4. Data and Methods

In order to answer the research questions, we conducted a simulation study based on the structure of IEA LSA data. Application of Monte Carlo simulation allowed us to draw samples with specific designs from an infinite population with particular features. For each specified scenario, we created 6,000 sample replicates. All samples displayed a two-level cluster design and mimicked the structure and particulars of typical datasets originating from educational surveys, that is, students nested in schools.[23] We then analyzed all sample replicates with four different two-level hierarchical linear models. The sampling errors—and therefore the precision—of all model parameters derived from 6,000 replicates can be regarded as outcomes of this research. We also examined the dependency of the sampling errors on varying sample designs (sample sizes at the student level and the school level, consideration of sampling weights) and on population parameters (i.e., the intraclass correlation and covariance structure).

4.1 Fixed Population and Sample Parameters

The data used to produce all outcomes were obtained from Monte Carlo simulations. We describe the different sampling scenarios in Section 4.2 below. The 6,000 sample replicates per scenario were selected from an infinite population that had the following characteristics:

- A two-level hierarchical structure, with individuals (e.g., students) at Level 1 and clusters (e.g., schools[24]) at Level 2;
- A normally distributed variable with a mean of 500 and a standard deviation of 100 to represent students' academic achievement;

23 Note that in many LSA whole classrooms are selected within schools (e.g., TIMSS, PIRLS, ICCS) instead of students selected across classrooms (e.g., PISA). In the former case, Level 2 is built up by the class rather than the school, or, more specifically, clustering effects come from both the school and the class level. However, the two effect levels cannot be disentangled using a three-level hierarchical model because the number of classes sampled per school—and even the number of classes available in each school—is clearly too small to allow setting up a meaningful respective model. Consequently, we decided not to address the disentangling effects from schools versus classes in our research.

24 Because this research mimics real hierarchical models applied in educational research, we use the terms "school" and "cluster" interchangeably in the text.

- The total variance of the achievement variable fixed at 10,000. The distribution of this variable into between-cluster (Level 2, schools) and within-cluster (Level 1, students) variance was determined by the considered ICC of the achievement variable (see Section 4.2.2). For instance, if the intraclass correlation coefficient was set to 0.1, then the within-cluster variance would be 9,000 and the between-cluster variance 1,000;
- A normally distributed variable with a mean of 0 and a standard deviation of 1 to reflect student SES;
- The within-cluster variance of the SES indicator set to 0.7 and the between cluster variance set to 0.3; and
- The covariance between the SES indicator and the achievement variable set to 30, which meant that the correlation between these two variables was 0.3.[25]

The decisions that we made when determining these parameters were based on the results of preliminary simulations and/or examination of TIMSS and PISA data, as described in the following sections.

4.1.1 Number of replicates

When investigating HLM model parameter estimation, several researchers have either repeatedly selected subsamples from a base sample with known properties or used a Monte Carlo simulation. These researchers include, amongst others, Maas and Hox (2005), Muthén and Muthén (2002), and Okumura (2007); refer also to Section 2.3.

In the current study, we examined a setting with fixed sample sizes (150 clusters, 10 individuals per cluster) in order to determine the number of sample replicates needed. The sampling errors of the parameters of Models 1 and 2 (see Section 4.3) were estimated for increasing numbers of replicates (1,000 to 30,000) and varying ICCs. With roughly 6,000 replicates (or with fewer numbers of replicates), the estimates of the sampling error stabilized sufficiently for all considered model parameters.

4.1.2 Achievement variable scale and socioeconomic status indicator

Most publicly available educational LSA datasets use an achievement outcome variable scaled to have a mean of 500 and a standard deviation of 100, as is the case with TIMSS, PIRLS,[26] ICCS,[27] TEDS, and PISA. It therefore seemed appropriate to adopt the same distribution for the achievement variable in the present study.

PISA provides an SES indicator with a mean of 0 and a standard deviation of 1 (OECD, 2006; Schulz, 2006). Also, many researchers drawing on other educational datasets have used a similar scale to calculate this indicator (Caro & Lehmann, 2009; Caro, McDonald, & Willms, 2009; Willms, 2003; Willms & Shields, 1996). We therefore used the same distribution for the SES indicator.

