

5. Results and Discussion

5.1 Coefficients of Variation of the Different Models' Parameters

In this chapter, we display and discuss the dependency of the sampling errors of the different parameters on varying sampling and population conditions. This dependency sometimes differs according to varying conditions and from one explored model to the next. Obviously, not every parameter is part of all four models. Table 5.1 provides an overview of the explored parameters, their model allocation, and whether they were measured between or within clusters.

Note here that the displayed graphs throughout this chapter depict only a purposive sample of the results obtained from the study and that we have illustrated only the most interesting findings graphically. The structure of the graphs therefore varies according to the message each needs to convey. Further associations are displayed graphically in the appendix.

Table 5.1: Explored model parameters and their model allocation*

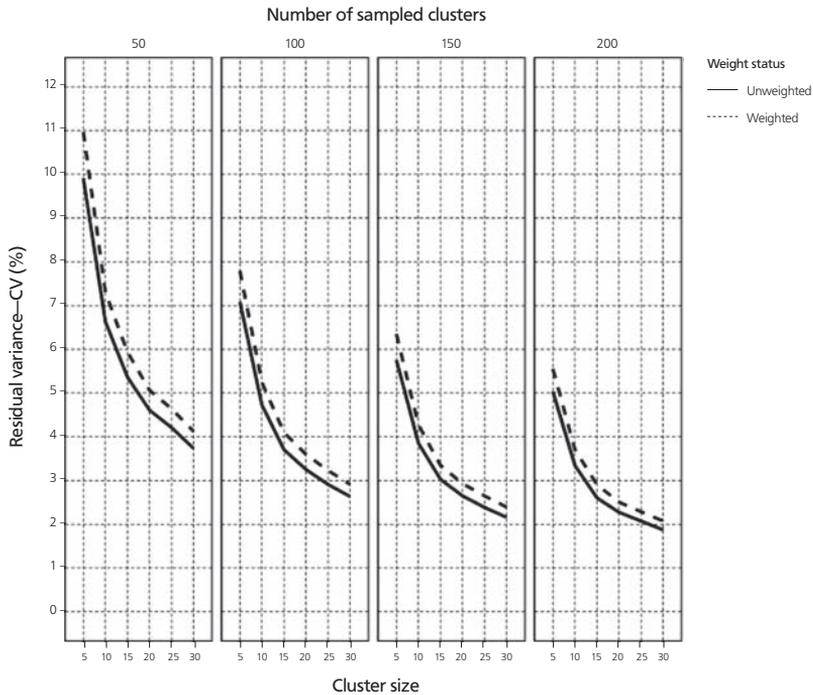
Parameter	Notation of parameter	Parameter is estimated in ...				Parameter is estimated ...	
		Model 1	Model 2	Model 3	Model 4	Between clusters	Within clusters
Residual variance	ε	x	x	x	x		x
Mean of random x intercepts	γ_{00}	x	x	x	x	x	
Variance of random intercepts	U_0	x	x	x	x	x	
Fixed slope	β_1		x	x			x
Slope of random intercepts	γ_{01}			x	x	x	
Mean of random slopes	γ_{10}				x	x	
Variance of random slopes	U_1				x	x	

Note: * Refer to Chapter 4.3 for definitions of the four models.

5.1.1 Residual variance

The residual variance in the hierarchical models represents the part of the total variance attributed to the within-group level. As already mentioned and graphically demonstrated in Section 4.4.1, the residual variance itself varies with the ICC but its coefficient of variation does not. The association between the coefficient of variation of the residual variance and the sample sizes on both levels, and whether the data were weighted or not, turned out to be completely independent of the type of model explored and of the covariance case being considered. The relationship is shown in Figure 5.1.

Figure 5.1: CV (%) of the residual variance by weight status and sample size at both levels



Clearly, the coefficient of variation increased exponentially as sample sizes on both levels decreased. Because the parameter is measured at the within-cluster level, only the increase of the total sample size matters; the level on which it is increased is less relevant. For example, selection of 100 clusters of size 20 results in the same error margins as selection of 200 clusters of size 10, a fact that could have relevance for cost discussions.

Overall, the sampling error assumed relatively low proportions compared with the sampling error of other model parameters, particularly other parameters measuring variances (explored in the sections below). Hence, the residual variance is a model parameter that can be estimated with comparatively high precision, even when the sample sizes are small.

Weights also had a slight but obvious enlarging effect on the coefficient of variation of the parameter of interest. Comparison of the coefficients of variation of unweighted and weighted data showed the latter increasing by a factor of 1.1 on average over the different settings. Note, however, that this effect decreased slightly with increasing sample sizes, on both Levels 1 and 2.

Tables A2 and A3 in the appendix provide the quadratic regression equations fitted to the displayed curves.

5.1.2 Mean of random intercepts

Introducing a random intercept in a model acknowledges the possibility that all clusters have their own mean. The term γ_{00} is the mean over the different group means.

Figures 5.2 and 5.3 show the dependency of the coefficient of variation of the mean of random intercepts (parameter γ_{00}) on the varied sample and population parameters. In general, we can see that the precision of this parameter is very high. Across all different sampling settings, the coefficient of variation of this parameter ranges from 0.4% to 2.2%. In fact, this parameter was the one that could be measured with the highest precision in all models.

Figure 5.2: CV (%) of the mean of random intercepts by weight status, ICC, and sample size at both levels: Models 1 and 2

