MULTIVARIATE BEHAVIORAL RESEARCH
2026, VOL. 61, NO. 4, 653–670

OPEN ACCESS

To Disaggregate or Not to Disaggregate: A Focus on Covariates in Multilevel Models

Remus Mitchell , Craig K. Enders, and Yi Feng

Department of Psychology, University of California, Los Angeles

ABSTRACT

It is routinely recommended that level-1 variables in multilevel models be disaggregated when they are of substantive importance. Yet, the consensus on the disaggregation of level-1 covariates is more mixed. Disaggregation clarifies interpretation and reduces bias in the covariate, though some methodologists argue that it is unnecessary when the covariate itself is not of substantive interest. Our study builds off recent work to explore the tradeoffs between bias and precision when choosing to disaggregate level-1 covariates when the primary interest lies in a level-2 predictor. Using a Monte Carlo simulation, we examine how factors such as the intraclass correlation, the magnitude of the contextual effect, the within- and between-level effect sizes, the correlation among level-2 effects, sample size at both levels, and the method of disaggregation (manifest versus latent) influence bias, precision, and power of a level-2 focal estimate. Our findings suggest that although disaggregation generally improves interpretability and reduces bias, there are conditions where a non-disaggregated approach may yield greater precision. These insights inform best practices for handling lower-level covariates in multilevel models.It is routinely recommended that level-1 variables in multilevel models be disaggregated when they are of substantive importance. Yet, the consensus on the disaggregation of level-1 covariates is more mixed. Disaggregation clarifies interpretation and reduces bias in the covariate, though some methodologists argue that it is unnecessary when the covariate itself is not of substantive interest. Our study builds off the work of Rights et al. to explore the tradeoffs between bias and precision when choosing to disaggregate level-1 covariates when the primary interest lies in a level-2 predictor. Using a Monte Carlo simulation, we examine how factors such as the intraclass correlation, the magnitude of the contextual effect, the within- and between-level effect sizes, the correlation among level-2 effects, sample size at both levels, and the method of disaggregation (manifest versus latent) influence bias, precision, and power of a level-2 focal estimate. Our findings suggest that although disaggregation generally improves interpretability and reduces bias, there are conditions where a non-disaggregated approach may yield greater precision. These insights inform best practices for handling lower-level covariates in multilevel models.

KEYWORDS

Multilevel models; disaggregation; blimp; covariates; random intercept model

Introduction

Multilevel models (MLMs) represent a powerful statistical framework designed to analyze nested or clustered data structures. Whereas single-level linear regression models assume there is only one source of variation in the data, MLMs acknowledge the hierarchical nature of data, allowing researchers to examine variation and associations at multiple levels (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Disaggregation of level-1 predictors in multilevel models involves separating the within-cluster and between-cluster variation of these variables. This is typically executed by centering predictors around the cluster means to distinguish the unique influence of individual differences from the collective influence of the cluster. The level-2 cluster means are then added as a distinct level-2 predictor.

Without disaggregating lower-level predictors, the level-specific slopes can be inadvertently combined,

© 2026 The Author(s).
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

yielding a single regression coefficient that conflates distinct within- and between-level effects into an uninterpretable blend (Enders & Tofighi, 2007; Hoffman, 2019; Preacher et al., 2010; Raudenbush & Bryk, 2002; Rights et al., 2020). Thus, many methodologists suggest taking appropriate action to disaggregate level-1 predictor variables when they are substantively interesting (Enders & Tofighi, 2007; Hoffman & Stawski, 2009; Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Moreover, disaggregating lower-level variables enables researchers to directly partition outcome variance in a manner that facilitates clear computation of effect sizes, offering clearer insight into the multilevel processes under study (Rights & Sterba, 2019). Disaggregation is especially critical when modeling random slopes and dynamic (lagged) effects, where accurate interpretation hinges on isolating within- and between-person components (Hamaker & Grasman, 2014; Hamaker & Muthén, 2020; Zyphur et al., 2020).

We explore two model specifications to make the concept of disaggregation more concrete: the non-disaggregated model, hereafter referred to as the *smushed model* (Hoffman & Walters, 2022)[1], and the parsed or disaggregated model. For consistency with another recent paper on this topic by Rights et al. (2020), we focus on a random-intercept model with a level-2 focal predictor and a level-1 covariate as a control. The disaggregated model is distinct in that the within and between effects are modeled with separate predictors. An example of the reduced form of this model specification is

$$Y_{ij} = \gamma_{00} + \gamma_Z Z_j + \gamma_{X_B}\overline{X}_{.j} + \gamma_{X_W}\left(X_{ij} - \overline{X}_{.j}\right) + u_{0j} + \varepsilon_{ij}$$
$$\varepsilon_{ij} \sim N\left(0, \sigma^2_\varepsilon\right) \text{ and } u_{0j} \sim N(0, \tau_{00}) \quad (1)$$

where Y_{ij} represents the continuous level-1 outcome, Z_j represents a continuous between-level focal predictor; X_{ij} is the level-1 covariate score for observation i in cluster j and $\overline{X}_{.j}$ is the average value of covariate X_{ij} for the jth cluster. Thus, γ_{X_B} represents the between-level effect of the covariate, γ_{X_W} is the within-level effect of the covariate, and γ_Z refers to the between-level effect of the focal predictor. The second line of Equation 1 indicates that the two residuals are assumed to be normally distributed. The cluster-level averages can be modeled as arithmetic means or latent group means (Lüdtke et al., 2008; 2011). The previous equation includes both the within- and between-cluster predictors. In cases where the level-2 predictor Z_j is of interest, it is only necessary to include the level-2 cluster means; yet, if a level-1 focal predictor is of interest, researchers should carefully consider which level-specific components of the predictor they aim to control for based on their hypotheses (Rights et al., 2020).

The smushed model below represents both the between and within effect of the covariate with a single regression coefficient, where γ_{X_C} represents this blended regression coefficient.

$$Y_{ij} = \gamma_{00} + \gamma_Z Z_j + \gamma_{X_C} X_{ij} + u_{0j} + \varepsilon_{ij} \quad (2)$$

This coefficient is conflated if the within- and between-cluster slopes in Equation 1 differ. A key issue for our paper is whether conflating the covariate's effects negatively impacts the focal slope at the second level. Rights et al. (2020) showed that γ_Z is a function of the covariance between Z_j and X_{ij} *via* X's level-2 cluster mean. When the conflated slope of the level-1 covariate differs from the covariate's between-cluster slope ($\gamma_{X_B} - \gamma_{X_C} \neq 0$) and the correlation between Z_j and $X_{.j}$ is nonzero, the estimation of the focal level- slope γ_Z will be biased. This bias grows as the absolute value of the correlation between Z_j and $X_{.j}$ increases.

While there is broad consensus regarding the need to disaggregate focal level-1 predictors, the literature is less consistent about whether researchers should also disaggregate covariates. Some sources suggest that the disaggregation of covariates may not be necessary (Enders & Tofighi, 2007). One rationale for this is that the conflated slopes matter less when the variable does not bear directly on the hypotheses being tested. Another context where researchers do not disaggregate is when cluster-robust standard errors are used instead of fitting a multilevel model (McNeish, 2023). Both these perspectives implicitly assume that the covariate exerts similar effects at the within- and between-levels, thereby minimizing the consequences of treating it as a single, undifferentiated predictor. However, this assumption is rarely justified, and prior work has demonstrated that equality of within- and between-level coefficients is more the exception than the rule (Raudenbush & Bryk, 2002). Furthermore, regardless of whether substantive researchers have included a variable as a statistical control, it is fairly common for these researchers to proceed with interpreting the parameter estimates associated with these covariates.

Another possible rationale for not disaggregating covariates is that the smallest standard error is

[1]The term "smushed model" for referring to the non-disaggregated model is consistent with Rights et al. (2020).

achieved when we don't disaggregate predictors, as initially demonstrated by Raudenbush and Bryk (2002) using Equation 3 below, which can also be found in Equation 5.39 within their book.

$$var(\hat{\gamma}_{X_C})^{-1} = \frac{\sum(\overline{X}_{.j} - \overline{X}_{..})^2}{\hat{\sigma}^2_{u_0} + \left(\frac{\hat{\sigma}^2_\varepsilon}{n}\right)} + \frac{\sum\sum(X_{ij} - \overline{X}_{.j})^2}{\hat{\sigma}^2_\varepsilon}$$
$$= var(\hat{\gamma}_{X_B})^{-1} + var(\hat{\gamma}_{X_W})^{-1} \quad (3)$$

Equation 3 is derived under the assumption that the within-cluster, between-cluster, and conflated effects of the level-1 predictor are equal ($\gamma_{X_C} = \gamma_{X_B} = \gamma_{X_W}$). The derivation additionally assumes a balanced design with equal cluster sizes and homoscedastic level-1 errors, which simplify the expressions for the sampling variances but are not central to the conceptual result. Under this data-generating model, the within- and between-cluster slopes estimate the same underlying parameter, and the conflated estimator is therefore the optimal estimator of that common effect.

In this setting, the sampling variance of the conflated estimator reflects information from both the within-cluster and between-cluster components. Accordingly, Equation 3 shows that the reciprocal of the sampling variance of $\hat{\gamma}_{X_C}$ equals the sum of the reciprocals of the sampling variances of $\hat{\gamma}_{X_B}$ and $\hat{\gamma}_{X_W}$. This relationship formalizes the intuition that, in the context of $\gamma_{X_C} = \gamma_{X_B} = \gamma_{X_W}$, combining two independent sources of information about the same parameter yields greater sampling precision than either source alone.

