Correcting for sampling error in between-cluster effects: An empirical Bayes cluster-mean approach with finite population corrections

Mark H. C. Lai[1], Yichi Zhang[1], and & Feng Ji[2, 3]

[1] Department of Psychology, University of Southern California

[2] Division of Biostatistics, University of California, Berkeley

[3] Department of Applied Psychology and Human Development, University of Toronto

Author Note

Mark H. C. Lai https://orcid.org/0000-0002-9196-7406

Yichi Zhang https://orcid.org/0000-0002-4112-2106

Feng Ji https://orcid.org/0000-0002-2051-5453

Feng Ji is now at the University of Toronto.

Full simulation results and additional online materials are openly available at https://github.com/marklhc/ebm-supp

The Version of Record of this manuscript has been published and is available in *Multivariate Behavioral Research* February 13, 2024 https://www.tandfonline.com/doi/abs/10.1080/00273171.2024.2307034

Correspondence concerning this article should be addressed to Mark H. C. Lai, Department of Psychology, University of Southern California, 3620 South McClintock Ave., Los Angeles, CA 90089-1061. E-mail: hokchiol@usc.edu

Abstract

With clustered data, such as where students are nested within schools or employees are nested within organizations, it is often of interest to estimate and compare associations among variables separately for each level. While researchers routinely estimate between-cluster effects using the sample cluster means of a predictor, previous research has shown that such practice leads to biased estimates of coefficients at the between level, and recent research has recommended the use of latent cluster means with the multilevel structural equation modeling framework. However, the latent cluster mean approach may not always be the best choice as it (a) relies on the assumption that the population cluster sizes are close to infinite, (b) requires a relatively large number of clusters, and (c) is currently only implemented in specialized software such as Mplus. In this paper, we show how using empirical Bayes estimates of the cluster means can also lead to consistent estimates of between-level coefficients, and illustrate how the empirical Bayes estimate can incorporate finite population corrections when information on population cluster sizes is available. Through a series of Monte Carlo simulation studies, we show that the empirical Bayes cluster-mean approach performs similarly to the latent cluster mean approach for estimating the between-cluster coefficients in most conditions when the infinite-population assumption holds, and applying the finite population correction provides reasonable point and interval estimates when the population is finite. The performance of EBM can be further improved with restricted maximum likelihood estimation and likelihood-based confidence intervals. We also provide an R function that implements the empirical Bayes cluster-mean approach, and illustrate it using data from the classic High School and Beyond Study.

Keywords: Multilevel modeling, contextual effect, centering, empirical Bayes estimates, finite population correction

Word count: 5,052

Correcting for sampling error in between-cluster effects: An empirical Bayes cluster-mean approach with finite population corrections

Multilevel modeling (MLM) is a popular approach to analyzing clustered data in social and behavioral sciences, such as data with students nested within schools or repeated measures nested within participants (Snijders & Bosker, 2012). However, modeling the effect of a within-cluster level predictor is not a trivial task, as only including the raw predictor variable may result in an estimated coefficient that conflates the effects at the within-cluster level and the between-cluster level. A standard approach to disentangle the between and the within effects is to compute the mean value of the within-cluster predictor for each cluster and include this cluster mean variable as a predictor (e.g., Enders & Tofighi, 2007; Kreft et al., 1995). The model should also include either the original within-level predictor, resulting in the so-called contextual model, or the cluster-mean-centered predictor from which the cluster means have been subtracted, resulting in the so-called between-within model.

As shown in Lüdtke et al. (2008), however, using the observed cluster mean variable—the *sample* mean predictor value of each cluster—may result in biases in the parameter estimates. This bias happens when the observed cluster mean is not a perfectly reliable measure of the true cluster mean, and has the most impact when the cluster sample size is small and is only a small fraction of the population cluster size (Asparouhov & Muthén, 2019; Shin & Raudenbush, 2010). For example, if a researcher computes the school-level achievement based on the mean score of five students in the sample, that sample mean likely contains much sampling error and is unreliable, and using this unreliable predictor leads to biased parameter estimation just like classical measurement error (e.g., Cole & Preacher, 2014). To adjust for such bias, Lüdtke et al. (2008) proposed using the latent cluster means, referred to as the *latent-means-as-covariate* (LMC) approach, by modeling the between-level cluster means as a latent variable under the multilevel structural equation modeling (MSEM) framework.

Although the LMC approach can remove the measurement-error-induced bias in estimating between-level coefficients, it has three major limitations. First, as shown in Lüdtke et al. (2008), LMC requires a relatively large sample size (with at least 100 clusters), and it results in less efficient estimates (in terms of root mean squared error [RMSE]) than using the observed means in small samples. Second, as LMC is based on the MSEM framework, it has added complexity in model specification (Hoffman, 2019) and requires specialized software (e.g., Mplus), which may not be familiar to researchers who regularly use MLM software. Third, the LMC approach assumes that the sample units in a cluster are drawn from an infinitely large population cluster; however, in some applications, such an assumption may not hold, like when researchers have surveyed all students in a classroom, in which case Lüdtke et al. has shown that using the observed cluster means results in less bias.

Given the bias produced by the observed cluster-mean approach (CM) and the limitations of LMC, in this paper, we aim to introduce researchers to the less well-known Empirical Bayes cluster-mean (EBM) approach for consistent estimation of between-level effects. While EBM was discussed in Shin and Raudenbush (2010), it has not been systematically evaluated and compared to LMC, to our knowledge. The contribution of the current paper is three-fold. First, we derive a bias-corrected estimator for the random intercept variance based on EBM, in addition to the fixed-effect coefficients. Second, using two simulation studies, we provide empirical evidence on how EBM compares to LMC, including conditions with finite population cluster sizes. Third, as EBM has not been implemented in commonly used software programs for multilevel modeling—a potential reason for its low usage in applied research—we provide an R function that uses EBM for corrected fixed and random effects.[1] The R function also allows researchers to specify the population cluster size(s) when the infinite population assumption in LMC is not tenable, as illustrated later using the classic High School and Beyond Survey data set (Raudenbush & Bryk, 2002).

Model Notations

Let X be a within-level predictor, and μ_X be a random variable of true cluster means. Following Lüdtke et al. (2008), we assume that μ_X is an error-free variable that is likely different from the observed cluster means, $\bar{X}_{\cdot}$. Let $n_{\text{pop},j}$ be the population size of the jth cluster and n_j is the sample cluster size. To the extent that the sample units in a cluster are considered a random

[1] The R function and the supplemental results can be found at https://anonymous.4open.science/r/ebm-supp-B7ED/

sample of all the units in that cluster, the sampling error of $\bar{X}_{.}$ as a representation of μ_X has variance

$$\mathrm{Var}(\bar{X}_{.j} - \mu_{Xj} | \mu_{Xj}) = \frac{\sigma_X^2}{n_j} \mathrm{fpc}_j,$$

where σ_X^2 is the within-cluster variance of X, which is assumed constant across clusters, and

$$\mathrm{fpc}_j = \frac{n_{\mathrm{pop},j} - n_j}{n_{\mathrm{pop},j} - 1}$$

is the finite population correction factor (FPC; e.g., Lai et al., 2018), which approaches one when $n_{\mathrm{pop},j}$ is large relative to n_j. When $\mathrm{fpc}_j = 1$, the measurement error variance becomes σ_X^2 / n_j as discussed in Lüdtke et al. (2008). However, in the case where all units in a cluster are included in the sample such that $n_{\mathrm{pop},j} = n_j$, the sampling error variance is 0, and Lüdtke et al. (2008) showed that one should use CM in this case.

In the general case where X and μ_X relate to an outcome variable Y differently, and the within-cluster slopes between Y and X vary across clusters, we have the following multilevel model:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \mu_{Xj}) + \gamma_{01}\mu_{Xj} + u_{0j} + u_{1j}(X_{ij} - \mu_{Xj}) + e_{ij}, \tag{1}$$

where γ_{00} is the grand intercept, γ_{10} is the average within-cluster slope, γ_{01} is the between-level slope, u_{0j} and u_{1j} are the cluster-specific deviations in the intercept and the slope, and e_{ij} is the within-cluster level error term. We apply the standard assumptions that u_{0j}, u_{1j}, and e_{ij} all have means zero, and that e_{ij} is independent to u_{0j} and u_{1j}. In addition, we assume that the random effects and errors are normally distributed and independent to X and μ_X.

For simplicity, we first consider the case where the sample and population cluster sizes are constant such that $n_j = n$ and $n_{\mathrm{pop},j} = n_{\mathrm{pop}}$ for all js. As shown in Lüdtke et al. (2008) and Grilli and Rampichini (2011), in CM, when the sample cluster mean $\bar{X}_{.}$ is used in place of the unobserved μ_X, the estimator for γ_{10} is still consistent, but the estimator for γ_{01} has a bias of magnitude $(\gamma_{10} - \gamma_{01})(1 - \lambda_X)$, where

$$\lambda_X = \frac{\tau_X^2}{\tau_X^2 + \sigma_X^2 \mathrm{fpc}_2 / n} \tag{2}$$

is the reliability of $\bar{X}_{\cdot}$, with $\tau_X^2 = \text{Var}(\mu_X)$. When n_j and/or $n_{\text{pop},j}$ are not constant, the bias is approximately $(\gamma_{10} - \gamma_{01})(1 - \bar{\lambda}_X)$, where

$$\bar{\lambda}_X = \frac{1}{J}\sum_{j=1}^{J}\frac{\tau_X^2}{\tau_X^2 + \sigma_X^2 \text{fpc}_j / n_j}$$

is the average reliability of the cluster means.