25 The correlation between two variables is equal to the covariance between these two variables, divided by the product of their standard deviations (in this case, 30/[100*1] = 0.3).

26 Progress in International Reading Literacy Study, conducted by IEA: http://pirls.bc.edu/

27 International Civics and Citizenship Study, conducted by IEA: http://www.iea.nl/iccs_2009.html

4.1.3 Covariance between the SES indicator and the achievement variable

In order to determine a default value for the covariance between the SES indicator and the achievement variable, we examined data from TIMSS 2007 (Grade 8 population).

The SES indicator variable derived from TIMSS data was built as a composite of home possessions, mother's education, and father's education, according to the approach proposed by Caro (2010).[28] Two major methods were used to calculate this measure of SES: IRT and principal component analysis (PCA). First, a home possessions index was estimated by means of a Rasch model (Masters & Wright, 1997; Rasch, 1980). Secondly, the first principal component was used to summarize the home possessions index and mother's and father's education into the single SES index. The final SES measure was standardized to have a mean of 0 and a standard deviation of 1 for the TIMSS 2007 Grade 8 student population.

On average across the full database (which included more than 200,000 surveyed students from 53 countries), the correlation was 0.294. For 60% of the participating countries, the correlation ranged from 0.2 to 0.4. On the basis of these results, we chose a covariance of 30, which corresponds to a correlation of 0.3, as the default value of the infinite population that we used as the starting point for the Monte Carlo simulation.

4.1.4 Within- and between-schools variance of the SES indicator

We again used the TIMSS 2007 data for the Grade 8 population to examine the within- and between-school variance of the SES indicator.[29] We used the average values for the variance across the examined countries as default values for the presented research (0.7 for the within- and 0.3 for the between-school variance of the SES indicator).

4.2 Varied Population and Sample Parameters

The following subsections describe which parameters we varied in order to examine different population and sampling scenarios. Overall, we examined 288 different sampling scenarios. Table A1 in the appendix provides an overview of these scenarios.

4.2.1 Sample size of clusters and within clusters

In order to examine the effects of cluster sample sizes on sampling errors of the studied HLM models, we set the number of sampled clusters (or schools) to 50, 100, 150, and 200. These cluster sample sizes are highly relevant in educational LSA.

The minimum total school sample size is generally set to 150 per participating country in these assessments. However, certain conditions, such as the following examples, make it necessary to select larger samples.

28 The index for this composite measure is similar to the SES index developed for PISA (Schulz, 2006).

29 An arbitrary sample of 13 participating countries was examined: Algeria, Bulgaria, Colombia, the Czech Republic, Ghana, Hungary, Indonesia, Iran, Italy, Korea, the Russian Federation, Tunisia, and the United States.

- The minimum sample size for students cannot be achieved with 150 schools due to small average school sizes.
- Large variances between schools with respect to the main subjects of interest cause high sampling errors. In such cases, the required precision of the estimates can only be achieved by increasing sample sizes.

Because research interest often focuses on single explicit strata and because the sample size within an explicit stratum is usually much smaller than in the whole sample, we also studied cluster sample sizes of 50 and 100.

In order to examine the influences of different within-cluster sample sizes, we considered 5, 10, 15, 20, 25, and 30 individuals per cluster for each case. While the higher values (≥ 20) naturally correspond to usual within-school sample sizes of students, we considered it would be interesting to determine if smaller sample sizes would satisfy certain precision requirements on estimates as well. We kept the group sizes within one sampling scenario equal so as to simplify the model conditions.[30]

4.2.2 Intraclass correlation coefficients (ICCs)

Intraclass correlation coefficients for student populations tend to range from 0.1 to 0.4. These values can be derived from publicly available LSA datasets, such as those from the various cycles of TIMSS and PISA. Only on rare occasions are higher coefficients found in the data from the different participating countries. We therefore set the ICC levels to be examined to 0.1, 0.2, 0.3, and 0.4.