Raudenbush and Bryk (2002) do not describe how Equation 3 behaves when the between- and within-level effects are not equal, because in that case the conflated slope no longer targets the same parameter as the within- and between-cluster coefficients and is generally biased. Nevertheless, the key implication of Equation 3 suggests that any precision advantage of the conflated estimator should be most pronounced when the contextual effect is small, in which case the variance relationship in Equation 3 may hold approximately. Yet, it is unknown whether this precision advantage in the covariate may extend to the focal predictor and compensate for the biases resulting from failing to disaggregate. In particular, any potential gain in precision at level-2 arises through the off-diagonal elements of the parameter covariance matrix, which are driven largely by the correlation between the level-1 and level-2 variables. Therefore, it is worthwhile to explore, in situations where the between- and within-estimates are equal in the covariate (or nearly so), whether the use of a single regression coefficient may optimize the power to detect an effect in the focal predictor.

The most authoritative paper on the disaggregation of covariates comes from Rights et al. (2020), who explored the consequences of failing to disaggregate a conflated level-1 covariate when the interest lies in a level-2 predictor. The researchers used the smushed and disaggregated random intercept models with a single level-2 predictor and a single level-2 covariate, in line with Equations 1 and 2 above. They focused on how four features influence bias of the level-2 slope: the magnitude of the difference between the within- and between-effects of the level-1 covariate, the sample size at each level, the correlation between level-2 variables (Z_j and $\overline{X}_{.j}$), and whether the focal predictor slope was zero or nonzero. Rights et al. (2020) convincingly demonstrated that there are negative consequences to *not* disaggregating. The take-home message from their article is that researchers should always disaggregate, perhaps selectively by choosing to control for the level-1, level-2 parts, or both. We don't disagree with this sentiment. However, several unanswered questions remain, particularly that of whether there is ever a downside to disaggregating covariates, given the potential loss of statistical power by disaggregating.

The purpose of this study is to explore both the costs and benefits of choosing to disaggregate level-1 covariates when interest lies with a level-2 predictor. As such, we extend the study from Rights et al. (2020) in several ways. One key extension of our work is that we include precision and power as outcomes. The rationale for this relates to Equation 3, which shows that precision may be optimized by keeping the covariate intact. Thus, even in cases where bias is expected, precision gains could counterbalance the bias-variance tradeoff and still lead to a more accurate estimate overall. Rights et al. (2020) did not systematically manipulate the size of the intraclass correlations (ICC), which is crucial to the issue of bias (Lüdtke et al., 2011). Using the population values reported in the study, we determined that the unconditional ICCs ranged from 0.25 to .57. Therefore, the lowest ICC in this work is higher than values that are typical of cross-sectional research (Baldwin et al., 2011; Glassman et al., 2015; Kivlighan et al., 2020; Masood & Reidpath, 2016; Sellström & Bremberg, 2006; Spybrook et al., 2011; Stockford, 2009). Therefore, another key extension of our work is to explicitly manipulate the ICC of the lower-level variables.

Another major extension of our work is that we distinguish between two ways of representing

between-cluster effects. Rights et al. (2020) defined $\overline{X}_{.j}$ in Equation 1 as the manifest mean in cluster j, which we refer to as the manifest aggregate approach. Whereas, the latent aggregate approach, as it is referred to in this paper, represents the unobserved cluster means as latent variables (Lüdtke et al., 2008, 2011). This distinction is important because the literature generally recommends the latent approach, as it tends to exhibit less bias than the manifest approach. However, prior work has shown that when bias does arise, specifically in the representation of the between-level component of a level-1 predictor, it can propagate to other between-level predictors in the model. In their simulation study, Zitzmann et al. (2016) demonstrated this by manipulating the degree of measurement error of a level-1 predictor and comparing latent versus manifest specifications: when the between-level effect of the predictor was biased, the estimated effects of other level-2 variables were also distorted, even though those variables themselves were measured without error. These methods for computing cluster means can have very different precision, especially when the ICC is low, or the sample size is small at either level. Whether these precision differences impact the focal slope at level-2 is an open question.

The purpose of this study is to further evaluate both the costs and benefits of disaggregating level-1 covariates when the primary inferential interest lies in a level-2 predictor. Building on and extending Rights et al. (2020), we use Monte Carlo simulations to identify the conditions under which disaggregation meaningfully improves estimation and when potential losses in precision or power might instead make a smushed approach more defensible. The remainder of the paper is organized as follows: we first describe the simulation design and analytic methods, then present results across the manipulated conditions. We next illustrate these issues using an applied example, and conclude with a discussion of implications, limitations, and directions for future research.

Methods

Monte Carlo data simulation was used to understand the effects of disaggregating covariates in multilevel models on the between-level focal predictor coefficient (γ_Z). We compared the performance of disaggregated and smushed models. Following Rights et al. (2020), we used a random intercept regression consisting of one level-2 focal predictor and a disaggregated level-1 covariate as the data-generating model, as represented by Equation 1.

Manipulated factors

We manipulated six between-subjects factors: the size of the contextual effect (i.e., the degree to which the level-specific effects differ) defined through a standardized metric that mimics Cohen's d benchmarks ($d = -0.5$, $-.2$, and 0.0); the R-squared effect size (Rights & Sterba, 2019) for within and between effects ($R^2_W = R^2_B = 0.02$ and 0.07); the intraclass correlation for both X and Y (ICC = 0.10, 0.20, and 0.50); number of clusters ($J = 50$, 100, and 200); number of units within each cluster ($n_j = 5$, 25, 50, and 100); and correlation between the level-2 focal predictor, Z, and the between-level portion of the covariate, $\overline{X}_{.j}$ ($\rho = 0.1$ and 0.3). Therefore, there are 864 combinations of between-group conditions. We use d throughout the paper to represent the size of the contextual effect on a standardized metric. These benchmarks are borrowed from Cohen's recommendations, although we use them differently. Finally, the simulation included one within-subject factor: the composition of the fitted model, which comprised a smushed model, a manifest aggregate model, and a latent aggregate model. Our online Open Science Framework (OSF) repository[2] includes more extreme conditions where $\rho = 0.5$ and $d = -0.8$.

The levels of these manipulated factors were meant to approximate values often encountered by social and behavioral researchers. The values of ICC condition were chosen to be representative of both cross-sectional (Baldwin et al., 2011; Kivlighan et al., 2020; Spybrook et al., 2011) and repeated measures designs (Bolger et al., 2013; Funder & Ozer, 2019; Singer & Willett, 2003). The myriad of combinations from the within- and between-level sample sizes were varied based on values representative of many applications of the social and behavioral sciences. Returning to Equation 1, we varied the correlation between the level-2 focal predictor, Z_j, and the between-level effect of the disaggregated covariate, $\overline{X}_{.j}$, to reflect effect size benchmarks (Cohen, 1988) and typical correlation values from published applications (Bosco et al., 2015; Funder & Ozer, 2019; Gignac & Szodorai, 2016).

Population parameters

The method described by Enders et al. (2025) was used to determine the multilevel population parameters. This approach uses expressions from Rights and Sterba (2019) that depend on the proportion of variance explained for the within-cluster predictors as well as the between-cluster predictors. The Rights and

Sterba (2019) equations have covariance matrices and multilevel model parameters as their inputs, and *R*-squared values are the outputs. In our approach, the *R*-squared values are the inputs, and multilevel model parameters are the outputs. Solving for the MLM parameters required the following inputs beyond the *R*-squared values: ICCs of all level-1 variables, means and variances of all variables, level-specific correlations among the predictors, and correlations among the random effects (Enders et al., 2025).

Interested readers may refer to the appendix in Enders et al. (2025) for a summary and their online supplementary material for additional information about the derivation of these formulas. The means of X and Z were both set to zero and each of their total variances were set to one. For the outcome variable Y, the mean was set to 50 and the variance was set equal to 100 (a larger outcome variance avoids fractional slopes that could be subject to rounding errors). These values were completely arbitrary. The within- and between-cluster variance in X and Y were then determined by the intraclass correlation. The formula for the population γ_{X_W} is

$$\gamma_{X_W} = \sqrt{\frac{\sigma_Y^2 \; R_W^2}{(1 - ICC_X)\sigma_X^2}} \tag{4}$$

where ICC_X is the intraclass correlation for the covariate, σ_Y^2 and σ_X^2 are the total variances described above, and R_W^2 is the proportion of the explained variation at level-1. Recall that R_W^2 has two levels and the ICC condition has three levels, resulting in six possible population values for γ_{X_W}.

