bias on $\beta_W$ is a decreasing function of the cluster size $n$). Secondly, the residual variances are inflated according to the formulae (6) and (7) of Section 3. However, if the target level 2 variance is not $\tau^2_{Y|X^B X^W}$ but $\tau^2_{Y|X}$ defined in (5), then the bias is downward. For example, when $\delta = 1$ the MC mean of the estimated level 2 variance is 1.50, compared to $\tau^2_{Y|X^B X^W} = 1$ and $\tau^2_{Y|X} = 2$. The existence of two meaningful level 2 variances such as $\tau^2_{Y|X^B X^W}$ and $\tau^2_{Y|X}$ is a source of ambiguity: for example, when Kim and Frees (2007) state that a consequence of endogeneity is a severe underestimation of the level 2 variance, they implicitly refer to $\tau^2_{Y|X}$.

Table 5: Reliability-corrected estimation of the contextual effect: Monte Carlo mean, s.e. and MSE of $\widehat{\delta}_m$ and $\widehat{\delta}_c$ for $J = 200$ clusters of size $n = 10$ and varying $\delta$ (data generated by a *Latent Cluster Mean* model with $\tau^2_X = 0.2$, $\sigma^2_X = 1$, $\alpha = 0$, $\beta_W = 1$, $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$).

| $\delta$ | $\widehat{\delta}_m$ | | $\widehat{\delta}_c$ | | | MSE | |
|---|---|---|---|---|---|---|---|
| $(\beta_B - \beta_W)$ | MC mean | MC s.e. | MC mean | MC s.e. | s.e.($\widehat{\delta}_c$)† | $\widehat{\delta}_m$ | $\widehat{\delta}_c$ |
| -1.50 | -0.995 | 0.152 | -1.510 | 0.251 | 0.239 | 0.2784 | 0.0631 |
| -1.00 | -0.669 | 0.145 | -1.014 | 0.229 | 0.223 | 0.1306 | 0.0527 |
| -0.50 | -0.337 | 0.139 | -0.510 | 0.213 | 0.212 | 0.0458 | 0.0455 |
| -0.25 | -0.168 | 0.138 | -0.256 | 0.214 | 0.212 | 0.0259 | 0.0457 |
| 0.00 | -0.003 | 0.139 | -0.005 | 0.213 | 0.210 | 0.0194 | 0.0452 |
| 0.25 | 0.172 | 0.141 | 0.262 | 0.216 | 0.211 | 0.0258 | 0.0468 |
| 0.50 | 0.332 | 0.137 | 0.501 | 0.209 | 0.212 | 0.0471 | 0.0437 |
| 1.00 | 0.667 | 0.143 | 1.010 | 0.226 | 0.224 | 0.1312 | 0.0512 |
| 1.50 | 1.003 | 0.143 | 1.520 | 0.239 | 0.239 | 0.2680 | 0.0576 |

† MC mean of the standard errors calculated by formula (18).

## 8.2 Performances of reliability-corrected and structural equation estimators

A corrected estimator of the contextual effect $\delta$, denoted with $\widehat{\delta}_c$, has been defined in equation (16) exploiting the fact that the attenuation of $\delta$ in the *Sample Cluster Mean* model equals the reliability of the covariate. In Section 5.1 we have shown that the reliability-corrected estimator is approximately unbiased, but its sampling variance is larger than that of the biased estimator $\widehat{\delta}_m$. Therefore, it is of interest to assess if the correction is convenient in terms of MSE. Table 5 reports the Monte Carlo means, standard errors and MSE of $\widehat{\delta}_m$ and $\widehat{\delta}_c$ from the *Sample Cluster Mean* model, using the same model parameters and data structure as in Table 3. In addition, Table 5 reports the Monte Carlo mean of the standard errors of $\widehat{\delta}_c$ calculated by means of formula (18), showing that the approximation performs well.