Latent Means as a Covariate

The LMC approach is based on the MSEM framework, available in software such as Mplus and the R package *OpenMx*.[2] For models without random slopes, it estimates parameters of equation (1) directly by treating μ_X as a latent variable and performs latent decomposition of X into the between-cluster and the within-cluster components (Asparouhov & Muthén, 2019). Lüdtke et al. (2008)'s simulation showed that, in terms of bias, LMC yielded unbiased contextual effect estimates (i.e., $\gamma_{10} - \gamma_{01}$) for conditions with sampling fraction (i.e., $n_j / n_{\text{pop},j}$) close to zero, 50 clusters or above, $n_j \geq 15$, and $\text{ICC}_X \geq .10$. However, in terms of efficiency (as measured by RMSE), estimates from LMC were less efficient than those from CM for conditions with 50 clusters (and $\text{ICC}_X \leq .20$). Similarly, Aydin et al. (2016) compared CM and LMC for two-level cluster-randomized trials with a covariate, and found that CM yielded better power and coverage rates for the between-level treatment effect while maintaining good Type I error rates.

Handling models with random slopes is more complex in MSEM. Before version 8.1, Mplus implemented LMC using the so-called "hybrid" method (Asparouhov & Muthén, 2019) with the model

$$Y_{ij} = \gamma_{00} + \gamma_{10}^* X_{ij} + \gamma_{01}\mu_{Xj} + u_{0j} + u_{1j}X_{ij} + e_{ij}, \tag{3}$$

which includes the latent μ_X and the uncentered X in the fixed effects, and the uncentered X for the random slope component. With the uncentered X in the model, γ_{01}^* corresponds to the contextual effect, while the between-cluster effect can be obtained as $\gamma_{10}^* + \gamma_{01}$. As pointed out in Asparouhov and Muthén (2019), this method conflates the level-1 and level-2 coefficients and may lead to biased estimates. More recently, Asparouhov and Muthén (2019) and the Mplus team

[2] Another popular R package for SEM, *lavaan*, currently only supports models without random slopes.

suggested directly estimating the model in equation (1) with μ_X treated as a latent variable, which was the same as the latent cluster-mean approach discussed in Shin and Raudenbush (2010). Bayesian estimation is needed, however, as the model involves the latent product term $u_{1j}\mu_X$. Asparouhov and Muthén (2019) found that Bayesian LCM had negligible bias with 500 clusters, but the coverage of the 95% interval was below the nominal level for some model parameters.

There are several advantages of the MSEM framework compared to standard MLM analysis (e.g., Preacher et al., 2010). First, unlike MLM, which requires the outcome variable to be at level 1, MSEM can incorporate outcome variables at upper levels. Second, while MLM assumes predictors to be perfectly reliable, MSEM incorporates measurement models for error-prone predictors (and outcomes) so that coefficients are adjusted. Third, MSEM is a multivariate technique that allows specifying a path model, such as a mediation model, with multiple outcome variables, whereas standard MLM only allows one outcome variable and requires burdensome steps to specify multivariate models (see e.g., Raudenbush & Bryk, 2002).

On the other hand, one limitation of MSEM, as compared to MLM, is that most MSEM software implementation uses maximum likelihood (ML) estimation, which gives biased estimates of random effect variances when the sample size is small relative to the number of predictors (McCulloch & Searle, 2001). This is in contrast to the ease of using restricted maximum likelihood (REML) estimation in MLM, which is theoretically unbiased with a correctly specified model (Snijders & Bosker, 2012).[3] Many MLM software programs also provide asymmetric profile likelihood CIs and small-sample adjustments (e.g., Kenward & Roger, 1997) that improve the accuracy of estimations and inferences, which may not be available in MSEM software. Although Bayesian MSEM, currently only implemented in Mplus among general-purpose software, can give more numerically stable parameter estimates and fewer estimation convergence problems in small samples (Depaoli & Clifton, 2015; Zitzmann et al., 2016), researchers more familiar with the MLM framework may find it a hurdle switching to MSEM software just to account for the unreliability of observed cluster means. Therefore, in what follows, we introduce an alternative that (a) gives

[3] Cheung (2013) discussed ways to implement REML in the SEM framework using a transformation matrix or a modified fitting function.

between-cluster effect estimates comparable to LMC and (b) can be easily implemented in standard MLM software.

Empirical Bayes Cluster-Mean Method (EBM) With Finite Population Correction

As demonstrated in Shin and Raudenbush (2010), an alternative method that avoids the bias in the between-level coefficient is to include in the model the empirical Bayes (EB) estimates of the cluster means of X, also called the best linear unbiased predictors.[4] When the model predicting Y contains no other between-level covariates, the EB cluster mean can be computed as

$$\hat{\mu}_{Xj}^{\text{EB}} = \hat{\lambda}_{Xj}\bar{X}_{.j} + (1 - \hat{\lambda}_{Xj})\hat{\gamma}_{00X}$$

where $\hat{\gamma}_{00X}$ is the sample grand mean of X. For models assuming normally distributed random effects and errors, the EB estimate discussed in Shin and Raudenbush (2010) can be obtained using standard MLM software, but it does not adjust for finite population cluster sizes. However, finite population correction can be incorporated by defining (Grilli & Rampichini, 2011, equation 32, p. 12)

$$\hat{\lambda}_{Xj} = \frac{\hat{\tau}_X^2}{\hat{\tau}_X^2 + \hat{\sigma}_X^2 \text{fpc}_j / n_j} \tag{4}$$

so that when $\text{fpc}_j \to 0$ or when n_j is large, the EB cluster means will be the same as the observed cluster means.

When the model predicting Y contains between-level covariates $\mathbf{C} = (C_1, C_2, \ldots)$, including cluster means of level-1 covariates other than X, the EB means could be obtained by fitting the multilevel model

$$X_{ij} = \gamma_{00X} + \mathbf{C}\boldsymbol{\gamma}_X + u_{0jX} + e_{ijX}, \tag{5}$$

where $\boldsymbol{\gamma}_X$ is a column vector of fixed effect coefficients of the covariates predicting X. An additional requirement, not discussed in Shin and Raudenbush (2010), is to also include random slope components of level-1 covariates $\mathbf{W} = (W_1, W_2, \ldots)$ to obtain the EB means, if those components

[4] Essentially the same procedure was proposed by Croon and van Veldhoven (2007), but in the context of predicting a between-level outcome.

will appear in the final model predicting Y. The model for obtaining the EB means thus becomes

$$X_{ij} = \gamma_{00X} + \mathbf{C}\boldsymbol{\gamma}_X + u_{0jX} + \sum_{s=1}^{q} u_{sjX} W_j + e_{ijX}. \tag{6}$$

The conditional reliability $\hat{\lambda}_{Xj}$ can be obtained using equation (4) but with the $\hat{\tau}^2_X$ and $\hat{\sigma}^2_X$ estimates from the model in (5), and

$$\hat{\mu}^{\text{EB}}_{Xj} = \hat{\lambda}_{Xj}\bar{X}_{.j} + (1 - \hat{\lambda}_{Xj})(\hat{\gamma}_{00X} + \mathbf{C}\hat{\boldsymbol{\gamma}}_X), \tag{7}$$

with $\hat{\gamma}_{00X}$ and $\hat{\boldsymbol{\gamma}}_X$ being the sample estimates.

One can either use $\hat{\mu}^{\text{EB}}_X$ in combination with X to estimate the contextual and the within-cluster effects, or $\hat{\mu}^{\text{EB}}_X$ in combination with $X - \hat{\mu}^{\text{EB}}_X$ to estimate the between- and within-cluster effects. The former has been demonstrated in a large data set to give very similar fixed effect estimates as LMC by Shin and Raudenbush (2010) without any finite population adjustment. Lüdtke et al. (2008) also conducted a simulation to compare EBM and LMC for estimating the between- and within-cluster effects without any random slopes and covariates, and found the two methods performed similarly in most conditions, but they implemented EBM under the MSEM framework with ML estimation, while we expect EBM using REML and likelihood-based CI will have better small sample performance.

Gottfredson (2019) proposed an alternative correction approach for obtaining point estimates of the between-level coefficients, using the reliability information discussed above. We expect EBM and Gottfredson (2019)'s approach would give similar results for models without other between-level coefficients. On the other hand, EBM is more general as it can also correct for the unreliability of cluster means of other between-level covariates, and automatically provides corrected standard errors and confidence intervals (CIs).

Correcting for Bias in Estimated Random Intercept Variance

Although EBM corrects for the bias in the estimated between-cluster coefficient due to measurement error in the observed cluster means, like CM, it overestimates τ^2_0. The reason is that

$\hat{\mu}_X^{\text{EB}}$, being a shrinkage estimate, has a variance that is systematically smaller than that of μ_X. Indeed, one can show that the naive estimate of τ_0^2 under EBM is the same as that under CM. As shown in the Appendix, a consistent estimate of τ_0^2 can be obtained as

$$\hat{\tau}_0^2 = \hat{\tau}_0^{2*} - (1 - \bar{\hat{\lambda}})(\hat{\gamma}_{01} - \hat{\gamma}_{10})^2 \hat{\tau}_X^2,$$

where $\hat{\tau}_0^{2*}$ is the naive estimate of τ_0^2 when using $\hat{\mu}_X^{\text{EB}}$ as a proxy of μ_X, and $\bar{\hat{\lambda}}$ is the average estimated reliability of cluster means.