As we define it, d was equivalent to standardizing γ_{X_B} and γ_{X_W} and then subtracting these standardized coefficients. Thus, γ_{X_B} coefficient was a function of the γ_{X_W} parameter and the size of the contextual effect condition, as defined by Cohen's d benchmarks ($d = -0.5, -0.2$, and 0.0).

$$\gamma_{X_B} = \gamma_{X_W} + d\frac{\sqrt{\sigma_Y^2}}{\sqrt{\sigma_X^2}} \tag{5}$$

When level-1 predictors have a nonzero ICC, a portion of the R_B^2 is attributable to their cluster means. While only level-1 predictors can contribute to within-level R_W^2, both the level-2 predictor Z and cluster aggregates $\overline{X}_{.j}$ contribute to the between-level model R_B^2.

$$\gamma_Z = \sqrt{\frac{\sigma_Y^2 \; R_B^2}{\sigma_Z^2 - \frac{(\sigma_{Z,X_B})^2}{ICC_X\sigma_X^2}}} \tag{6}$$

The σ^2 values represent the variances of designated portions of the model, and ICC_X is the intraclass correlation for the covariate. Additionally, σ_{Z,X_B} represents the covariance of the between-level predictors. Furthermore, the intercept for this model was set to be equal to μ_Y, which was preset to 50. The random intercept variance was determined by calculating the total between-level variance of Y then subtracting the portion of this variance that was explained by the between-level effects, as shown in Equation 7 below.

$$\sigma_{u_0}^2 = ICC_Y \times \sigma_Y^2 - \boldsymbol{\gamma}_B^T \; \boldsymbol{\Sigma}_B \; \boldsymbol{\gamma}_B \tag{7}$$

Where $\boldsymbol{\gamma}_B = \begin{pmatrix} \gamma_{X_B} \\ \gamma_Z \end{pmatrix}$, representing the vector of between-level effects, and $\boldsymbol{\Sigma}_B$ represents the covariance matrix for the between level effects. Likewise, the residual variance was determined by subtracting the within-cluster variance in Y that is explained by the within-cluster predictor from the total within-cluster variance of Y.

$$\sigma_\varepsilon^2 = (1 - ICC_Y) \times \sigma_Y^2 - \sigma_{X_W}^2 \times \gamma_{X_W} \tag{8}$$

Data generation and model fitting

After determining the population parameters, we utilized R version 4.2.2 to simulate data for our study. Specifically, we generated 2,000 independent samples for each of the 864 unique combinations of between-group design conditions. These combinations were defined by variations in key study parameters, ensuring a comprehensive assessment of how different design factors influence our results. Based on the population parameters, we first simulated the within-cluster deviations of the level-1 predictor. Next, we drew the cluster-level means of the level-1 predictor and the level-2 predictor from a bivariate normal distribution. The means were both 0, the variance of cluster means of X was determined by the ICC, the variance of Z was set at 1, and the covariance was a manipulated feature. The between cluster means of X were generated as a latent variable because it was treated as a random variable with its own distribution rather than deterministic feature of the level-1 scores. We then generated the random intercept residuals as well as the individual-level residuals (i.e., the u_{0j} and ε_{ij} terms). A predicted outcome score was generated using the simulated predictors, population slopes, and simulated random intercepts, and adding simulated within-cluster residuals completed simulation of the outcome.

We then used Markov Chain Monte Carlo (MCMC) estimation, a Bayesian approach, in Blimp version 3.2.1 (Keller & Enders, 2021) to fit the three model specifications. Blimp is a freely available latent-

variable modeling program that uses the flexible power of Bayesian estimation in a user-friendly application that requires minimal scripting and no deep-level knowledge about Bayesian statistics. An initial pilot simulation comparing MCMC to a maximum likelihood within lavaan revealed that the difference between these estimators was trivial. Thus, as the results from each were comparable, we opted to use MCMC within our simulation as recent advancements in the multilevel model literature (i.e., the modeling of interaction effects, dynamic lagged effects, and missing data) rely on MCMC estimation. Although the models we explore within our simulation do not necessitate this specific estimation method, we argue that MCMC connects better to these recent innovations in the multilevel model literature. Another major reason to employ MCMC is that Bayesian estimation has been shown to stabilize parameter estimates (Asparouhov & Muthén, 2021b; Zitzmann et al., 2016).

The number of burn-in iterations and the number of iterations after the burn-in period (5,000 and 10,000, respectively) were determined from a preliminary simulation that collected the potential scale reduction (Gelman & Rubin, 1992) and number of effective MCMC samples (Gelman et al., 2014) diagnostics from a small number of replications in selected design cells. Finally, to mimic how researchers would typically use the software, we adopted the default diffuse prior distributions described in Enders et al. (2020) and Keller and Enders (2021). The posterior medians and standard deviations served as the point estimates and 'Bayesian standard errors', respectively. Given the diffuse priors described above, these quantities are numerically equivalent to the maximum-likelihood estimates and their standard errors.

Outcomes

We evaluated the following features of the focal level-2 regression slope, γ_Z: percent bias, mean square error, and power. Percent bias is the difference between an average estimate and its true parameter value expressed as a percentage of the true value, $(\bar{\hat{\gamma}}_Z - \gamma_Z)/\gamma_Z \times 100$. Mean square error (MSE), a measure of the overall accuracy that combines squared bias with sampling variance, is the averaged squared difference between the parameter value and the estimates. To facilitate interpretation, we calculated MSE ratios by dividing the MSE for the disaggregated models by the MSE of the reference model, which was the smushed model. Ratios greater than 1.0 indicate lower accuracy for the disaggregated models, whereas ratios below 1.0 indicate lower accuracy for the smushed model. Finally, power is the probability of rejecting a false null hypothesis (Aberson, 2015). Power was estimated by determining the proportion of replications in each design cell where the estimate was statistically significant at an alpha level of 0.05, taking advantage of the Bayesian Wald statistics. Some employ MCMC estimation due to their general preference for the Bayesian theoretical framework, while others take advantage of MCMC's flexibility in estimation and maintain a frequentist stance, referred to as a computational frequentist perspective (Levy & McNeish, 2023). Although referred to as Bayesian Wald statistics, the test itself is not Bayesian: the procedure simply plugs the MCMC-generated point estimates and corresponding 'Bayesian standard errors' into the frequentist Wald formula to obtain a p-value from the central χ^2 distribution. Blimp takes on each of these perspectives and includes Bayesian Wald statistics within its output (Asparouhov & Muthén, 2021a).

Data analysis procedure

We conducted a series of factorial analyses of variance (ANOVAs) to identify the most influential sources of variation for the percent bias outcome. These analyses enabled us to assess which factors were most impactful, guiding us in simplifying the full factorial design and presenting the results more clearly in tabular form. For each outcome, we fit three models: the first model examined the main effects of each condition, the second model included all possible two-way interactions, and the third extended to include all possible three-way interactions. To quantify the influence of each condition and interaction, effect sizes were calculated by dividing the sum of squares for the effect by the total sum of squares. Conditions and interactions with effect sizes greater than 0.01 were considered salient and informed the development of our graphical displays, ensuring that only the most meaningful patterns were highlighted.

Results

The results section is organized into five subsections, each examining key interactions affecting estimation accuracy and power. The first two subsections assess the three outcomes in relation to a three-way interaction between model type, level-2 correlation, and contextual effect size. The next two subsections explore the three outcomes as a function of a three-

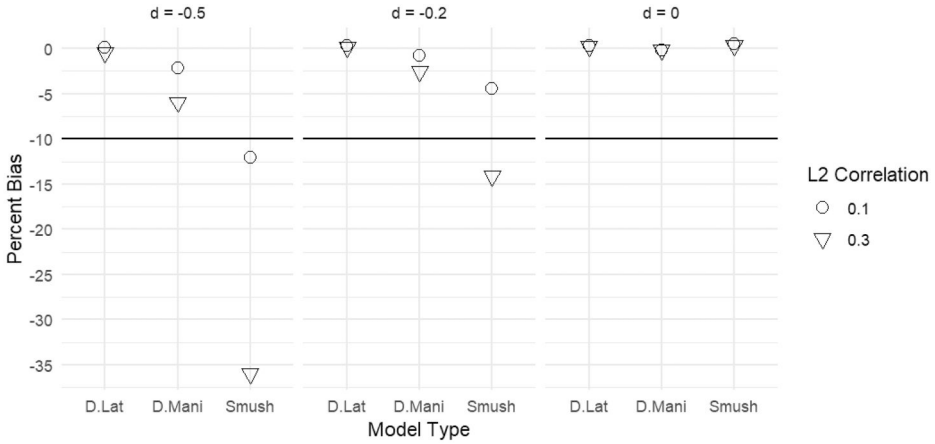


Figure 1. Plot displays the interaction between model type, between-level correlation, and contextual effect size in explaining percent bias. "D.Mani" refers to the manifest aggregate model, "Smush" refers to the smushed/conflated model, and "D.Lat" refers to the latent aggregate model. The solid black line marks the 10% bias cut off.

way interaction between model type, intraclass correlation, and contextual effect size.

Three-way interaction between model type, level-2 correlation, & contextual effect

Percent bias

Rights et al. (2020) determined that the correlation between level-2 part of the covariate and focal predictor, the degree to which the covariates level-specific effects differed, and choice of model specification were most impactful on percent bias of the focal estimate. The results of the simulation were consistent with that of the Rights paper by indicating that the interaction between model type, between-level correlation, and the size of the contextual effect accounted for about 4% of the variance in percent bias. Figure 1 displays how percent bias of the focal predictor slope changed as a function of the interaction of these conditions. Each facet of Figure 1 displays the two-way interaction between the between-level correlation and model type for a specific level of the contextual effect condition. The solid lines in Figure 1 represent the 10% percent bias, which is commonly used as a cutoff for an "acceptable" amount of bias.