Both $MSE(\widehat{\delta}_c)$ and $MSE(\widehat{\delta}_m)$ increase with the absolute value of $\delta$, but $MSE(\widehat{\delta}_c)$ grows at a much lower rate. $MSE(\widehat{\delta}_c)$ is lower than $MSE(\widehat{\delta}_m)$ for values of $|\delta|$ greater than 0.5, suggesting that the proposed correction is worthwhile in many situations. The minimum value of $\delta$ for which the correction is convenient decreases as the number of clusters $J$ increases. For example, a simulation not reported here shows that with the design of Table 4, where $J = 1000$, the correction is worthwhile even for $|\delta| = 0.25$.

The performance of the reliability-corrected estimator $\widehat{\delta}_c$ deteriorates as the number of clusters $J$ diminishes. Table 6 presents the results in the case $\delta = 1$, for varying number of clusters $J$, while the cluster size is kept constant at $n = 10$ and the reliability is thus $\lambda_X = 2/3$.

When the number of clusters is small, say $J \leq 30$, there is a non negligible proportion of samples yielding a low estimated cluster variance $\widehat{\tau}^2_X$ of the covariate and thus a low estimated reliability $\widehat{\lambda}_X$ (see the column labelled $\%trunc$ in Table 6). This is a problem, since a small $\widehat{\lambda}_X$ gives a large $\widehat{\delta}_c$, with the consequence that the MC mean of $\widehat{\delta}_c$ is substantially higher than the true $\delta$ when $J < 50$.

Table 6: Alternative estimators of the contextual coefficient: Monte Carlo mean and MSE for $n = 10$ and varying $J$ (data generated by a *Latent Cluster Mean* model with $\tau^2_X = 0.2$, $\sigma^2_X = 1$, $\alpha = 0$, $\beta_W = 1$, $\delta = 1$,$\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$).

| | $\widehat{\delta}_m$ | | $\widehat{\delta}_c$ | | $\widehat{\delta}_{c,trunc}$ | | | $\widehat{\delta}_s$ | |
|---|---|---|---|---|---|---|---|---|---|
| $J$ | mean | MSE | mean | MSE | mean | MSE | % trunc | mean | MSE |
| 20 | 0.676 | 0.332 | 1.187 | 1.749 | 1.091 | 0.702 | 18.5 | 1.135 | 1.071 |
| 30 | 0.687 | 0.248 | 1.152 | 0.731 | 1.101 | 0.465 | 13.0 | 1.083 | 0.525 |
| 50 | 0.674 | 0.182 | 1.060 | 0.236 | 1.052 | 0.224 | 4.8 | 1.051 | 0.265 |
| 75 | 0.667 | 0.163 | 1.038 | 0.147 | 1.036 | 0.144 | 2.0 | 1.021 | 0.152 |
| 100 | 0.662 | 0.161 | 1.013 | 0.115 | 1.012 | 0.114 | 1.1 | 1.008 | 0.102 |
| 200 | 0.667 | 0.131 | 1.010 | 0.051 | 1.010 | 0.051 | 0.0 | 1.001 | 0.049 |

To avoid the overestimation of $\delta$ one could use a truncated version of $\widehat{\lambda}_X$. For example, the estimator $\widehat{\delta}_{c,trunc}$ in Table 6 is defined as $\widehat{\delta}_{c,trunc} = \widehat{\delta}_c$ if $\widehat{\lambda}_X > 0.5$ and $\widehat{\delta}_{c,trunc} = \widehat{\delta}_m/0.5$ if $\widehat{\lambda}_X \leq 0.5$. Indeed, the $\widehat{\delta}_{c,trunc}$ estimator is approximately unbiased and considerably reduces the MSE when the number of clusters $J$ is small.

Table 6 reports also the results for the estimator of the contextual coefficient $\widehat{\delta}_s$ obtained with the structural equation approach outlined in Section 5.2. The structural equation model is fitted by means of the ML estimator implemented in M*plus* (Muthén and Muthén, 2007). A detailed simulation study on the properties of the structural estimator is carried out by Lüdtke *et al.* (2008).