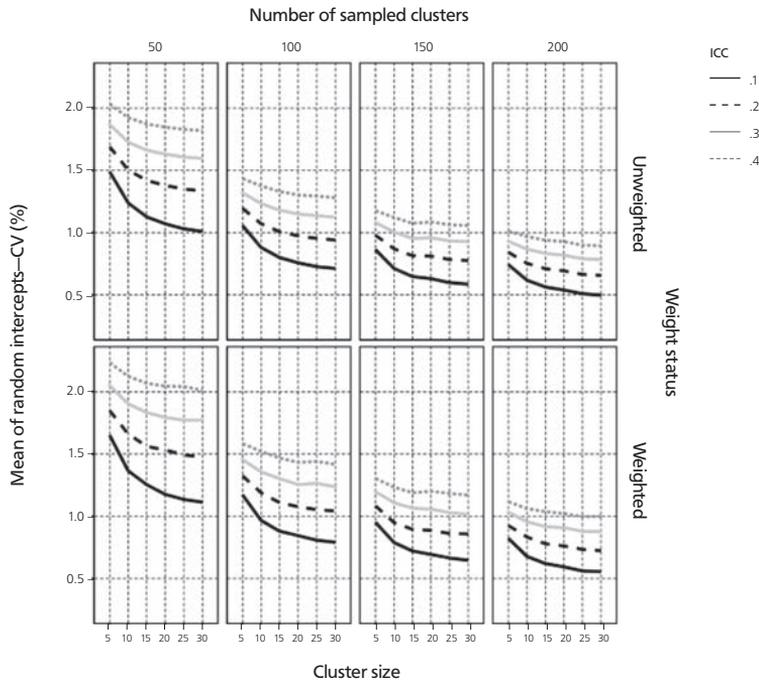
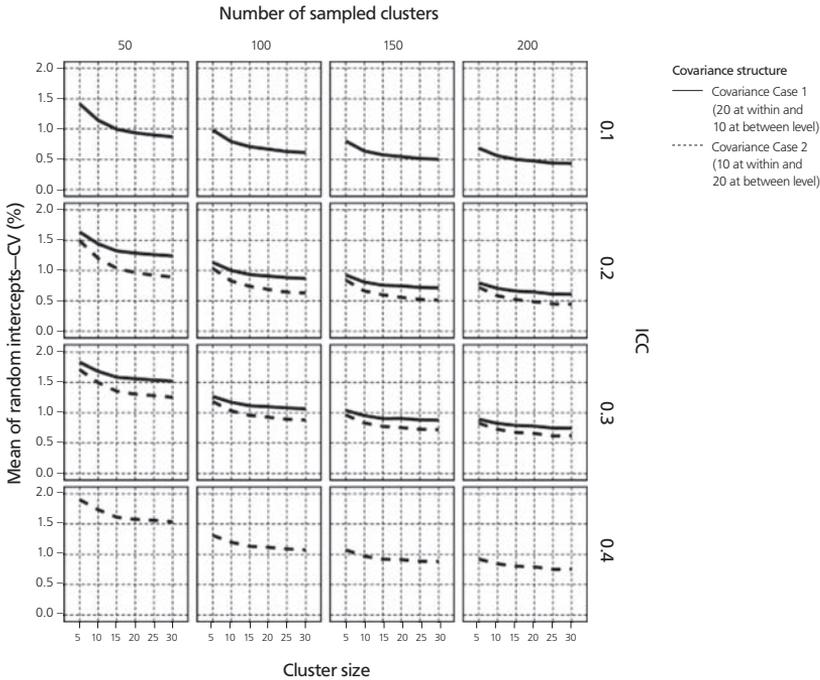


Figure 5.3: CV (%) of the mean of random intercepts by covariance distribution, ICC, and sample size at both levels: Models 3 and 4, unweighted data



As sample sizes increased, the coefficient of variation decreased, with the decrease following the format of a quadratic function within the explored ranges. This held true for the sample sizes on both levels. Also evident here is the fact that this parameter can be measured with much higher precision if—assuming constant total-sample sizes—the Level 2 sample size is favored over the Level 1 sample size. For example, if the total sample size is 500, the coefficient of variation is smaller when 100 clusters, each of size 5, are sampled than when 50 clusters, each of size 10, are sampled. This pattern aligns with findings from various authors (e.g., Cohen, 1998; Mok, 1995; Snijders, 2005).

Introducing weights and increasing ICC levels induced higher sampling errors. On average over the different settings, weights enlarged the coefficients of variation by a factor of 1.1 to 1.2, while increasing ICCs enlarged the coefficients of variation by up to 0.2%, stepping from one ICC level to the next.

The same effect of the ICC on sampling errors of population total means is well known. For example, the PISA technical report (OECD, 2006) illustrates the relationship between ICC and sampling errors of mean estimates, here dependent on total sample sizes. The more similar the individuals are within clusters (high ICC), the less precise the estimates are, assuming the sample size is stable. However, according to our findings, the effect of the ICC on the coefficient of variation of γ_{00} amplified with

increasing within-cluster sample sizes but it remained stable for increasing Level 2 sample sizes. This can be seen by comparing the graphs in Figure 5.2 above with Figure A4 in the appendix: the gaps between the lines widen as cluster sizes increase (Figure 5.2), but they barely widen as the numbers of sampled clusters increase (Figure A4, appendix).

The relationship between the coefficient of variation of the explored parameter and the sample settings was uniform for Models 1 and 2. For these two models, the considered cases of covariance distribution had no influence on this relationship.⁴⁴ Figure 5.2 depicts the explored association graphically for the first two models. The respective quadratic equations estimated to describe the curves can be found in Tables A4 and A5 of the appendix.

The Figure 5.2 graph makes it possible to easily reconstruct the general minimum sample sizes applied in many LSA, where the minimum sample size is often set to 150 schools, with one class per school. This rule is based on the precision requirement for the main outcome of these studies, which is usually a scale score with an overall mean of 500 and a standard deviation of 100. The sampling error of this score should be below 5 (the coefficient of variation would consequently be below 1%). If we assume a typical ICC of 0.3 and a medium class size of 25, and data originating from complex samples (requiring weights to be applied), the required sample size at Level 2 would indeed be approximately 150. This outcome explains why the sample size in some countries needs to be increased when, for example, the countries have larger ICCs or smaller classes.

In Models 3 and 4, the covariance distribution had an effect on the coefficient of variation of γ_{00} . For comparable ICC levels, the coefficient of variation was smaller in the case where the covariance was stronger between rather than within clusters. The behavior of the coefficient of variation of the parameter of interest was uniform for these two models. Figure 5.3 displays the respective curves.

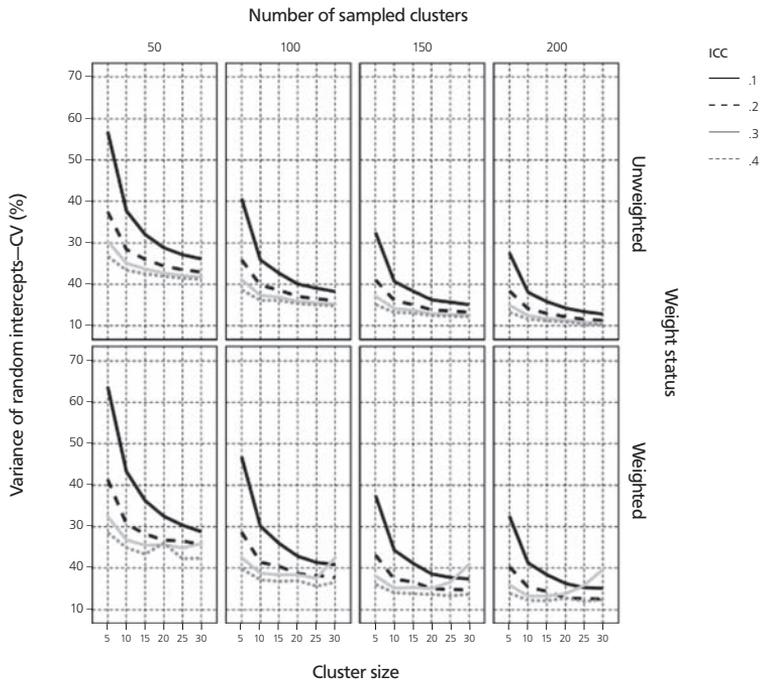
Because the effect of weights in Models 3 and 4 was similar to the effect in Models 1 and 2, we decided not to display the effect graphically. The graph represents results for unweighted data. Equations fitted to the curves (separated for weighted/unweighted data) are presented in Appendix Tables A6 and A7, which are accompanied by additional graphs. The shape of the curves looks identical to the shape presented in the preceding figure. The scale, however, has shifted. The coefficient of variation has become—for the comparable settings—slightly smaller in Models 3 and 4, meaning that the estimates of the mean of random intercepts became slightly more precise, especially when an explanatory variable was added at Level 2.

