



Estimation Quality and Required Sample Sizes in Three-Level Contextual Analysis Models

Denny Kerkhoff^{1,2} , Fridtjof W. Nussbeck¹ 

[1] *Department of Psychology, University of Konstanz, Konstanz, Germany.* [2] *Department of Psychology, Bielefeld University, Bielefeld, Germany.*

Methodology, 2023, Vol. 19(2), 133–151, <https://doi.org/10.5964/meth.9775>

Received: 2022-06-23 • **Accepted:** 2023-05-19 • **Published (VoR):** 2023-06-30

Handling Editor: Shahab Jolani, Maastricht University, Maastricht, The Netherlands

Corresponding Author: Denny Kerkhoff, Department of Psychology, Bielefeld University, Universitätsstraße 25, P.O. Box 100131, 33501 Bielefeld, Germany. E-mail: denny.kerkhoff@uni-bielefeld.de

Supplementary Materials: Materials [see [Index of Supplementary Materials](#)]



Abstract

In multilevel analysis, Level-1 predictors that also explain variance at a higher level are called contextual predictors. In the multilevel manifest covariate model, the Level-2 component is modeled as the average of the Level-1 predictor scores within a cluster. In the multilevel latent covariate model, the predictor is decomposed into two latent variables at Level-1 and Level-2. Performance conditions of these modeling approaches for three-level models are largely unexplored. We investigate the two approaches' performance with respect to bias, coverage, and power in a three-level random intercept model. Results reveal differences in estimation quality and required sample sizes. We provide sampling recommendations for both approaches.

Keywords

hierarchical linear model, three-level model, sample sizes, parameter estimation bias, power, coverage, contextual variable

Three-level clustered data have become increasingly prevalent in behavioral research. As clustered observations are no longer independently sampled, applying standard regression analysis yields biased estimates and standard errors (e.g., [Chen, 2012](#); [Hox, 1998](#)). Three-level modeling accounts for dependencies in the data by decomposing the outcome variance across data levels. It also allows for the inclusion of predictors at each level. While most predictors are assumed to explain variance at a specific level, higher-level predictors might also be obtained by aggregating lower-level scores, such as



the proportion of boys (Level-1: L1) in a class (Level-2: L2) or school (Level-3: L3), which are considered contextual variables (e.g., [Enders, 2013](#); [Raudenbush & Willms, 1995](#); cf. also: [Stapleton et al., 2016](#)). Oftentimes, the predictor independently explains variance at the level it is measured at and at the aggregate level, and both components should therefore be modeled as separate variables (e.g., [Diez-Roux, 1998](#)). Models including these kinds of predictors are called contextual analysis or compositional models ([Harker & Tymms, 2004](#); [Lüdtke et al., 2008](#)).

Within the context of contextual analysis models, it is important to distinguish between formative and reflective measurement processes ([Lüdtke et al., 2008](#); cf. also [Bollen & Lennox, 1991](#)): Formative constructs are directly caused by the aggregate (e.g., the proportion of boys in a class). Hence, the number of observations is finite. For reflective constructs, lower-level observations are manifest realizations of the construct. Hence, the number of observations to measure a reflective construct is potentially infinite. The modeling approach needs to account for the nature of the higher-level construct, especially by correcting for sampling error in the case of reflective constructs.

While there is a strong body of research on contextual predictors in two-level analysis (e.g., [Korendijk et al., 2011](#); [Lüdtke et al., 2008, 2011](#); [Marsh et al., 2009](#)), contextual analysis models are frequently applied to three-level data (e.g., [Chen & Cui, 2020](#); [Rathmann et al., 2020](#)). However, information on estimation quality in three-level contextual models is still sparse, especially with respect to reflective measurement processes.

In this study, we first describe the two common modeling approaches for contextual variables in three-level models. We then analyze and compare estimation quality for the two approaches by means of Monte Carlo simulations. Finally, we derive concise sampling recommendations.

The Multilevel Manifest Covariate (MMC) Model for Three-Level Data

In contextual analysis models, the most widely used approach to obtain higher-level predictor variables is to compute the average scores of all L1-units in a L2-subcluster or L3-cluster. Modeling averages at higher levels requires centering procedures at the lower levels. We follow the notation by [Brincks et al. \(2017\)](#) and differentiate between grand-mean centering (GMC) and centering-within-context (CWC). For a linear model with $k = 1, \dots, n_3$ L3-clusters, each with $j = 1, \dots, n_2$ L2-subclusters, each with $i = 1, \dots, n_1$ L1-units, outcome Y_{ijk} , L1-predictor X_{ijk} , L2-predictor $\bar{X}_{.jk}$ (i.e., the subcluster means), and L3-predictor $\bar{X}_{..k}$ (i.e., the cluster means), the MMC model can be formulated as:

$$Y_{ijk} = \gamma_{000} + \gamma_{100} \cdot (X_{ijk} - \bar{X}_{.jk}) + \gamma_{010} \cdot (\bar{X}_{.jk} - \bar{X}_{..k}) + \gamma_{001} \cdot (\bar{X}_{..k} - \bar{X}_{...}) + \nu_{00k} + u_{0jk} + e_{ijk} \quad (1)$$

On Level-1, $(X_{ijk} - \bar{X}_{.jk})$ is the CWC predictor, obtained by subtracting the subcluster-mean from each L1-score. As a result, the regression coefficient γ_{100} addresses only the L1-specific influence of the predictor on the outcome. $e_{ijk} \sim N(0, \sigma_e^2)$ is a random effect (prediction error) with variance component σ_e^2 .

On Level-2, $(\bar{X}_{.jk} - \bar{X}_{..k})$ is the CWC contextual predictor, obtained by subtracting the respective cluster mean from the subcluster-means, with respective L2-specific regression coefficient γ_{010} . $u_{0jk} \sim N(0, \sigma_{u_0}^2)$ is a random effect with variance $\sigma_{u_0}^2$. Note that $\bar{X}_{..k}$ refers to the average L1-scores within a cluster (not to the average of the subcluster averages, which yields different results for unbalanced samples).

On Level-3, $(\bar{X}_{..k} - \bar{X}_{...})$ is the grand-mean centered (GMC) L3-predictor, with respective L3-specific regression coefficient γ_{001} . $v_{00k} \sim N(0, \sigma_{v_0}^2)$ is a random effect containing the unexplained cluster-level variance component.