4.2.3 Distribution of covariance between the SES indicator and the achievement variable between the hierarchical levels

As we explained in Section 4.1.3, we set the overall covariance of the SES indicator and the achievement variable to 30. Two different distributions of the covariance over Level 1 and Level 2 were considered in this study. In the first case, the covariance was determined to be stronger at the within level (covariance = 20 within and 10 between clusters). In the second case, the covariance was determined to be stronger at the between level (covariance = 10 within and 20 between clusters).

The latter case is evident with higher ICCs and is often typical for students in highly tracked education systems: the influence of SES on achievement is stronger across schools because students within schools are more similar. This second case was examined in connection with ICC levels 0.2, 0.3, and 0.4.[31] In the first case, the ICC was low: the clusters were more similar to one another but the connection between SES and achievement appeared stronger within the cluster. We examined this case in connection with ICC levels 0.1, 0.2, and 0.3.

30 Corrections for clustering based on the design effect assume equal group sizes (Kish, 1965); multilevel analysis does not. However, Maas and Hox (2005) found no discernible effect of unbalance on multilevel estimates or their standard errors even in extreme unbalanced designs. This outcome is also supported by work conducted by Grilli and Pratesi (2004).

31 If an ICC of 0.1 is considered, the maximum value for the covariance between clusters is 10. Therefore, this ICC could not be considered in the second case.

4.2.4 Weights

All sampling scenarios were first analyzed (see Section 4.3) using unweighted data. As we pointed out in the literature review, LSA data are usually collected from surveys with complex sampling designs. This means that individuals and clusters may have different selection probabilities. Data collected by means other than a simple random sample should therefore be analyzed with caution. If the complexity of the sample designs is overlooked, the estimates can be severely biased. Rutkowski, Gonzalez, Joncas, and von Davier (2010) outline the correct use of sampling weights in hierarchical modeling of data drawn from LSA.

In order to achieve self-weighted samples in LSA,[32] the primary sampling units (here, schools/clusters) are generally selected with probabilities proportional to their sizes (see, in this regard, Joncas, 2008). This selection method results in school design weights that follow the character of a Poisson distribution.[33] Figure 4.1 illustrates this fact with regard to the TIMSS 2007 Grade 8 population. Because the base weights of Level 1 units (here, students) in many LSA are all identical within a cluster, we disregarded them in this research. Consequently, we created Level 2 design weights as random variables that followed a Poisson distribution:

$$f(k,\lambda) = \frac{\lambda^k e^{-\lambda}}{k!}.$$

Here, $\lambda = 2$ and k is a positive integer, attached to all 6,000 datasets in each of the 288 different sampling scenarios. We analyzed all sampling scenarios a second time, using weighted data.

4.3 Hierarchical Models

We analyzed, for each sample scenario, four different hierarchical models, each of which we describe below. In order to bring meaning to the abstract equations, we provide an exemplary research question for each model.

- *Model 1—the empty (or null) model:* This model does not contain an explanatory variable and the intercept is random.

$$\begin{cases} y = \beta_0 + \varepsilon \\ \beta_0 = \gamma_{00} + U_0 \end{cases} \quad (1)$$