⁴⁴ For Model 1, this is due to the design of the model: no explanatory variable is included.

5.1.3 Variance of random intercepts

The variance of the random intercepts (parameter U_0) represents the proportion of the total variance attributed to the between-group level. Figure 5.4 displays the relationship between the explored sample settings and the coefficient of variation of the targeted parameter for the empty model (Model 1). The interrelations again followed quadratic courses. The respective equations can be obtained from Appendix Tables A8 and A9.

Figure 5.4: CV (%) of the variance of random intercepts by weight status, ICC, and sample size at both levels: Model 1



The coefficient of variation of the variance of random intercepts is clearly much larger here than it was for the model parameters discussed in the previous sections. Across all explored settings, the coefficient ranged from 10% up to 93%, with an average of 22%. This finding is in agreement with Afshartous's (1995) findings because it indicates the need to have significantly larger sample sizes when the main focus of interest is estimation of variance components rather than of fixed effects.

The effect of the ICC on the coefficient of variation of this parameter is inverted compared to the effect on the coefficient of variation of γ_{00} . The lower the ICC, the higher the coefficient of variation of the variance of random intercepts. This pattern means that this parameter can be measured more precisely when the ICC is higher, an outcome that is intuitively understood. Because the total variance was fixed,⁴⁵ the

⁴⁵ The total variance was fixed to 10,000 (refer to Section 4.1).

parameter itself increased with increasing ICC levels, making it “easier” to measure it precisely (refer to Section 5.2.2 for further discussion).

The use of weights again increased the coefficient of variation of this parameter by, on average, a factor of approximately 1.1.

Note that the change in the coefficient of variation of the considered parameter seems to be notably large when stepping from 5 to 10 units sampled per cluster, especially for low ICCs. In fact, the gain in precision is not so much larger when, for example, doubling sample sizes at Level 1 than when doubling sample sizes at Level 2. This finding might also be of particular interest with respect to cost considerations.

Finally, we can see from Figure 5.4 that the estimates become a little unstable for weighted data when the within-cluster sample sizes are large.

When we look at the results for the other models, it is apparent that the curves are the same shape as in Model 1 but that they have shifted on the scale: the coefficients of variation have increased slightly with the increasing complexity of the model (refer to Figures 5.5 and 5.6). This is especially true for low ICCs. Introducing an explanatory variable at Level 2 has thus made it harder to estimate, with high precision, the variance of the random intercept. The effect of weights is similar to Model 1, so we again elected not to display this effect graphically but to compare the models in illustrative ways instead.

Figure 5.5: CV (%) of the variance of random intercepts by ICC, model, and sample size at both levels: Covariance Distribution Case 1 (20 at within and 10 at between level), unweighted data

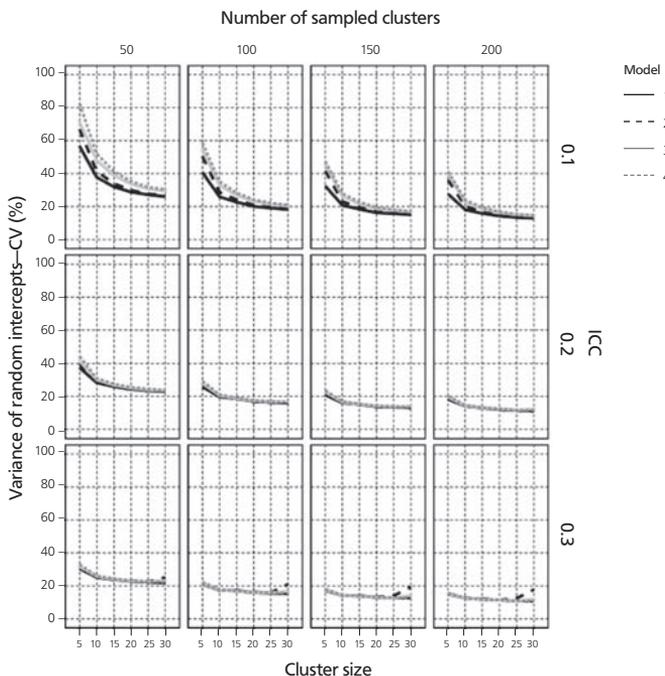
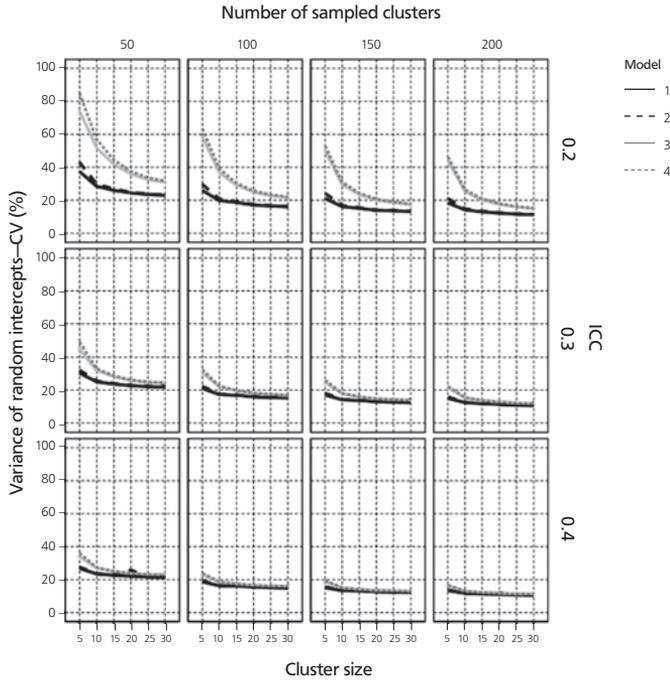


Figure 5.6: CV (%) of the variance of intercepts by ICC, model, and sample size at both levels: Covariance Distribution Case 2 (10 at within and 20 at between level), unweighted data



Figures 5.5 and 5.6 show the results of our analysis of unweighted data. The considered cases of covariance distributions have a barely noticeable effect for Model 2 but become significant in Models 3 and 4. For example, the explored coefficient of variation is approximately twice as large for ICC = 0.2 and small cluster sizes in the case where the covariance is stronger between than within groups. In general, differences in the coefficients of variation of the variance of random intercepts between the different models and between the different ICC levels become marginal for increasing sample sizes at both levels.

