METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

# Obtaining Sound Intraclass Correlation and Variance Estimates in Three-Level Models: The Role of Sampling-Strategies

Denise Kerkhoff[1] , Fridtjof W. Nussbeck[1]

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

Methodology, 2022, Vol. 18(1), 5–23, https://doi.org/10.5964/meth.7265

**Received:** 2021-08-03 • **Accepted:** 2022-02-03 • **Published (VoR):** 2022-03-31

**Corresponding Author:** Denise Kerkhoff, University of Konstanz, Department of Psychology, Universitätsstraße 10, P.O. Box 14, 78464 Konstanz, Germany. E-mail: denise.kerkhoff@uni-konstanz.de

**Supplementary Materials:** Data [see Index of Supplementary Materials]

## Abstract

Three-level clustered data commonly occur in social and behavioral research and are prominently analyzed using multilevel modeling. The influence of the clustering on estimation results is assessed with the intraclass correlation coefficients (ICCs), which indicate the fraction of variance in the outcome located at each higher level. However, ICCs are prone to bias due to high requirements regarding the overall sample size and the sample size at each data level. In Monte Carlo simulations, we investigate how these sample characteristics influence the bias of the ICCs and statistical power of the variance components using robust ML-estimation. Results reveal considerable underestimation on Level-3 and the importance of the Level-3 sample size in combination with the ICC sizes. Based on our results, we derive concise sampling recommendations and discuss limits to our inferences.

## Keywords

hierarchical linear modeling, Monte Carlo simulation, statistical power, sample size, bias

In the behavioral research and related fields, researchers increasingly employ multilevel modeling on three-levels to analyze clustered data. These data structures consist of measurement objects at Level-1, which are nested within higher-order Level-2 subclusters, which are nested in Level-3 clusters, such as patients within therapists within clinics (e.g., Firth et al., 2019).

Common estimates of the influence of the clustering are the intraclass correlation coefficients (ICCs), which depict the fraction of variance in the criterion located at Level-2 ($ICC_2$) or Level-3 ($ICC_3$). They thus indicate the extent to which the criterion is shaped by superordinate clusters, which may reveal, e.g., at which level(s) predictors or random effects should be included to reduce unexplained variance, or they inform the sampling procedure, ensuring high estimation quality in cluster-randomized experiments (Hedges & Hedberg, 2013). ICCs further indicate the degree to which standard errors are biased in standard regression analysis (two-level models: e.g., Kreft & de Leeuw, 1988; three-level models: e.g., Cunningham & Johnson, 2016). Therefore, ensuring that a nested structure is accurately characterized is a central issue, and reporting ICCs is a standard analysis step in multilevel modeling (see Hox et al., 2017).

However, research has shown that unbiased estimation and sufficient power in multilevel models depend on an interplay of adequate sample sizes on each level, the effect sizes, and the levels at which the predictors are modelled (more recently Cox & Kelcey, 2019; Kerkhoff & Nussbeck, 2019; LaHuis et al., 2020). To our knowledge, no study has systematically investigated required sample sizes to ensure sufficient estimation quality of the ICCs in three-level models. In this study, we shed light on this relationship and derive sampling recommendations for a wide range of ICC sizes. We limit our analyses to models with continuous responses, since the computation of ICCs in multilevel models with discrete and continuous responses differ (Goldstein et al., 2002; Leckie et al., 2020), resulting in limited comparability of estimation results and interpretation of the ICC sizes.

## Variance Decomposition in the Linear Three-Level Model

The linear three-level model with Level-1 units $i$ nested within Level-2 subclusters $j$ nested within Level-3 clusters $k$ decomposes the values of the criterion $Y_{ijk}$ as follows:

$$Y_{ijk} = \gamma_{000} + v_{00k} + u_{0jk} + e_{ijk} \tag{1}$$

The coefficient $\gamma_{000}$ is the grand mean of $Y$, $v_{00k} \sim N(0, \sigma^2_{v_0})$ is a random variable reflecting the cluster-specific deviations from the mean, $u_{0jk} \sim N(0, \sigma^2_{u_0})$ is the random variable of subcluster-specific deviations from the respective cluster mean, and $e_{ijk} \sim N(0, \sigma^2_e)$ is the random variable of deviations of the Level-1 unit values from the respective subcluster mean. The total variance of $Y$ is given by $\sigma^2_Y = \sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e$. The ICCs express the fraction of variance at the respective level in relation to the total variance:

$$\text{Level 3:} \quad ICC_3 = \frac{\sigma^2_{v_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \tag{2}$$

$$\text{Level 2:} \quad ICC_2 = \frac{\sigma^2_{u_0}}{\sigma^2_{v_0} + \sigma^2_{u_0} + \sigma^2_e} \tag{3}$$

For the $ICC_2$, an alternative approach is to include both higher-level variances in the enumerator ($\sigma^2_{v_0} + \sigma^2_{u_0}$). The value then expresses how strongly any two Level-1 units from the same subcluster correlate (Hox et al., 2017). Typical ICC values are most comprehensively reported for educational research, ranging from 0.1 to 0.3 (Dong et al., 2016; Hedges & Hedberg, 2013), and research has demonstrated that even small ICCs can result in meaningful bias if the clustering is ignored (Lai & Kwok, 2015).

## Assessing Estimation Quality of the Variance Components

When using Monte Carlo (MC) simulations to investigate estimation quality in multilevel models across many generated samples, the most common measures of estimation quality are the parameter estimation bias (*PEB*) to evaluate the accuracy of point estimates, and measures of significance to evaluate statistical power.

### Parameter Estimation Bias (PEB)

The *PEB* reflects the percentage of over- or underestimation of a population parameter. We refer to this commonly reported *PEB* (e.g., Muthén & Muthén, 2002) as *relative PEB.* For a population parameter $\theta$, estimated by $\widehat{\theta}_i$ in $i = 1, \ldots, n$ replications, it is defined as

$$rPEB_\theta = \frac{\sum_{1 \le i \le n} \frac{\widehat{\theta}_i - \theta}{\theta}}{n} \tag{4}$$

We argue that the relative *PEB* alone is not sufficient for evaluating bias: By design, under- and overestimation balance out when computing the relative bias across replications. This can result in a relative *PEB* close to zero, although each replication might be either over- or underestimated. The magnitude of bias for a single sample can more precisely be approximated by the absolute *PEB*, given by:

$$aPEB_\theta = \frac{\sum_{1 \le i \le n} \frac{|\widehat{\theta}_i - \theta|}{\theta}}{n} \tag{5}$$

For the absolute *PEB*, only the strength of the bias is averaged across replications, and the direction of bias for each sample is ignored. Together, the relative and absolute *PEB* reflect the expected strength and overall direction of bias.