The performances of reliability corrected and structural equation estimators are similar in both mean and MSE, except in designs with few clusters ($J \leq 30$), where $\widehat{\delta}_s$ is better than $\widehat{\delta}_c$. However, in such cases both estimators have a poor performance since they are upward biased and have an high MSE.

In summary, when the number of clusters is small ($J \leq 30$), the utility of the reliability-based correction is doubtful, since the increase of the MSE, due to the large sample variance of $\widehat{\lambda}_X$, is not compensated by the bias reduction. In such cases, as pointed out by the good performance of the truncated version $\widehat{\delta}_{c,trunc}$, it could be worthwhile to correct the contextual coefficient using a more reliable value of the reliability of the covariate obtained from external sources or previous studies.

## 8.3 Unbalanced case

To evaluate how the measurement error correction based on the reliability of $X_{ij}$ works in unbalanced cases, we perform some simulations with varying cluster sizes $n_j$. In particular we consider a balanced design with $J = 200$ and $n = 10$ and three unbalanced designs with the same average cluster size, i.e. $\overline{n} = 10$. Table 7 reports the Monte Carlo means of the estimates of the contextual effect obtained with the *Sample Cluster Mean* model. Moreover, Table 7 shows the estimated reliabilities $\lambda_X(\overline{n})$ and $\overline{\lambda}_X$ defined in Section 5.1, and the corresponding corrected estimators of $\delta$.

The attenuation of the contextual coefficient due to measurement error increases with the degree of unbalancedness, while the reliability at the average cluster size $\lambda_X(\overline{n})$ is obviously constant. On the contrary, the average reliability $\overline{\lambda}_X$ decreases with the degree of unbalancedness and it is close to the true attenuation factor, except in the last case.

Table 7: Reliability-corrected estimation of the contextual effect in the unbalanced case: Monte Carlo mean of estimators of the reliability and the contextual coefficient for $J = 200$ and $\overline{n} = 10$ (data generated by a *Latent Cluster Mean* model with $\tau_X^2 = 0.2$, $\sigma_X^2 = 1$, $\alpha = 0$, $\beta_W = 1$, $\delta = 1$, $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$).

| Cluster size $n_j$ | | | $\widehat{\lambda}_X$ with $\overline{n}$ | | average $\widehat{\lambda}_X$ | |
|---|---|---|---|---|---|---|
| $j$=1,$\cdots$,100 | $j$=101,$\cdots$,200 | $\widehat{\delta}_m$ | $\widehat{\lambda}_X(\bar{n})$ | $\widehat{\delta}_c$ | $\widehat{\overline{\lambda}}_X$ | $\widehat{\delta}_c$ |
| 10 | 10 | 0.66 | 0.66 | 1.00 | 0.66 | 1.00 |
| 7 | 13 | 0.65 | 0.66 | 0.98 | 0.65 | 1.00 |
| 4 | 16 | 0.57 | 0.66 | 0.86 | 0.60 | 0.95 |
| 1 | 19 | 0.34 | 0.66 | 0.51 | 0.48 | 0.71 |

Table 8: Reliability-corrected estimation of the contextual effect when sampling $X_{ij}$ from clusters of finite size: Monte Carlo mean and MSE of $\widehat{\delta}_m$, $\widehat{\delta}_c$ and $\widehat{\delta}_c^f$ when sampling $n = 10$ values from $J = 200$ clusters of varying size $N$ (true values: $\tau_X^2 = 0.2$, $\sigma_X^2 = 1$, $\alpha = 0$, $\beta_W = 1$, $\delta = 1$, $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$)