Tables A8 to A15 of the appendix give the quadratic equations separately for all different models and sample scenarios. Each of these tables is accompanied by figures that give diagrammatic form to the tables' contents.

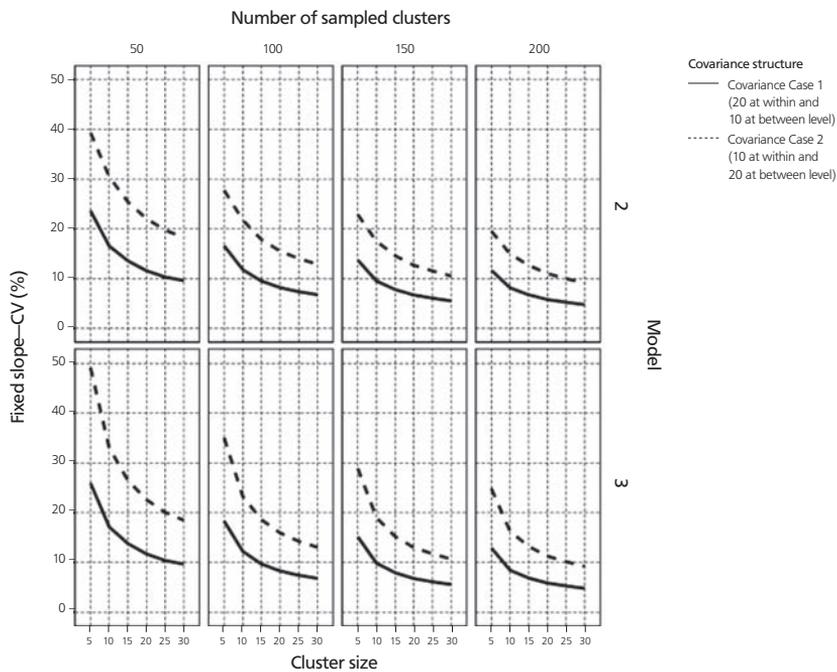
5.1.4 Fixed slope

The parameter β_1 in Models 2 and 3 indicates the association between the dependent and explanatory variables. In our example, the parameter is an estimate of the simulated relationship between SES and achievement at the individual level. In Models 2 and 3, this slope is "fixed"—the relationship is assumed to be the same in each cluster, or school.

Figure 5.7 displays how precisely this parameter can be measured under varying population and sample conditions. As we expected, the coefficient of variation again decreased with increasing sample sizes. The covariance distribution also played an important role: when the covariance between SES and the achievement variable was stronger between than within groups (covariance is 10 within and 20 between groups), the coefficient of variation of the fixed slope was approximately double in all sampling settings. Note that the influence of the ICC on the coefficient of variation of this parameter, although present, was negligible and so is not considered in the graphs.

The effect of weights on the coefficient of variation of this model parameter was no different from what it was for all other explored coefficients of variation: all increased by roughly a factor of 1.1. Tables A16 to A19 of the appendix (with accompanying figures) present the respective quadratic equations.

Figure 5.7: CV (%) of the fixed slope by model, covariance distribution, and sample size at both levels: unweighted data



Similar to the finding for the model parameter ε (residual variance), the increase in total sample size determines the gain in precision, while the level (group or within group) on which the sample sizes are increased is of minor importance. For example, doubling the total sample from 1,000 to 2,000 individuals barely matters when the group level sample size is doubled from 100 to 200 and the within-group sample size is kept at 10, or when the number of sampled clusters is kept at 100 and the within-group sample size is doubled from 10 to 20.

5.1.5 Slope of random intercepts

The parameter γ_{01} , here referred to as the “slope of random intercepts,” introduces an explanatory variable at the group level in a hierarchical model. In our example, the mean SES level in a school served as the group-level explanatory variable. In other examples, the parameter captures any contextual effects that can be measured at the group level.

Figures 5.8 and 5.9 present the association between the coefficient of variation of parameter γ_{01} and the various sample and population parameters, separated by models and the considered cases of covariance distribution.

The figures immediately make clear that it is much harder to estimate this parameter than all previously discussed parameters (under most settings) with a high degree of precision. First, we can readily see that increasing the group-level sample size (e.g., the number of schools) leads to higher precision gains than does increasing the sample size within groups (e.g., students within schools); this finding is in line with the discussed literature (Cohen, 1998; Mok, 1995; Snijders, 2005). Secondly, we can see that the covariance distribution plays a very important role. In Model 3 (Figure 5.8), the variation coefficients almost double when the covariance is stronger within groups than between groups. As a reminder, we consider two cases of the split of the covariance at the within- and the between-cluster level. In the first case, the within-level covariance was set to 20 and the between-level covariance was set to 10. The distribution over Levels 1 and 2 was reversed in the second case.

Figure 5.8: CV (%) of the slope of random intercepts by ICC, covariance distribution, and sample size at both levels: Model 3, unweighted data

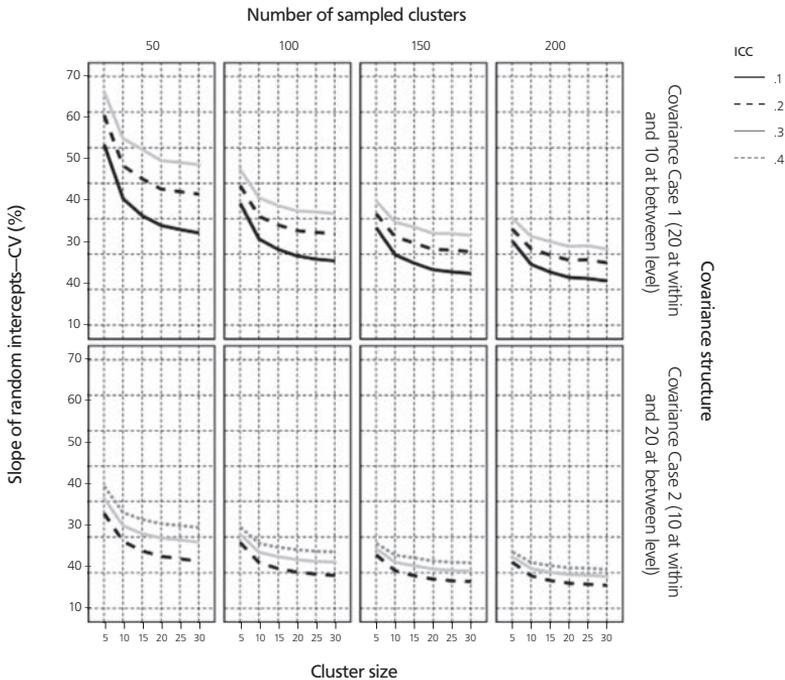
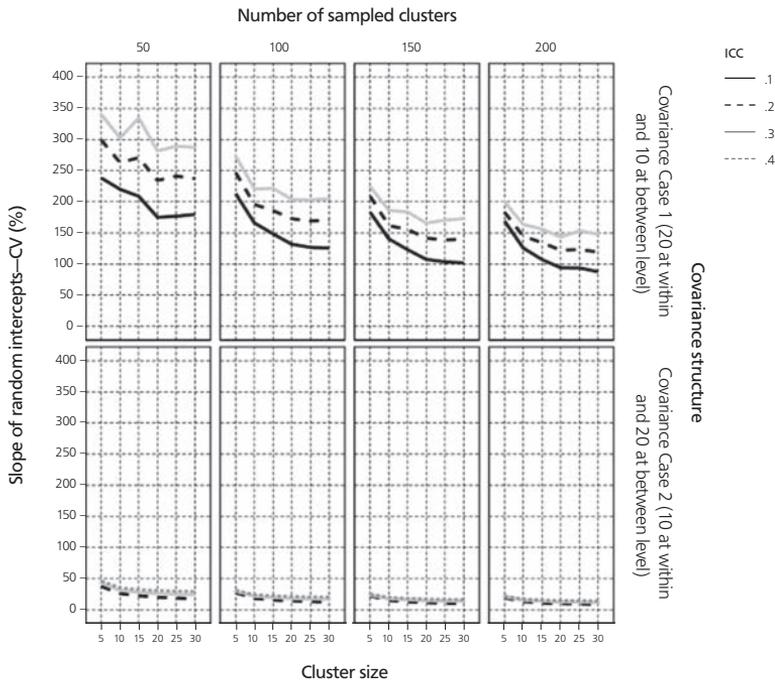