Despite the simplicity of EBM compared to LMC and its improvement over conventional CM, CM remains the dominant method in MLM.[5] The implication is that, when estimating between-level or contextual effects, researchers have to assume either (a) no sampling error in the observed cluster means as in CM, which only happens when the within-cluster units are completely homogeneous (i.e., $\sigma_X^2 = 0$) or when all units in a cluster have been sampled (i.e., $n_{\text{pop}} = n$), or (b) infinitely large population cluster sizes as in LMC, which does not hold when clusters have finite sizes (e.g., students in schools or classrooms). On the other hand, using EBM with FPC allows one to incorporate information on population cluster sizes, which conceptually subsumes LMC ($\text{fpc}_2 = 1$) and CM ($\text{fpc}_2 = 0$) as special cases.

Current Studies

In this paper, we present the designs and results of two simulation studies to examine the performance of EBM. The main manipulated factors are sampling fraction within clusters, random effect variances of X (τ_X^2), and average cluster size. We expect LMC, which assumes an infinite population, to have the best performance when the sampling fraction is 0, but have increasingly biased estimates of the between-cluster effects when the sampling fraction increases. On the other hand, we expect EBM with FPC to maintain similar performance across different sampling fractions. As shown in equation (2), smaller τ_X^2 and average cluster size correspond to lower reliability of cluster means, so we expect all methods to perform worst in those conditions, and the bias would be enlarged for LMC and EBM without FPC when the assumption of an infinite

[5] For example, a quick survey of recent MLM textbooks used in social and behavioral sciences (Heck & Thomas, 2020; Hox et al., 2018; Luke, 2020; Snijders & Bosker, 2012) found only discussions of CM, but not EBM.

population does not hold.

To increase the generalizability of our simulation results, we also vary the number of clusters, random intercept variance of Y (τ_0^2), and imbalance of cluster sizes. We expect estimation with EBM and other methods to be more challenging for conditions with fewer clusters and unbalanced cluster sizes, as well as when τ_0^2 is small as it means limited information in the between-cluster level.

In Study 1, we compare CM, EBM, and LMC using a model with one level-1 (within-cluster) predictor with its cluster means; the level-1 coefficients vary across clusters (i.e., random slopes), and the sample units are drawn from clusters with finite sizes so that we can evaluate the incorporation of FPC into EBM. In Study 2, to imitate the model complexity in typical MLMs, we add one within-cluster and one between-cluster covariates into the model, and evaluate how EBM recovers the parameters associated with the predictors and the cross-level interaction.

Study 1

In Study 1, we compare CM, EBM, and LMC approaches in terms of parameter bias, the accuracy of statistical inference, and efficiency for estimating the between-level effect. As suggested by an anonymous reviewer, in order to isolate the impact of using different approaches for cluster means, we should compare the methods using the same estimation methods and CI procedures as much as possible. Given that different software programs are used for CM and EBM (*lme4* in R) and for LMC (Mplus), and REML and likelihood-based CIs are not implemented in Mplus, we use ML estimation and Wald confidence intervals for all three approaches.[6] We discuss how the use of REML and likelihood-based CIs can improve upon these simulation results later in the paper.

We simulate data with both infinite and finite population cluster sizes with varying sampling fractions (i.e., the ratio of sample cluster size to population cluster size). Previously, for models without random slopes, Lüdtke et al. (2008) and Grilli and Rampichini (2011) showed that CM outperformed LMC when the sampling fraction is large, so we expect similar results here with random slopes.

[6] However, this does not control for different software using different numerical algorithms and convergence criteria to find ML solutions.

The simulation data are generated using equation (1). For all conditions, we set the mean of X to 1, $\gamma_{00} = 0$ and $\sigma_X^2 = \sigma^2 = 1$ without loss of generality. We set γ_{01} = -0.3 and $\gamma_{10} = 0.7$ for a large discrepancy between the two coefficients, which is similar to the well-studied big-fish-little-pond effect (Marsh & Parker, 1984). We also simulate cluster sizes to be unbalanced: the J clusters are divided into five strata, each with $J/5$ clusters, and the cluster sizes are $\bar{n}/5$, $3\bar{n}/5$, $\bar{n}$, $7\bar{n}/5$, $9\bar{n}/5$, respectively, so that the ratio of the largest to the smallest cluster sizes is 9 to 1. For example, when $J = 100$ and $\bar{n} = 25$, the cluster sizes are $n = 5$, 15, 25, 35, 45, each for 20 clusters. The other design factors for data generation are described below.

Design Conditions

Random Intercept Variance of Y (τ_0^2) and Random Slope Variance (τ_1^2)

The conditional random intercept variance of Y is set to either 0.10 or 0.40. Thus, the conditional intraclass correlation (ICC) is either .09 or .28, which are on the low and high ends of values typically seen in cross-sectional data (Hedges & Hedberg, 2007). The random slope variance is $\tau_1^2 = \tau_0^2/4$, similar to some other simulation studies (e.g., Kwok et al., 2007).

Random Intercept Variance of X (τ_X^2)

The random intercept variance of X is set to 0.05, 0.25, and 1.0, so the corresponding ICCs for X are .05, .20, .50. Note that $\text{ICC}_X = .50$ is larger than the maximum value (.30) used in Lüdtke et al. (2008), and we expect that the between-level effect estimates will be more stable when the predictor has more variance at the between level.

Number of Clusters (J)

Previous simulations on LMC have relied on large numbers of clusters, with J between 50 and 500 in Lüdtke et al. (2008) and $J = 500$ in Asparouhov and Muthén (2019). Lüdtke et al. (2008) found that LMC showed biases generally for conditions with $J = 50$, which could be due to the sample size requirement for LMC (see also Kelcey et al., 2021). We expect that EBM will yield more stable estimates in small J conditions common in MLM. Thus, we simulate data with $J = 20$, 50, or 100. With frequentist analyses, we expect to see downward biases in estimates of τ_0^2 when $J = 20$, based on previous literature (e.g., Maas & Hox, 2005).

Average Cluster Size ($\bar{n}$)

We set $\bar{n}$ to either 5 or 25, which covers a similar range used in Lüdtke et al. (2008).

Sampling Fraction (SF)

We assume that the population size is constant across clusters, so with unbalanced cluster sizes, the sampling fraction is not constant across clusters. Instead, we define SF as the ratio of $\bar{n}$ to the population cluster size. The conditions are 0 (infinite population), .2, and .5.

Data Generation and Analyses

The Monte Carlo simulation is structured using the R package *SimDesign* (Chalmers & Adkins, 2020). For all conditions, we simulated the between- and the within-level components of X and all error terms from independent normal distributions. For conditions with SF > 0, we first simulated 20 sets of finite populations; the finite population size was $\bar{n}$/SF for each cluster. The sample units in the simulated data were drawn without replacement. Therefore, at the cluster level, the sampling fractions ranged from SF/5 (when $n_j = \bar{n}\ /\ 5$) to $9 \times$SF/5 (when $n_j = 9\ \bar{n}\ /\ 5$). For each finite population, we simulated 100 replication data sets, so the number of replications was 2,000 per condition.

We analyzed each simulated data set using CM, EBM, EBM-FP (i.e., EBM with FPC), and LMC. Including EBM without FPC allows us to evaluate the impact of incorporating FPC. For CM, EBM, and EBM-FP, we used the R package *lme4* (Bates et al., 2015) to obtain ML estimates for γ_{01}, γ_{10}, and τ_0^2, as well as the corresponding Wald CIs. For LMC, we used Mplus 8.8 to fit a two-level multilevel SEM model with ML estimation using the "hybrid" approach, and obtained 95% Wald CIs (i.e., estimate $\pm$ 1.96 $\widehat{\text{SE}}$) for the same three parameters. We used the `MODEL CONSTRAINT` routine to obtain estimates of the between-level coefficient (γ_{01}) by adding together the estimated contextual effect and the estimated within-level effect.

For each method in each replication, we computed (empirical) bias, root mean squared error (RMSE), and the coverage rates of 95% CIs. However, from an initial summary of the results, we found that the parameter estimates were highly unstable for conditions with small τ_0^2 or τ_X^2, and reporting the mean across 2,000 replications may result in biases of $>$ 10,000 for some conditions due to a few extreme outliers. To avoid the influence of extreme outliers, we instead computed robust versions of bias and RMSE using 20% trimmed means (Wilcox, 2017), which was a good compromise between the arithmetic mean (or 0% trimmed mean, which is highly sensitive to outliers) and the median (or 100% trimmed mean, which is robust but inefficient for normally distributed data).

For a sample estimate $\hat{\theta}$ estimating parameter θ, the bias was computed as $\bar{\hat{\theta}} - \theta$, where $\bar{\hat{\theta}}$ is the 20% trimmed mean of the $\hat{\theta}$ estimates across replications. The robust RMSE was computed as $\sqrt{\widehat{\text{Bias}}^2 + [MAD(\hat{\theta})]^2}$, where $MAD(\hat{\theta})$ was the sample median absolute deviation (from the median with a scale factor of 1.4826) of the 2,000 $\hat{\theta}$ estimates. The RMSE indicated the typical distance of $\hat{\theta}$ from the generated value of θ, and methods that yield smaller RMSEs should be preferred.

To evaluate the performance of the CIs, we computed the coverage rate as the proportion of replications where θ was inside the sample CI.

Results

We first consider the proportion of outliers when estimating γ_{01} (the between-cluster coefficient) as an indicator of the numerical stability of the three methods. Outliers were identified based on the boxplot method (Chambers et al., 1983/2018). The proportions of outlying $\hat{\gamma}_{01}$ estimates were 0.98% for CM, 3.53% for EBM, 3.10% for EBM-FP, and 2.83% for LMC, respectively. Extreme estimates were more common with EBM and LMC when the reliability of the cluster means, $\hat{\lambda}_{Xj}$, was small (i.e., when $\tau_X^2 = .05$ and $\bar{n} = 5$), in which case the proportion of outliers were up to 9.55% to 10.15% for EBM, EBM-FP, and LMC, compared to 1.90% for CM.