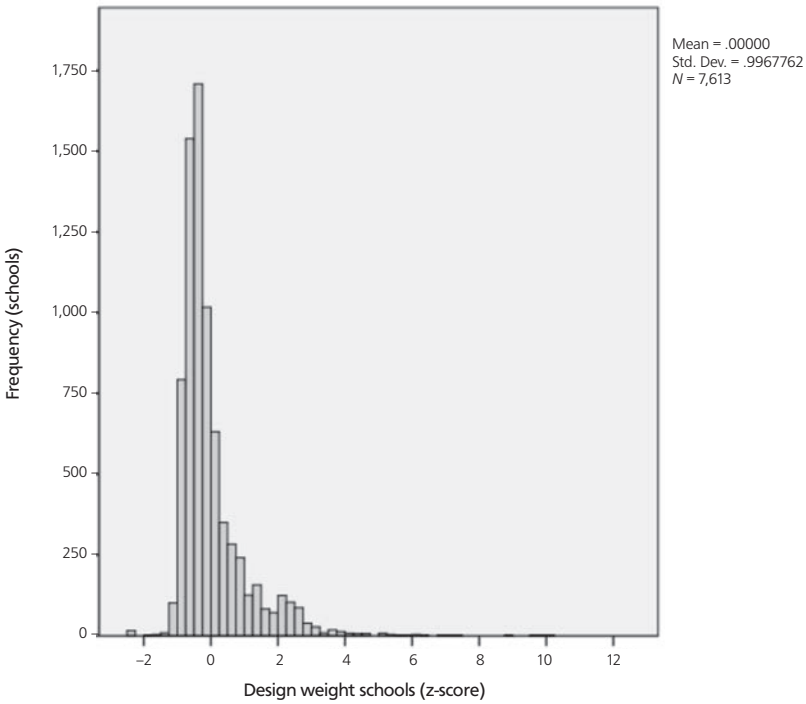
Example research question: To what extent are students within schools more alike than students between schools in terms of their academic achievement?

The question could be answered by measuring, as an outcome of this model, the intraclass correlation coefficient of a given country.

32 Study targets have similar estimation weights.

33 Weights are inversely proportional to the selection probability. The sampling method applied in most LSA (sampling with selection probabilities proportional to size) leads to similar distributions of selection probabilities (and consequently weights), as evident with Poisson sampling (see, for example, Sarndal, Swenson, & Wretman, 1992).

Figure 4.1: Distribution of design weights of schools (over all participating countries, after z-transformation at country level)



Source: TIMSS 2007, Grade 8.

- *Model 2:* This model has one explanatory variable at Level 1. The intercept is random and the slope is fixed.

$$\begin{cases} y = \beta_0 + \beta_1 x_{ij} + \varepsilon \\ \beta_0 = \gamma_{00} + U_0 \\ \beta_1 = \gamma_{10} \end{cases} \quad (2)$$

Example research question: What is the association between family SES and academic achievement at the individual level, after controlling for school-level effects?

This association can be measured by β_1 and its significance.

- *Model 3:* Here, there is one explanatory variable at Level 1 and one explanatory variable at Level 2. The intercept is random and the slope is fixed.

$$\begin{cases} y = \beta_0 + \beta_1 x_{ij} + \varepsilon \\ \beta_0 = \gamma_{00} + \gamma_{01} x_j + U_0 \\ \beta_1 = \gamma_{10} \end{cases} \quad (3)$$

Example research question: Is there evidence for contextual SES influences?

Or, what is the difference in academic achievement between two students with comparable SES levels but who attend schools that differ in terms of school SES?

This can be captured by γ_{01} when x_{ij} is the individual SES variable and x_j is the averaged school SES. γ_{01} captures contextual effects if SES at Level 1 is grand-mean centered. If the mean is group centered, then contextual effects are approximated by $\gamma_{01} - \gamma_{10}$.

- *Model 4:* This model has one explanatory variable at Level 1 and one explanatory variable at Level 2. The intercept and the slope are random.

$$\begin{cases} y = \beta_0 + \beta_1 x_{ij} + \varepsilon \\ \beta_0 = \gamma_{00} + \gamma_{01} x_j + U_0 \\ \beta_1 = \gamma_{10} + U_1 \end{cases} \quad (4)$$

Example research question: Does the influence of SES on achievement vary between schools, that is, does SES affect students' achievement in different schools to different magnitudes or even directions?

This construct is measured by U_1 and its significance.

For each model, the variables are defined as:

y Achievement variable
x_{ij} SES indicator at Level 1
x_j SES indicator at Level 2
ε Residual variance
β_0 Random intercept
γ_{00} Mean of random intercepts
U_0 Variance of random intercepts
γ_{01} Slope of random intercepts
β_1 Fixed or random slope (SES indicator)
γ_{10} Mean of random slopes (SES indicator)
U_1 Variance of random slopes (SES indicator)

$$\text{where } \begin{cases} \varepsilon \sim N(0, \sigma^2) \\ \begin{bmatrix} U_0 \\ U_1 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{01} & \tau_{11} \end{bmatrix}\right) \end{cases} .$$

The hierarchical models that we chose are ones that are commonly used in educational research.