Figure 1 shows that the degree of percent bias was fairly consistent across both disaggregated models (as both remained under the 10% threshold), but was far more pronounced in the smushed models. Based on bias alone, the smushed model appeared to be appropriate for use when the contextual effect size was −0.20, and the correlation was .10. Consistent with Rights et al. (2020), bias increased as the correlation between level-2 predictors increased. The trend described in the previous sentence was minimal for the disaggregated models but was far more pronounced in the smushed model. The size of this two-way effect of model type and between-level correlation depends on the size of the contextual effect. When the contextual effect was zero, there was virtually no bias (the maximum value was less than one half of one percent). As the size of the contextual effect increased, the impact of increasing the between-level correlation was exacerbated across each model type, especially for the smushed model. The figure suggests that even a weak correlation between level-2 predictors can be problematic when contextual effect difference is moderate, with the smushed model reaching −12% bias in this context. In contrast, the latent aggregate model showed minimal bias across all conditions, with the highest level of bias equal to −0.44%. Relating these findings back to our main research question, the smushed model may still be appropriate when the contextual effect size and the between-level correlation are small, but as these values increase, disaggregation is preferred. In regard to our secondary question, the latent aggregate model tended to perform better than the manifest aggregate model with larger contextual effect sizes and larger between-level correlations.

Mean square error

Bias doesn't tell the entire story because the smushed and disaggregated models had different precision; when the contextual effect is zero, analytic expressions predict that the smushed model had the smallest variance (Raudenbush & Bryk, 2002, Equation 5.39), but it is unknown whether this precision advantage in the covariate may extend to the focal predictor and compensate for the biases from the previous section. One of the major goals of this study was to investigate

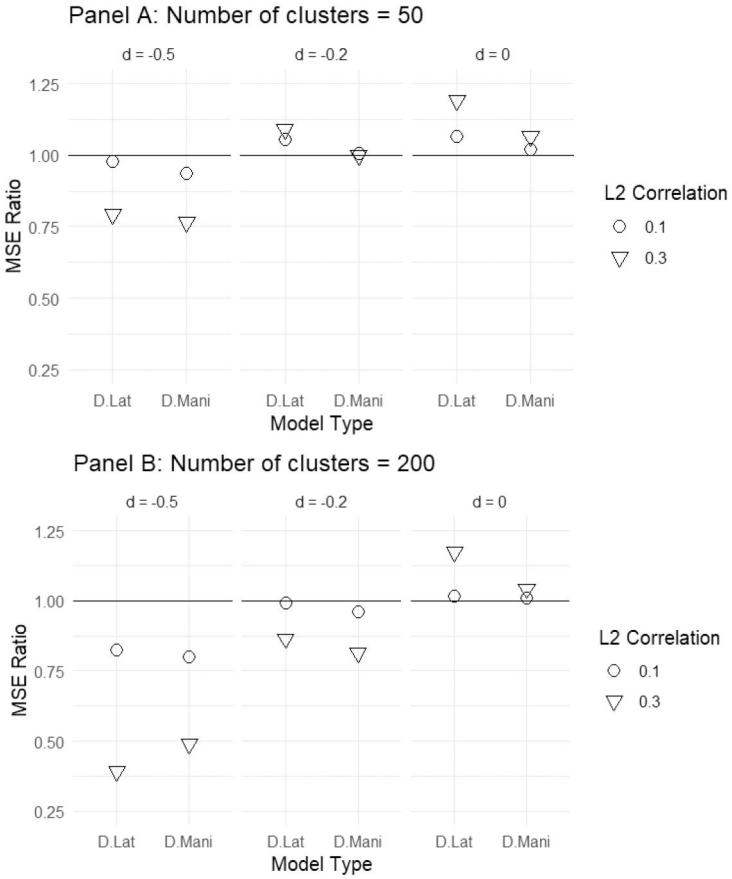


Figure 2. Plot displays the interaction between model type, between-level correlation, and contextual effect size in explaining MSE. Values on this plot represent the MSE ratio with the smushed model as the reference model, which is why the points for this model are stable at 1.0. "D.Mani" refers to the manifest aggregate model, and "D.Lat" refers to the latent aggregate model. The solid line indicates the point where the MSE ratio for the smushed and comparative model are equivalent.

whether precision differences could make the smushed model preferable to disaggregation.

Figure 2 below displays the MSEs ratios for the same combination of effects as Figure 1. Because the MSE of the level-2 slope depends on the number of level-2 units, the figure displays ratios for 50 and 200 clusters. Panel A of this figure shows how the MSE ratios changed for the smallest number of clusters in this study, whereas panel B shows this relationship for conditions with the largest number of clusters. The size of within-cluster units was held constant at 50, as there were no differences in the MSE ratio of the focal predictor slope across this condition. Recall that MSE ratios were calculated relative to the smushed model; MSE ratios larger than 1.0 indicate that the disaggregated estimates were less accurate. As well, MSE ratio values less than 1.0 indicate that the smushed model estimates were less accurate.

First, we tried to understand the difference in performance between the smushed model and the two disaggregated models where there is no contextual effect. Both panels of Figure 2 show this to be the case, and there were not any meaningful differences across level-2 sample sizes. Recall from Figure 1 that all models were free of bias in these conditions. Thus, these differences in MSE reflect sampling error alone. In evaluating these MSE ratios, we viewed a value exceeding 1.25 as a substantial difference between the smushed model and the comparative model. In the context of a single-level model, an MSE ratio of 1.25 would suggest that we would need to increase our sample size by 25% to achieve the precision of the reference model. Although the MSE ratio's relation to sample size is more complex in a multilevel model, 1.25 roughly implies a 25% increase in the number of level-2 clusters.

One of the major goals for this paper was to examine whether the type of level-2 group mean (manifest or latent) impacts disaggregation. The contextual effects literature suggests that latent group means are unbiased, but noisy (Lüdtke et al., 2008). In our simulation, this imprecision also propagated to the level-2 focal slope, where MSE ratios were close to 1.20 in both level-2 sample size conditions. In contrast, the

largest MSE ratio for the manifest means was 1.07. Although this value is not very large, it could be practically significant because a 0.10 increase in MSE could translate into significantly more resources needed to increase the level-2 sample size. As a quality control check on these MSEs, we compared the empirical standard deviation of the estimates across iterations to the average of the estimated standard deviations (Bayesian standard errors). The two estimates of precision were virtually indistinguishable, which suggests that the smushed model was not benefiting from a systematic distortion in its variability. As a part of the supplemental material, we provided two tables (one for the smushed model and one for the disaggregated models) comparing the average posterior standard deviations and the standard deviation of the posterior medians is available in the supplemental material for this study.

Next, we evaluated the MSE for nonzero contextual effect sizes, in which case ratios were driven by bias and sampling error. When the level-2 sample was small ($j = 50$) and the contextual effect size was at its smallest ($d = -0.20$), the smushed model was equally as accurate or very slightly more accurate than the disaggregated models. There were slight differences in the MSE ratios across the different levels of the between-level correlations and the methods of disaggregation, but the MSE ratios remained well under 1.25. When the level-2 sample size was large ($j = 200$) and the between-level correlation was 0.10, the smushed model was nearly as precise as the disaggregated model with manifest means (MSE ratio was 0.96) as well as with the latent aggregates (MSE ratio was 0.99). However, the disaggregated models were clearly advantageous in conditions where the between-level correlation was 0.30, with MSE ratios equal to 0.81 for the manifest aggregate model and 0.86 for the latent aggregate model. When the contextual effect size was −0.50, the disaggregated models were uniformly more precise than the smushed model. This trend was especially pronounced with 200 clusters. The MSE ratios for the manifest aggregate model was as low as 0.49, and those for the latent aggregate model were as low as 0.39. Overall, the bias associated with larger contextual effects substantially outweighed any precision gains associated with the smushed model.

As a secondary question, we wanted to determine which of the disaggregated models—manifest or latent group means—performed better. Lüdtke et al. (2008) showed that there were large differences in the precision of the between-level slope of the level-1 predictor

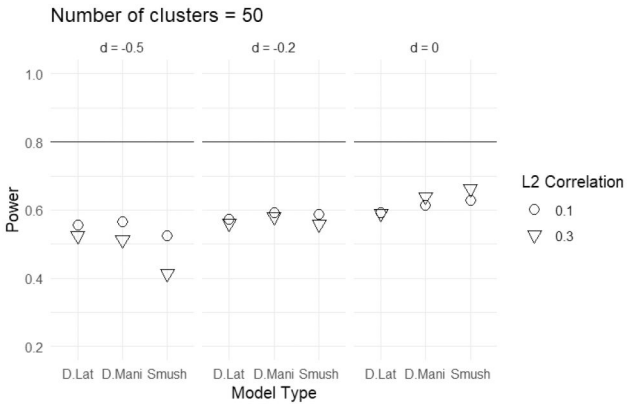


Figure 3. Plot displays the interaction between model type, between-level correlation, and contextual effect size in explaining power to detect the focal predictor. "Smush" refers to the smushed/conflated model, "D.Mani" refers to the manifest aggregate model, and "D.Lat" refers to the latent aggregate model. The solid black line marks the point where the average estimate achieves 80% power.

between the manifest and latent aggregate models. However, our study showed that these differences did not propagate to the level-2 focal slope, except when the between-level correlation was larger. The difference between these models was most pronounced when the contextual effect size was −0.20, although this difference was still small. With a between-level correlation of 0.30, the manifest aggregate model produced an MSE ratio of 0.82 while the latent aggregate model produced a ratio of 0.86. This is expected give that, in situations where the latent group means are very noisy, the focal slope inherits this noise when it is most strongly correlated with the latent means. Next, we considered how the MSE ratio changes as the with as the number of clusters increases. Unlike with the small level-2 sample size, the difference between the two disaggregated models diminished when the contextual effect was small, as can be seen in the center of panel B. For example, when the contextual effect and the between-level correlation were at their smallest, the latent aggregate model produced an MSE ratio of 0.86 and the manifest aggregate model produced an MSE ratio of 0.81. In general, as the number of clusters increased, the latent aggregate model was preferable, especially considering that the latent aggregate model exhibited slightly less bias than the manifest aggregate model.