For LMC, EBM, and EBM-FP, estimation was more challenging for conditions with $\bar{n} = 5$ and $\tau_X^2 = 0.05$, where the reliability of the cluster means was low. Therefore, we present results for these conditions first in Table 1. When the EB cluster means could not be computed due to the REML/ML estimates of τ_X^2 being zero, results are inadmissible for EBM and EBM-ML. Proportions of inadmissible results were especially high (> 80%) for conditions with few clusters and large SF. When considering only the admissible results, when SF = 0, EBM-FP had slightly smaller bias than LMC when $J = 20$; all methods gave severely biased estimated between-level coefficients in other conditions, and CM was the most stable when SF = 0.5.

Figure 1 compares the parameter bias ($\hat{\gamma}_{01}$, $\hat{\tau}_0^2$, and $\hat{\tau}_1^2$) from all four methods for the remaining conditions. As expected, CM yielded biased estimates of γ_{01} when the reliability of the cluster means was small, with magnitudes close to the analytic results (i.e., $[\gamma_{10} - \gamma_{01}]\hat{\lambda}_{Xj}$).[7] When SF = 0, EBM (bias between -0.18 and 0.01; RMSE $\leq$ 1.34) and LMC (bias between -0.2 and 0.01; RMSE $\leq$ 1.40) showed smaller biases than CM for most conditions. Consistent with the results by Lüdtke et al. (2008), when SF > 0, LMC and EBM, which assumed infinte population cluster sizes, underestimated γ_{01} especially when $\hat{\lambda}_{Xj}$ was small (with magnitudes up to 1); EBM-FP, which used finite population corrections, showed much less bias (magnitudes up to 0.24). Also, EBM showed better estimates of $\hat{\tau}_0^2$ and $\hat{\tau}_1^2$ than LMC.

For coverage, as shown in Figure 2, EBM-FP generally gave CIs closed to nominal coverage for γ_{01} for conditions with either $\bar{n} = 25$ or $\tau_X^2 \geq 0.25$, but it had suboptimal coverage rates of around 80 to 90% for τ_0^2 and τ_1^2 in smaller samples (i.e., $J \leq 50$ or $\bar{n} = 5$), which is likely due to the use of Wald CIs and can be improved with likelihood-based CIs as shown later in the paper. LMC showed suboptimal coverage for γ_{01} with nonzero SF due to the parameter bias, but had better coverage rates than EBM-FP when SF = 0.

Study 2

In Study 2, we compare the performance of CM, EBM, and LMC when the data-generating model also contains a between-level covariate and a within-level covariate (Z and W, respectively), and a cross-level interaction between μ_X and W. The data-generating model is

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \mu_{Xj}) + \gamma_{01}\mu_{Xj} + \gamma_{02}Z_j + \gamma_{20}W_{ij} + \gamma_{21}\mu_{Xj}W_{ij} + u_{0j} + u_{2j}W_{ij} + e_{ij},$$

where $\text{Var}(u_{0j}) = \tau_0^2 = 0.4 - \gamma_{01}^2$ and $\text{Var}(u_{2j}) = \tau_2^2 = .05$. We manipulated $\{\gamma_{10}, \gamma_{01}\}$ to be either $\{0.4, -0.2\}$ or $\{0.1, 0.3\}$. The other manipulated variables were J and $\bar{n}$, each with the same levels as in Study 1. In addition, we also simulated data to have balanced or unbalanced cluster sizes as in Study 1. For all conditions we set γ_{02} to 0.5, γ_{20} to 0.3, and γ_{21} to 0.2. Both W and Z had variance

[7] For example, when $\tau_X^2 = 0.05$ and $\bar{n} = 25$, $\lambda_{Xj} = 0.56$, so the expected bias is 0.56.

of 1.0. We also allowed $X^{(w)} = X - \mu_X$ to covary with W and μ_X to covary with Z by simulating

$$\mu_{Xj} = 0.5 + 0.3Z_j + u_{X0j},$$
$$X_{ij}^{(w)} = 0.5W_{ij} + e_{Xij},$$

where the conditional variances of u_{X0j} and e_{Xij} were $.91\tau_X^2$ and .75, so that the total variances of μ_{Xj} and $X_{ij}^{(w)}$ were the same as in Study 1.

The added complexity makes the data-generating model better resemble the multilevel models used in applied research, compared to the models used in Lüdtke et al. (2008) and Lüdtke et al. (2011), which contained only the between- and within- components of X with no other covariates.

Results

Like in Study 1, all methods run into issues in conditions with small cluster mean reliability (i.e., $\bar{n} = 5$ and $\tau_X^2 = 0.05$), so we first presented parameter bias for those conditions in Figure 3. We only presented results for conditions with $\{\gamma_{10}, \gamma_{01}\} = \{0.4, \text{-}0.2\}$ in the main text, as the bias pattern was similar (but in the opposite direction) for conditions with $\{\gamma_{10}, \gamma_{01}\} = \{0.1, 0.3\}$, which can be found in the supplemental material. The parameters include the between-cluster effect of X ($\hat{\gamma}_{01}$), the coefficient of the level-2 covariate ($\hat{\gamma}_{02}$), the cross-level interaction ($\hat{\gamma}_{21}$), and the variance components ($\hat{\tau}_0^2$ and $\hat{\tau}_1^2$). As shown in the figure, CM produced biased estimates for the fixed-effect coefficients; while EBM and LMC gave less biased estimates, the bias was still substantial. Also, like in Study 1, LMC provided biased estimates of τ_0^2 and τ_1^2.

Figure 4 shows the bias of parameter estimates for conditions with either $\bar{n} = 25$ or $\tau_X^2 >$ 0.05. In summary, EBM and LMC gave mostly unbiased estimates except for a few conditions with a small population τ_0^2. Figure 5 further shows that EBM and LMC generally yielded reasonable CI coverage for the fixed effect parameters, but similar to Study 1, the coverage rates for EBM with τ_0^2 and τ_1^2 were suboptimal, which could again be due to the use of Wald CIs. We investigated this in the supplemental simulations, as described in the next section.

Better Estimation and CIs with EBM and LMC

As noted before, the EBM and LMC results might be improved by using different estimation and/or CI construction methods. In the case of EBM, switching from ML to REML estimation likely reduces bias in small samples (Hox et al., 2018), and using likelihood-based (LB) CIs instead of Wald CIs likely improves coverage rates (Bates et al., 2015). For LMC, the sandwich estimator for the standard errors is used by default in Mplus (with "ESTIMATOR=MLR"), which might improve CI coverage when normality does not hold. More recently, Asparouhov and Muthén (2019) suggested using Bayesian estimation with LMC to improve estimation when random slopes are present.

Boundary-Avoiding EBM

As shown in the results, an issue of EBM is that $\hat{\mu}_X^{\mathrm{EB}}$ depends on $\hat{\tau}_X^2$, which is often estimated to be 0 in situations with small sample sizes (e.g., Snijders & Bosker, 2012). When $\hat{\tau}_0^2 = 0$, $\hat{\lambda}_{Xj}$ becomes 0, $\hat{\mu}_{Xj}^{\mathrm{EB}}$ becomes linearly dependent on $\mathbf{C}$ (or becomes a constant if there are no $\mathbf{C}$ in the model), and the model is not estimable. One solution, suggested by Chung et al. (2013), is to use a penalized likelihood estimator for the variance components to avoid a zero estimate. This estimator is equivalent to one using the Bayesian posterior mode of τ_0^2 with a weakly informative gamma prior distribution, and is implemented in the R package *blme* (Chung et al., 2013). Given that the penalized likelihood estimator has not been widely studied in the MLM literature but is useful for the EBM approach, in our simulation studies, we include a version of EBM that estimates $\hat{\mu}_{Xj}^{\mathrm{EB}}$ with penalized likelihood, and label this approach boundary-avoiding EBM (EBM-BA).

Additional Results

To examine whether using alternative estimation CI construction methods improves EBM and LMC, we also compared four additional methods: (a) EBM-REML-FP, EBM with REML, LB CI, and finite population correction, (b) EBM-BA-FP, same as (a) but with boundary-avoiding priors when obtaining EB cluster means, (c) LMC-MLR, and (d) LMC-BAYES, using the same conditions as in Studies 1 and 2. As expected, results showed that EBM-REML-FP improved over EBM with ML and Wald CIs in terms of parameter bias and CI coverage rates, although it had similar convergence issues in conditions with low cluster-mean reliability. The coverage rates with

EBM-REML-FP were close to 95% for most conditions and parameters, except for γ_{01} and τ_0^2 in conditions with $\tau_X^2 = 0.05$. On the other hand, EBM-BA-FP had 100% convergence rates in all conditions and performed similarly to EBM-REML-FP in conditions with large samples and high cluster-mean reliability. More importantly, it showed less bias than other EBM and LMC methods, including LMC-BAYES, for estimating the γ_{01} (between-cluster effects) in conditions with low cluster-mean reliability when SF ≥ 0.2. We also found EBM-BAYES generally had less bias than LMC-ML and LMC-MLR in estimating γ_{01}, but it generally overestimated τ_0^2 and τ_1^2. The exception for the improved performance of EBM-BA-FP and LMC-BAYES is in conditions with $\bar{n} = 5$, $\tau_X^2 = 0.05$, $J = 20$, and SF $= 0$, as they showed more bias than other methods for estimating γ_{01}.

Overall, the supplemental results showed that using EBM with REML estimation, LB CI, and FPC generally gave satisfactory results. When there are convergence problems, we suggest computing the EB means using boundary-avoiding priors, which is also available in the R function discussed below.