To summarize, we applied four different hierarchical models to analyze 288 different sampling scenarios, each with 6,000 replicates.

4.4 Outcomes

4.4.1 Coefficients of variation

The sampling error for each of the model parameters in the four different models was the main outcome of this research. The Monte Carlo simulation offered two possible ways of retrieving these sampling errors:[34]

1. The sampling error could be estimated as the standard deviation of the sample distribution over the 6,000 sample replicates per sample scenario. This method provides an unbiased and (given the sample size of 6,000) very reliable and precise estimate of the true sampling error. Sampling errors obtained by this method are further referred to as SE (Method I).
2. The sampling error could be estimated by using a sandwich estimator (the standard Huber-White procedure;[35] Muthén, 2008) for each sample replicate. With this method, the average of the 6,000 sampling error estimates should also provide a good estimate of the real sampling error for a given model parameter. Sampling errors obtained by this method are further referred to as SE (Method II).

As the results of our research show, the two methods gave almost identical sampling error estimates for most model parameters and under most different sampling scenarios. An example of this similarity is shown in Figure 4.2 where both lines flow almost congruently. However, the sampling error of the mean, the variance, and the slope of the random intercepts are systematically underestimated by Method II if the number of sampled clusters is small (i.e., < 100, see Figure 4.3 for an example). Note that other authors (Maas & Hox, 2005; Van der Leeden, Busing, & Meijer, 1997) report similar observations with regard to macro-level variance estimates. In addition, and to an even greater extent, the sampling error of the variance of the random slope (evaluated in Model 4) is strongly overestimated by Method II (see Figure 4.4).[36] We consequently decided to use throughout our research only those SEs estimated by Method I.[37]

34 See also Muthén and Muthén (2002).

35 The Huber-White sandwich estimator is calculated using a Taylor series expansion.

36 We refer interested readers to Maas and Hox (2005), who evaluated bias in the estimation of standard errors in hierarchical models under certain conditions. Also, as Muthén and Muthén (2002) point out, sampling error in hierarchical modeling might be over- or underestimated depending on the situation.

37 Although this matter is not the focus of this research, users of Mplus should be aware of the possible over/underestimation of sampling errors of specific model parameters when exploring similar hierarchical models.

Figure 4.2: Exemplary comparison of two methods of SE estimation: Model 1, SE of residual variance, mean over four ICC levels

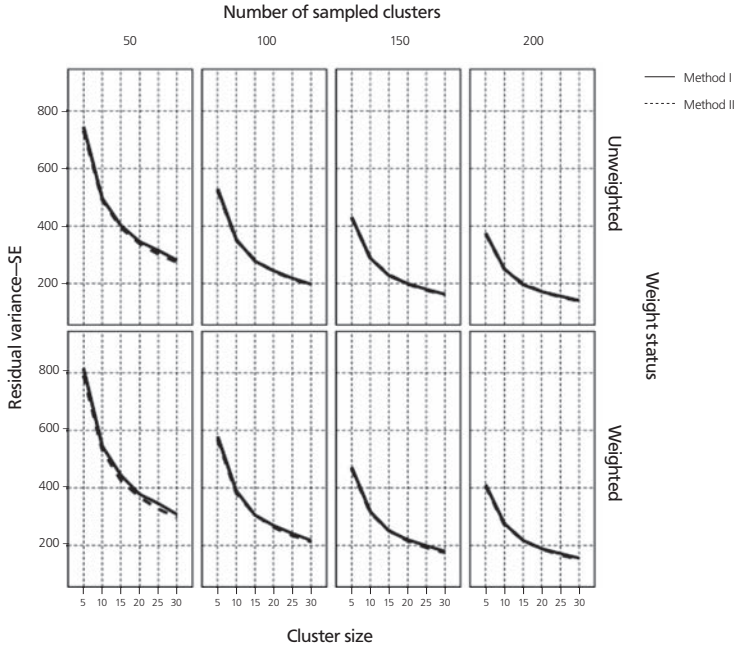


Figure 4.3: Exemplary comparison of two methods of SE estimation: Model 3, SE of slope of random intercepts, mean over two cases of covariance distribution and four ICC levels