Lastly, γ_{000} is the intercept, and coefficients γ_{100} , γ_{010} , and γ_{001} are the regression coefficients quantifying the level-specific influence of the respective predictor on the outcome (Brincks et al., 2017).

This approach has been criticized as insufficient for reflective constructs since it assumes a finite population of L1-units. Sampling a finite set of interchangeable indicators for an unobservable construct (e.g., repeated measures of wellbeing in students) disregards unreliability due to sampling error and can result in considerable bias and underestimated standard errors (e.g., Grilli & Rampichini, 2011; Harker & Tymms, 2004). Furthermore, the assumption of perfect reliability in the MMC approach is also violated for formative constructs if the sampling rate, i.e., the rate of units sampled from the total number of units in a (sub-)cluster, is small (e.g., inhabitants in cities).

The Multilevel Latent Covariate (MLC) Model for Three-Level Data

The alternative approach is the multilevel latent covariate (MLC) approach. It treats the observations at L1 as manifest realizations of an underlying latent variable with variance at each level (Lüdtke et al., 2008). Extending the two-level notation, the decomposition of observed predictor X_{ijk} and outcome Y_{ijk} with means μ_X and μ_Y takes the form of:

$$X_{ijk} = \mu_X + V_{Xk} + U_{Xjk} + R_{Xijk} \quad (2)$$

$$Y_{ijk} = \mu_Y + V_{Yk} + U_{Yjk} + R_{Yijk}, \text{ with relationships} \quad (3)$$

$$\text{Level-3: } V_{Yk} = \beta_{L3} \cdot V_{Xk} + \tau_k \quad (4)$$

$$\text{Level-2: } U_{Yjk} = \beta_{L2} \cdot U_{Xjk} + \delta_{jk} \quad (5)$$

$$\text{Level-1: } R_{Yijk} = \beta_{L1} \cdot R_{Xijk} + \varepsilon_{ijk} \quad (6)$$

V_{Xk} , U_{Xjk} , and R_{Xijk} are independent, latent representations of predictor X . V_{Xk} has a mean μ_X and a variance expressing the L3-specific deviations from μ_X in the predictor. U_{Xjk} and R_{Xijk} each have mean zero and variances expressing the L2- and L1-specific deviations. Regression coefficients β_{L1} , β_{L2} , and β_{L3} express the level-specific effects of the predictor. ε_{ijk} , δ_{jk} and τ_k are the residual and random intercepts, respectively. Integrating Equations 3 to 6 yields the combined equation:

$$Y_{ijk} = \mu_Y + \beta_{L3} \cdot V_{Xk} + \beta_{L2} \cdot U_{Xjk} + \beta_{L1} \cdot R_{Xijk} + \tau_k + \delta_{jk} + \varepsilon_{ijk} \quad (7)$$

The MLC approach accounts for sampling error by treating the measurements as potentially biased realizations of the latent construct and has therefore been shown to produce higher estimation quality for reflective constructs or formative constructs with a low (20%) sampling rate (Lüdtke et al., 2008, 2011).

In this study, we compare the approaches regarding their estimation quality in samples drawn from an infinite population (reflective process), since research fields that commonly employ multilevel modeling oftentimes investigate constructs with a conceptually infinite number of observations. Additionally, correctly specifying a reflective construct might pose a challenge for researchers due to high sampling requirements to obtain sound estimation results for latent variables in three-level models.

Evaluating Estimation Quality

The estimation quality of contextual predictors can be assessed in Monte Carlo simulations. Most commonly, estimation quality examinations are based on the point estimates and standard errors.

Parameter Estimation Bias (PEB)

For a true parameter θ and estimates $\hat{\theta}_1, \dots, \hat{\theta}_n$ in n samples, the relative parameter estimation bias (rPEB) is defined as:

$$rPEB_{\theta} = \frac{\sum_{1 \leq i \leq n} \frac{\hat{\theta}_i - \theta}{\theta}}{n} \quad (8)$$

It is interpreted as the average bias rate. Since across samples, negative and positive bias values are averaged, the rPEB expresses the direction of bias across samples, but not the average strength of bias, which is expressed using an absolute bias measure. In this study, we use the absolute PEB (aPEB):

$$aPEB_{\theta} = \frac{\sum_{1 \leq i \leq n} \left| \frac{\hat{\theta}_i - \theta}{\theta} \right|}{n} \quad (9)$$

The aPEB measures the rate of absolute misestimation across samples as an alternative to common measures such as the root mean squared error (RMSE). Both the RMSE and aPEB capture mean bias and variability in estimates, since conditions producing a larger variance in estimates (given the same mean estimate) result in increased RMSE and aPEB values. We argue, however, that the aPEB is better suited for this study, since it is scaled relative to the true value θ , making bias rates more easily comparable across parameter sizes. Additionally, interpretation of the aPEB (average strength of bias) well complements the interpretation of the relative bias (average rate/direction of bias).

Statistical Power and Coverage

Standard errors are commonly evaluated by means of statistical power and coverage rates (e.g., [Muthén & Muthén, 2002](#)). For fixed effects, power of an estimate is the rate of statistically significant Wald-tests ([Wald, 1943](#)) across all analyzed samples. The coverage rate expresses the rate of samples with a 95%-confidence interval (CI) that includes θ :

$$[\hat{\theta} - SE(\hat{\theta}) \cdot 1.96; \hat{\theta} + SE(\hat{\theta}) \cdot 1.96] \quad (10)$$

Sample Size Recommendations

Sampling advice for three-level models is still sparse (see [Kerkhoff & Nussbeck, 2019](#), for an overview), and research on the estimation quality in three-level contextual models is still sparse. [Usami \(2017\)](#) derived power formulas for regression coefficients in three-level contextual analysis models with manifest means. Comparisons between derived and observed power in simulations reveal that observed power may be biased due to unreliability of the mean values, and that increasing both L1 and L2 sample sizes reduces differences between derived and observed power.

Regarding three-level models, research has shown that estimation quality is mostly determined by the number of clusters (L3-sample size) and the sample size at the level the coefficient of interest is measured at ([de Jong et al., 2010](#); [Dong et al., 2018](#); [Kerkhoff & Nussbeck, 2019, 2022](#); [Lee & Hong, 2021](#); [Li & Konstantopoulos, 2016](#)). Regarding contextual predictors for reflective constructs in two-level models, [Lüdtke et al. \(2008, 2011\)](#) found that for the MLC approach, bias remains within 10% in most conditions with at least 50 clusters of cluster size 5, while the MMC approach is more heavily biased. Due to narrow CIs, the MMC approach suffers from low coverage rates. In contrast, the absolute bias is higher for the MLC approach due to high variance in estimates.