Power

Next, we explored how the choice to disaggregate level-1 covariates impacts the power to detect an effect of the level-2 focal predictor slope. We again consider the same three-way interaction between model

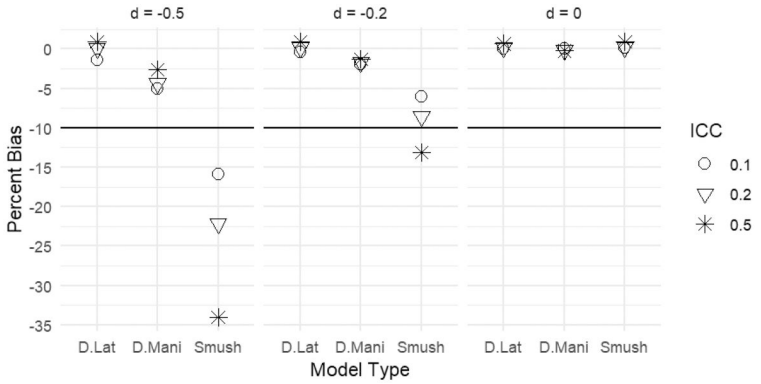


Figure 4. Plot displays the interaction between model type, between-level correlation, and contextual effect size in explaining percent bias. "D.Mani" refers to the manifest aggregate model, "Smush" refers to the smushed/conflated model, and "D.Lat" refers to the latent aggregate model. The solid black line marks the 10% bias cut off.

specification, contextual effect size, and the level-2 correlation. Figure 3 shows this interaction, holding the number of clusters constant at 50, as power quickly approached one with increased level-2 sample size. First, we evaluated the differences between the smushed and the disaggregated models. Recall that when the contextual effect was zero, there was no bias across the three model specifications, as shown in Figure 1. Figure 2a showed that, with a small sample size, the smushed model had a slight advantage in accuracy compared to the manifest aggregate model, but the latent model was less precise. Figure 3 shows that the smushed model had a very slight power advantage relative to the manifest model for the zero contextual effect condition, with the greatest difference occurring with the highest between-level correlation (0.64 for the manifest model and 0.66 for the smushed). The difference between the latent aggregate and smushed models was more distinct, with the largest difference also occurring with the highest between-level correlation (0.59 for the latent model).

When the contextual effect was −0.20, the smushed model's advantage began to disappear, as the negative bias pushed the coefficient closer to zero. This was a feature of our design rather than a substantive finding; if we had made Z's slope negative, the bias direction may have changed, and power could have increased. Regardless, the difference between the smushed and manifest aggregate model was negligible for the low between-level correlation, as both were about .59. The manifest approach performed minimally better for the largest between-level correlation (0.58 vs. 0.56 for the manifest and smushed models, respectively). Additionally, the difference between the latent and the smushed models was dependent on the corresponding between-level correlation value at this level of the contextual effect. Both these models achieved power of about 0.56 when the between-level correlation was 0.10, but slight differences were observed when the between-level correlation was 0.30 (0.57 for the latent model and 0.59 for the smushed model). When the $d = -0.50$, there was substantial bias in the smushed model, eliminating this model specification as a realistic option in these circumstances. Thus, we do not discuss the difference in power between the smushed and the disaggregated models at this level of the contextual effect.

In relation to our second research question, the manifest and latent approach generally tended to be consistent with one another. The largest difference between the two specifications occurred when the contextual effect was zero and the between-level correlation was 0.30, where the manifest aggregate model achieved 0.63 power and the latent aggregate model achieved power of .59. This was fairly consistent when the contextual effect was zero and the between-level correlation was .10. Otherwise, the power differences were minimal. For instance, the manifest aggregate model achieved 0.51 power when the contextual effect was −0.50 and the between-level correlation was 0.30 whereas the latent aggregate model achieved 0.52 power in this set of conditions. Thus, the manifest model may be more appropriate in certain conditions, but power differences were generally slight.

Three-way interaction between model type, intraclass correlation, & contextual effect

Percent bias

The modeling procedure described above indicated that the interaction between model type, the size of the contextual effect, and ICC accounted for about 1.3% of the variance in percent bias of the focal predictor slope. Each facet of Figure 4 displays the two-

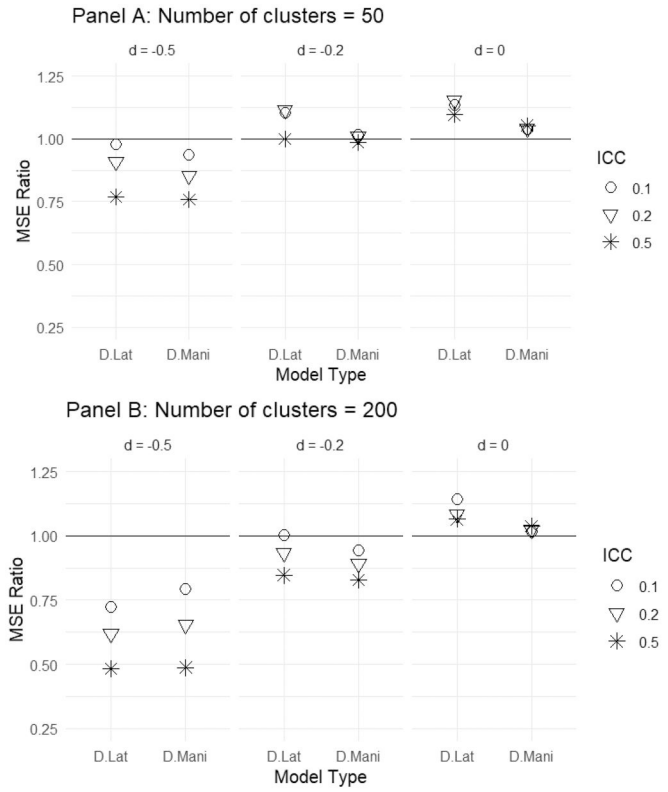


Figure 5. Plot displays the interaction between model type, ICC, and contextual effect size in explaining MSE. Values on this plot represent the MSE ratio with the smushed model as the reference model, which is why the points for this model are stable at 1.0. "D.Mani" refers to the manifest aggregate model, and "D.Lat" refers to the latent aggregate model. MSE ratio values greater than 2.0 were excluded from the figure. The solid line indicates the point where the MSE ratio for the smushed and comparative model are equivalent.

way interaction between ICC and model type for a specific level of the contextual effect condition. The solid lines in Figure 4 again represent the 10% percent bias, which is commonly used as a cutoff for an "acceptable" amount of bias. Figure 4 shows the degree of percent bias was generally consistent across both disaggregated models but was more pronounced in the smushed models, especially as the contextual effect increased in magnitude. As the size of the contextual effect increased, the relationship between ICC and model type changed noticeably.

When the ICC is 0.10, bias within the smushed model was within acceptable range when $d = -0.20$. It was only at $d = -0.50$ that bias became extreme, where the percent bias in the smushed model was −15.91%. Conversely, when ICC is 0.50, bias was problematic at $d = -0.20$. Figure 4 shows that the smushed model could result in biased estimates even when the size of the contextual effect was small ($d = -0.20$), as shown by the average percent bias of −13.11% when the ICC was .50. Relating these findings back to our main research question, the smushed model may still be appropriate when the contextual effect size and the ICC are small, but as these values increase, disaggregation is preferred. Regarding our secondary question, the latent aggregate model tended to perform better with larger contextual effect sizes. When the contextual effect was −0.50, the manifest model resulted in −5.06% bias when the ICC was 0.10, whereas an ICC of 0.50 for this model resulted in only −2.64% bias. There was virtually no bias in the latent aggregate model across the various sizes of the contextual effect, with the highest degree of bias (−1.40%) corresponding to when the ICC was at its largest and the contextual effect size was −0.50.

Mean square error

Figure 5 below displays the MSEs ratios for the same combination of effects as Figure 4. Because MSE of the level-2 slope depends on the number of level-2 units, the figure displays ratios for 50 and 200 clusters in panels A and B, respectively. The size of within-cluster units was held constant at 5 as there were no differences in the MSE ratio of the focal predictor slope across the within sample size condition.