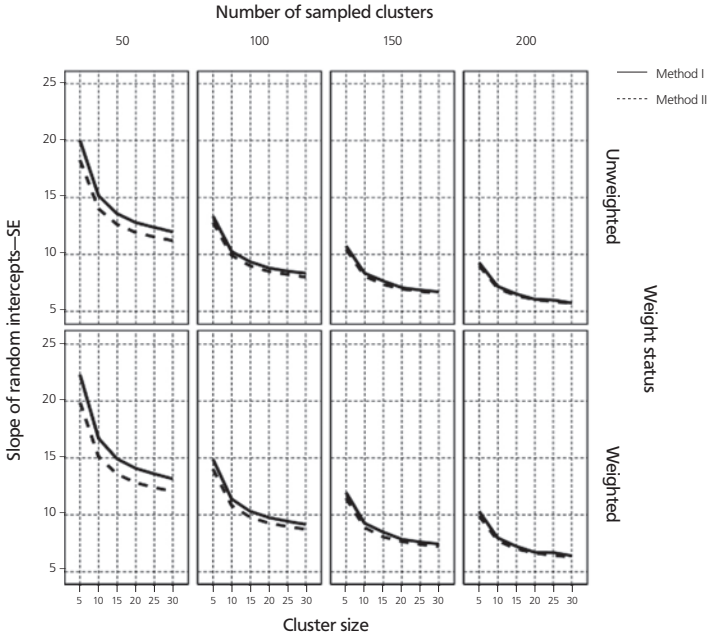
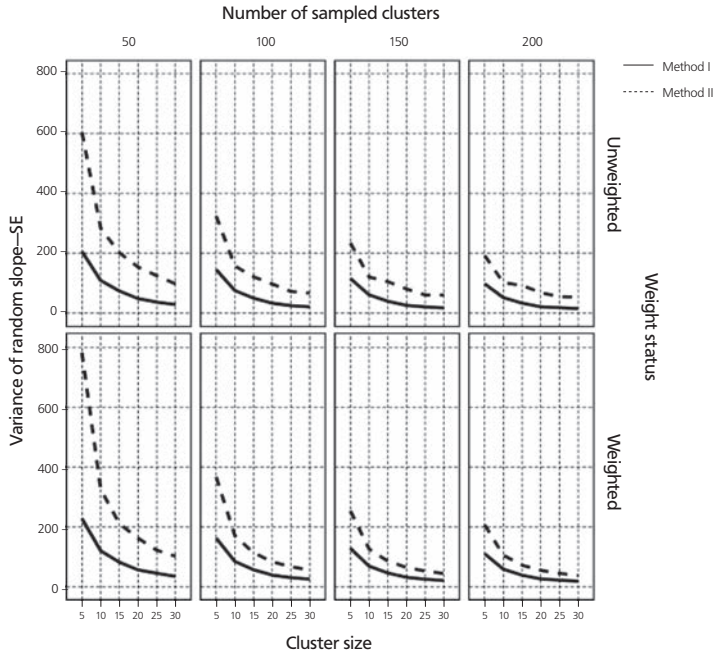


Figure 4.4: Comparison of two methods of SE estimation: Model 4, SE of variance of random slope, mean over two cases of covariance distribution and four ICC levels