Empirical Illustration

To demonstrate EBM, we revisit the classic example from Raudenbush and Bryk (2002) based on a subset of the High School and Beyond Study of 1980. The data, which has 7,185 students from 160 schools, was also used for demonstration in Shin and Raudenbush (2010). Specifically, we consider the following model:

$$\begin{aligned}\text{MATH}_{ij} &= \gamma_{00} + \gamma_{10}(\text{SES}_{ij} - \overline{\text{SES}}_j) + \gamma_{01}\overline{\text{SES}}_j + \gamma_{02}\text{SECTOR}_j \\ &\quad + u_{0j} + u_{1j}(\text{SES}_{ij} - \overline{\text{SES}}_j) + e_{ij},\end{aligned}$$

where SECTOR was the school sector (0 = public, 1 = Catholic) and SES was a standardized composite variable representing students' socioeconomic status. The data set also contains a SIZE variable that indicates school enrollment. Therefore, the data of each school can be considered a sample from a finite population. The sample school sizes ranged between 14 and 67 with a mean of 44.90, while the school enrollment ranged between 100 and 2713; the sampling fractions ranged between 0.01 and 0.41 across clusters, with an overall sampling fraction of 0.04, so the need for finite population corrections is small. A snapshot of the data is shown in Figure 6. Because the cluster sizes were relatively large, the reliability of the cluster means of SES was high (median reliability = .94, range = .81 to .95), so the observed cluster means were similar to the EB cluster means.

We first fit a CM model with observed cluster means of SES (for between-level prediction) and cluster-mean centered SES (for within-level prediction) with the R package *lme4*, and then compared the results to those using EBM (with and without finite population corrections) and LMC with Mplus. To run EBM, researchers can use the `lmer_ebm()` R function in the supplemental material, with the following sample syntax:

```
lmer_ebm(MATHACH ~ SES_ebm + SES_ebmc + SECTOR + (SES_ebmc | ID),
  data = hsb, formulax = SES ~ SECTOR + (1 | ID),
  pop_clus_size = hsb$SIZE)
```

where the variables `SES_ebm` and `SES_ebmc` are not in the original data but are EB estimates of cluster means and the EB mean-centered variables created by the function. In the input for `formulax`, we specified `SES` as the variable to have the EB cluster means computed across clusters (the membership of which is named `ID` in the data set), with any cluster-level covariates included (`SECTOR` in this case). Therefore, if researchers are interested in the between-level effect of a level-1 predictor named `pred` in the data, they can specify `pred_ebm` and `pred_ebmc` in the model formula. The function also returns a corrected estimate of the intercept variance (τ_0^2). Based on the simulation results, we expect EBM and LMC to give similar results and CM to give slightly biased results. As shown in Table 2, CM gave the smallest estimate for the between-level coefficient for SES (γ_{01}), which also led to the largest estimate for the coefficient of SECTOR (γ_{02}). It also resulted in the largest estimate of τ_0^2 due to the downward bias in the between-level coefficient for SES. Such results are consistent with our simulations showing CM to be biased. On the other hand, EBM gave a larger estimate of γ_{01} as it corrected for the measurement error in the cluster means of SES, but a smaller estimate of γ_{02}. With FPC, the estimate of γ_{01} was slightly smaller while that of γ_{02} was slightly larger. LMC also gave a larger γ_{01} estimate, and consistent with our simulation results, the estimates of τ_0^2 and τ_1^2 with LMC were smaller and likely underestimates.

While the difference between CM and EBM was relatively small in this example, as demonstrated in our simulations and in Lüdtke et al. (2008), the difference could be substantial when the cluster sizes are small. Similarly, the effect of correcting for finite population sizes would be more prominent when the sampling fraction is relatively large, such as when a majority of students in a school are sampled. Indeed, if all units in a cluster are sampled, CM is a better choice as the sample cluster means are also the population cluster means.

Discussion

As multilevel modeling (MLM) has become a standard technique in researchers' toolboxes, it is important to ensure that researchers are aware of different analytic issues, including the best practices for separating between and within effects and estimating contextual effects. However, a recent review of research in organizational science and applied psychology (Antonakis et al., 2021) showed that only about half (106 out of 204) of the reviewed articles included the cluster means of level-1 predictors in a multilevel analysis. While recent research has proposed using latent cluster means (e.g., Asparouhov & Muthén, 2019; Lüdtke et al., 2008) with a multilevel structural equation modeling framework—as opposed to the observed group means traditionally used in MLM, there have been limited empirical studies on the performance of using latent cluster means in small samples and in models with random slopes and covariates. Also, researchers may not be aware of the assumption underlying the latent cluster mean approach, namely that the sampled units of a cluster represent a small fraction of the population units of that cluster, which may not always be appropriate in applied research.

Drawing from the existing methodological literature, we propose the use of empirical Bayes cluster means (EBM) with finite population corrections to obtain consistent estimates of between-level effects (with centering of the level-1 predictor) and contextual effects (without centering). A correction on the estimated level-2 variance is also discussed. The EBM approach takes into account the population cluster sizes and thus subsumes both the case of negligible (as in latent cluster means) and non-negligible (as in observed group means) sampling fractions. In a series of simulation studies, it is shown that EBM, like the latent cluster-mean approach, gives consistent estimates (with respect to increasing numbers of clusters) of between-level effects when the ratio of sample cluster size to the population cluster size is large. The estimation and inferences with EBM can be improved by using restricted maximum likelihood and likelihood-based confidence intervals. It is also found that for models with random slopes, when cluster size is five or fewer and the ICC of the predictor is $< .05$, all approaches examined in this paper lead to highly unstable and biased parameter estimates. While the boundary-avoiding EBM approach helped mitigate the bias, the bias was still substantial. Future research can explore multilevel bootstrap methods (Lai, 2020; van der Leeden et al., 2008) as alternatives for correcting biases in coefficients.

It is shown that the need for finite-population correction is highest when the population cluster size is small and the sampling fraction is high. To facilitate the use of the proposed method, we also provide an R function `lmer_ebm()` that automates the computational steps for using EBM, and provide a real-data example using the classic High School and Beyond survey data set. While the provided function only works with the R package *lme4*, one can obtain EBM using equation (7) with any multilevel software.

There are several limitations of the current study that deserve attention in future studies. First, while we dealt with the basic case where only one between-level effect or contextual effect is of interest, which is fairly common in practice, future research can explore how the proposed method can be extended to handle multiple such effects. Second, the present paper only concerns the error due to approximating the population cluster means with the sample means, which happens in standard multilevel modeling applications. However, as shown in Lüdtke et al. (2011), the latent means approach with multilevel structural equation modeling can also handle measurement error on the individual predictor scores. Theoretically, the empirical Bayes estimate can also incorporate unreliability due to such measurement error (e.g., Zitzmann, 2018) Lai, 2021, assuming that an estimate of the reliability of the individual scores is known. Future research can further explore this extension and compare it with the latent means approach.

In addition, our discussion is limited to two-level models; there is additional complexity for defining cluster means in three-level and crossed designs (Brincks et al., 2017); Lai, 2019, and the potential need for finite population corrections at more than one level. Finally, the proposed method can be extended to cluster means of binary predictors, with which the cluster-mean reliability depends not only on the cluster size but is also a function of the cluster mean estimate, as well as to generalized linear mixed models with nonnormal outcome variables.

References

Antonakis, J., Bastardoz, N., & Rönkkö, M. (2021). On ignoring the random effects assumption in multilevel models: Review, critique, and recommendations. *Organizational Research Methods*, *24*(2), 443–483. https://doi.org/10.1177/1094428119877457

Asparouhov, T., & Muthén, B. (2019). Latent variable centering of predictors and mediators in multilevel and time-series models. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(1), 119–142. https://doi.org/10.1080/10705511.2018.1511375

Aydin, B., Leite, W. L., & Algina, J. (2016). The effects of including observed means or latent means as covariates in multilevel models for cluster randomized trials. *Educational and Psychological Measurement*, *76*(5), 803–823. https://doi.org/10.1177/0013164415618705

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research*, *52*(2), 149–163. https://doi.org/10.1080/00273171.2016.1256753

Chalmers, R. P., & Adkins, M. C. (2020). Writing effective and reliable Monte Carlo simulations with the SimDesign package. *The Quantitative Methods for Psychology*, *16*(4), 248–280. https://doi.org/10.20982/tqmp.16.4.p248

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (2018). *Graphical methods for data analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9781351072304 (Original work published 1983)

Cheung, M. W.-L. (2013). Implementing restricted maximum likelihood estimation in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(1), 157–167. https://doi.org/10.1080/10705511.2013.742404

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A Nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*(2), 300–315. https://doi.org/10.1037/a0033805

Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, *12*(1), 45–57. https://doi.org/10.1037/1082-989X.12.1.45

Depaoli, S., & Clifton, J. P. (2015). A Bayesian approach to multilevel structural equation modeling with continuous and dichotomous outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(3), 327–351. https://doi.org/10.1080/10705511.2014.937849

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121–138. https://doi.org/10.1037/1082-989X.12.2.121

Gottfredson, N. C. (2019). A straightforward approach for coping with unreliability of person means when parsing within-person and between-person effects in longitudinal studies. *Addictive Behaviors*, *94*, 156–161. https://doi.org/10.1016/j.addbeh.2018.09.031

Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *7*(4), 121–133. https://doi.org/10.1027/1614-2241/a000030

Heck, R. H., & Thomas, S. L. (2020, April). *An introduction to multilevel modeling techniques: MLM and SEM approaches* (4th edition). Routledge.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hoffman, L. (2019). On the interpretation of parameters in multivariate multilevel models across different combinations of model specification and estimation. *Advances in Methods and Practices in Psychological Science*, *2*(3), 288–311. https://doi.org/10.1177/2515245919842770

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (Third edition). Routledge.