### Statistical Power

Statistical power of a parameter in the context of simulation studies is the rate of replications with a statistically significant point estimate. However, the commonly used

Wald-test (Wald, 1943) is unreliable for variance components (Berkhof & Snijders, 2001; Molenberghs & Verbeke, 2004; Raudenbush & Bryk, 2002).

An alternative procedure to test statistical significance of parameters in multilevel models is the $\chi^2$-test-based comparison between the full model and a nested model where the parameter of interest is constrained to be zero (Hox et al., 2017). Based on the scaled $\chi^2$-test statistic (Satorra & Bentler, 1994), the modified test statistic $\tilde{T}_d$ proposed by Satorra and Bentler (2010, henceforth: SB-test) is computed using the unscaled maximum likelihood $\chi^2$-values of the nested ($T_0$) and full ($T_1$) models, the respective scaled $\chi^2$-test statistics of $\bar{T}_0$ and $\bar{T}_1$, with degrees of freedom $r_0$ and $r_1$, and scaling correction factors $\hat{c}_0 = T_0/\bar{T}_0$ and $\hat{c}_1 = T_1/\bar{T}_1$:

$$\tilde{T}_d := \frac{\bar{T}_0 * \hat{c}_0 - \bar{T}_1 * \hat{c}_1}{\tilde{c}_d}, \quad with \quad \tilde{c}_d = \frac{r_0 * \hat{c}_0 - r_1 * \hat{c}_1}{r_0 - r_1} \tag{6}$$

The score is compared to a $\chi^2$-distribution with $r_0 - r_1$ degrees of freedom. As an alternative, the likelihood-ratio test (LRT) follows the same procedural logic but uses the log-likelihood to compute the test statistic (see Herzog et al., 2007, for an overview). To assess the statistical significance of the Level-2 (Level-3) variance, the empty model (Eq. 1) is compared to a nested model where the Level-2 (Level-3) variance is constrained to be zero (Greven et al., 2008) using a one-sided test, since significantly negative variance values are inadmissible (Berkhof & Snijders, 2001). However, this approach has drawbacks. Particularly, both the LRT and SB-test can result in inadmissible, negative test-values, requiring a more complex approach using an auxiliary model (Bryant & Satorra, 2012; Satorra & Bentler, 2010). Further, the distribution of the test statistic has been shown to more closely follow a $\chi^2$-mixture distribution, but nevertheless, research suggests that $\chi^2$-based tests for the variances still produce sufficiently accurate results (Dominicus et al., 2006; LaHuis & Ferguson, 2009).

## Sampling Recommendations

Despite its importance as a measure of cluster influences, there are no comprehensive recommendations for unbiased estimation of the ICC values, and only few studies report on the estimation quality of the variance components. In Kerkhoff and Nussbeck (2019), a small Level-3 intercept variance is heavily overestimated in samples with up to 55 clusters with 5 or 15 units per cluster. For ten or less sampled Level-3 clusters, McNeish and Wentzel (2017) found that the Level-2 variance is underestimated by up to 5% and the Level-3 variance is underestimated by 30% or more. Although these studies do not provide information regarding accurate estimation of the ICCs, they demonstrate that the estimation of variances is not trivial in three-level models, and that substantial bias can occur if the sampling process is not optimized.

## Aim of This Study

By means of extensive Monte Carlo simulations, we evaluate how estimation bias and statistical power relate to the overall sample size, the allocation of units on each level, and the ICC sizes. We are particularly interested in minimum required samples sizes and advantageous sampling-strategies for overall sound estimation.

# Method

## Simulation Setup

For the simulations, we used the empty three-level model (Eq. 1), with $\gamma_{000} = 0$ and $Y_{ijk} \sim N(0, \sigma^2_Y)$. In total, we generated 1,125 simulation conditions: 5 ($n_3$ = 5, 10, 50, 100, or 200) × 5 ($n_2$ = 2, 5, 10, 20, or 30) × 5 ($n_1$ = 2, 5, 10, 20, or 30) × 3 ($\sigma^2_{v_0}$ = 0.1, 0.2, or 0.6) × 3 ($\sigma^2_{u_0}$ = 0.1, 0.2, or 0.6) × 1 ($\sigma^2_e$ = 1). The combinations of variances yielded nine different combinations of $ICC_3$ and $ICC_2$, calculated as in Eq. 2 and 3 (see Table 1). We chose ICC values and sample sizes based on empirical findings in the social and behavioral research (e.g., Dong et al., 2016; Kerkhoff & Nussbeck, 2019).

**Table 1**

*Variance Sizes, Resulting ICC Combinations, and Notation*

| $\sigma^2_{v_0}$ | $\sigma^2_{u_0}$ | $ICC_3$/$ICC_2$ | Notation |
|---|---|---|---|
| 0.1 | 0.1 | .083/.083 | S/S |
| 0.1 | 0.2 | .077/.154 | S/M |
| 0.1 | 0.6 | .059/.353 | S/L |
| 0.2 | 0.1 | .154/.077 | M/S |
| 0.2 | 0.2 | .143/.143 | M/M |
| 0.2 | 0.6 | .111/.333 | M/L |
| 0.6 | 0.1 | .353/.059 | L/S |
| 0.6 | 0.2 | .333/.111 | L/M |
| 0.6 | 0.6 | .273/.273 | L/L |

*Note.* $\sigma^2_{v_0}$ = Level-3 variance, $\sigma^2_{u_0}$ = Level-2 variance, Level-1 residual variance $\sigma^2_e = 1$ in all conditions; ICC = Intraclass correlation coefficients for Level-3 ($ICC_3$) and Level-2 ($ICC_2$).