The actual value of the sampling error is meaningful only in connection to the value of the parameter for which it is calculated. For example, a sampling error of 5 has no meaning unless it is considered as the sampling error for a particular mean, say, 500. Also, we were interested not so much in the mere magnitude of the sampling errors as in the "behavior" of these errors under changing sampling conditions. Therefore, we decided to present coefficients of variation, calculated as

$$CV\,(\%) = \frac{SE\,(Parameter)\ x\ 100}{Parameter} \qquad (5)$$

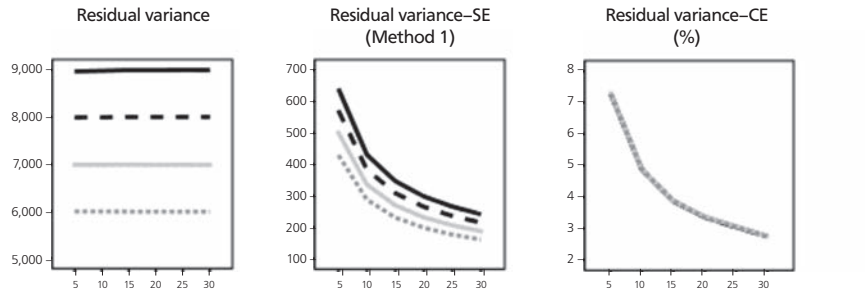
in order to display the sampling error as a percentage of the parameter it was estimated for. We used SE estimated by Method I to calculate this ratio. The following example illustrates this concept (refer also to Figure 4.5).

Consider the residual variance ε in Model 1. Not surprisingly, this parameter differs if the population is modeled with different ICCs. As can be seen in the first graph of Figure 4.5, the value of the residual variance differs by 1,000 across the four different ICC levels. The value does not differ, however, according to the number of sampled clusters or their size. If the effect of the ICCs on the sampling error is considered by merely examining the sheer value of the sampling error, one may conclude wrongly that the behavior of the SE of the residual variance depends on the ICC level (shown in the second graph of Figure 4.5). But this is not the case if we consider instead the proportion between the parameter itself and its SE. As is evident in the third graph in

Figure 4.5, the lines for the different ICC levels flow congruently, which means that the ICC level has no influence on the SE of the residual variance.

The ratio can also be used to determine if a specific parameter is significant: dividing the respective coefficient by its sampling error gives the respective *t* statistic. For example, if the coefficient of variation is 40% and the coefficient has a value of 50, its sampling error will be 20 and the *t*-value will equal 2.5, which would be considered significantly different from zero.[38]

Figure 4.5: Residual variance, its SE and the CV (%) (y-axes) by cluster size (x-axes): Model 1, average over all sampling scenarios



Note: The different lines display different ICC levels.

4.4.2 Curve estimation and equations

A glance at the graphs displayed above suggests that the curves describing the coherence between coefficients of variation and the different sample scenario settings seem to follow a curvilinear course. In fact, fitting quadratic functions to the curves arose as the best method of describing any of the outcome curves mathematically.[39] For most of these quadratic regression models, the *R* squares are above 0.95, which means that the equations fit the curves extremely well. Therefore, for each setting, we fitted a quadratic function and made cluster sample sizes and number of clusters the independent variables, thereby producing this format:

$$y = b_0 + b_1 z + b_2 z^2. \qquad (6)$$

Here, y is the coefficient of variation of the explored model parameter, b are the estimated curve parameters, and z is the cluster sample size or the number of clusters.

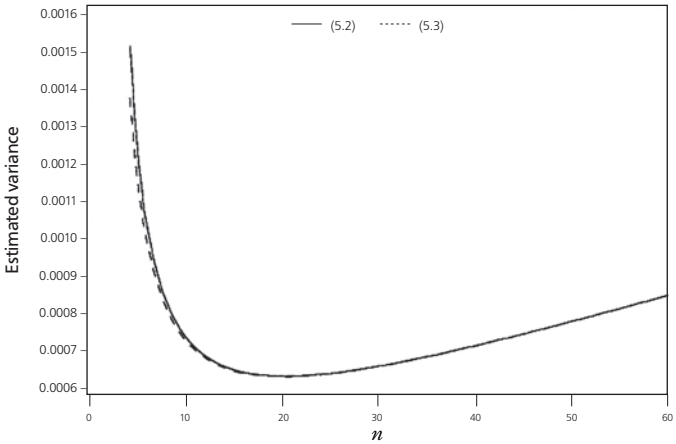
We could argue that the curves might be better described as exponential functions because a quadratic function (with a negative slope) would have a fixed minimum and would then increase (which is counter-intuitive). However, exponential equations did not fit the curves as well as quadratic ones—R^2 was smaller for exponential equations. We do acknowledge, though, that the equations should not be used outside the ranges of the explored sample sizes and population conditions.