Kelcey, B., Cox, K., & Dong, N. (2021). Croon's bias-corrected factor score path analysis for small- to moderate-sample multilevel structural equation models. *Organizational Research Methods*, *24*(1), 55–77. https://doi.org/10.1177/1094428119879758

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983. https://doi.org/10.2307/2533558

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), 1–21. https://doi.org/10.1207/s15327906mbr3001_1

Kwok, O.-m., West, S. G., & Green, S. B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A monte carlo study. *Multivariate Behavioral Research*, *42*(3), 557–592. https://doi.org/10.1080/00273170701540537

Lai, M. H. C. (2019). Correcting fixed effect standard errors when a crossed random effect was ignored for balanced and unbalanced designs. *Journal of Educational and Behavioral Statistics*, *44*(4), 448–472. https://doi.org/10.3102/1076998619843168

Lai, M. H. C. (2020). Bootstrap confidence intervals for multilevel standardized effect size. *Multivariate Behavioral Research*, 1–21. https://doi.org/10.1080/00273171.2020.1746902

Lai, M. H. C. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. https://doi.org/10.1037/met0000287

Lai, M. H. C., Kwok, O.-m., Hsiao, Y.-Y., & Cao, Q. (2018). Finite population correction for two-level hierarchical linear models. *Psychological Methods*, *23*(1), 94–112. https://doi.org/10.1037/met0000137

Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A 2 × 2 taxonomy of multilevel latent contextual models: Accuracy–bias trade-offs in full and partial error correction models. *Psychological Methods*, *16*(4), 444–467. https://doi.org/10.1037/a0024376

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, *13*(3), 203–229. https://doi.org/10.1037/a0012869

Luke, D. A. (2020). *Multilevel modeling* (Second edition). SAGE Publishing.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*(3), 86–92. https://doi.org/10.1027/1614-2241.1.3.86

Marsh, H. W., & Parker, J. W. (1984). Determinants of student self-concept: Is it better to be a relatively large fish in a small pond even if you don't learn to swim as well? *Journal of Personality and Social Psychology*, *47*(1), 213–231. https://doi.org/10.1037/0022-3514.47.1.213

McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models* (1st ed.). Wiley. https://doi.org/10.1002/0471722073

Preacher, K. J., Zyphur, M. J., & Zhang, Z. (2010). A general multilevel SEM framework for assessing multilevel mediation. *Psychological Methods*, *15*(3), 209–233. https://doi.org/10.1037/a0020141

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

Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, *35*(1), 26–53. https://doi.org/10.3102/1076998609345252

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In *Handbook of multilevel analysis* (pp. 401–433). Springer.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th edition). Elsevier, Academic Press.

Zitzmann, S. (2018). A computationally more efficient and more accurate stepwise approach for correcting for sampling error and measurement error. *Multivariate Behavioral Research*, *53*(5), 612–632. https://doi.org/10.1080/00273171.2018.1469086

Zitzmann, S., Lüdtke, O., Robitzsch, A., & Marsh, H. W. (2016). A Bayesian approach for estimating multilevel latent contextual models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(5), 661–679. https://doi.org/10.1080/10705511.2016.1207179

Table 1

Inadmissible Solutions and Bias of Between-Level Coefficients in Low Cluster-Mean Reliability Conditions of Study 1.

					% Inadmissible[a]	Bias for γ_{01}			
$\bar{n}$	τ_X^2	J	SF	τ_0^2	EBM	CM	EBM	EBM-FP	LMC
5	0.05	20	0	0.1	36.25	0.83	0.03	0.03	0.19
				0.4	36.05	0.85	0.00	0.00	0.02
			0.2	0.1	54.40	0.80	-0.77	-0.41	-0.20
				0.4	53.45	0.82	-0.81	-0.46	-0.48
			0.5	0.1	88.25	0.71	-3.91	-0.56	-1.16
				0.4	87.55	0.76	-4.47	-0.74	-1.91
		50	0	0.1	18.55	0.83	-0.07	-0.07	0.03
				0.4	18.55	0.85	-0.07	-0.07	-0.22
			0.2	0.1	42.75	0.79	-1.82	-1.18	-0.63
				0.4	42.80	0.82	-1.80	-1.19	-1.33
			0.5	0.1	93.45	0.70	-9.78	-1.44	-1.70
				0.4	93.55	0.75	-9.53	-1.40	-3.83
		100	0	0.1	8.55	0.84	-0.14	-0.14	-0.06
				0.4	8.60	0.86	-0.13	-0.13	-0.12
			0.2	0.1	34.60	0.80	-2.03	-1.32	-0.84
				0.4	34.60	0.82	-1.91	-1.27	-1.58
			0.5	0.1	98.15	0.71	-17.58	-2.52	-1.98
				0.4	98.20	0.76	-17.33	-2.67	-4.78

Note. [a]Results are admissible for all replications in CM and LMC. True $\gamma_{01} = 0.7$ in the data generating model.

Table 2

A Comparison of Different Estimation Approaches for the Empirical Illustration.

term	CM	EBM	EBM (FPC)	LMC
Intercept	12.06 (0.20)	12.09 (0.20)	12.09 (0.20)	12.08 (0.21)
SES (Between)	5.25 (0.37)	5.47 (0.40)	5.45 (0.39)	5.58 (0.39)
SES (Within)	2.19 (0.13)	2.20 (0.13)	2.20 (0.13)	2.20 (0.13)
SECTOR	1.37 (0.31)	1.31 (0.31)	1.31 (0.31)	1.34 (0.38)
τ_0^2	2.39	2.30	2.31	2.26
τ_1^2	0.70	0.70	0.70	0.46
σ^2	36.71	36.71	36.71	36.78

Note. CM = Observed cluster mean approach. EBM = Empirical Bayes mean approach. FPC = with finite population correction. LMC = Latent mean centering (hybrid approach in Mplus).

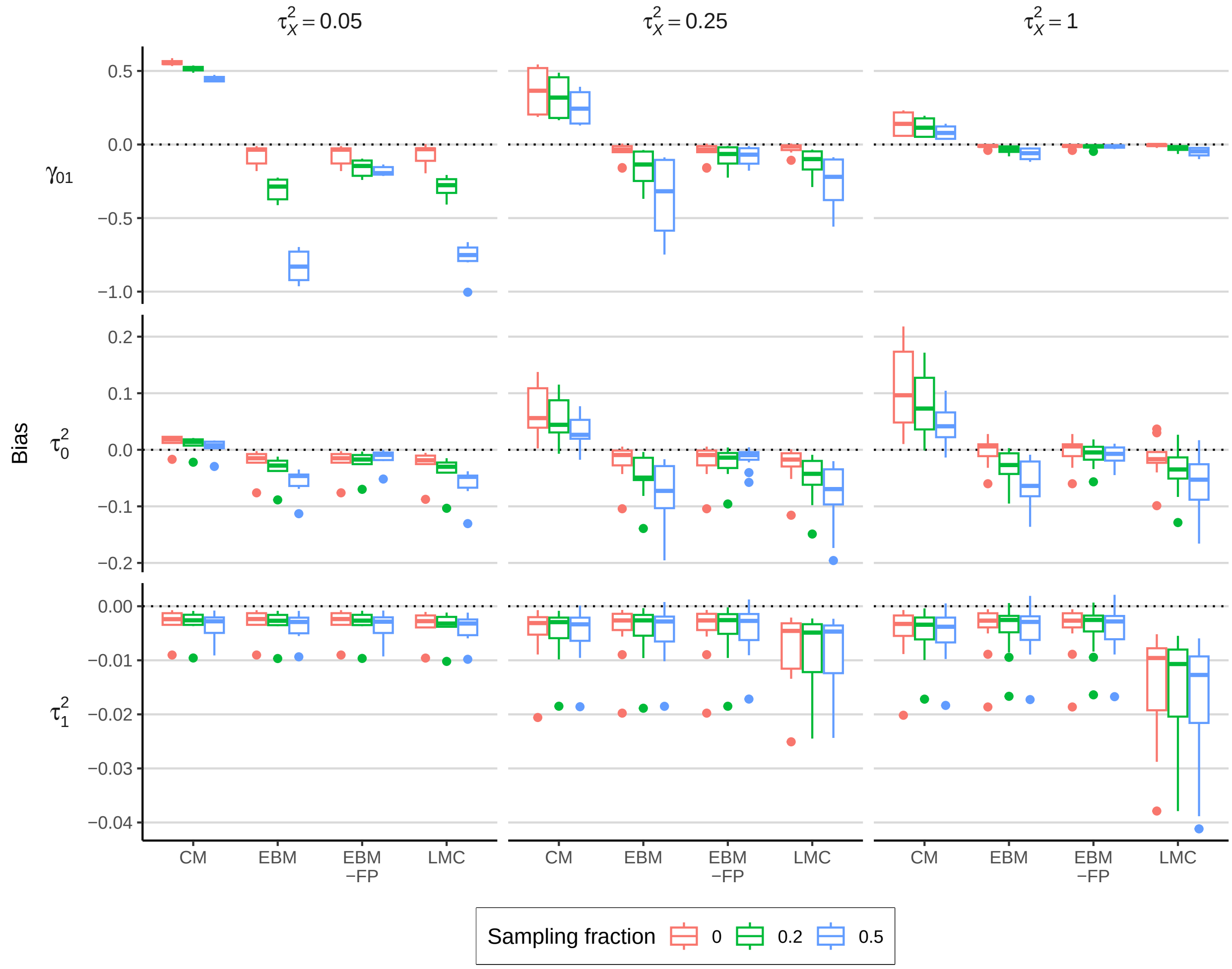

Figure 1

Bias for parameter estimates in Study 1. The panels show, from top to bottom, the between-cluster effect, the conditional random intercept variance of the outcome, and the random slope variance. CM, EBM, and LMC represents analyses with observed, Empirical Bayes, and latent means as covariate. EBM-FP = EBM with finite population corrections. Conditions with average cluster size = 5 and τ_X^2 = 0.05 are not shown (see Table 1).