For every condition, we generated 1,000 samples. For each sample, we fitted the empty model using robust maximum likelihood estimation with the expectation maximization algorithm and 500 admissible iterations. Data generation and model estimation was done in Mplus Version 8 (Muthén & Muthén, 1998-2017), and results were imported to R 6.3.1 (R Core Team, 2019) using the MplusAutomation package (Hallquist & Wiley, 2018) for subsequent analyses.

We refer to the total number of observations in a condition ($n_3 \times n_2 \times n_1$) as NOBS. We further refer to the specific combination of $n_3$, $n_2$, and $n_1$ in a condition as "allocation" and abbreviate the allocation by $n_3$-size/$n_2$-size/$n_1$-size. For example, 100/20/5 subsumes samples with 100 clusters, each with 20 subclusters, which in turn contain 5 Level-1 units. 100/5/• comprises all conditions with 100 clusters, each with 5 subclusters and any number of Level-1 units. ICC sizes are summarized where appropriate by L = large (ICC = .273, .333, or .353), M = medium (ICC = .111, .143, or .154), and S = small (ICC = .059, .077, or .083). ICC sizes of a condition are abbreviated by $ICC_3$-size/$ICC_2$-size (see Table 1).

## Outcome Measures

We report convergence rates, but computed coefficients only across runs that converged normally. To explore the influence of sample sizes on bias (*rPEB* and *aPEB* for $ICC_2$ and $ICC_3$), we ran four analyses of variance, and report the partial omega-squared ($\omega_p^2$) effect size. We further computed the median estimate, the upper and lower quartiles, and the relative and absolute *PEB* of the ICCs and the Level-1 variance. We considered -0.10 < *rPEB* < 0.10 as sufficiently unbiased (see Flora & Curran, 2004; Muthén & Muthén, 2002). There are no established thresholds to indicate absolute unbiasedness, so we considered *aPEB* < 0.15 to be unbiased, since we argue that in practice, less than 15% over- or underestimation of an ICC does not considerably change statistical inferences. We further calculated the rate of biased runs, which is the percentage of replications in a condition with at least 15% over- or underestimation, as an indicator for the risk of producing a biased estimate.

For the Level-2 and Level-3 variance components, we assessed statistical power by the rate of significant one-sided SB-tests as in Eq. 6[1]. We considered a power of 80% or higher as sufficient (see Cohen, 1992; Muthén & Muthén, 2002).

# Results

First, since Level-1 residuals were estimated accurately in most conditions (relative unbiasedness in NOBS > 50, absolute unbiasedness in NOBS > 100), we do not provide detailed information about the estimation quality on Level-1. Complete results are tabulated in the supplementary dataset.

Out of the 1,125 conditions, 299 conditions (all $n_3 \leq 10$, 26.58% of all conditions) produced at least one replication that did not converge normally. Moreover, convergence problems occurred more frequently for conditions with small $ICC_3$ (see also Table 2).

1) While we initially computed both LRTs and SB-tests, we chose not to report results for the LRT, since the rate of inadmissible test values was considerably higher.

**Table 2**

*Sampling Conditions With a Convergence Rate of 95% or Less*

| $n_3/n_2/n_1$ | $ICC_3/ICC_2$ | | |
|---|---|---|---|
| 5/2/2 | All | | |
| 5/2/5 | All | | |
| 5/2/10 | S/• | M/• | L/S |
| 5/2/20 | S/• | M/• | L/L |
| 5/2/30 | S/• | M/M | M/L |
| 5/5/2 | S/• | M/• | |
| 5/5/5 | S/• | M/M | M/L |
| 5/5/10 | S/• | M/L | |
| 5/5/20 | S/M | S/L | M/L |
| 5/5/30 | S/M | S/L | M/L |
| 5/10/2 | S/• | M/M | M/L |
| 5/10/5 | S/M | S/L | |
| 5/10/10 | S/L | M/L | |
| 5/10/20 | S/L | M/L | |
| 5/10/30 | S/L | | |
| 5/20/2 | S/L | | |
| 10/2/2 | S/• | M/S | |
| 10/2/5 | S/S | M/S | |

*Note.* ICC = Intraclass correlation coefficients for Level-3 ($ICC_3$) and Level-2 ($ICC_2$). $n_1$, $n_2$, $n_3$ indicate the number of clusters ($n_3$), subclusters per cluster ($n_2$), and Level-1 units per subcluster ($n_1$). S, M, L indicate small, medium, and large ICC sizes, respectively.

## Results Across Allocations

Median estimates, quartiles of ICC estimates, *PEB* and power across all sampling conditions are presented in Table 3. Results show that the $ICC_3$ tended to be underestimated, while the $ICC_2$ tended to be overestimated. Analyses of variance (Table 4) show that on Level-3, the number of clusters was most influential on both, relative and absolute *PEB*. On Level-2, relative bias was mostly influenced by the total sample size, while absolute bias was mostly influenced by the number of clusters.

**Table 3**

*Estimates and Bias of the ICCs and Power of the Variances Across Sampling Conditions*

| | Level-3 | | | | | Level-2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $ICC_3$ estimate | | $rPEB_{ICC_3}$ | $aPEB_{ICC_3}$ | Power of $\sigma^2_{v_0}$ | $ICC_2$ estimate | | $rPEB_{ICC_2}$ | $aPEB_{ICC_2}$ | Power of $\sigma^2_{u_0}$ |
| **ICC** | ***Mdn*** | ***[Q1; Q3]*** | ***Mdn*** | ***Mdn*** | ***Mdn*** | ***Mdn*** | ***[Q1; Q3]*** | ***Mdn*** | ***Mdn*** | ***Mdn*** |
| .059 | .058 | [.053; .059] | -.016 | .420 | .865 | .060 | [.059; .065] | .021 | .201 | 1.000 |
| .077 | .075 | [.067; .076] | -.024 | .275 | .992 | .078 | [.077; .080] | .013 | .166 | 1.000 |
| .083 | .081 | [.072; .082] | -.026 | .240 | .982 | .084 | [.084; .085] | .008 | .163 | 1.000 |
| .111 | .109 | [.097; .110] | -.020 | .273 | 1.000 | .113 | [.112; .118] | .016 | .155 | 1.000 |
| .143 | .139 | [.123; .141] | -.028 | .183 | 1.000 | .144 | [.143; .146] | .007 | .125 | 1.000 |
| .154 | .150 | [.131; .152] | -.026 | .163 | 1.000 | .154 | [.154; .155] | .004 | .120 | 1.000 |
| .273 | .266 | [.233; .270] | -.026 | .145 | 1.000 | .275 | [.274; .285] | .009 | .107 | 1.000 |
| .333 | .325 | [.288; .330] | -.024 | .117 | 1.000 | .334 | [.333; .336] | .002 | .081 | 1.000 |
| .353 | .345 | [.308; .349] | -.023 | .111 | 1.000 | .353 | [.350; .354] | .001 | .073 | 1.000 |