38 Given sufficient degrees of freedom and a significance level of $p < 0.05$.

39 Afshartous (1995) explored the dependency of sampling errors on different sampling settings for hierarchical models with fixed effects. He mentioned that there is a "somewhat" quadratic connection between the Level 2 sample size and the sampling errors.

The claim of describing the coherence between coefficients of variation and the different sample scenario settings by quadratic functions receives support from the findings of other researchers. Cohen (1998) and Longford (1993) described the maximum likelihood estimates of variance components (parameters ε, U_0, and U_1) of hierarchical models as having asymptotic sampling variances.[40] Cohen (1998), for example, displayed estimated sampling variances of school-level variance components depending on the within-cluster sample size (see Figure 4.6). As the sample within clusters exceeds 20, the curve does indeed increase. So, what we see in our explored conditions might be only the first part of such a function, which does follow a quadratic course.

Figure 4.6: Estimated variance of the school-level variance component by within-cluster sample size



Note:
The school-level variance component in our models is parameter U_0. See Cohen (1998, p. 272).

The equations resulting from the curve estimations appear in the appendix. In addition to the curve parameters, the appendix tables contain goodness-of-fit measures (R^2) and *p*-values. (In Section 5.3, we explain in detail how to use these equations in order to retrieve required sample sizes for practical use.) The appendix tables are also accompanied by figures (Appendix Figures 1 to 41) that give diagrammatic form to the equations.[41]

Researchers interested in including the sample sizes for both levels in the equations can do so fairly easily by replacing the terms b_0, b_1, and b_2 in Equation (6) by other quadratic terms derived from the displayed equations. We decided not to conduct this step in order to keep the results simple and "user friendly."

40 The sampling error is the square root of the sampling variance.

41 Readers should assume that reference to the appendix tables includes references to the figures accompanying the tables.

Application of the relevant equation makes it possible to estimate the expected coefficient of variation of the respective model parameter under specific sampling conditions. Or, in turn, the minimum sample size can be derived by solving the equation for z if certain precision levels are required. The graphs can be utilized similarly, providing the requested information in a more handy way, but offering less precision.

4.5 Software Used

Although many statistical packages exist that provide tools for the appropriate analysis of multilevel data (e.g., HLM, SAS, MLwiN), we chose the statistical software package Mplus (Muthén & Muthén, 2008) to create all replicated datasets and to conduct the HLM analyses. Our choice was based on two main reasons. First, Mplus can be used as a powerful tool for Monte Carlo simulation. This is because the software makes it possible to automate the selection of subsamples from a predetermined artificial population with specific features. Secondly, the HLM tool of the software enables users to apply sampling weights and to use PML (pseudo maximum likelihood)[42] as a contemporary method of parameter estimation that is approximately unbiased.[43] In short, most of the steps described above can be performed within a single software package.

We used SAS 9.1 to create weights as random variables following a Poisson distribution. We used the graphic tool PASW 1.0 to develop the presented figures.

42 If Level 1 units are selected with unequal selection probabilities at the second sampling stage, an extended method—MPML (multilevel pseudo maximum likelihood)—is applied in Mplus (Mutthén & Muthén, 2008).

43 The currently available estimation methods are called "approximately unbiased" because various simulation studies indicate that parameter estimates can be biased, especially if cluster sample sizes are small (Graubard & Korn, 1996; Korn & Graubard, 2003; Pfeffermann et al., 1998, 2006; Stapleton, 2002). The results of our research, during which we used PML as the estimation method, support these findings.