| | | | MC Mean | | | MSE | | |
|---|---|---|---|---|---|---|---|---|
| $N$ | $n/N$ | $\lambda_X^f$ | $\widehat{\delta}_m$ | $\widehat{\delta}_c$ | $\widehat{\delta}_c^f$ | $\widehat{\delta}_m$ | $\widehat{\delta}_c$ | $\widehat{\delta}_c^f$ |
| 10 | 1.00 | 1.000 | 1.003 | 2.247 | 1.003 | 0.0265 | 1.6950 | 0.0265 |
| 20 | 0.50 | 0.792 | 0.804 | 1.361 | 1.031 | 0.0595 | 0.2062 | 0.0376 |
| 40 | 0.25 | 0.722 | 0.725 | 1.163 | 1.016 | 0.0965 | 0.0896 | 0.0441 |
| 100 | 0.10 | 0.688 | 0.689 | 1.058 | 1.009 | 0.1153 | 0.0554 | 0.0434 |
| 200 | 0.05 | 0.677 | 0.678 | 1.032 | 1.010 | 0.1231 | 0.0511 | 0.0475 |
| 1000 | 0.01 | 0.669 | 0.669 | 1.030 | 1.003 | 0.1297 | 0.0477 | 0.0492 |

To summarize, the average reliability $\overline{\lambda}_X$ tends to under-correct the estimate of $\delta$, but the correction is satisfactory in most cases.

## 8.4 Sampling $X_{ij}$ from clusters of finite size

Let us now evaluate the performance of the estimator of the contextual effect based on the adjusted reliability (21) when the cluster mean is a formative measure and the level 1 observations are randomly sampled from a population with clusters of finite size (see Section 7).

To this end, we generate six finite populations with $J = 200$ clusters and varying cluster size $N \in \{10, 20, 40, 100, 200, 1000\}$. For each finite population, the values of the covariate $X_{ij}$ are generated such as the level 2 variance is $\tau_X^2 = 0.2$ and the level 1 variance is $\sigma_X^2 = 1$. At each replication of the MC simulation, a sample of 2000 observations is drawn as follows: first, we sample without replacement $n = 10$ values of the covariate $X_{ij}$ from every cluster of the finite population; next, we generate the corresponding responses $Y_{ij}$ according to the random intercept model (3).

Table 8 reports the results for the uncorrected estimator of the contextual coefficient $\widehat{\delta}_m$ and the corrected estimator $\widehat{\delta}_c^f = \widehat{\delta}_m / \widehat{\lambda}_X^f$, where $\widehat{\lambda}_X^f$ is the estimate of the reliability (21) for simple random sampling from clusters of finite size, using the ANOVA estimates (22) and (23) defined in Section 7.

The first row of Table 8 reports the results when the within-cluster sampling fraction is 1, i.e. the values $X_{ij}$ are not sampled and thus the measurement error is not an issue. On the contrary, the last row refers to a tiny within-cluster sampling fraction ($n/N = 0.01$), so $\lambda_X^f \simeq \lambda_X = 2/3$ and thus the attenuation due to measurement error is very close to the case of sampling $X_{ij}$ from clusters of infinite size (see Table 5 at the row $\delta = 1$). In intermediate cases with $n/N \in \{0.5, 0.25, 0.10, 0.05\}$, the simulation results show that the adjusted reliability $\lambda_X^f$ is a good approximation of the attenuation of the contextual coefficient due to measurement error, thus the corrected estimator $\widehat{\delta}_c^f$ has a good performance. On the contrary, the estimate $\widehat{\delta}_c$ based on the standard reliability $\lambda_X$ yields an overcorrection that is remarkable for within-cluster sampling fractions of $0.25$ or more.

As the within-cluster sampling fraction becomes larger, $MSE(\widehat{\delta}_c)$ increases, while both $MSE(\widehat{\delta}_m)$ and $MSE(\widehat{\delta}_c^f)$ decrease. Except for extreme sampling fractions, $MSE(\widehat{\delta}_c^f)$ is the smallest one and thus the reliability $\lambda_X^f$ adjusted for finite-size clusters proves to be an effective way to correct the contextual coefficient.