Aim of This Study

Since contextual analysis in three-level models is of increasing relevance, we investigate the estimation quality for both the MMC and MLC approach to derive answers to the following research questions:

1. How do sample sizes relate to bias, coverage, and power for each modeling approach, i.e., (a) what are influential sample characteristics and (b) what patterns emerge between sampling conditions and estimation quality indicators?
2. What are (a) minimum required sample sizes and (b) advantageous sampling strategies to achieve sound estimation quality for each approach?

Method

Data was drawn from a population with a single predictor variable explaining variance at each level. We varied the simulation conditions regarding the number of clusters, $n_3 = 15, 30, 50, 100, 150,$ or 200 , the number of subclusters per cluster, $n_2 = 5, 10, 15, 20, 25,$ or 30 , and the number of L1-units per subcluster, $n_1 = 5, 10, 15, 20, 25,$ or 30 . We specifically included conditions with a low number of clusters ($n_3 = 15$) to assess if estimation quality can be ensured by compensating few clusters with high cluster and subcluster sizes. We further varied the higher-level regression coefficient sizes (β_{L2}, β_{L3}) and kept $\mu_Y = 0$ and $\beta_{L1} = 0.1$ constant. Group-level variance components are most commonly available for educational research constructs and typically range between 10% and 30% (Dong et al., 2016; Hedges & Hedberg, 2013). Therefore, we specified $X \sim N(0, 1 + \frac{1}{3} + \frac{1}{3})$, i.e., 20% of predictor variance at each higher level. Similarly, we computed $\sigma_e^2, \sigma_{u0}^2,$ and σ_{v0}^2 for the full model such that $Y \sim N(0, 1 + \frac{1}{3} + \frac{1}{3})$, i.e., $ICC_3 = ICC_2 = 0.2$ in the empty model.

We selected unstandardized regression weights such that the resulting standardized L2 and L3 effect sizes are either small ($\beta_{std} = 0.1$), medium-sized ($\beta_{std} = 0.3$), or large ($\beta_{std} = 0.5$), and of different size, such that most of the resulting contextual effects, e.g., $\beta_{L3} - \beta_{L1}$, are not zero. We abbreviate these conditions as shown in Table 1.

Table 1

Population Parameters and Notation

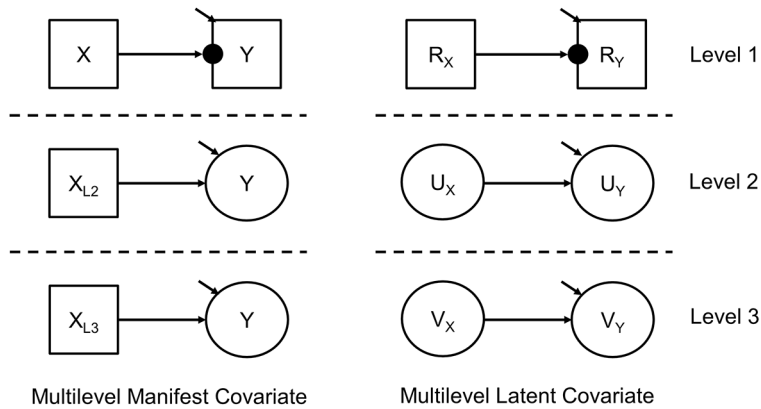
| β_{L1} | β_{L2} | β_{L3} | $\beta_{L1.std}$ | $\beta_{L2.std}$ | $\beta_{L3.std}$ | σ_e^2 | σ_{u0}^2 | σ_{v0}^2 | Notation |
|--------------|--------------|--------------|------------------|------------------|------------------|--------------|-----------------|-----------------|------------------|
| 0.1 | 0.289 | 0.058 | 0.1 | 0.5 | 0.1 | 0.99 | 0.25 | 0.33 | SLS |
| 0.1 | 0.173 | 0.289 | 0.1 | 0.3 | 0.5 | 0.99 | 0.303 | 0.25 | SML |
| 0.1 | 0.058 | 0.173 | 0.1 | 0.1 | -0.3 | 0.99 | 0.33 | 0.303 | SSM ⁺ |
| 0.1 | -0.058 | -0.173 | 0.1 | -0.1 | -0.3 | 0.99 | 0.33 | 0.303 | SSM ⁻ |

Note. $\beta_{L1}, \beta_{L2},$ and β_{L3} are population fixed effects on Level-1, Level-2, and Level-3, respectively. $\beta_{L1.std}, \beta_{L2.std},$ and $\beta_{L3.std}$ are standardized effects. $\sigma_{v0}^2, \sigma_{u0}^2,$ and σ_e^2 are population variance components at Level-1, Level-2, and Level-3 in the full model.

We also included one effect size combination with negative regression weights to explore differences between positive and negative coefficients on a small scale (see Table 1, con-

Figure 1

Conceptual Models



Note. X_{L2} refers to the subcluster means, X_{L3} refers to the cluster means. R , U , and V correspond to the latent predictor and outcome components as in Equations 4 to 6. Data levels are arranged vertically and separated by dashed lines, similarly to figures used in Lüdtke et al. (2008, 2011) and the Mplus manual (Muthén & Muthén, 1998–2017).

dition SSM $\bar{}$). We furthermore evaluated additional effect size combinations, sample sizes, and unbalanced designs, but respective results are only reported in the [Supplementary Materials](#) since they do not meaningfully impact inferences reported below. The calculus to obtain conditional variances and regression weights can be found in the R-script in the [Supplementary Materials](#). For each generated sample, models were fitted according to Equation 1 and Equation 7. Figure 1 shows the conceptual models.

In total, we analyzed 1,728 conditions, each with 1,000 generated samples. Data generation and model estimation were done in Mplus 8 (Muthén & Muthén, 1998–2017) using maximum likelihood estimation with robust standard errors (MLR). Results were imported to R 4.0.5 (R Core Team, 2021) for subsequent analyses. To distinguish between conditions, we abbreviate each combination of sample sizes by n_3 -size/ n_2 -size/ n_1 -size. For example, 100/5/5 encompasses samples with 100 clusters, each containing 5 subclusters, which in turn contain 5 L1-units. 200/2/* subsumes conditions with 200 clusters with 2 subclusters each and any number of L1-units.