Figure 5.9: CV (%) of the slope of random intercepts by ICC, covariance distribution, and sample size at both levels: Model 4, unweighted data



In Model 4, the differences of the coefficient of variation between the two covariance cases also reached much higher levels (up to a factor of 13), as can be seen in Figure 5.9. This result suggests that when the covariance is strong at the within-group level, the slope of random intercepts can barely be measured with sufficient confidence for models with explanatory variables at both the within- and the between-group level.

The ICC also played a significant role. With increasing ICCs, the coefficient of variation of γ_{01} also increased. Similar to observations of parameter γ_{00} , the effect became larger as cluster sample sizes increased but remained relatively stable as the numbers of sampled clusters increased.

Weights influenced the coefficient of variation of the slope of random intercepts in similar vein to the other explored parameters (increasing by a factor of approximately 1.1). Tables A20 to A23 in the appendix display the quadratic equations, listed by model, covariance distribution case, weights, ICC, and number of sampled clusters. Each table is accompanied with a graph depicting the table's contents. Note that for Model 4, the slope of the coefficients of variation for Case 1 of the covariance distribution is not as smooth as it usually is for all settings with small group-level sample sizes (upper left-hand graphs of Figure 5.9). As a consequence, the fit of the quadratic curves to the slope leads to lower R squares.

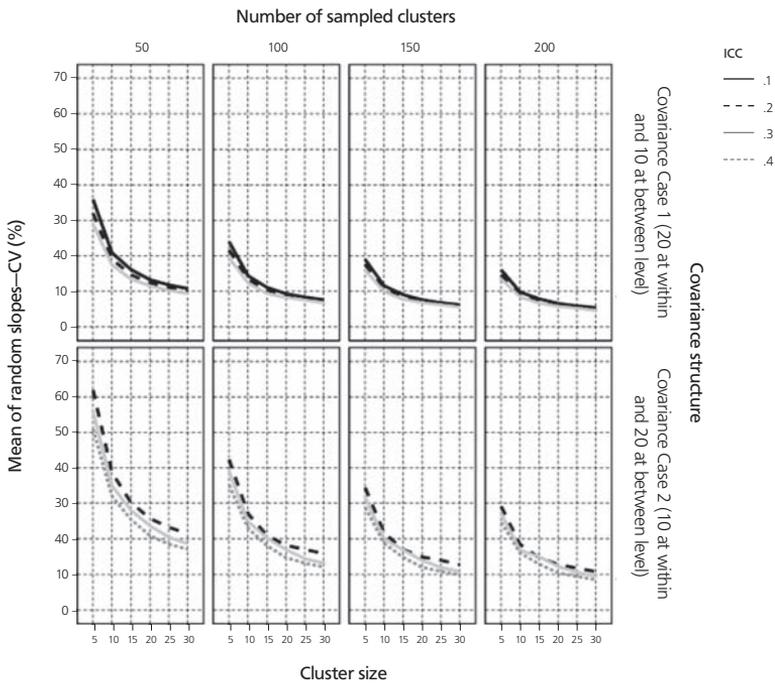
Given the extreme gradient in sampling error between the two examined cases of covariance distribution, more research involving other covariance distribution settings is required to explore the behavior of SEs.

5.1.6 Mean of random slopes

An individual-level explanatory variable can influence the outcome variable in different ways (i.e., magnitudes or even directions). This influence can be modeled by introducing a further random term into the model ($\beta_1 = \gamma_{10} + U_1$). This was the scenario considered in Model 4, where the term γ_{10} stands for the mean of the random slopes. But how precisely can we measure this term when using sample data?

Figure 5.10 provides a partial answer to this question. As the graphs suggest, the case of covariance distribution had a significant influence on the sampling error of this parameter as well. The coefficient of variation approximately doubled when the covariance between the explanatory and the outcome variable was stronger at the between-group level. With the latter case, minimum sample sizes of > 5 within clusters are indicated when the parameter itself needs to be significant (i.e., different from zero) and when few clusters (< 100) are sampled. Note also that the ICC has an influence—albeit a relatively small one—on the coefficient of variation. With increasing ICC levels, the coefficient of variation decreased slightly. When weights were used, the coefficient of variation again increased by a factor of approximately 1.1.

Figure 5.10: CV (%) of the mean of random slopes by covariance distribution, ICC, and sample size at both levels: Model 4, unweighted data



Appendix Tables A24 and A25 display the respective quadratic equations, derived separately for different covariance distribution cases, weight status, ICCs, and sample size settings.