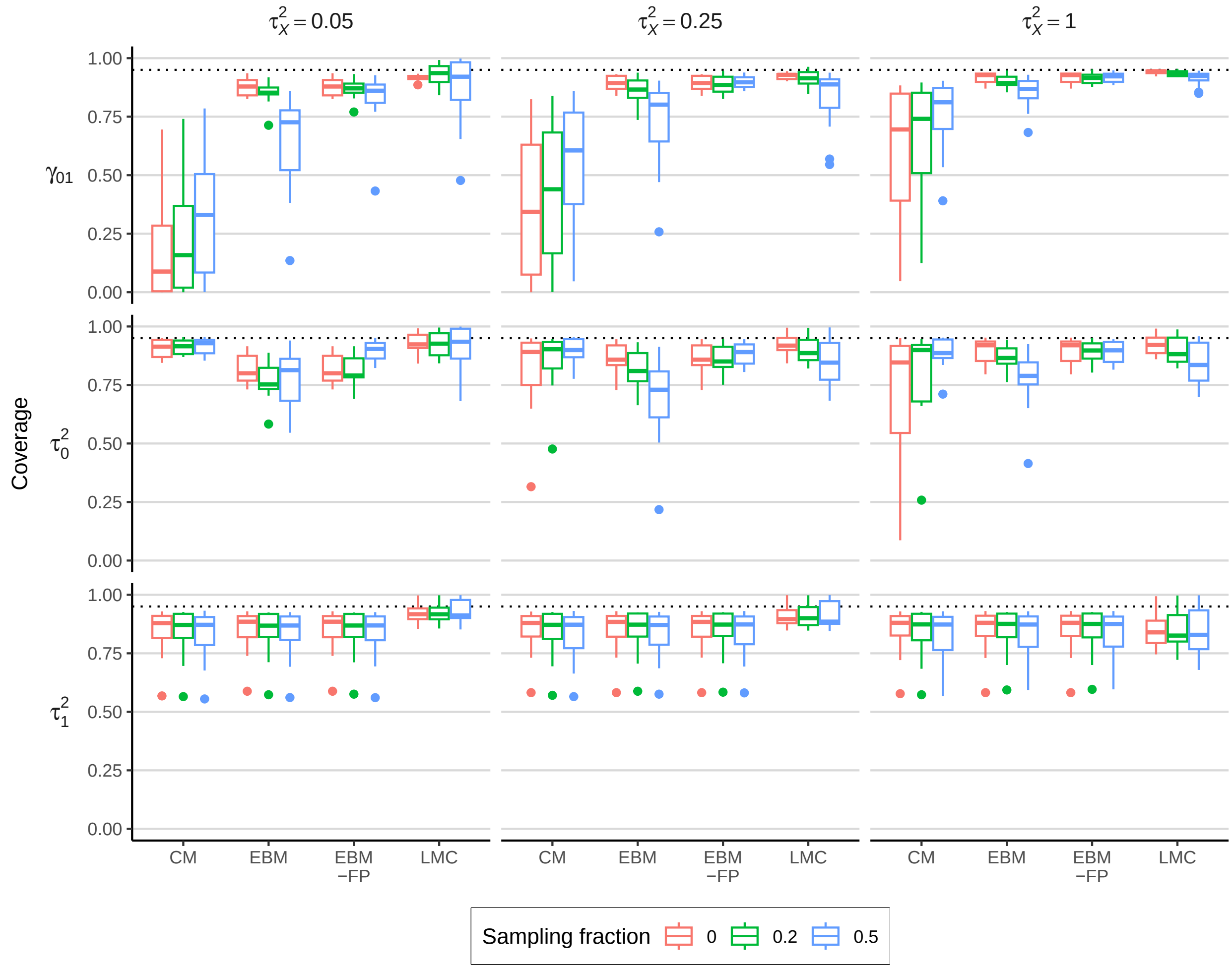

Figure 2

Empirical coverage for Study 1 (all conditions). The panels show, from top to bottom, the between-cluster effect, the conditional random intercept variance of the outcome, and the random slope variance. CM, EBM, and LMC represents analyses with observed, Empirical Bayes, and latent means as covariate. EBM = EBM with maximum likelihood estimation and 95% Wald intervals. EBM-FP = EBM with finite population corrections. The dashed line represents the 95% reference rate.

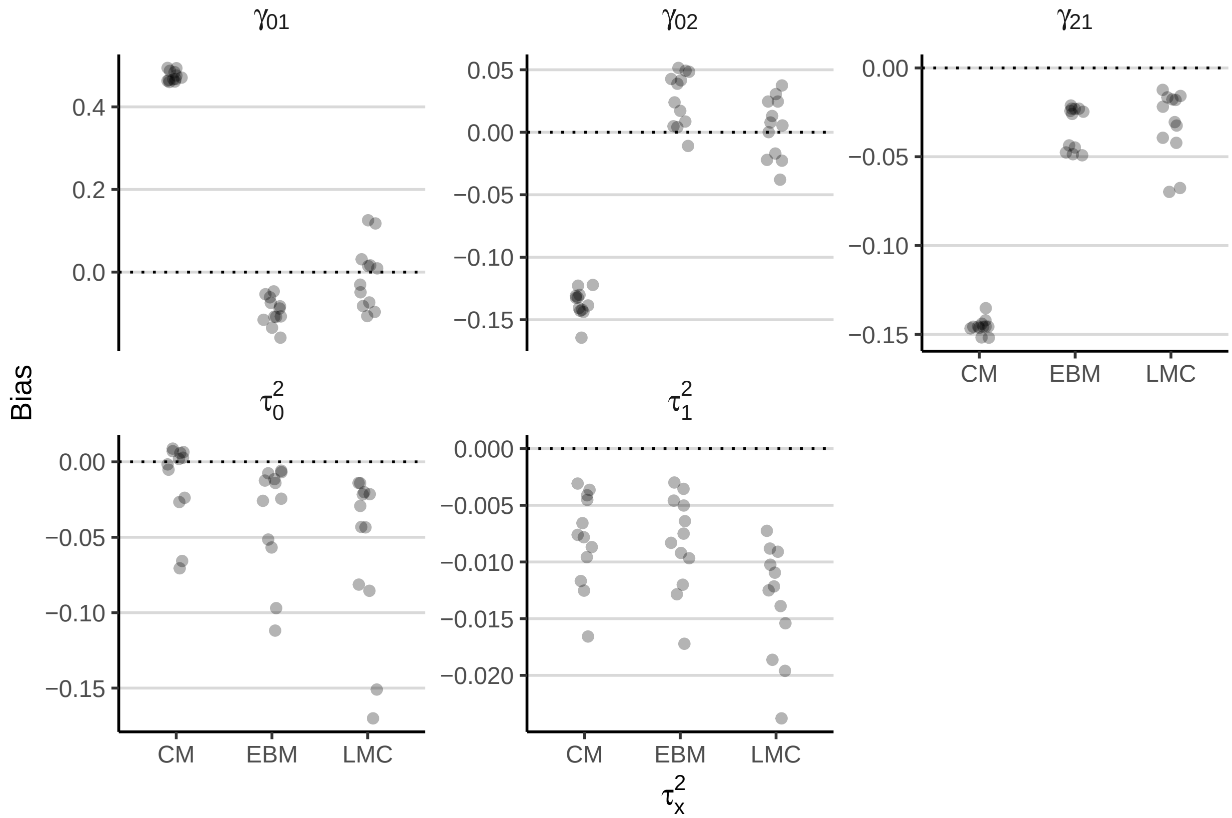

Figure 3

Bias of parameter estimates in Study 2 for conditions with low cluster mean reliability (i.e., average cluster size = 5 and τ_X^2 = 0.05). The panels show the between-cluster effect (γ_{01}), the effect of the level-2 covariate (γ_{02}), the cross-level interaction (γ_{21}), the conditional random intercept variance of the outcome (τ_0^2), and the random slope variance (τ_1^2). CM, EBM, and LMC represents analyses with observed, Empirical Bayes, and latent means as covariate.