Evaluation Strategy

For each condition separately, we computed the mean estimates across samples, the rPEB and aPEB, power, and coverage rates for the level-specific estimates (MMC: \hat{Y}_{100} , \hat{Y}_{010} , \hat{Y}_{001} ; MLC: $\hat{\beta}_{L1}$, $\hat{\beta}_{L2}$, $\hat{\beta}_{L3}$). In line with common recommendations (Flora & Curran,

2004; Muthén & Muthén, 2002), we consider $|rPEB| < 0.10$, power ≥ 0.8 , and $0.91 \leq$ coverage ≤ 0.98 to indicate sufficient estimation quality. We further computed analyses of variance (ANOVA) to evaluate how the simulation conditions influence estimation quality, using bias, coverage, or power as outcomes and effect sizes as well as sampling conditions as factors. Due to the large sample size, we only report partial effect sizes η^2 . To keep analyses concise, we primarily focus on samples with a maximum of 10,000 observations (see the [Supplementary Materials](#) for full results).

Results

Three conditions in the MMC approach and 13 conditions in the MLC approach showed convergence issues for at least one sample. Convergence rates did not drop below 99.7% (as observed in 15/5/5 with effect sizes SLS in the MLC approach).

Overview of Estimation Quality

[Table 2](#) lists median estimates and standard errors, averaged across values for conditions with up to 10,000 observations. Due to overall high estimation quality of the L1-effect, only results for the higher-level effects are comprehensively reported. On L1, all conditions are unbiased, and power and coverage rates are insufficient only in the smallest sample sizes, e.g., 15/5/10 (see the [Supplementary Materials](#) for full results).

Table 2

Median Estimates and Standard Errors Across Sampling Conditions

| Level | β -size | MMC approach | | MLC approach | |
|---------|---------------|--------------|-------|--------------|-------|
| | | Median | SE | Median | SE |
| Level-1 | 0.1 | 0.100 | 0.015 | 0.100 | 0.015 |
| Level-2 | -0.058 | -0.023 | 0.050 | -0.058 | 0.067 |
| | 0.058 | 0.067 | 0.050 | 0.058 | 0.066 |
| | 0.173 | 0.158 | 0.049 | 0.173 | 0.064 |
| | 0.289 | 0.246 | 0.046 | 0.289 | 0.060 |
| Level-3 | -0.173 | -0.155 | 0.168 | -0.176 | 0.190 |
| | 0.058 | 0.068 | 0.173 | 0.056 | 0.198 |
| | 0.173 | 0.172 | 0.167 | 0.173 | 0.191 |
| | 0.289 | 0.276 | 0.153 | 0.290 | 0.175 |

Note. β -size = size of the population coefficient. SE = standard error, MMC = multilevel manifest covariate, MLC = multilevel latent covariate.

Table 3*Effect Sizes (Partial η^2) Based on Analyses of Variance for Bias, Coverage, and Power*

| Level | Factor | df | MMC approach | | | | MLC approach | | | |
|---------|------------------|-----|--------------|-------|-------|----------|--------------|------|-------|----------|
| | | | rPEB | aPEB | Power | Coverage | rPEB | aPEB | Power | Coverage |
| Level-2 | β -size | 3 | .788 | .760 | .832 | .453 | .037 | .696 | .827 | .002 |
| | n_3 | 5 | <.001 | .501 | .435 | .263 | .002 | .588 | .564 | .895 |
| | n_2 | 5 | <.001 | .397 | .321 | .180 | .002 | .474 | .437 | .172 |
| | n_1 | 5 | .277 | .184 | .032 | .461 | .003 | .126 | .122 | .156 |
| | $n_3 \times n_2$ | 25 | <.001 | .173 | .003 | .063 | .004 | .177 | .007 | .253 |
| | $n_3 \times n_1$ | 25 | <.001 | .014 | .011 | .175 | .005 | .035 | .001 | .175 |
| | $n_2 \times n_1$ | 25 | <.001 | .008 | .005 | .092 | .014 | .027 | <.001 | .454 |
| | NOBS | 125 | <.001 | .002 | .010 | .022 | .014 | .010 | .003 | .560 |
| Level-3 | β -size | 3 | .711 | .806 | .830 | .051 | .116 | .777 | .812 | .012 |
| | n_3 | 5 | <.001 | .674 | .743 | .951 | .011 | .663 | .694 | .965 |
| | n_2 | 5 | .023 | .001 | .015 | .295 | .068 | .051 | .076 | .198 |
| | n_1 | 5 | .002 | <.001 | .001 | .052 | .006 | .003 | .002 | .025 |
| | $n_3 \times n_2$ | 25 | .009 | .001 | .001 | .316 | .064 | .031 | .005 | .187 |
| | $n_3 \times n_1$ | 25 | .003 | <.001 | .001 | .110 | .014 | .002 | <.001 | .121 |
| | $n_2 \times n_1$ | 25 | .023 | .001 | .002 | .266 | .073 | .003 | .002 | .331 |
| | NOBS | 125 | .024 | .002 | .003 | .386 | .103 | .005 | .002 | .453 |

Note. β -size = size of the population coefficient, rPEB = relative bias, aPEB = absolute bias. n_1 , n_2 , n_3 indicate the number of clusters, subclusters per cluster, and Level-1 units per subcluster. NOBS = total number of observations. MMC = multilevel manifest covariate, MLC = multilevel latent covariate. Due to heterogeneity of variances, values are likely to have positive bias.

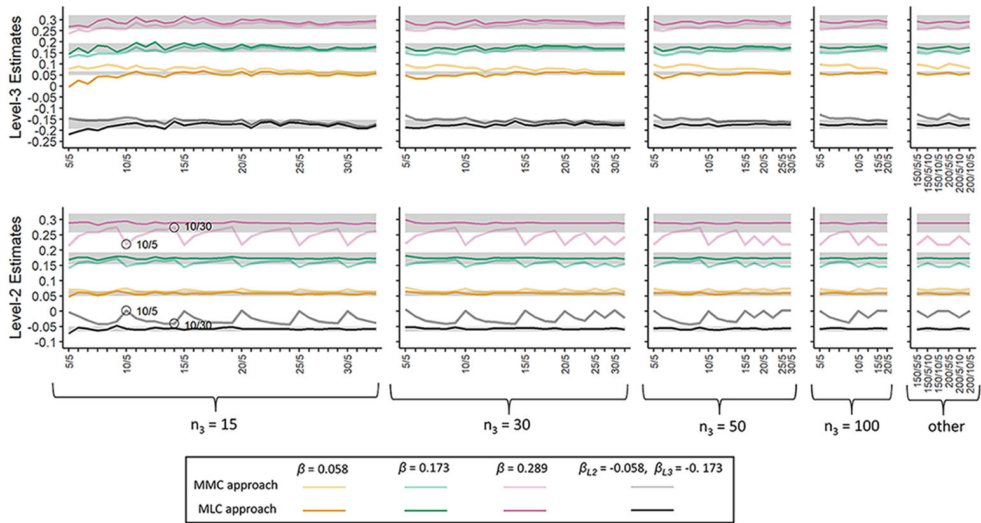
With respect to research question (1a), ANOVA results in Table 3 show that the number of clusters is the most relevant sampling factor for estimation quality in the MLC approach. Similarly, in the MMC approach, the number of clusters is the most important factor to achieve estimation quality, except for L2 relative bias and coverage. Notably, differences in L2 relative bias are almost exclusively determined by the number of sampled L1-units per subcluster.