# 9 Implications for effectiveness evaluation

A relevant use of the *Latent Cluster Mean* model (3) is for the assessment of the relative effectiveness of a set of institutions, such as hospitals or schools (Grilli and Rampichini, 2009). To illustrate the point, we focus on the school effects framework of Raudenbush and Willms (1995), where the level 2 units are schools and the level 1 units are pupils. In the basic value-added specification, $Y_{ij}$ is a measure of pupil's final attainment and $X_{ij}$ is a measure of prior attainment. Thus $X_j^B$ is the school component of prior attainment and its slope $\delta$ is the contextual coefficient, whose estimate is usually positive in the educational setting.

The total effect of school $j$, called *Type A* effect, is $A_j = \delta X_j^B + u_j$, which is the sum of the effects of context $\delta X_j^B$ and school practice $u_j$. The effect of the school practice is called *Type B* effect: $B_j = u_j$. Therefore, $\tau^2_{Y|X^B X^W}$ is the variance of *Type B* effects, while $\tau^2_{Y|X} = \delta^2 \tau^2_X + \tau^2_{Y|X^B X^W}$, defined in (5), is the variance of *Type A* effects. Students and their families are interested in *Type A* effects, while evaluation agencies and school staffs are interested in *Type B* effects.

In the applications the unobservable school component of prior attainment $X_j^B$ is replaced with the sample cluster mean $\overline{X}_j$, so the *Sample Cluster Mean* model (11) is adopted. The standard estimators of *Type A* and *Type B* effects are:

$$\widehat{A}_j = \overline{Y}_j - \widehat{\alpha} - \widehat{\beta}_W \overline{X}_j \tag{24}$$

$$\widehat{B}_j = \overline{Y}_j - \widehat{\alpha} - \widehat{\beta}_B \overline{X}_j = \widehat{A}_j - \widehat{\delta}\overline{X}_j. \tag{25}$$

The measurement error involved in using $\overline{X}_j$ instead of $X_j^B$ is usually ignored in the school evaluation framework, since the reliability $\lambda_X$ is often over 0.90 (Raudenbush and Willms, 1995). However, in order to deal with cases where the reliability $\lambda_X$ is far from one, it is essential to examine the consequences of the measurement error on the assessment of *Type A* and *Type B* effectiveness.

First note that the measurement error concerns $\beta_B$ but not $\beta_W$, so the estimator (25) of the *Type B* effects is biased, while the *Type A* effects are unbiasedly estimated up to a constant. Indeed the constant $\alpha$ is estimated with bias, as shown in (14), but this is irrelevant for comparison purposes.

As for the variance of the effects, the estimable level 2 variance from the *Sample Cluster Mean* model is $\tau^2_{Y|X^B X^W,m}$ defined in (15), which is higher than the variance of *Type B* effects,

$$\tau^2_{Y|X^B X^W,m} - \tau^2_{Y|X^B X^W} = (1-\lambda_X)\tau^2_X\delta^2 = \left(\frac{1}{\lambda^2_X} - \frac{1}{\lambda_X}\right)\tau^2_X\delta^2_m \tag{26}$$

and lower than the variance of *Type A* effects,

$$\tau^2_{Y|X} - \tau^2_{Y|X^B X^W,m} = \lambda_X\tau^2_X\delta^2 = \frac{1}{\lambda_X}\tau^2_X\delta^2_m \ . \tag{27}$$

Therefore, the variances of *Type B* and *Type A* effects can be estimated by correcting the level 2 variance from the *Sample Cluster Mean* model using (26) and (27), respectively. Note that for increasing cluster size $n$ the reliability $\lambda_X$ tends to 1, so the difference (26) vanishes, while the difference (27) tends to $\tau^2_X\delta^2$.