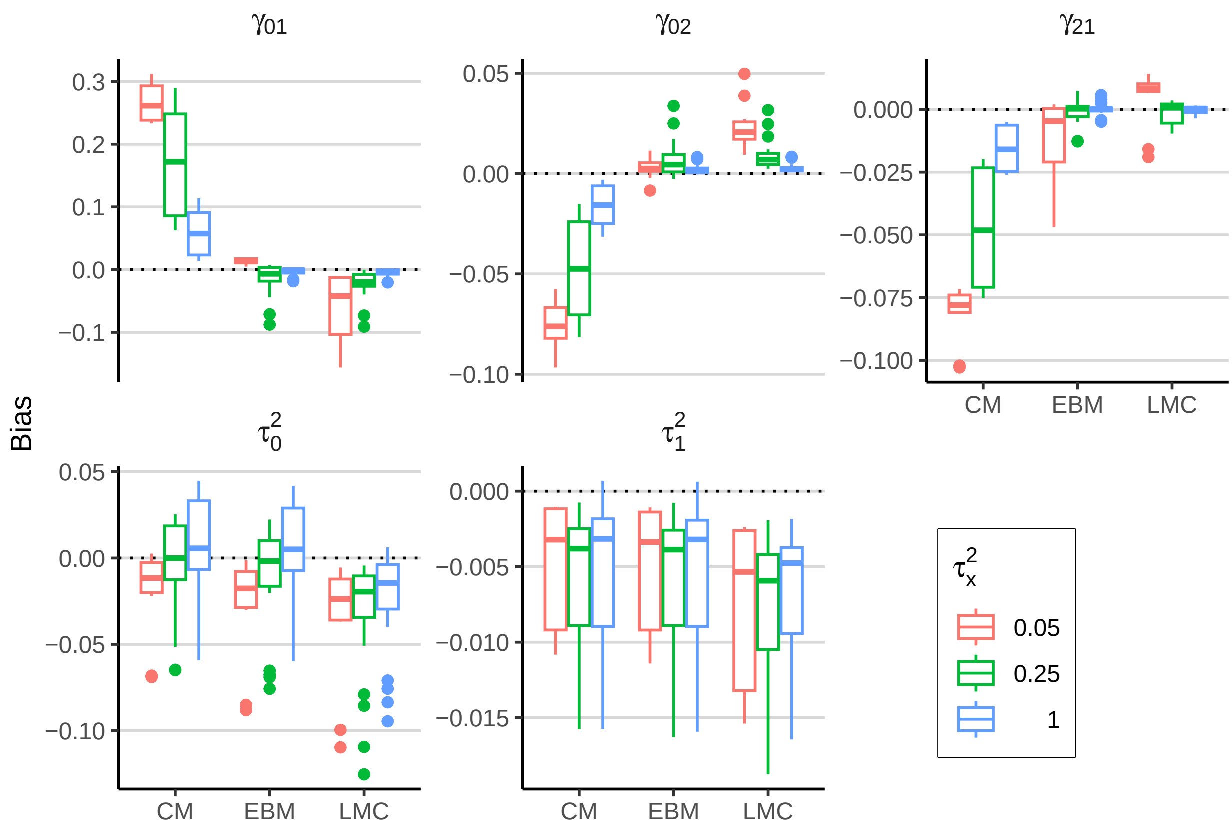

Figure 4

Bias of parameter estimates in Study 2 for conditions with average cluster size = 25 or $\tau_X^2 \geq 0.25$. The panels show the between-cluster effect (γ_{01}), the effect of the level-2 covariate (γ_{02}), the cross-level interaction (γ_{21}), the conditional random intercept variance of the outcome (τ_0^2), and the random slope variance (τ_1^2). CM, EBM, and LMC represents analyses with observed, Empirical Bayes, and latent means as covariate.

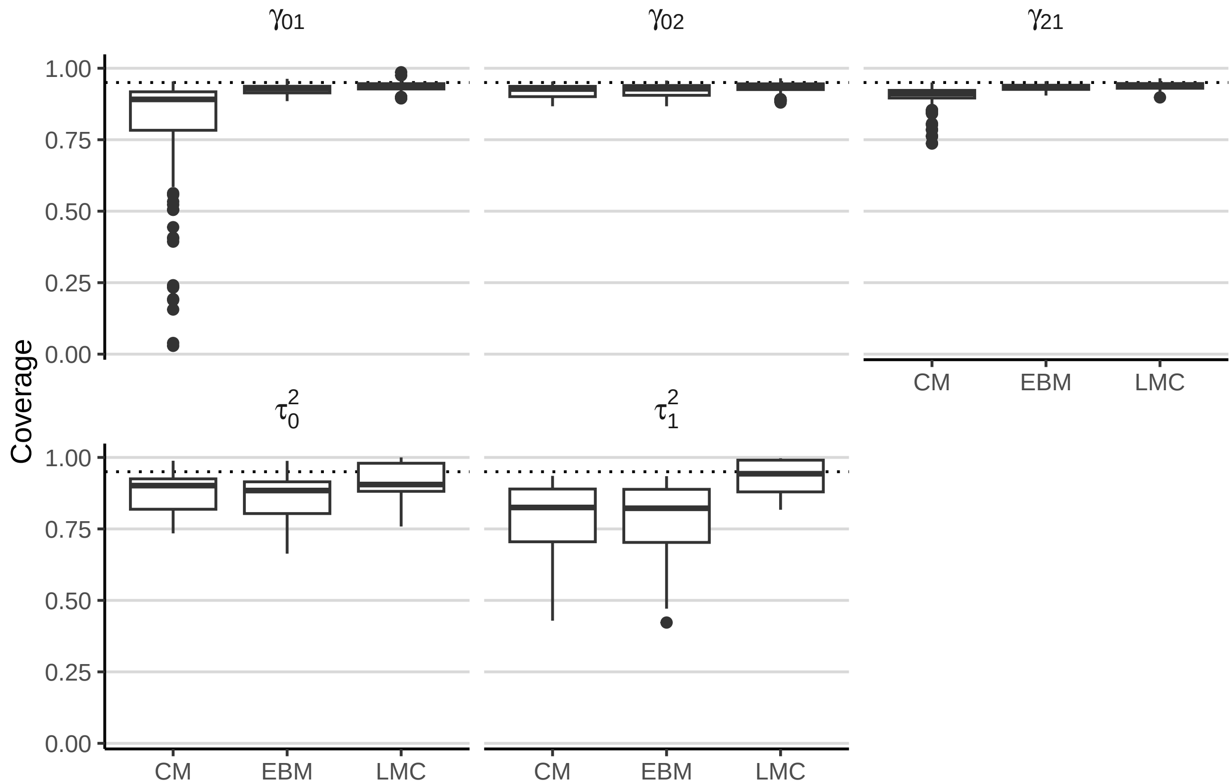

Figure 5

Empirical coverage for Study 2. The panels show the between-cluster effect (γ_{01}), the effect of the level-2 covariate (γ_{02}), the cross-level interaction (γ_{21}), the conditional random intercept variance of the outcome (τ_0^2), and the random slope variance (τ_1^2). The dashed line represents the 95% reference rate. CM, EBM, and LMC represents analyses with observed, Empirical Bayes, and latent means as covariate.

	ID	SCHOOL SIZE	CLUSTER SIZE	SES	SES CM	SES EBM	REL	REL (FPC)
1	2305	485	67	-0.368	-0.6280	-0.6033	0.960	0.965
2	2305	485	67	-0.588	-0.6280	-0.6033	0.960	0.965
3	2768	1680	25	0.332	-0.0536	-0.0488	0.900	0.901
4	2768	1680	25	-1.958	-0.0536	-0.0488	0.900	0.901
5	4410	100	41	-0.528	0.0964	0.0899	0.937	0.961
6	4410	100	41	0.962	0.0964	0.0899	0.937	0.961
7	5761	215	52	-1.238	-0.3230	-0.3069	0.949	0.961
8	5761	215	52	-1.368	-0.3230	-0.3069	0.949	0.961
9	7688	1410	54	0.792	0.1859	0.1765	0.951	0.953
10	7688	1410	54	0.432	0.1859	0.1765	0.951	0.953
11	8367	153	14	-0.228	0.0256	0.0204	0.835	0.847
12	8367	153	14	-0.048	0.0256	0.0204	0.835	0.847

Figure 6

A snapshot of the data for the empirical illustration, including the observed and empirical Bayes cluster means and the cluster-mean reliability.

Appendix

Deriving a Consistent Estimate of τ_0^2 Under EBM

Consider a random intercepts model at the population level

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \mu_{Xj}) + \gamma_{01}\mu_{Xj} + u_{0j} + e_{ij}$$

where u_0 and e are assumed independent and independent to μ_{Xj} and both have zero means, and the variance of u_0 is τ_0^2. This between-within model can be reparameterized as an equivalent contextual model

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + (\gamma_{01} - \gamma_{10})\mu_{Xj} + u_{0j} + e_{ij}.$$

Let τ_X^2 be the variance of μ_X. The above model implies that the partial variance of Y accounted for by the group mean, μ_X, is $\tau_X^2(\gamma_{01} - \gamma_{10})^2$, after conditioning on X_{ij}.

In EBM, when the empirical Bayes estimate of the group mean, $\hat{\mu}_{Xj}^{\text{EB}}$, is used in place of μ_{Xj}, the proportion of variance of Y it accounts for is attenuated to the extent that $\hat{\mu}_{Xj}^{\text{EB}}$ is not a perfectly reliable measurement of μ_X (i.e., $\lambda_j < 1$). Because the variance of $\hat{\mu}_{Xj}^{\text{EB}}$ is $\lambda_j\tau_X^2$, it follows that the partial variance of Y accounted for by $\hat{\mu}_{Xj}^{\text{EB}}$ is $\lambda_j\tau_X^2(\gamma_{01} - \gamma_{10})^2$, which is smaller than that by μ_X. The difference, $(1 - \lambda_j)\tau_X^2(\gamma_{01} - \gamma_{10})^2$, will be added to the random intercept variance of Y. Therefore, the random intercept variance estimate of Y under EBM converges to

$$\tau_0^{2*} = \tau_0^2 + (1 - \lambda)(\gamma_{01} - \gamma_{10})^2\tau_X^2.$$

As under EBM, the sample ML and REML estimates $\hat{\lambda}$, $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$, and τ_X^2 are consistent, a consistent estimator of τ_0^2 can be obtained as

$$\hat{\tau}_0^{2*} - (1 - \hat{\lambda})(\hat{\gamma}_{01} - \hat{\gamma}_{10})^2\hat{\tau}_X^2.$$