*Note.* $rPEB_\theta$ = relative *PEB* of θ; $aPEB_\theta$ = absolute *PEB* of θ; *Mdn* = median; *Q1* = lower quartile; *Q3* = upper quartile; ICC = Intraclass correlation coefficients for Level-3 ($ICC_3$) and Level-2 ($ICC_2$); $\sigma^2_{v_0}$ = Level-3 variance; $\sigma^2_{u_0}$ = Level-2 variance.

**Table 4**

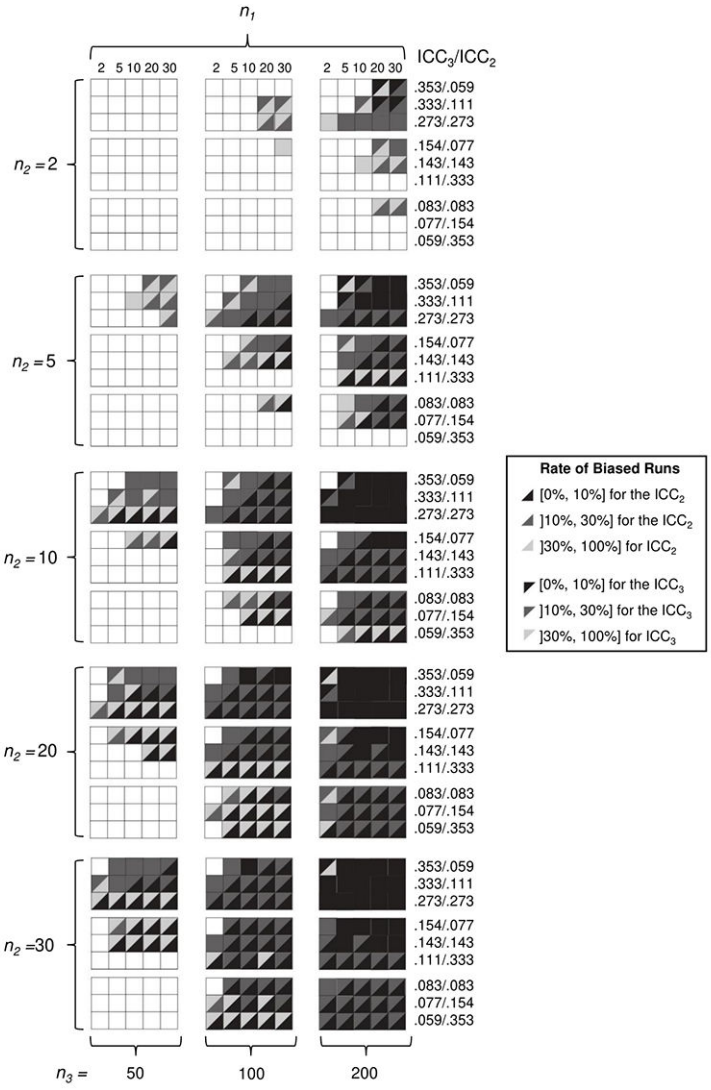
*Effect Sizes (Partial $\omega^2$) Resulting From Analyses of Variance in Relative and Absolute Bias of the ICCs.*

| | | Level-3 | | Level-2 | |
|---|---|---|---|---|---|
| | | $rPEB_{ICC_3}$ | $aPEB_{ICC_3}$ | $rPEB_{ICC_2}$ | $aPEB_{ICC_2}$ |
| **Factor** | ***df*** | $\omega^2_p$ | $\omega^2_p$ | $\omega^2_p$ | $\omega^2_p$ |
| Population $\sigma^2_{v_0}$ | 2 | .043 | .427 | .072 | .042 |
| Population $\sigma^2_{u_0}$ | 2 | .033 | .198 | .087 | .363 |
| $n_3$ | 4 | .462 | .814 | .074 | .651 |
| $n_2$ | 4 | .074 | .508 | -.002 | .565 |
| $n_1$ | 4 | .002 | .106 | .093 | .526 |
| NOBS | 34 | .098 | .177 | .142 | .456 |

*Note.* $rPEB_\theta$ = relative *PEB* of θ; $aPEB_\theta$ = absolute *PEB* of θ; $\omega^2_p$ = partial omega squared; ICC = Intraclass correlation coefficients for Level-3 ($ICC_3$) and Level-2 ($ICC_2$); $\sigma^2_{v_0}$ = Level-3 variance; $\sigma^2_{u_0}$ = Level-2 variance. $n_1$, $n_2$, $n_3$ indicate the number of clusters ($n_3$), subclusters per cluster ($n_2$), and Level-1 units per subcluster ($n_1$). NOBS = total number of observations. Residual *df* = 1074.

**Figure 1**

*Conditions With Sufficient Power, Relative Unbiasedness, and Absolute Unbiasedness*




*Note.* Each square represents a condition, ordered by $n_3$ (grouped columns), $n_2$ (grouped rows), $n_1$ (single columns within an $n_3$ group), and ICC sizes (single rows within an $n_2$ group). Shaded diagonals show the rate of runs with at least 15% under- or overestimation of the $ICC_3$ (upper diagonal) and $ICC_2$ (lower diagonal). White squares indicate that the condition did not yield unbiased estimates or sufficient power at all levels.

In total, 384 conditions (34.13%) resulted in sufficient power and relative and absolute unbiasedness on all levels. These conditions can be identified in Figure 1 as squares with gray areas, where the shades of the triangles indicate the rate of biased runs for the $ICC_3$ (upper triangle) and $ICC_2$ (lower triangle). For example, in 200/2/2, only the condition with large ICCs was estimated with sufficient power and without bias, yet both ICCs were still biased in more than 30% of replications in this condition. Conditions with 5 or 10 clusters did not result in accurate estimation across levels. Overall, conditions with at least 200/20/2 or 100/20/5 resulted in sufficient estimation quality across levels.