Estimation Bias

To answer research question (1b) for the relative and absolute bias, Figures 2 and 3 plot the mean regression estimates with grey areas indicating $|rPEB| \leq 0.10$ (Figure 2) and aPEB (Figure 3). While regression estimates are rather unaffected by different sample sizes in the MLC approach, estimates in the MMC approach tend to be biased for condi-

Figure 2

Mean Estimates According to Sampling Condition



Note. Plots show mean estimates for each Level-3 (upper plot) and Level-2 (lower plot) regression coefficient, in sample sizes with up to 10,000 observations. Shaded areas indicate relative unbiasedness. Plots are grouped according to n_3 , each x-axis is sorted according to n_1 within n_2 , but only the first condition per n_2 is labelled on the x-axis to visually differentiate between n_2 -sizes, i.e., for $n_3 = 15$ (leftmost plot), 10/5 indicates $n_2 = 10$ with $n_1 = 5$, which is followed by 10/10, 10/15 etc. MLC = multilevel latent covariate, MMC = multilevel manifest covariate.

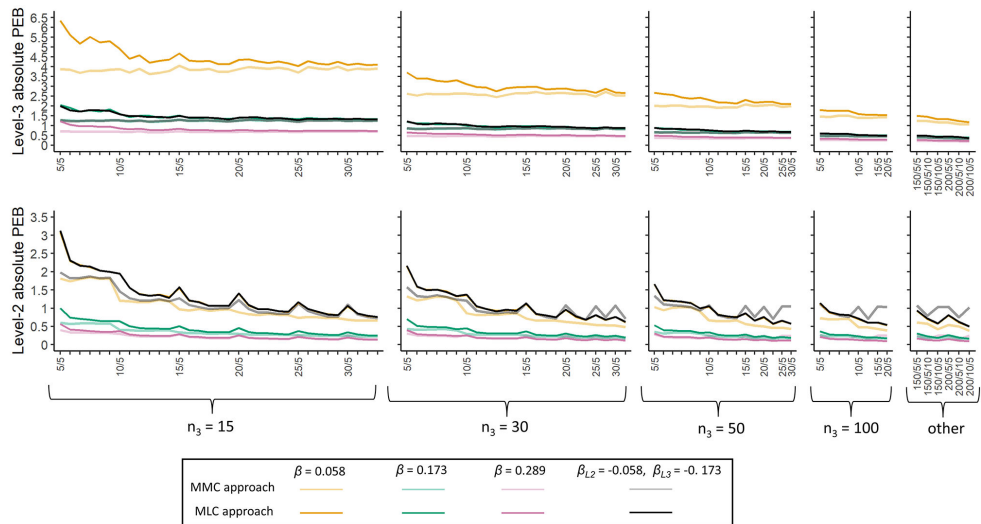
tions with fewer L1-units. Moreover, the MMC approach more heavily overestimates the small negative L2-effect than the small positive L2-effect. In contrast, while the relative bias is higher in the MMC approach, the absolute bias tends to be higher in the MLC approach.

Power and Coverage Rates

To answer research question (1b) for coverage and power rates, **Figures 4 and 5** plot power (**Figure 4**) and coverage rates (**Figure 5**). **Figure 4** indicates that power tends to be higher in the MMC approach. Notably, due to overestimation, statistical power is higher for the small positive effects than for small negative effects on L2 in the MMC approach. Strikingly, coverage rates for (negative) small and large effects on L2 decrease drastically as n_2 and n_3 increase. This is due to smaller CIs around consistently biased estimates (cf. **Table 2**).

Figure 3

Absolute Bias of Estimates



Note. Plots show absolute bias for each Level-3 (upper plot) and Level-2 (lower plot) regression coefficient, in sample sizes with up to 10,000 observations. Plots are grouped according to n_3 , each x-axis is sorted according to n_1 within n_2 , but only the first condition per n_2 is marked on the x-axis. MLC = multilevel latent covariate, MMC = multilevel manifest covariate.

Required Sample Sizes

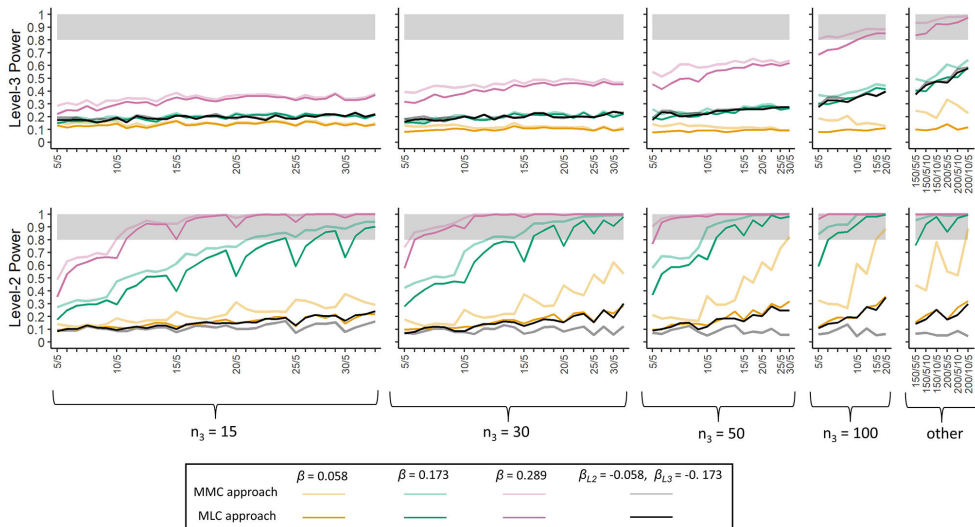
To answer the research questions (2a) and (2b), Tables 4 and 5 show quartiles of absolute bias and required sample sizes for relative unbiasedness, sufficient coverage, and sufficient power for the MMC approach (Table 4) and MLC approach (Table 5).