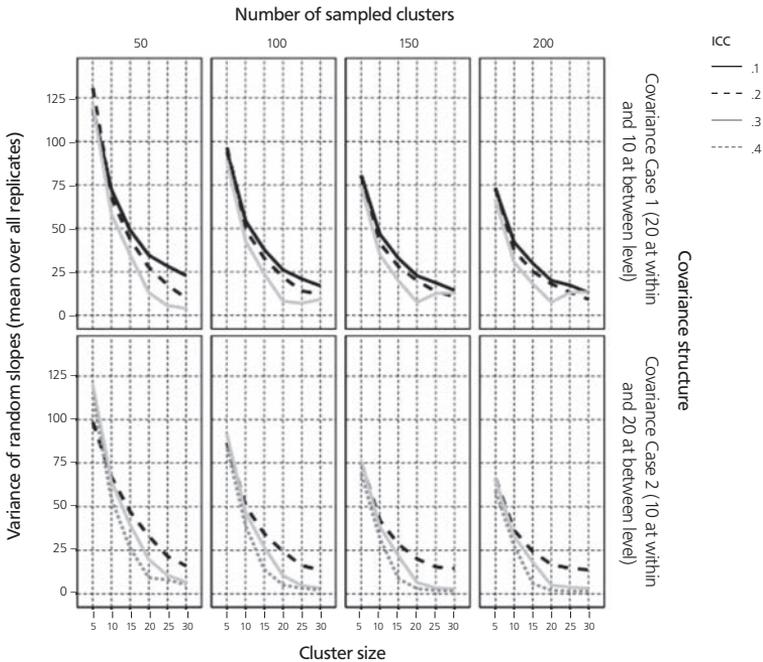
5.1.7 Variance of random slopes

The second—and often more interesting—term in a model with random slopes is U_1 , which represents the variance of the random slopes. If this term is significantly different from zero, it can be inferred that the explanatory variable is connected to the outcome variable in different magnitudes or even directions in different clusters. In our example, it could mean, for instance, that students’ SES may be positively correlated to achievement in some schools but not in others.

Unfortunately, the estimation of the parameter itself seems to be affected with significant bias, which must be caused by the estimation procedure used in Mplus. As displayed in Figure 5.11, the parameter seems to depend on sample sizes even though this cannot be the case. Therefore, the obtained sampling errors also cannot be relied on either, and for this reason we have not provided graphs or equations for the coefficients of variation of this parameter.

As part of our ongoing research, we intend to conduct an in-depth exploration of the discovered parameter estimation bias.

Figure 5.11: Variance of random slopes (means over all replicates) by covariance distribution, ICC, and sample size at both levels: Model 4



5.2 Effects of Variable Population and Sample Parameters

5.2.1 Sample sizes

As expected, the coefficients of variation of all explored parameters decreased when the sample sizes increased, regardless of whether the increase was within clusters or pertained to the number of sampled clusters (or both). The dependency between sample sizes and coefficients of variation always followed a quadratic curve progression within the explored settings. For example, increasing sample size decreased the diminishing effect on the coefficient of variation. The curves could be approximated with quadratic equations that fitted extremely well to the observed curves for most parameters (R squares mostly > 0.95). This general observation was affected neither by the intraclass correlation coefficients, the weight status, and the covariance distribution, nor by the complexity of the explored model.

The magnitude of this decrease, and whether the effect is more pronounced with sample size increases on one or the other hierarchical level, can, however, depend on all these factors and so were different for the explored model parameters. We detailed this matter in Section 5.1.

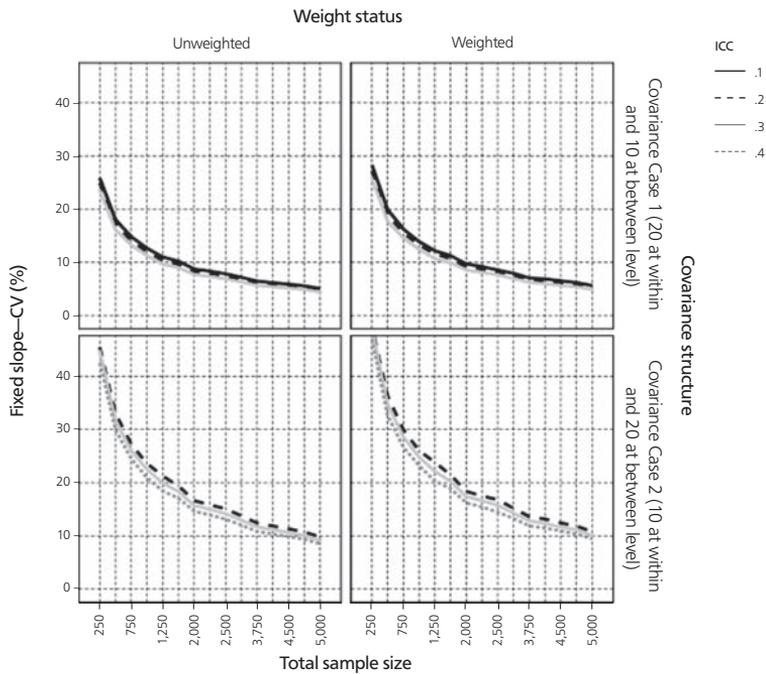
In contradiction to general rules of thumb recommended in the literature (e.g., Hox, 1995; Kreft, 1996; Mok, 1995), our findings suggest that the required sample sizes depend heavily on the parameter of interest. In particular, sample size requirements were very different if the focus of interest was the estimation of fixed model parameters or, rather, the estimation of variances. Inferences from another simulation study (Afshartous, 1995) support these findings. Also, and in agreement with the literature (Cohen, 1998; Mok, 1995; Snijders, 2005), it appears that it is more effective to increase the number of sampled clusters than the cluster sample size if the research interest concerns macro-level regression coefficients. If the focus is on variance estimates, however, the level on which the sample size is increased appears to be of less importance.

In Section 5.1, curves were displayed according to their dependency on the cluster size. It is possible, of course, to look at the interrelations between parameter precision and sample sizes from different perspectives. Figures 5.12 and 5.13 illustrate this concept. They show, for one of the explored parameters, how the curves would look if the total sample sizes or the number of sampled clusters respectively were set as the explanatory variable. Interested readers can utilize the appendix equations to produce such graphs for other parameters themselves.

Although the following matters are not part of the main scope of this paper, we encourage readers to keep two related issues in mind when determining required sample sizes:

1. Currently, available estimation methods can produce biased parameter estimates if the sample size (at either level) is small. The literature provides a variety of articles on this topic (refer to, for example, Asparouhov et al., 2006; Bell et al., 2010; Graubard & Korn, 1996; Korn & Graubard, 2003; Kovacevic & Rai, 2003; Rabe-Hesketh & Skrondal, 2006). Simulation studies, however, indicate that as

Figure 5.12: CV (%) of the fixed slope by covariance distribution, ICC, and total sample size, averaged over Models 2 and 3 and weight status