Raudenbush and Willms (1995) and Rettore and Martini (2001) tackle the problem of estimating the variance of *Type A* effects from the *Raw Covariate* model in presence of level 2 endogeneity. To this end, they both suggest to fit the *Sample Cluster Mean* model and correct the estimated level 2 variance by adding the term $\delta^2_m Var(\overline{X}_j)$, which is taken as an estimate of the term $\delta^2\tau^2_X$ in (5). In both papers, the authors assume that the measurement error is negligible, so no attempt to correct $\delta_m$ is made. Nevertheless, since $Var(\overline{X}_j) = \tau^2_X/\lambda_X$, the proposed correction term turns out to coincide with the correction term (27), derived under an explicit treatment of measurement error. However, ignoring the measurement error entails assuming that the level 2 variance from the *Sample Cluster Mean* model is equal to the variance of *Type B* effects, which is not the case, as shown in expression (26).

# 10 Concluding remarks

In many applications of multilevel analysis the between and within slopes are different, namely there is a contextual effect. In such cases, the omission of the cluster mean from the model equation generates level 2 endogeneity. However, the inclusion of the sample cluster mean yields a model that is still affected by level 2 endogeneity which is due to the measurement error caused by the substitution of the unobservable population cluster mean of the covariate with the observable sample cluster mean. Focusing on the random intercept model with a single covariate, in the paper we studied the effects of the measurement error on the contextual coefficient and also on the variance components, an aspect usually neglected. The attenuation factor of the contextual coefficient is the reliability of the covariate. On the other hand, the level 2 variance is inflated by a quantity that depends on several entities, namely the reliability of the covariate, the cluster variance of the covariate, and the contextual coefficient. Our analysis focused on balanced designs, but we showed that in unbalanced designs the average reliability is a good approximation of the attenuation factor. We also addressed the issue of sampling from clusters of

finite size, showing that the attenuation is substantially weaker when the within-cluster sampling fraction is high.

We suggested a simple procedure that yields unbiased estimates of the parameters of interest. In particular, the correction of the contextual coefficient through the reliability is straightforward and is carried out after fitting the multilevel model, so the task can be easily performed using standard software for multilevel analysis. We derived an approximate formula for the standard error of the corrected contextual coefficient and showed that the correction is worthwhile in terms of MSE for moderate or large values of the contextual coefficient. The correction can be applied even to the estimates obtained by other researchers. Moreover, with good prior information on the ICC of the covariate, the amount of attenuation can be evaluated when planning the sampling design.

An alternative approach for fitting random effects models with endogeneity is based on Instrumental Variable (IV) estimators, proposed by Hausman and Taylor (1981) and extended by Kim and Frees (2007). The key idea is that centering a covariate with respect to the sample cluster mean yields an instrument for amending the effects of level 2 endogeneity. Contrary to standard IV applications, the centered covariate is an internal instrument, namely it is derived without external data. This approach allows to estimate only the within slope, so the measurement error on the contextual coefficient is not an issue. Obviously, the IV method is not useful when the contextual effects of level 1 covariates are of interest. Instead of enhancing the estimators via instrumental variables, we prefer to solve the level 2 endogeneity by expanding the model with the cluster means: beyond the possibility to estimate the contextual effects, in this way the mechanism underlying the endogeneity is made explicit and the parameters have a clear interpretation that facilitates the connection with the theory.

The approach based on the reliability described in this paper is useful to understand the consequences of the measurement error induced by sample cluster means and yields a straightforward and effective correction when the model is simple. In a linear model with several covariates the correction via the reliability is still feasible: the formulas become complex, but they can be derived, e.g. following the lines of Croon and van Veldhoven (2007). In non linear models the reliability approach leads to intractable formulas and it can be useful only as a raw approximation.