## Estimation of Variance Components

In Figure 2, we present the median, upper and lower quartiles for $\sigma^2_{v_0}$ (upper plot) and $\sigma^2_{u_0}$ (lower plot) across allocations. We do not provide information for conditions with more than 10,000 observations because estimation quality did not further improve for larger samples. Since statistical power of all variance sizes was sufficient for a variety of conditions, we do not plot power results, but report requirements for sufficient power in conjunction with Table 5. Comprehensive estimation results are tabulated in columns 13 to 24 in the supplementary dataset.

### Level-3 Variance Component

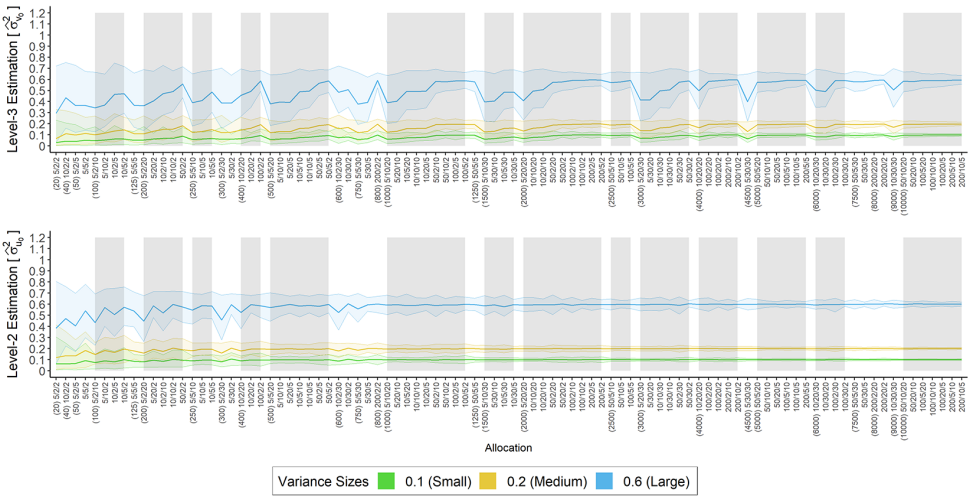
In general, $\sigma^2_{v_0}$ was consistently underestimated in most conditions, and estimates fluctuated strongly across replications for the smallest samples. Within conditions with varying allocations but equal NOBS (gray areas in Figure 2), conditions with large $n_3$ produced less biased estimates and less fluctuations. Sufficient power for the estimation of Level-3 variances was ensured in most conditions with 7,500 or more observations and some conditions with smaller samples, such as 10/30/•, 50/10/•, or 100/5/•. Sufficient power for medium and large Level-3 variance components was obtained with at least 2,500 observations, and conditions •/30/5, 10/20/•, 50/5/•, •/20/10, or 10/10/5 (see Table 5).

### Level-2 Variance Component

Estimates of $\sigma^2_{u_0}$ tended to be underestimated for less than 1,250 observations and, similar to Level-3 findings, estimates fluctuated more heavily across replications with small samples (especially NOBS ≤ 1000). Power was sufficient for most allocations with at least 500 observations and some conditions with less observations, such as 10/•/20. For medium and large Level-2 variances, power was sufficient for conditions with at least 500 observations, or allocations with at least •/5/10, •/10/5, or $n_1 \geq 20$ (see Table 5).

**Figure 2**

*Median Estimates, Upper and Lower Quartiles of Higher-Level Variances Across Allocations*




*Note.* Plotted are medians, upper and lower quartiles of $\sigma^2_{v_0}$ (upper plot) and $\sigma^2_{u_0}$ (lower plot) estimates across sample size conditions with up to 10,000 observations. Lines represent the median estimates, colored ribbons encompass upper and lower quartiles. Colors represent large (blue: $\sigma^2_{v_0}$, $\sigma^2_{u_0}$ = 0.6), medium (yellow: $\sigma^2_{v_0}$, $\sigma^2_{u_0}$ = 0.2), and small (green: $\sigma^2_{v_0}$, $\sigma^2_{u_0}$ = 0.1) population variance sizes. On the x-axis, numbers in parentheses indicate the overall number of observations in this and—if applicable—the following conditions. Shaded areas mark multiple conditions with the same number of observations but differing allocations. Values for each variance component's population parameter have been averaged (e.g., estimates for $\sigma^2_{v_0}$ = 0.1 are average scores across $\sigma^2_{u_0}$ = 0.1, 0.2 and 0.6)

## Intraclass Correlation Coefficients

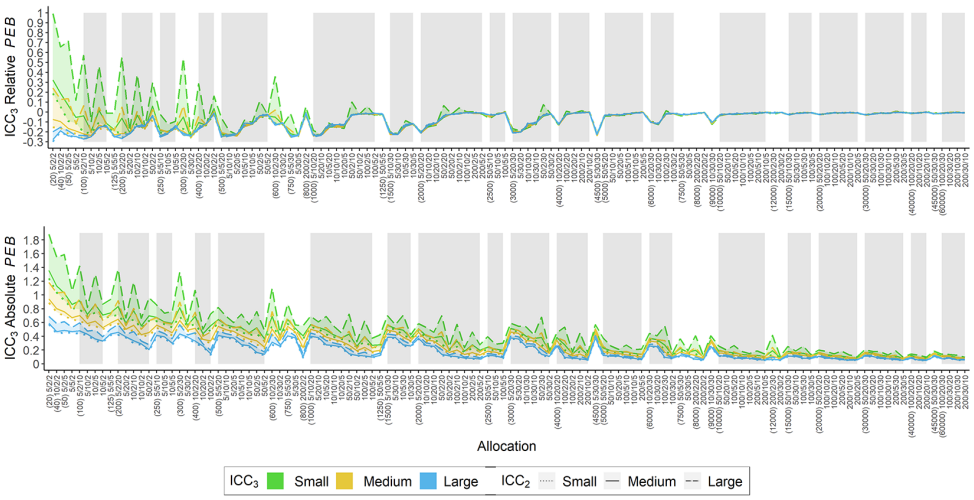
Allocations to achieve unbiasedness for the ICCs are presented in Table 5. Results for relative and absolute bias are plotted in Figure 3 (Level-3) and Figure 4 (Level-2). Conditions with large NOBS did not result in meaningfully biased estimates and are not presented. Comprehensive bias results are tabulated in columns 25 to 33 in the supplementary dataset.