Model-level MSE patterns are driven by the covariate's intraclass correlation (ICC), and the strength of that ICC interacts with the magnitude of the

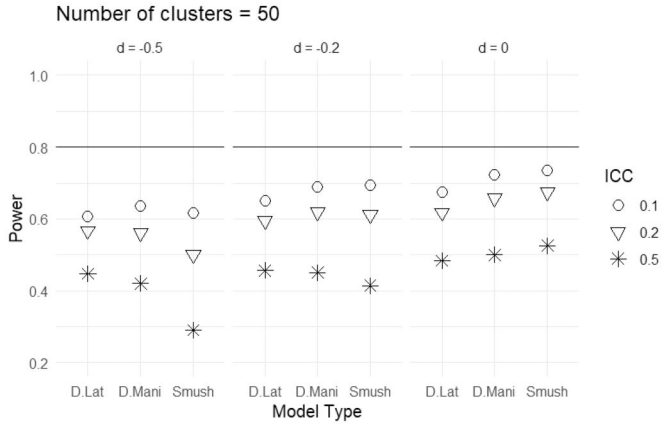


Figure 6. Plot displays the interaction between model type, ICC, and contextual effect size in explaining power to detect the focal predictor. "D.Mani" refers to the manifest aggregate model, "Smush" refers to the smushed/conflated model, and "D.Lat" refers to the latent aggregate model. The solid black line marks the point at which the average estimate achieves 80% power.

contextual effect, as predicted. When the contextual effect is absent, any gap between a disaggregated model and the smushed benchmark is due to additional sampling variability as there is no bias across the three model specifications. Once again, the comparison between the average posterior standard deviation and the standard deviation of the posterior medians ruled out the possibility that the smushed model performed better because it is underrepresenting the sampling variability. Across both panels of Figure 5, ratios roughly fall between 1.00 and 1.15 regardless of the between-level sample size. The MSE ratios is most pronounced when the ICC is smallest, as the ratios for the latent aggregate model with ICC of 0.10 were 1.13 and 1.14 when the between-level sample was 50 and 200, respectively. As well, for the manifest aggregate model with an ICC of 0.10, the MSE ratios were 1.03 and 1.01 when the between-level sample was 50 and 200, respectively. Differences in MSE ratio across ICC levels were minimal across the manifest aggregate model, regardless of sample size. However, differences across ICC for the latent aggregate model across sample size. For instance, the latent aggregate model with an ICC of 0.50 had 1.09 and 1.07 when the between-level sample was 50 and 200, respectively.

Once a modest contextual effect ($d = -0.20$) was introduced, bias becomes influential and its influence depends on the ICC. At the lowest ICC of 0.10 the smushed and manifest-aggregate models were almost indistinguishable as MSE ratios hovered around 1.0, whereas the latent-aggregate model was slightly less precise with a smaller between-level sample size (MSE ratio = 1.10). As ICC climbed to 0.50 the latent aggregate MSE ratio decreased to 1.00 and the manifest aggregate MSE ratio fell to 0.98, signaling a marginal precision gain. These patterns are visible in both sample-size panels but are most obvious when the between-level sample size was 50, where finite-sample noise amplifies small bias differences. With a stronger contextual effect, the role of ICC became more decisive. Both disaggregated specifications outperformed the smushed model across every sample size, and the advantage grows monotonically with ICC: at ICC of.10 and small between-level sample size, the manifest and latent aggregate ratios sit near 0.94 and 0.98, respectively; by ICC =0.50 they drop to roughly 0.76 and 0.77. The gains were especially striking with the large between-level sample, reflecting the fact that a larger number of clusters let the disaggregated models capitalized on the contextual information while the smushed model remains biased.

Differences between the manifest and latent aggregate models are modest compared with their shared contrast to the smushed model specification. When $d = 0$, the latent version pays a small sampling-error penalty, especially at low ICC, pushing its ratio slightly above that of the manifest aggregate model. Once the contextual effect size departs from zero, however, the latent aggregate model quickly closes the gap and eventually becomes the most precise option when the ICC is 0.50 and the contextual effect is large ($d = -0.50$). In short, lower ICC values favor the manifest-aggregate slope, higher ICC values and stronger contextual effects favor the latent-aggregate slope, and both disaggregated strategies dominate the smushed model whenever a non-trivial contextual effect is present.

Power

Finally, we explored how the choice to disaggregate level-1 covariates impacts the power to detect an effect

in the level-2 focal predictor slope in the context of the three-way interaction between model specification, contextual effect size, and ICC. Figure 6 shows this interaction, holding the number of clusters constant at 50 as power quickly approached one as the level-2 sample size increased. First, we evaluate the differences between the smushed and the disaggregated models. Recall that when there was no difference in the level-specific contextual effects, there was no bias across the three model specifications as shown in Figure 4. Figure 5a showed that, with a small sample size, the smushed model had a slight advantage in accuracy compared to the manifest aggregate model and the latent model had more sampling variance. Figure 6 indicates that the smushed model had a very slight power advantage relative to the manifest model, with the greatest difference occurring at the highest ICC value (0.50 for the manifest model and 0.53 for the smushed). The difference between the latent aggregate and the smushed models was more distinct, with the largest difference occurring with the lowest ICC (0.67 for the latent model and 0.74 for the smushed). Relating these findings back to our main research question, disaggregation may still be appropriate when the contextual effect size is zero as long as the manifest approach is employed instead of the latent.

When the contextual effect was small ($d = -0.20$), the slight advantage of the smushed model began to disappear as the negative bias is pushing the coefficient closer to zero. Again, this is an arbitrary feature of our design rather than a substantive finding. Regardless, the difference between the smushed and manifest aggregate model was negligible for low ICCs, both were about 0.69 when the ICC was 0.10, and the manifest approach performed somewhat better for large ICCs (0.45 vs. 0.41). Additionally, the difference between the latent and the smushed model was dependent on the corresponding ICC value at this level of the contextual effect. When $d = -0.50$, there was substantial bias in the smushed model, eliminating this model specification as a realistic option in these circumstances. Thus, we do not discuss the difference in power between the smushed and the disaggregated models at this level of the contextual effect.

In relation to our second research question, the differences between the manifest and latent approach tended to be most prominent at the more extreme levels of the contextual effect size and ICC conditions. The manifest model performed better than the latent for zero to small contextual effects as well as small ICCs. For example, the manifest and latent models achieved power levels of 0.72 and 0.67, respectively, when the contextual effect was zero and the ICC of the covariate was .10. However, as the ICC increased to 0.50, the manifest model achieved 0.50 in power and the latent model achieved .48. This trend of higher ICCs relating to lower power was consistent for larger contextual effect sizes. For instance, when the contextual effect was −0.50, the manifest and latent models achieved power levels of 0.64 and 0.61 for ICC of 0.10 and achieved power of 0.42 and 0.45 when the ICC was .50. Thus, there appears to be slight power differences between the two disaggregated models. The latent approach performs somewhat better than the other specifications when the ICC and the contextual effect size are relatively large. Thus, the latent model may be appropriate based on its power behavior in some cases, but power differences are generally slight.

Applied example

In this section, we employ data from substantive research to exemplify the statistical concepts and methodologies discussed throughout this paper. The illustrative dataset comes from the Adolescent Medicine Trials Network for HIV/AIDS Interventions (ATN), comprising 66 daily measurements from 61 HIV-infected men between the ages of 16 and 24 (Cherenack et al., 2016) In a multilevel context, daily measurements are at the lower level and the between-level represents individual differences. We relied on a hypothetical research scenario where we are interested in the influence of baseline HIV stressors (*HIV*) on individual's daily depression (*DEP*) controlling for the degree of daily stressors (*DAILY*). Here, we are interested in controlling for the person-level effect of daily stress and the daily-level effect of this variable is substantively interesting. The disaggregated equation for this illustrative example is included below. The three versions of the model were fit in Blimp, and the results for each appear in Table 1.

$$DEP_{ij} = \gamma_{00} + \gamma_Z HIV_j + \gamma_{X_B}\overline{DAILY}_{.j} + \gamma_{X_W}\left(DAILY_{ij} - \overline{DAILY}_{.j}\right) + u_{0j} + \varepsilon_{ij} \quad (7)$$

In an intensive repeated measures design like this, researchers are increasingly interested in autoregressive effects involving lagged predictors (dynamic structural equation models; DSEM; McNeish & Hamaker, 2020). Blimp's DSEM functionality would accommodate lagged effects for depression or daily stressors (or both). However, because the only way to get unbiased estimates of dynamic effects is to center

Table 1. Comparison of results from each model specification.

	Smushed					
	Est.	SD	95% CI		χ^2	*p*-value
			LCL	UCL		
Intercept	5.04	0.68	3.78	6.40	55.34	< .001
Daily Stress	0.42	0.03	0.36	0.47	197.33	< .001
HIV Stress	0.18	0.09	−0.01	0.36	3.47	.063
Intercept Variance	3.92	0.80	2.72	5.84		
Residual Variance	6.34	0.14	6.08	6.64		
	Latent Aggregate					
	Est.	SD	95% CI		χ^2	*p*-value
			LCL	UCL		
Intercept	5.36	0.65	4.04	6.59	67.99	< .001
Daily Stress (*within*)	0.41	0.03	0.35	0.47	183.80	< .001
Daily Stress (*between*)	0.66	0.28	0.07	1.20	5.61	.018
HIV Stress	0.14	0.09	−0.04	0.32	2.34	.126
Intercept Variance	3.90	0.81	2.71	5.86		
Residual Variance	6.35	0.14	6.08	6.63		
	Manifest Aggregate					
	Est.	SD	95% CI		χ^2	*p*-value
			LCL	UCL		
Intercept	4.89	0.71	3.41	6.29	47.59	< .001
Daily Stress (*within*)	0.41	0.03	0.35	0.47	185.42	< .001
Daily Stress (*between*)	0.69	0.25	0.16	1.16	7.39	.007
HIV Stress	0.13	0.10	−0.06	0.32	1.90	.168
Intercept Variance	3.90	0.81	2.72	5.91		
Residual Variance	6.34	0.14	6.08	6.63		

lagged predictors at their latent group means (Zyphur et al., 2020), including such effects here could predispose methods to behave differently for reasons other than those considered in the simulation. To facilitate comparisons, we thus use the same models as Equations 1 and 2, acknowledging that unmodeled carryover effects could distort all three sets of estimates (Fleiss, 1989; Senn, 1992). We point interested readers to Enders and Keller (2025) for a DSEM example that uses latent group means for disaggregation.