Level-2 MMC Estimates

Small positive effects are unbiased (rPEB) for $n_1 \geq 25$ with $n_2 \geq 10$. Medium and large effects are unbiased for most samples with $n_1 \geq 20$. Conditions with large n_1 , such as 30/5-10/25-30, ensure sufficient coverage irrespective of effect size. Power for small effects is achieved in large samples, such as 100/25/•. For large effects, power is sufficient (80% or higher) in most conditions. For medium effects, power is sufficient for most conditions with $n_3 \geq 100$ or $n_2 \geq 20$. The average absolute bias for small effects exceeds 50% even for larger samples.

Figure 4

Power Rates of Estimates



Note. Plots show power rates for each Level-3 (upper plot) and Level-2 (lower plot) regression coefficient, in sample sizes with up to 10,000 observations. Shaded areas indicate ranges for sufficient power. Plots are grouped according to n_3 , each x-axis is sorted according to n_1 within n_2 , but only the first condition per n_2 is marked on the x-axis. MLC = multilevel latent covariate, MMC = multilevel manifest covariate.

Level-3 MMC Estimates

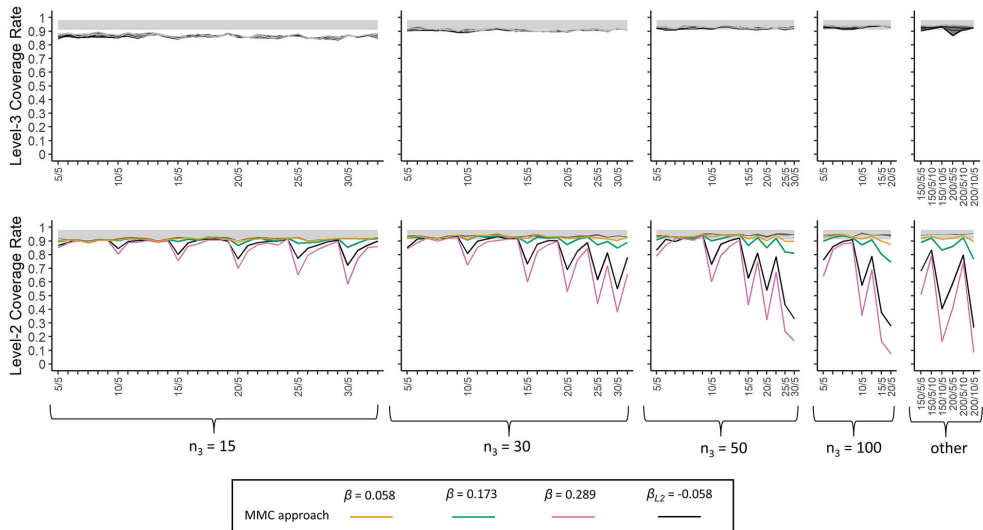
Relative unbiasedness for small effect requires $n_2 \geq 30$. Medium-sized effects are unbiased in most conditions with $n_2 \geq 25$ or both n_1 and $n_2 \geq 10$. Large effects are unbiased for most conditions with $n_2 \geq 10$. Coverage rates are mostly sufficient by sampling at least 50 clusters. Statistical power is only sufficient for large effects in conditions with at least 100 clusters. Average absolute bias is considerably high even for large effects.

Level-2 MLC Estimates

Medium and large effects are generally unbiased (rPEB). Small effects have sufficiently low relative bias in most conditions with $n_3 \geq 30$. Coverage rates are sufficient in most conditions. Power of small effects is only sufficient in extremely large samples, such as 200/20/•. For medium-sized effects, power is sufficient in most conditions with at least 5,000 observations. Large effects mostly have sufficient power even in small samples. Average absolute bias exceeds 50% for small effects even in larger samples.

Figure 5

Coverage Rates of Estimates



Note. Plots show coverage rates for each Level-3 (upper plot) and Level-2 (lower plot) regression coefficient, in sample sizes with up to 10,000 observations. Shaded areas indicate ranges for sufficient coverage. Due to very similar values, all lines in the upper plot and lines resulting from the MLC approach in the lower plot are encompassed by a dark ribbon. Plots are grouped according to n_3 , each x-axis is sorted according to n_1 within n_2 , but only the first condition per n_2 is marked on the x-axis. MLC = multilevel latent covariate, MMC = multilevel manifest covariate.

Level-3 MLC Estimates

Large and medium-sized effects have sufficiently low relative bias in most conditions. For small effects, most conditions with $n_3 = 50$ in combination with $n_2 \geq 10$, or $n_3 = 30$ in combination with $n_2 \geq 15$ are unbiased. Samples with at least 50 clusters have sufficient coverage. Power is only sufficient for large effects in samples with $n_3 \geq 150$, or $n_3 = 100$ with $n_2 \geq 10$. Average absolute bias is high for all effect sizes and exceeds 50% even for large effects.

Table 4

Absolute Bias and Minimum Sample Sizes to Achieve Sufficient Estimation Quality in the Multilevel Manifest Covariate Approach

| Level | β -size | aPEB Q1; Q3 | $ rPEB \leq 0.1$ | $.91 \leq \text{Coverage} \leq .98$ | Power $\geq .80$ |
|---------|---------------|--------------|--|-------------------------------------|---------------------------------|
| Level-2 | -0.058 | 0.836; 1.190 | not achieved | $n_1 = 30$ with $n_3 \leq 50$ | 150/30/30; 200/30/20; 200/25/25 |
| | 0.058 | 0.602; 0.993 | $n_2 \geq 10/n_1 \geq 25; n_2 \geq 25/n_1 \geq 20$ | $n_1 \geq 10$ | 150/15/•; 100/25/• |
| | 0.173 | 0.213; 0.329 | $n_1 \geq 10$ | $n_1 \geq 20$ | $n_3 \geq 100$ or $n_2 \geq 20$ |
| | 0.289 | 0.164; 0.249 | $n_1 \geq 20$ | $n_1 = 30$ with $n_3 \leq 50$ | all but 15/5/• and 30/5/5 |
| Level-3 | 0.058 | 1.985; 3.838 | $n_2 \geq 30$ | $n_3 \geq 50$ | not achieved |
| | -0.173 | 0.646; 1.245 | $n_2 \geq 25$ or $n_1 \geq 10$ with $n_2 \geq 10$ | $n_3 \geq 50$ | not achieved |
| | 0.173 | 0.637; 1.237 | $n_2 \geq 10$ | $n_3 \geq 50$ | not achieved |
| | 0.289 | 0.352; 0.685 | $n_2 \geq 10$ or $n_1 \geq 20$ | $n_3 \geq 50$ | $n_3 \geq 100$ |

Note. β -size = size of the population coefficient, rPEB = relative bias, aPEB = absolute bias with lower quartile Q1 and upper quartile Q3, n_3 = number of clusters, n_2 = subclusters per clusters, n_1 = Level-1-units per subcluster.