### Results for the $ICC_3$

In most conditions with $n_3 \geq 100$ and minimum sample sizes of 200/20/• or 100/30/•, we found less than 15% absolute bias. For medium and large $ICC_3$ sizes, sampling less Level-2 units (200/10/• and 100/20/•) or emphasizing the number of Level-3 units (200/5/5) also resulted in absolute unbiasedness. Further, conditions with $n_3 = 50$, even conditions with large NOBS, such as 50/30/30, were considerably biased for small and most medium sized

**Figure 3**

*Absolute and Relative PEB for the $ICC_3$ Across Allocations*




*Note.* Plotted are the relative (upper plot) and absolute (lower plot) *PEB* of the $ICC_3$ in conditions with up to 60,000 observations. $ICC_3$ sizes are differentiated by color (green = small, yellow = medium, blue = large), $ICC_2$ sizes are differentiated by line type (dotted = small, line = medium, dashed = large). Areas are colored to facilitate readability. On the x-axis, numbers in parentheses show the number of observations in this and—if applicable—the following conditions. Shaded areas mark multiple conditions with the same number of observations but differing allocations.

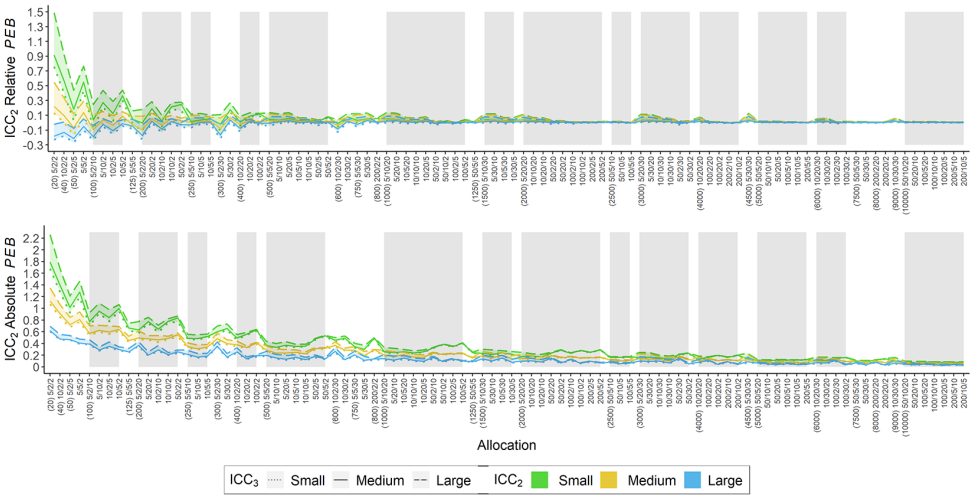
$ICC_3$. Notably, bias of the $ICC_3$ was higher in conditions where the complementing $ICC_2$ was large rather than medium or small.

## Results for the $ICC_2$

$ICC_2$ relative bias remained within 10% over-/underestimation for a variety of allocations, such as small $n_3$ (e.g., 10/5/10, 10/10/5) or small $n_2$ and $n_1$ (e.g., 200/2/2, 50/2/5, 50/5/2). Absolute unbiasedness was achieved in most conditions with at least 5,000 observations (e.g., 10/30/20, 50/10/10, 100/5/10, 200/5/5). For medium or large $ICC_2$, conditions with smaller NOBS, such as 100/5/5 or 50/5/10 also resulted in absolute unbiasedness.

**Figure 4**

*Absolute and Relative PEB for the $ICC_2$ Across Allocations*




*Note.* Plotted are the relative (upper plot) and absolute (lower plot) *PEB* of the $ICC_2$ in conditions with up to 10,000 observations. $ICC_2$ sizes are differentiated by color (green = small, yellow = medium, blue = large), $ICC_3$ sizes are differentiated by line type (dotted = small, line = medium, dashed = large). Areas are colored to facilitate readability. On the x-axis, numbers in parentheses show the number of observations in this and—if applicable—the following conditions. Shaded areas mark multiple conditions with the same number of observations but differing allocations.

Methodology
2022. Vol. 18(1), 5–23
https://doi.org/10.5964/meth.7265

**Table 5**

*Minimum Allocations to Achieve Unbiasedness and Sufficient Power, Differentiated by $n_3$.*