Before discussing the example, it is important to note that, since we use Bayesian estimation, the table includes MCMC standard deviations rather than frequentist standard errors. As well, we report the Bayesian Wald statistics from the Blimp output (i.e., χ^2 in Table 1). These MCMC-generated quantities are frequentist significance tests. We include these for researchers who adopt a computational frequentism perspective to use MCMC for its computational advantages rather than favoring its philosophical framework (Levy & McNeish, 2023). These are the tests we used to compute power within the simulation.

This example highlights that the choice of specification of the covariate does make a difference in the interpretation of the focal slope (e.g., HIV-related stressors in this example). The largest difference across focal slope estimates is about 0.05, representing about half of a "Bayesian standard error" unit corresponding to the effect in the smushed model. The effect of HIV stress is closest to significance in the smushed model, as the lower confidence interval was −0.01 opposed to −0.04 or −0.06. Furthermore, the effect of HIV stress for the smushed model exhibited the smallest *p*-value and the largest χ^2 statistic. However, it's important to note that fitting these models to a single sample does not allow us to determine which version is the most "correct." That said, the fact that the HIV stress slope in the smushed model is near the significance threshold, while the others are not, highlights the potential impact of analytic choices on results and reinforces the need for researchers to carefully consider their modeling decisions.

The difference in significance of the focal predictor slopes is not attributable to differences in the posterior standard deviations of the HIV stress slope, as this value only differed by 0.01 across the three versions of the model. Rights et al. (2020) demonstrated that the conflated slope is a weighted average of the within-cluster and between-cluster slopes of the level-1 predictor, and that this weight depends on how much variance in the level-1 predictor lies at level-2, as a higher ICC pulls the conflated slope closer to the between-cluster slope. They also derive an expression for the level-2 bias, demonstrating that the direction and magnitude of the bias depend on two components: (1) the difference between the true between-cluster slope and the conflated slope, and (2) the covariance/correlation between the cluster means and the level-2 predictor.

In our example, we estimated the difference between the level-specific effects of daily stress and found the contextual effect to be small ($d = 0.13$) and the ICC of the level-1 covariate was .39. Thus, the between-cluster slope was larger than the within-cluster slope, the conflated slope must fall between them and therefore ends up smaller than the between-cluster slope, which is what we observed in this example. Additionally, the correlation between individuals' average level of daily stress and their baseline HIV stress was $r = 0.31$, which aligns with the simulation conditions. Since the cluster means of the level-1 predictor and the level-2 predictor are positively correlated, combining this positive correlation with the fact that the conflated slope is too small produces a positive bias in the level-2 slope. In practical terms, the level-2 effect will appear stronger than it truly is. Again, we caution readers from attempting to verify the data-generating process from fitting a single

sample to a set of models. However, it is worth emphasizing that the slope difference, ICC, and correlation provide the necessary conditions for positive bias (and thus a *p*-value that is moving in the direction of significance).

Discussion

Rights et al. (2020) conducted an authoritative study on the disaggregation of covariates in multilevel models (MLMs), examining the consequences of failing to disaggregate level-1 covariates when interest lies in a level-2 predictor. While the researchers concluded that disaggregating covariates is essential to avoid biased estimates, the question of whether there are potential downsides of disaggregating covariates remained. Our study shared much in common with the Rights paper, and we replicated their main findings. However, we attempted to clarify the conditions under which bias occurs by systematically varying the ICC and the size of the contextual effect (both key determinants of bias). Our study extended Rights et al. (2020) in three important ways. First, the smushed model is known to exhibit the highest precision (smallest standard error) in some cases. Thus, a major goal of our study was to determine whether there is a bias-variance tradeoff that favors the smushed model, even when its estimates are biased. Second, the level-2 group means play a vital role in disaggregation. Recent methodological literature seems to favor latent group means over arithmetic group means. We investigated both forms of disaggregation. Finally, we sought to document power differences between the disaggregated and smushed model.

The first question we sought to address was whether the disaggregated models were always more appropriate when accounting for level-1 covariates or whether the smushed model could exhibit acceptable performance in certain circumstances. The results of this simulation did not provide a definitive answer regarding whether researchers should always disaggregate their covariates in MLMs or not. Rather, the results indicated that this decision is more nuanced and dependent on certain features of the data. Consistent with Equation 3 (Raudenbush & Bryk, 2002, Equation 5.39), the smushed model slope had a lower MSE and higher power when there was no contextual effect (the within and between slope were the same). However, these precision differences tended to translate into very subtle power differences as, when the contextual effect size was zero, the typical power increase was from the disaggregated models to the smushed model was still less than 2%. For example, the MSE ratio was 1.19 when $J = 100$, $n_j = 50$, $R^2_W = R^2_B = 0.02$, $\rho = 0.30$, and ICC = 0.20, which would suggest that there is potentially a sizable gain in precision from using the smushed model. Yet, even in this context, the largest power difference was only about 3%, and the added precision did not necessarily translate into higher power. This slight precision advantage extended to select scenarios with a small contextual effect size, but it mostly vanished due to the influx of bias. In general, there was not a meaningful variance-bias tradeoff in most cases beyond those predicted by Equation 3.

Although the smushed model exhibits desirable properties when there is no contextual effect, it was also preferable in some cases where the contextual effect was small. This finding gives us pause to conclude that disaggregating is always statistically superior. The smushed model can produce accurate estimates when the contextual effect size, the correlation between the level-2 components, and the ICC of the covariate are relatively small. To reiterate, in cases where the smushed model offered a precision advantage, that advantage typically translated into only modest gains in power. Because researchers are unlikely to assess this specific constellation of conditions, it is reasonable to conclude that the smushed model's limited benefits do not outweigh its broader risks across situations in which precision gains fail to offset bias.

We next sought to address whether there were substantial differences in the latent and manifest approaches when disaggregating the level-1 covariate. The literature largely favors the latent aggregate approach over manifest cluster means (Asparouhov & Muthén, 2019; Lüdtke et al., 2008; 2011). However, it was not clear from previous work whether the purported advantages of latent group mean would propagate to the level-2 focal slope when the latent means were serving as a covariate. Our study showed that key differences in the degree of bias present in the focal slope only occurred when the covariate's contextual effect was large. These differences became more pronounced as the size of the level-2 correlation increased and as the size of the covariate's ICC decreased, in line with Lüdtke et al. (2008), with the latent aggregate model outperforming the manifest aggregate model. Yet, the latent approach can be noisy when the contextual effect is small, particularly when the ICC of level-1 covariate is below 0.20 or when there are a small number of clusters. Thus, we generally recommend the use of the latent aggregate model

when the contextual effect is large and the manifest approach when the contextual effect is small.

There are several notable limitations in this study. We focused on what some fields may consider high ICC values. For instance, it is not uncommon to observe ICC values less than 0.10 in educational research (Hedges & Hedberg, 2007; Sellström & Bremberg, 2006) as well as in a public health context (Katzmarzyk et al., 2022; Masood & Reidpath, 2016). Although our effect size-based approach to defining the population parameters prevented us for examining very small ICCs, this is an important case to consider. For one, it is not uncommon for models to include covariates with very small ICCs less than 0.05 (e.g., a background variable with similar frequency across clusters). We would hesitate to disaggregate in this edge case for two reasons: a very low ICC implies that the predictor's effect is primarily at the within level to begin with, and estimating latent group means from such a small pot of variance is often challenging unless the sample size is very large. Because the ICC is such an important feature, we strongly recommend that researchers estimate predictor ICCs before starting their modeling.

In addition to low ICCs, our study also did not consider small level-2 sample sizes. For example, our lowest sample size condition was considerably larger than recommendations such as the 30/30 rule (Kreft, 1996). With smaller samples, we generally saw that smushed approach tended to perform better as far as precision. In contexts with sample sizes smaller than what was observed within our simulation, it is reasonable to assume that the differences between the smushed model and the two disaggregated model would become more pronounced. As well, we did not explicitly explore the impact of disaggregating covariates on the estimation of a level-1 focal predictor. Given that the covariate could impact the estimation of the level-1 focal predictor at either level, the choice to disaggregate may become an even more critical question. Future research should explore how the choice to disaggregate covariates may change in this context. Finally (and most notably), the model we evaluated was overly simple. The model only included a single focal predictor and a single level-1 covariate. Although we chose this for consistency with Rights et al. (2020), researchers are likely to include multiple predictors for both substantive hypotheses and statistical control. Furthermore, covariate sets usually include discrete variables, which can also be disaggregated (Yaremych et al., 2023). The one covariate case we reviewed in this simulation is somewhat extreme. When there are multiple covariates in the model, some will not strongly correlate with the focal predictor. Therefore, the negative impacts of the correlation that drives the bias should not scale linearly when you add additional predictors. Otherwise put, there is nothing to suggest that simply adding more covariates should result in a greater biasing effect. That being said, future research should still explore how different features of sets of covariates may impact the interpretation of the focal predictor.