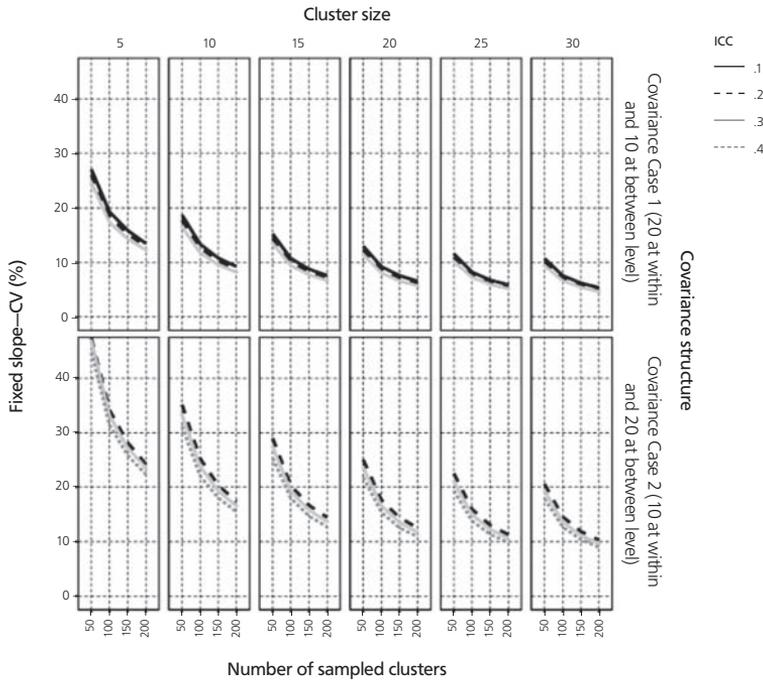


the number of clusters and the cluster sample size increase, the parameter bias is generally eliminated. Estimation methods that possess this property are referred to as approximately unbiased. All currently available software packages that accommodate hierarchical modeling apply such estimation methods. The degree of parameter bias depends on the model itself, the model parameter of interest, and the sample conditions. As a result of data collected as a side product of this study, we do not recommend dropping below a within-cluster sample size of 10 because doing so risks biasing the parameter estimation.

2. The estimation of the sampling errors can be biased. For the explored parameters and conditions, our results suggest that the bias can be substantial⁴⁶ if the number of sampled clusters is below 100. Further details on this topic can be found in Section 4.4.1.

⁴⁶ Depending on the parameter of interest, the SE was overestimated by up to a factor of 4.

Figure 5.13: CV (%) of the fixed slope by covariance distribution, ICC, and sample size at both levels, averaged over Models 2 and 3 and weight status



5.2.2 Intraclass correlation coefficients

The intraclass correlation coefficient (ICC) had no influence on the coefficient of variation of the residual variance (parameter ϵ). The coefficients of variation of the parameters γ_{00} (mean of random intercepts) and γ_{01} (slope of random intercepts) grew with increasing ICC levels. Preliminary findings suggest that the same association was in effect for the coefficient of variation of the variance of random slopes (parameter U_1). For the former parameters, the effect of ICC diminished as sample sizes within clusters increased.

When we explored the impact of the ICC on the coefficients of variation of the remaining parameters— U_0 , variance of random intercepts; γ_{10} , mean of random slopes, and β_1 , fixed slope—we found the inverse effect: as the ICC levels increased, the coefficients of variation decreased. The effect for the latter parameter was, however, minimal.

We ask readers, when considering these results, to keep the design of this study in mind. In general, variance and regression coefficients are easier to measure precisely if they are bigger, while within- and across-cluster means are harder to measure precisely when the variance is higher. This pattern can easily be seen when looking at extreme cases. If the variance is 0 at any level, then all the individuals or all the clusters are alike and any sample will estimate the true mean correctly. If, however, the variance is very

big, the estimates of the mean can be very different from one sample to another and, even though unbiased, can be very imprecise (high sampling error).

This pattern holds true for both levels—individuals and clusters. The regression coefficients link to the variances and covariances as follows:⁴⁷

$$\text{regression coefficient} = \frac{\text{cov}(\text{achievement}, \text{ses})}{\text{variance}(\text{achievement})}$$

The pattern is also valid at both levels: smaller regression coefficients are harder to measure with high precision. Because the overall variance in our simulation study was fixed, a higher ICC caused greater variance between clusters, leaving U_0 measured more precisely and the mean γ_{00} less precisely. Also, because the covariance is fixed for each case of the covariance distribution, larger ICCs and higher variance between clusters induce smaller regression coefficients for the random slopes γ_{01} along with higher coefficients of variation of γ_{01} . However, larger ICCs also mean lower variance within clusters, which leads to higher regression coefficients β_1 and lower coefficients of variation.

Our efforts to compare these findings with earlier research proved fairly unproductive. Although a few authors have explored the effect of ICC on bias in HLM parameter estimation (Asparouhov et al., 2006; Kovacevic & Rai, 2003) and bias in the estimation of sampling errors (Maas & Hox, 2005), we could not find explicit statements about the influence of ICC on the sampling errors of different parameters in hierarchical models.

5.2.3 Covariance distributions

During this research, we explored only two cases of covariance distribution. As such, we could not make general inferences about the gradual effects of varying covariance distributions. Comparisons could only be made between the two considered cases.

The covariance distribution had no effect on the coefficients of variation of any of the explored parameters in Models 1 and 2, except for the fixed slope. For Model 1, this was caused by the design of the model (no explanatory variable was introduced). Also, no effect was observed on parameter ε in any model.

The first considered case of covariance distribution (20 at within-group and 10 at between-group level) was connected to higher coefficients of variation compared to the second considered case (10 at within-group and 20 at between-group level) for the parameters γ_{00} and γ_{01} . For the latter parameter, the differences in the coefficients of variation were extreme, particularly when explanatory variables on both levels were introduced to the model (Model 4; refer to section 5.1.5).

⁴⁷ This formula refers to simple ordinary least squares (OLS) estimation. Note that different procedures (maximum likelihood estimation) are used to estimate coefficients in HLM.

On examining the effect on the coefficients of variation of the parameters U_0 , β_1 , and γ_{10} , we found that the ratios were higher when the covariance distribution was stronger between groups (Covariance Case 2). We emphasize, though, that the effect of the covariance distribution on the coefficients of variation of the parameters β_1 , and particularly γ_{01} , was even more pronounced than the effect of varying sample sizes.

In the simulation studies conducted so far (refer to Section 2.3), the covariance distribution was always fixed. For this reason, the results cannot be compared to previous related research.

5.2.4 Weights

The weights applied in this research increased the coefficients of variation of all explored parameters consistently by a factor of approximately 1.1. But concluding that not using weights is preferable because this practice increases the sampling error would be a serious mistake: using sampling weights is the only way to prevent bias when estimating parameters from data collected with a complex sample design.

The method used to simulate the weights should be kept in mind when evaluating this result (refer to Section 4.2.4). Preliminary evaluations of the findings with real data⁴⁸ showed that the aforementioned factor only held true if the actual Level 2 weights followed a Poisson distribution. This is what happens when the implemented sample design fulfills the following conditions:

1. The clusters (e.g., schools) are selected with probabilities proportional to their size;
2. The sizes of the schools in the respective country are close to a Poisson distribution; and
3. No oversampling is performed in any explicit stratum.

If these conditions are not fulfilled or, in other words, the Level 2 weights deviate from a Poisson distribution, the weights may have larger (or sometimes smaller) effects on the coefficients of variation of the different model parameters.