| | Level-3 | | | | Level-2 | | | |
|---|---|---|---|---|---|---|---|---|
| $n_3$ | $ICC_3$ | $rPEB_{ICC_3}$ | $aPEB_{ICC_3}$ | Power of $\sigma^2_{v_0}$ | $ICC_2$ | $rPEB_{ICC_2}$ | $aPEB_{ICC_2}$ | Power of $\sigma^2_{u_0}$ |
| 5 | S | none[a] | none | 20/10[a] | S | 5/5[a] | none[a] | $n_2 \times n_1 \geq 100$ |
| | M | none[a] | none | $n_2 \times n_1 \geq 60$, if $n_2 \geq 10$ | M | $n_2 \times n_1 \geq 20$[a] | none[a] | $n_2 \times n_1 \geq 40$, if $n_1 \geq 5$ |
| | L | none | none | $n_2 \times n_1 \geq 20$, if $n_2 \geq 5$ | L | $n_2 \geq 5$ | 20/5[a] or 10/20[a] | $n_2 \times n_1 \geq 20$ |
| 10 | S | none[a] | none | $n_2 \times n_1 \geq 40$, if $n_2 \geq 10$ | S | $n_2 \times n_1 \geq 20$, except 10/2 | $n_2 \times n_1 \geq 300$, except 10/30 | $n_2 \times n_1 \geq 40$, if $n_1 \geq 5$ |
| | M | none[a] | none | $n_2 \times n_1 \geq 10$ | M | any[a] | $n_2 \times n_1 \geq 150$, if $n_2 \geq 10$[a] | $n_2 \times n_1 \geq 25$, except 20/2 |
| | L | none | none | $n_2 \times n_1 \geq 10$, except 2/5 | L | $n_2 \times n_1 \geq 10$, except 2/5 | $n_2 \times n_1 \geq 40$, if $n_2 \geq 10$ | $n_2 \times n_1 \geq 10$ |
| 50 | S | any[a] | none | $n_2 \geq 5$ | S | $n_2 \times n_1 \geq 10$ | $n_2 \times n_1 \geq 100$ | $n_2 \times n_1 \geq 20$, except 10/2 |
| | M | any | 30/5[a] or 20/20[a] | $n_2 \times n_1 \geq 10$ | M | any | $n_2 \times n_1 \geq 50$, if $n_2 \geq 5$ | $n_2 \times n_1 \geq 10$ |
| | L | any | any except 2/2, 2/5 | any | L | any | $n_2 \geq 5$ or $n_1 \geq 20$ | any |
| 100 | S | any | 20/2 or 10/10[a] | $n_2 \times n_1 \geq 10$, except 2/5 | S | $n_2 \times n_1 \geq 10$ | $n_2 \times n_1 \geq 50$, if $n_1 \geq 5$ | $n_2 \geq 10$ or $n_1 \geq 5$ |
| | M | any | $n_2 \times n_1 \geq 20$, if $n_2 \geq 5$ | $n_2 \times n_1 \geq 10$ | M | any | $n_2 \times n_1 \geq 25$ | $n_2 \times n_1 \geq 10$ |
| | L | any | any | any | L | any | $n_2 \times n_1 \geq 10$ | any |
| 200 | S | any | $n_2 \times n_1 \geq 20$, if $n_2 \geq 5$[a] | $n_2 \times n_1 \geq 10$ | S | any | $n_2 \times n_1 \geq 25$ | $n_2 \times n_1 \geq 10$ |
| | M | any | $n_2 \times n_1 \geq 10$,[a] except 2/5 | any | M | any | $n_2 \times n_1 \geq 20$ | any |
| | L | any | any | any | L | any | any | any |

*Note.* $rPEB_0$ = relative *PEB* of θ; $aPEB_0$ = absolute *PEB* of θ. -0.1 < *rPEB* < 0.1 and *aPEB* < 15% indicate relative and absolute unbiasedness, power ≥ 0.8 indicates sufficient power. ICC = Intraclass correlation coefficients for Level-3 ($ICC_3$) and Level-2 ($ICC_2$); $\sigma^2_{v_0}$ = Level-3 variance; $\sigma^2_{u_0}$ = Level-2 variance. S, M, L indicate small, medium, and large ICC sizes, respectively. $n_1$, $n_2$, $n_3$ indicate the number of clusters ($n_3$), subclusters per cluster ($n_2$), and Level-1 units per subcluster ($n_1$). $n_2 \times n_1$ refers to the multiplied sample size, e.g., $n_2 \times n_1 = 10$ is realized with $n_2 = 2$ and $n_1 = 5$, or with $n_2 = 5$ an $n_1 = 2$.

[a]with exceptions, e.g.: 20/20[a] indicate that most conditions with $n_2 = n_1 = 20$ are unbiased or with sufficient power.

# Discussion

Our findings extend our knowledge on the estimation quality in three-level modeling by showing that moderate to large samples and an advantageous allocation are needed for overall good estimation quality of the ICCs, and that the size of the ICCs and the number of available clusters greatly influences required sample sizes.

## The Role of the Sampling-Strategies

Results demonstrate that required $n_2/n_1$ depend on the available number of clusters and ICC scores (see Table 5). Consequently, even large samples may result in biased estimates if allocations are suboptimal. Although the number of clusters ($n_3$) is most important to ensure estimation quality, statistical power and relative unbiasedness on Level-2 can be achieved even with a small number of clusters. Specifically, to achieve sufficient power, it is advantageous to sample $n_1 \geq 5$ to increase power at Level-2, and $n_2 \geq 5$ to increase power at Level-3.

Interestingly, we found that the variance components are consistently underestimated. Since $\sigma^2_{v_0}$ is more strongly underestimated than $\sigma^2_{u_0}$, the resulting $ICC_2$ is consistently overestimated, while the $ICC_3$ is underestimated, which can result in misinterpretation of the three-level structure. Similar patterns of underestimation are visible in previous research (McNeish & Wentzel, 2017) but have, to our knowledge, not yet been systematically investigated.

Further, convergence rates for the smallest samples are considerably low. Research suggests that restricted maximum likelihood (REML) may improve convergence and reduce bias in small samples (McNeish & Wentzel, 2017). However, additional analyses (not reported) show mixed results for our data: Since REML is not implemented in Mplus, we compared convergence rates and ICC bias resulting from REML and regular maximum likelihood (ML) estimation in R using the lme4-package (Bates et al., 2015) for the allocations listed in Table 2 (all ICC-sizes, i.e., 162 conditions). We considered all issues resulting in non-computable ICC values as convergence issues. Across conditions, REML estimation improved convergence rates by $Mdn$ = 3.75 percentage points. REML performed better for medium and large $ICC_2$ ($Mdn_{ML}$ = -0.44, $Mdn_{REML}$ = -0.33, for $ICC_2 \geq .143$), but worse for small $ICC_2$ ($Mdn_{ML}$ = 1.28, $Mdn_{REML}$ = 1.68, for $ICC_2 \leq .111$). For the $ICC_3$, differences between estimation methods were small ($Mdn_{ML}$ = 0.04, $Mdn_{REML}$ = 0.01). Absolute bias was comparably high for both ICCs and estimation methods.

## The Role of ICC-Sizes

Results show that smaller variance components require considerably larger samples for sufficient estimation quality. For example, small ICCs require at least twice (four times) as many observations as medium (large) ICCs for a given number of clusters for absolute unbiasedness. Similarly, in samples with 5 or 10 clusters, required $n_2/n_1$ sample sizes to

achieve sufficient power are at least two times (four times) higher for small variance components compared to medium (large) components.