In summary, the aim of this study was to encourage researchers to think critically about the disaggregation of level-1 covariates in MLMs, rather than to promote a one-size-fits-all prescription. Our findings largely reinforce the central conclusion of Rights et al. (2020): failure to disaggregate covariates can introduce meaningful bias in the estimation of level-2 effects when contextual effects are present. At the same time, we show that the magnitude and practical importance of this bias depend systematically on features of the data, including: the size of the contextual effect, the ICC of the covariate, and the strength of between-level correlations. In the relatively narrow set of conditions in which these features are small, the smushed model can yield slightly more precise estimates, though such gains rarely translate into substantively meaningful improvements in power and are easily offset once bias increases.

Taken together, these results suggest that disaggregation should remain the default strategy for explanatory modeling in MLMs, particularly when researchers seek unbiased estimates of level-2 effects and cannot rule out nontrivial contextual influences. However, our findings also highlight that thoughtful departures from this default may be defensible in specific settings, such as predictive modeling or small-sample contexts where precision is paramount and contextual effects are negligible. By clarifying when and why different approaches perform well or poorly, this study aims to move the discussion beyond blanket recommendations and toward a more nuanced, transparent practice of multilevel modeling.

Article information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human

participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was not supported by a grant.

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institution is not intended and should not be inferred.

References

Aberson, C. L. (2015). Statistical power analysis. In *Emerging Trends in the Social and Behavioral Sciences*. (pp. 1–14) John Wiley & Sons, Ltd. https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118900772.etrds0319 https://doi.org/10.1002/9781118900772.etrds0319

Asparouhov, T., & Muth´
tering 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

Asparouhov, T., & Muthén, B. (2021a). Advances in Bayesian model fit evaluation for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 1–14. https://doi.org/10.1080/10705511.2020.1764360

en, B. (2019). Latent variable cen-

Asparouhov, T., & Muthén, B. (2021b). Bayesian estimation of single and multilevel models with latent variable interactions. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(2), 314–328. https://doi.org/10.1080/10705511.2020.1761808

Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., Andersson, G., Atkins, D. C., Carlbring, P., Carroll, K. M., Christensen, A., Eddington, K. M., Ehlers, A., Feaster, D. J., Keijsers, G. P. J., Koch, E., Kuyken, W., Lange, A., Lincoln, T., … Watson, J. (2011). Intraclass correlation associated with therapists: Estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy*, *40*(1), 15–33. https://doi.org/10.1080/16506073.2010.520731

Bolger, N., Laurenceau, J.-P., & Kenny, D. A. (2013). Intensive longitudinal methods: An introduction to diary and experience sampling research.

Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology*, *100*(2), 431–449. https://doi.org/10.1037/a0038047

Cherenack, E. M., Wilson, P. A., Kreuzman, A. M., & Price, G. N. (2016). The Feasibility and Acceptability of Using Technology-Based Daily Diaries with HIV-Infected Young Men Who have Sex with Men: A Comparison of Internet and Voice Modalities. *AIDS AND Behavior*, *20*(8), 1744–1753.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. (2 ed.) Routledge.

Enders, C. K., Du, H., & Keller, B. T. (2020). A model-based imputation procedure for multilevel regression models with random coefficients, interaction effects, and nonlinear terms. *Psychological Methods*, *25*(1), 88–112. https://doi.org/10.1037/met0000228

Enders, C. K., & Keller, B. T. (2025). Factored structural equation modeling in blimp. *PsyArXiv*.

Enders, C. K., Keller, B. T., & Woller, M. P. (2025). A simple monte carlo method for estimating power in multilevel designs. *Psychological Methods*, *30*(5), 980–996. https://doi.org/10.1037/met0000614

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

Fleiss, J. L. (1989). A critique of recent research on the two-treatment crossover design. *Controlled Clinical Trials*, *10*(3), 237–243. https://doi.org/10.1016/0197-2456(89)90065-2

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168. https://doi.org/10.1177/2515245919847202

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, *24*(6), 997–1016. https://doi.org/10.1007/s11222-013-9416-2

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472. https://doi.org/10.1214/ss/1177011136

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

Glassman, J. R., Potter, S. C., Baumler, E. R., Coyle, K. K., Glassman, J. R., P, S. C., Baumler, E. R., & Coyle, K. K. (2015). Estimates of intraclass correlation coefficients from longitudinal group-randomized trials of adolescent HIV/STI/pregnancy prevention programs. *Health Education & Behavior: The Official Publication of the Society for Public Health Education*, *42*(4), 545–553. https://doi.org/10.1177/1090198114568308

Hamaker, E. L., & Grasman, R. P. P. P. (2014). To center or not to center? Investigating inertia with a multilevel autoregressive model. *Frontiers in Psychology*, *5*, 1492. https://doi.org/10.3389/fpsyg.2014.01492

Hamaker, E. L., & Muthén, B. (2020). The fixed versus random effects debate and how it relates to centering in multilevel modeling. *Psychological Methods*, *25*(3), 365–379. https://doi.org/10.1037/met0000239

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

Hoffman, L., & Stawski, R. S. (2009). Persons as contexts: Evaluating between-person and within-person effects in longitudinal analysis. *Research in Human Development*, *6*(2-3), 97–120. https://doi.org/10.1080/15427600902911189

Hoffman, L., & Walters, R. W. (2022). Catching up on multilevel modeling. *Annual Review of Psychology*, *73*(Volume 73(1), 659–689. https://doi.org/10.1146/annurev-psych-020821-103525

Katzmarzyk, P. T., Denstel, K. D., Martin, C. K., Newton, R. L., Apolzan, J. W., Mire, E. F., Horswell, R., Johnson, W. D., Brown, A. W., & Zhang, D, PROPEL Research Group (2022). Intraclass correlation coefficients for weight loss cluster randomized trials in primary care: The PROPEL trial. *Clinical Obesity*, *12*(4), e12524. https://doi.org/10.1111/cob.12524

Keller, B. T., Enders, C. K. (2021). Blimp user's guide (Version 3). www.appliedmissingdata.com/blimp

Kivlighan, D. M., Aloe, A. M., Adams, M. C., Garrison, Y. L., Obrecht, A., Ho, Y. C. S., Kim, J. Y. C., Hooley, I. W., Chan, L., & Deng, K. (2020). Does the group in group psychotherapy matter? A meta-analysis of the intraclass correlation coefficient in group treatment research. *Journal of Consulting and Clinical Psychology*, *88*(4), 322–337. https://doi.org/10.1037/ccp0000474

Kreft, I. G. G. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. California State University. eric.ed.gov

Levy, R., & McNeish, D. (2023). Perspectives on Bayesian inference and their implications for data analysis. *Psychological Methods*, *28*(3), 719–739. https://doi.org/10.1037/met0000443

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

Masood, M., & Reidpath, D. D. (2016). Intraclass correlation and design effect in BMI, physical activity and diet: A cross-sectional study of 56 countries. *BMJ Open*, *6*(1), e008173. https://doi.org/10.1136/bmjopen-2015-008173

McNeish, D. (2023). A practical guide to selecting and blending approaches for clustered data: Clustered errors, multilevel models, and fixed-effect models. *Psychological Methods*, 1-29. 10.1037/met0000620

McNeish, D., & Hamaker, E. L. (2020). A primer on two-level dynamic structural equation models for intensive longitudinal data in Mplus. *Psychological Methods*, *25*(5), 610–635. https://doi.org/10.1037/met0000250

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 Publications.

Rights, J. D., Preacher, K. J., & Cole, D. A. (2020). The danger of conflating level-specific effects of control variables when primary interest lies in level-2 effects. *The British Journal of Mathematical and Statistical Psychology*, *73 Suppl 1*(S1), 194–211. https://doi.org/10.1111/bmsp.12194

Rights, J. D., & Sterba, S. K. (2019). Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures. *Psychological Methods*, *24*(3), 309–338. https://doi.org/10.1037/met0000184

Sellström, E., & Bremberg, S. (2006). Is there a "school effect" on pupil outcomes? A review of multilevel studies. *Journal of Epidemiology and Community Health*, *60*(2), 149–155. https://doi.org/10.1136/jech.2005.036707

Senn, S. J. (1992). Is the 'simple carry-over' model useful? *Statistics in Medicine*, *11*(6), 715–726. https://doi.org/10.1002/sim.4780110603

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. 2. ed

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. W. (2011). Optimal design plus empirical evidence: Documentation for the "Optimal Design" Software. Version 3.0. In.

Stockford, S. M. (2009). *Meta-analysis of intraclass correlation coefficients from multilevel models of educational achievement* [Ph.D., Arizona State University]. United States. https://www.proquest.com/docview/304845471/abstract/92B57E7198C84889PQ/1

Yaremych, H. E., Preacher, K. J., & Hedeker, D. (2023). Centering categorical predictors in multilevel models: Best practices and interpretation. *Psychological Methods*, *28*(3), 613–630. https://doi.org/10.1037/met0000434

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

Zyphur, M. J., Voelkle, M. C., Tay, L., Allison, P. D., Preacher, K. J., Zhang, Z., Hamaker, E. L., Shamsollahi, A., Pierides, D. C., Koval, P., & Diener, E. (2020). From data to causes II: Comparing approaches to panel data analysis. *Organizational Research Methods*, *23*(4), 688–716. https://doi.org/10.1177/1094428119847280