Table 5

Absolute Bias and Minimum Sample Sizes to Achieve Sufficient Estimation Quality in the Multilevel Latent Covariate Approach

| Level | β -size | aPEB Q1; Q3 | $ rPEB \leq 0.1$ | $.91 \leq \text{Coverage} \leq .98$ | Power $\geq .80$ |
|---------|---------------|--------------|--|-------------------------------------|--|
| Level-2 | -0.058 | 0.777; 1.282 | most | most | 200/20/•; 150/25/• |
| | 0.058 | 0.782; 1.327 | most | most | 200/20/•; 150/25/• |
| | 0.173 | 0.249; 0.431 | all | most | $n_3 \geq 100$; 50/10/•; 30/15/• |
| | 0.289 | 0.137; 0.239 | all | most | most |
| Level-3 | 0.058 | 2.260; 4.256 | $n_3 \geq 50$; or $n_3 = 30$ with $n_2 \geq 15$ | $n_3 \geq 50$ | not achieved |
| | -0.173 | 0.734; 1.362 | most | $n_3 \geq 50$ | not achieved |
| | 0.173 | 0.728; 1.366 | most | $n_3 \geq 50$ | not achieved |
| | 0.289 | 0.398; 0.751 | all | $n_3 \geq 50$ | $n_3 \geq 150$; or $n_3 = 100$ with $n_2 \geq 10$ |

Note. β -size = size of the population coefficient, rPEB = relative bias, aPEB = absolute bias with lower quartile Q1 and upper quartile Q3, n_3 = number of clusters, n_2 = subclusters per clusters, n_1 = Level-1-units per subcluster.

Discussion

In this study, we investigated the estimation quality of the MMC and MLC approaches in three-level models in order to (1) evaluate how bias, coverage and power rates relate to the sample size at each data level and (2) derive advantageous sampling strategies to achieve sound estimation quality (see Tables 4 and 5). Overall, sampling 100/10/• or

150 clusters ensures overall sound estimation quality for large effects and additionally medium-sized effects on L2 in the MLC approach. For the MMC approach, sampling 100/•/20 ensures sound estimation quality for large L3-effects and medium-sized L2-effects. Extending our knowledge regarding required sample sizes in multilevel modeling, our results may help researchers make informed decisions regarding required sample sizes. Most notably, for the MMC approach, tendencies to over- or underestimate effects (relative bias) depend on the (sub-)cluster sizes. Since for this approach, estimation bias does not generally improve as the overall sample size increases, but standard errors become smaller, coverage rates deteriorate as the samples get larger. It is therefore highly important to sample a sufficiently high number of lower-level units to avoid biased estimates at the higher levels. In contrast, the MLC approach has higher absolute bias, indicating higher variance in estimates, and slightly lower statistical power, but estimation quality for MLC estimates can be reliably improved by sampling more clusters. Naturally, our recommendations are based on specified thresholds indicating sufficient estimation quality (esp. [Muthén & Muthén, 2002](#)), and might therefore differ if stricter or less strict thresholds are used. For example, [Burton et al. \(2006\)](#) recommend basing coverage rate thresholds on the number of simulation replications. For our analyses, this translates to an admissible coverage range of 93.5% to 96.4%, such that additional conditions (mainly $n_3 = 15, 30, 50$) result in insufficiently low coverage.

Limitations and Future Prospects

Most importantly, our analyses are limited by the simulation conditions. For example, in some research contexts, only two L2 subclusters per L3 cluster might be available. Such samples limit admissible model complexity, but contextual effects might still be reliably estimated. Additional analyses (see the [Supplementary Materials](#)) to explore estimation quality for such samples show that—in comparison to conditions with $n_2 = 5$ —rPEB is at least twice as high, except for nearly unchanged rPEB values for L2 estimates in the MMC approach. Similarly, $n_2 = 2$ results in at least 30% less power than $n_2 = 5$, except for power on L3 in the MMC approach, which has only about 5% less power.

Moreover, previous studies argue that the variance distribution of the predictor variable across levels influences estimation quality ([Lüdtke et al., 2008, 2011](#); [Zitzmann et al., 2015](#)). For reflective constructs measured by the MMC in particular, reliability of the predictor in two-level models is a function of the predictor variance at the respective level (predictor ICC) and the group size (cf. [Lüdtke et al., 2008](#)). Our results confirm the importance of the subcluster size for unbiasedness of the MMC approach on L2 and demonstrate the role of the cluster size ($n_2 \times n_1$ and n_2) for estimation quality on L3. However, to focus on the interplay between sample size, effect size, and analysis approach, we kept the variance fractions constant with 20% of variance on each higher level. To illustrate how the distribution influences results, we ran additional simulations with 60% of variance at either higher level (see the [Supplementary Materials](#) for full

results). Results show that coverage rates and power are not meaningfully affected, while absolute bias decreases at the level with the higher variance. For the relative bias, differences between approaches are considerable: While for the MLC approach, differences are marginal, we find that for the MMC approach, the relative bias is consistently smaller at the level with the high variance proportion. These additional analyses suggest that the variance distribution of the predictor needs to be considered in future research to develop more specific sampling recommendations for three-level contextual models.

Lastly, we limited our analyses to samples generated from an assumed infinite population (reflective process). For two-level models, Lüdtke et al. (2008) showed that for finite samples (formative process), the MMC approach results in smaller bias than the MLC approach with a sampling rate of 50% or higher. We hence consider it promising extending our research to three-level contextual analysis for finite populations, especially since for three-level data, the sampling rate at both L1 and L2 must be considered.