In order to deal with measurement error in complex models, more general approaches are preferable. In particular, the missing data method of Shin and Raudenbush (2010) can be used with linear random slope models with several covariates, while the structural equation approach (Lüdtke *et al.*, 2008) can be applied to a wide range of linear and non-linear models. Both approaches, however, assume that the cluster mean is a latent variable and thus they may have a poor performance when sampling from clusters of finite size. In this case the reliability correction can be easily adjusted as shown in Section 7, while further research is needed to develop general methods.

# Acknowledgements

This research was partially supported by MIUR funds PRIN 2008 (2008WKHJPK_003) "Latent class and multilevel models: methodology and applications in evaluation and causal inference".

# References


Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition, Boca Raton, FL: Chapman & Hall/CRC.

Casella, G. and Berger, R.L. (2001). *Statistical Inference, 2nd Edition*. Pacific Grove, CA: Duxbury Press.

Croon, M.A. and van Veldhoven, M.J.P.M (2007). Predicting Group-Level Outcome Variables From Variables Measured at the Individual Level: A Latent Variable Multilevel Model. *Psychological Methods*, 12, 45–57.

Ebbes, P., Bockenholt, U. and Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, 58, 161–178.

Ferrão, M. E. and Goldstein, H. (2009). Adjusting for measurement error in the value added model: evidence from Portugal. *Quality and Quantity*, 43, 951–963.

Fielding, A. (2004). The Role of the Hausman Test and whether Higher Level Effects should be treated as Random or Fixed. *Multilevel Modelling Newsletter*, 16, 3–9.

Grilli, L. and Rampichini, C. (2009). Multilevel models for the evaluation of educational institutions: a review. In: Monari, P.; Bini, M.; Piccolo, D.; Salmaso, L. (Eds.) *Statistical Methods for the Evaluation of Educational Services and Quality of Products*, pp 61-80. Heidelberg: Physica-Verlag.

Hausman, J.A. and Taylor, W.E. (1981). Panel data and unobservable individual effects. *Econometrica*, 49, 1377–1398.

Hutchison, D. (2004). The effect of measurement errors on apparent group-level effects in educational progress. *Quality and Quantity*, 38, 407–424.

Kim, J.S. and Frees, E.W. (2007). Multilevel Modeling with Correlated Effects. *Psychometrika*, 72, 505–533.

Lüdtke O., Marsh H.W., Robitzsch A., Trautwein U., Asparouhov T. and Muthn B. (2008). The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies. *Psychological Methods*, 13, 203–229.

Lüdtke O., Marsh H.W., Robitzsch A., Trautwein U., Asparouhov T., Muthn B. and Nagengast B. (2009). Doubly-Latent models of School Contextual Effects: Integrating Multilevel and Structural Equation Approaches to Control Measurement and Sampling Error. *Multivariate Behavioral Research*, 44, 764–802.

Mundlak, Y. (1978). On the pooling of time series and cross-sectional data. *Econometrica*, 46, 69–86.

Muthén, L.K. and Muthén, B.O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Neuhaus, J. M. and Kalbfleish, J. D. (1998). Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, 54,638–645.

Raudenbush, S.W. and Willms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307–335.

Rettore, E. and Martini, A. (2001). Constructing league tables of service providers when the performance of the provider is correlated to the characteristics of the clients. In *Processi e metodi statistici di valutazione*, Proceedings of the Conference of the Italian Statistical Society, Roma.

Shin, Y. and Raudenbush, S.W. (2010). A Latent Cluster-Mean Approach to the Contextual Effects Model with Missing Data. *Journal of Educational and Behavioral Statistics*, 35, 26–53.

Snijders, T.A.B. and Berkhof, J. (2008). Diagnostic Checks for Multilevel Models. In J. de Leeuw and E. Meijer (Editors), *Handbook of Multilevel Analysis*. New York: Springer.

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

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

Woodhouse, G., Yang, M., Goldstein, H. and Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society A*, 159, 201–212.

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

van Mierlo, H., Vermunt, J.K. and Rutte, C.G. (2009). Composing group-level constructs from individual-level survey data. *Organizational Research Methods*, 12, 368–392.