Interestingly, the bias of an ICC estimate is higher if the ICC at the other level is larger. As an example, the $ICC_3$ in M/L was more heavily biased than in M/M or M/S. In additional simulations, we tested if this is a direct consequence of the simulation setup, since, for example, the $ICC_3$ was slightly smaller in S/L ($ICC_3$ = .059) than S/M and S/S ($ICC_3$ = .077, .083, respectively). These additional analyses (100 replications each for $n_3$ = 50, 100; $n_2$ = 5, 10; $n_1$ = 5, 10; $ICC_2$ = .077, .143, .333, $ICC_3$ = .143) indicated, however, that this pattern is still present if the $ICC_3$ size does not vary for different $ICC_2$ sizes. Furthermore, this pattern is similar for statistical power, where small Level-3 variances required larger samples for sufficient power if the respective Level-2 variance was large. This pattern remains unexplained and requires more detailed research on the relationship between Level-3 and Level-2 bias.

## Evaluation of Estimation Quality Indicators

Most importantly, our results demonstrate that relative unbiasedness of a simulation condition does not imply that a sample generated from this condition produces unbiased estimates, as indicated by the rate of biased runs and the absolute *PEB*.

Further, our inferences regarding statistical power are based on the one-sided SB-test. Our findings may therefore not be directly compared to previous research, since there is no single established coefficient assessing the power of variance estimates in multilevel research. Hence, approaches incorporating auxiliary models for the SB- or LRT-test or differing test distributions might suggest different sampling requirements, and we suggest that future studies include and compare different power measures in their simulations.

## Concluding Recommendations

As a rule of thumb, overall estimation quality is achieved if samples ensure absolute unbiasedness of Level-3 estimates. If there is no information about ICC sizes, large samples with an emphasis on the number of clusters, such as 200/10/5, or 100/20/5, are recommended. If both ICCs are at least of medium size, required sample sizes reduce to e.g., 100/20/2 or 200/5/5. Achieving sufficient estimation quality with 50 clusters is still possible with at least $n_2$ = 10 in populations with large ICC sizes.

In conclusion, our findings reveal that correctly characterizing a three-level structure through ICC estimates requires an advantageous sampling-strategy, where the number of achievable clusters determines the required numbers of subclusters and Level-1 units. Particular attention must be paid to the $ICC_3$, which will most likely be slightly underestimated, even with moderate sample sizes. Researchers should take advantage of

previously reported ICC sizes in their domain to identify a most likely adequate sampling strategy for a feasible overall sample size.

**Funding:** The authors have no funding to report.

**Acknowledgments:** The authors thank Zoran Kovacevic for his support on creating the figures for this work.

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

**Data Availability:** For this article, data are freely available (Kerkhoff & Nussbeck, 2022).

## Supplementary Materials

The Supplementary Materials contain the research data and codebook with estimation results for all conditions and variables analyzed in the article (for access see Index of Supplementary Materials below).

### Index of Supplementary Materials

Kerkhoff, D., & Nussbeck, F. W. (2022). *Supplementary materials to "Obtaining sound intraclass correlation and variance estimates in three-level models: The role of sampling-strategies"* [Research data and codebook]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5418

## References

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

Berkhof, J., & Snijders, T. A. (2001). Variance component testing in multilevel models. *Journal of Educational and Behavioral Statistics, 26*(2), 133–152. https://doi.org/10.3102/10769986026002133

Bryant, F. B., & Satorra, A. (2012). Principles and practice of scaled difference chi-square testing. *Structural Equation Modeling, 19*(3), 372–398. https://doi.org/10.1080/10705511.2012.687671

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Cox, K., & Kelcey, B. (2019). Optimal sample allocation in group-randomized mediation studies with a group-level mediator. *Journal of Experimental Education, 87*(4), 616–640. https://doi.org/10.1080/00220973.2018.1496060

Cunningham, T. D., & Johnson, R. E. (2016). Design effects for sample size computation in three-level designs. *Statistical Methods in Medical Research, 25*(2), 505–519. https://doi.org/10.1177/0962280212460443

Dominicus, A., Skrondal, A., Gjessing, H. K., Pedersen, N. L., & Palmgren, J. (2006). Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior Genetics, 36*(2), 331–340. https://doi.org/10.1007/s10519-005-9034-7

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

Firth, N., Saxon, D., Stiles, W. B., & Barkham, M. (2019). Therapist and clinic effects in psychotherapy: A three-level model of outcome variability. *Journal of Consulting and Clinical Psychology, 87*(4), 345–356. https://doi.org/10.1037/ccp0000388

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

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1*(4), 223–231. https://doi.org/10.1207/S15328031US0104_02

Greven, S., Crainiceanu, C. M., Küchenhoff, H., & Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics, 17*(4), 870–891. https://doi.org/10.1198/106186008X386599

Hallquist, M. N., & Wiley, J. F. (2018). MplusAutomation: An R package for facilitating large-scale latent variable analyses in Mplus. *Structural Equation Modeling, 25*(4), 621–638. https://doi.org/10.1080/10705511.2017.1402334

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

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling, 14*(3), 361–390. https://doi.org/10.1080/10705510701301602

Hox, J. J., Moerbeek, M., & van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge. https://doi.org/10.4324/9781315650982

Kerkhoff, 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

Kreft, G., & de Leeuw, E. D. (1988). The see-saw effect: A multilevel problem? *Quality & Quantity, 22*(2), 127–137. https://doi.org/10.1007/BF00223037

LaHuis, D. M., & Ferguson, M. W. (2009). The accuracy of significance tests for slope variance components in multilevel random coefficient models. *Organizational Research Methods, 12*(3), 418–435. https://doi.org/10.1177/1094428107308984

LaHuis, D. M., Jenkins, D. R., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2020). The effects of misspecifying the random part of multilevel models. *Methodology, 16*(3), 224–240. https://doi.org/10.5964/meth.2799

Lai, M. H., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *Journal of Experimental Education, 83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Leckie, G., Browne, W. J., Goldstein, H., Merlo, J., & Austin, P. C. (2020). Partitioning variation in multilevel models for count data. *Psychological Methods, 25*(6), 787–801. https://doi.org/10.1037/met0000265

McNeish, D., & Wentzel, K. R. (2017). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research, 52*(2), 200–215. https://doi.org/10.1080/00273171.2016.1262236

Molenberghs, G., & Verbeke, G. (2004). Meaningful statistical model formulations for repeated measures. *Statistica Sinica, 14*(3), 989–1020.

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

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

R Core Team. (2019). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed., Vol. 1). Sage.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Satorra, A., & Bentler, P. M. (2010). Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika, 75*(2), 243–248. https://doi.org/10.1007/s11336-009-9135-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