In conclusion, our results suggest that the MLC approach tends to be advantageous for research where the number of sampled clusters can be more easily increased than the (sub-)cluster sizes. The MMC approach, however, has the advantage of higher power and lower absolute bias (i.e., lower variability in estimates), especially for samples with less than 50 clusters. Thus, the MMC approach might be preferable for research where (sub-)cluster sizes can be readily increased for a limited number of clusters.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Data Availability: Data from the supplemental analyses is freely available at [Supplementary Materials](#).

Supplementary Materials

The supplementary materials provided are the Parametervalue.R file for computation of unstandardized regression coefficients, PDF file describing the supplementary analyses evaluated, csv files containing results for the main as well as for the supplemental analyses, and can be accessed in the [Index of Supplementary Materials](#) below.

Index of Supplementary Materials

Kerckhoff, D., & Nussbeck, F. W. (2023). *Supplementary materials to "Estimation quality and required sample sizes in three-level contextual analysis models"* [R file, tabulated csv analyses results, codebook]. PsychOpen GOLD. <https://doi.org/10.23668/psycharchives.12944>

References

- Bollen, K., & Lennox, R. (1991). Conventional wisdom on measurement: A structural equation perspective. *Psychological Bulletin*, *110*(2), 305–314. <https://doi.org/10.1037/0033-2909.110.2.305>
- Bruncks, 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>
- Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine*, *25*(24), 4279–4292. <https://doi.org/10.1002/sim.2673>
- Chen, F., & Cui, Y. (2020). Investigating the relation of perceived teacher unfairness to science achievement by hierarchical linear modeling in 52 countries and economies. *Educational Psychology*, *40*(3), 273–295. <https://doi.org/10.1080/01443410.2019.1652248>
- Chen, Q. (2012). The impact of ignoring a level of nesting structure in multilevel mixture model: A Monte Carlo study. *SAGE Open*, *2*(1). <https://doi.org/10.1177/2158244012442518>
- de Jong, K., Moerbeek, M., & van der Leeden, R. (2010). A priori power analysis in longitudinal three-level multilevel models: An example with therapist effects. *Psychotherapy Research*, *20*(3), 273–284. <https://doi.org/10.1080/10503300903376320>
- Diez-Roux, A. V. (1998). Bringing context back into epidemiology: variables and fallacies in multilevel analysis. *American Journal of Public Health*, *88*(2), 216–222. <https://doi.org/10.2105/AJPH.88.2.216>
- Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses for moderator effects in three-level cluster randomized trials. *Journal of Experimental Education*, *86*(3), 489–514. <https://doi.org/10.1080/00220973.2017.1315714>
- Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review*, *40*(4), 334–377. <https://doi.org/10.1177/0193841X16671283>
- Enders, C. K. (2013). Centering predictors and contextual effects. In M. A. Scott, J. S. Simonoff, & B. D. Marx (Eds.), *SAGE handbook of multilevel modeling* (pp. 89–108). SAGE. <https://doi.org/10.4135/9781446247600.n6>
- Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods*, *9*(4), 466–491. <https://doi.org/10.1037/1082-989X.9.4.466>
- Grilli, L., & Rampichini, C. (2011). The role of sample cluster means in multilevel models: A view on endogeneity and measurement error issues. *Methodology*, *7*(4), 121–133. <https://doi.org/10.1027/1614-2241/a000030>
- Harker, R., & Tymms, P. (2004). The effects of student composition on school outcomes. *School Effectiveness and School Improvement: An International Journal of Research, Policy and Practice*, *15*(2), 177–199. <https://doi.org/10.1076/sesi.15.2.177.30432>

- Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review, 37*(6), 445–489. <https://doi.org/10.1177/0193841X14529126>
- Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147–154). Springer.
- Kerckhoff, D., & Nussbeck, F. W. (2019). The influence of sample size on parameter estimates in three-level random-effects models. *Frontiers in Psychology, 10*, Article 1067. <https://doi.org/10.3389/fpsyg.2019.01067>
- Kerckhoff, D., & Nussbeck, F. W. (2022). Obtaining sound intraclass correlation and variance estimates in three-level models: The role of sampling-strategies. *Methodology, 18*(1), 5–23. <https://doi.org/10.5964/meth.7265>
- Korendijk, E. J., Hox, J. J., Moerbeek, M., & Maas, C. J. (2011). Robustness of parameter and standard error estimates against ignoring a contextual effect of a subject-level covariate in cluster-randomized trials. *Behavior Research Methods, 43*(4), 1003–1013. <https://doi.org/10.3758/s13428-011-0094-8>
- Lee, E., & Hong, S. (2021). Adequate sample sizes for a three-level growth model. *Frontiers in Psychology, 12*, Article 685496. <https://doi.org/10.3389/fpsyg.2021.685496>
- Li, W., & Konstantopoulos, S. (2016). Class size effects on fourth-grade mathematics achievement: Evidence from TIMSS 2011. *Journal of Research on Educational Effectiveness, 9*(4), 503–530. <https://doi.org/10.1080/19345747.2015.1105893>
- 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>
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research, 44*(6), 764–802. <https://doi.org/10.1080/00273170903333665>
- Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed). Muthén & Muthén. https://www.statmodel.com/download/usersguide/MplusUserGuideVer_8.pdf
- Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling, 9*(4), 599–620. https://doi.org/10.1207/S15328007SEM0904_8
- R Core Team. (2021). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rathmann, K., Herke, M., & Richter, M. (2020). The role of class-level composition and schools' contextual characteristics for school-aged children's life satisfaction: A three-level multilevel

- analysis. *International Journal of Adolescence and Youth*, 25(1), 12–24.
<https://doi.org/10.1080/02673843.2019.1573743>
- Raudenbush, S. W., & Willms, J. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20(4), 307–335. <https://doi.org/10.3102/10769986020004307>
- Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, 41(5), 481–520.
<https://doi.org/10.3102/1076998616646200>
- Usami, S. (2017). Generalized sample size determination formulas for investigating contextual effects by a three-level random intercept model. *Psychometrika*, 82(1), 133–157.
<https://doi.org/10.1007/s11336-016-9532-y>
- Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54(3), 426–482.
<https://doi.org/10.1090/S0002-9947-1943-0012401-3>
- Zitzmann, S., Lüdtke, O., & Robitzsch, A. (2015). A Bayesian approach to more stable estimates of group-level effects in contextual studies. *Multivariate Behavioral Research*, 50(6), 688–705.
<https://doi.org/10.1080/00273171.2015.1090899>



Methodology is the official journal of the European Association of Methodology (EAM).



leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.