These conditions are reasonably standard assumptions for the presented research because they apply to many LSA. However, the sample designs for (e.g.) particular countries frequently deviate from this ideal condition for many reasons. It is well known that, in most instances, sampling weights increase the sampling error and hence the coefficients of variation. The effect could, however, depend on the distribution of these weights and maybe even their correlation with the dependent variable. Further research is needed to give a more exact understanding of the effect of weights and thereby avoid making assumptions about their distribution.

⁴⁸ Informal evaluations of the findings were conducted with data from the TIMSS 2007 Grade 8 population. Nine countries were examined.

We were unable to identify a single article in the literature focusing on the effects of weights in HLM on the sampling variance of the estimates. However, in reference to a side product of their research, Grilli and Pratesi (2004) and Pfeffermann et al. (1998) point out that using weights increases the sampling variance and provides less biased parameter estimates. When we compared the data presented in the tables in Grilli and Pratesi's article with the findings of our research, we found similar degrees of sampling variance. We also found when scouring the literature with respect to the debate on the role of sampling weights in multilevel models that Zaccarin and Donati (2008) agreed that weights have relevant effects on parameter estimates and their sampling errors. However, the two authors did not investigate the subject in more detail.

5.2.5 Model complexity

The four models considered in this research were built with increasing complexity.⁴⁹ This complexity influenced the coefficients of variation of all observed parameters except one. The influence varied, however, with the parameter of interest as well as with the considered case of covariance distribution.

Model complexity did not appear to influence the coefficient of variation of parameter ϵ (residual variance).

As was demonstrated in Section 5.1.2, the coefficients of variation of parameter γ_{00} (mean of random intercepts) behaved uniformly for Models 1 and 2, and also for Models 3 and 4. However, the coefficient of variation was smaller for the latter two models than for the simpler Models 1 and 2. We can infer, therefore, that the introduction of the Level 2 explanatory variable induces a gain in the precision of the estimation of parameter γ_{00} . This effect was more pronounced when the covariance was stronger between clusters (see Figure 5.14), a finding that aligns with observations made by Raudenbush (1997), who proposed using covariates in order to determine the optimal design of cluster randomized trials.

If we look at parameter U_0 (variance of random intercepts), we can see that the effect of the model complexity is inverted. The model's coefficient of variation has increased as the model has become more complex. The differences are marginal, however, as long as the covariance is strong at the within-cluster level, but when the covariance strengthens between groups, Models 1 and 2 show significantly smaller coefficients of variation than do Models 3 and 4 for the discussed parameter. The effect decreases with increasing ICC levels (refer to Figure 5.15).

⁴⁹ Note that even the most complex model explored in this research is still a relatively simple one.

Figure 5.14: CV (%) of the mean of random intercepts by ICC, model, and cluster size: average over weight status and numbers of sampled clusters

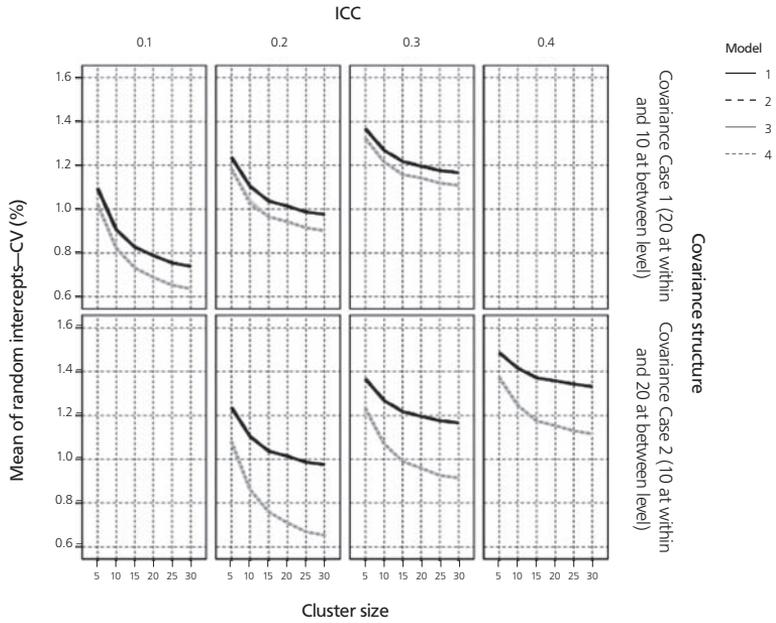
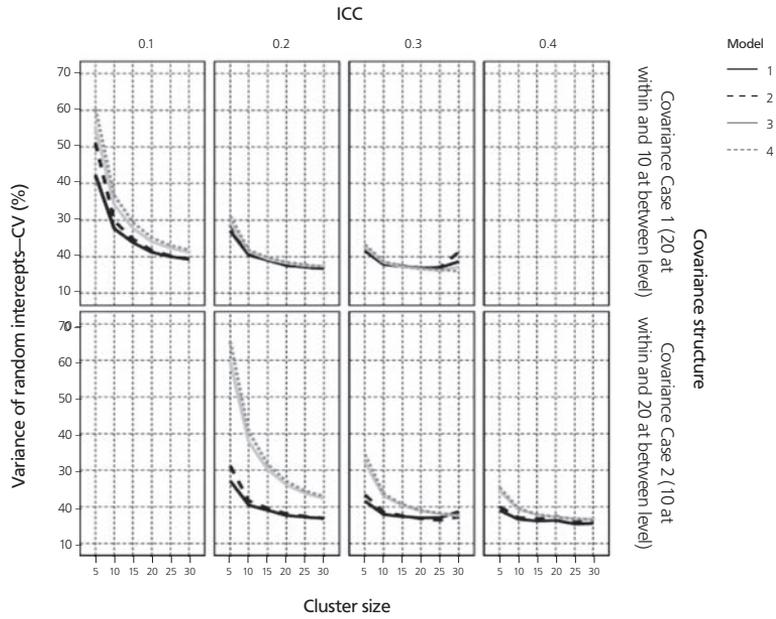


Figure 5.15: CV (%) of the variance of random intercepts by ICC, model, and cluster size: average over weight status and numbers of sampled clusters



A look at the coefficients of variation of model parameter β_1 (fixed slope, explored in Models 2 and 3) reveals very little difference between the two models when the covariance is strong at the within-cluster level, as illustrated in Figure 5.16. For the other considered covariance case, however, the parameter is estimated with less precision in Model 3, especially when the within-cluster sample size is small.

Finally, we found that the coefficient of variation of the slope of random intercepts (parameter γ_{01}) also increased with model complexity (it was higher in Model 4 than in Model 3). This finding held true for both covariance cases. However, as illustrated in Figure 5.17, the effect became much more pronounced when the covariance was stronger within clusters (the coefficient of variation increased by up to factor 6).

Figure 5.16: CV (%) of the fixed slope by ICC, model, and cluster size: average over weight status and numbers of sampled clusters

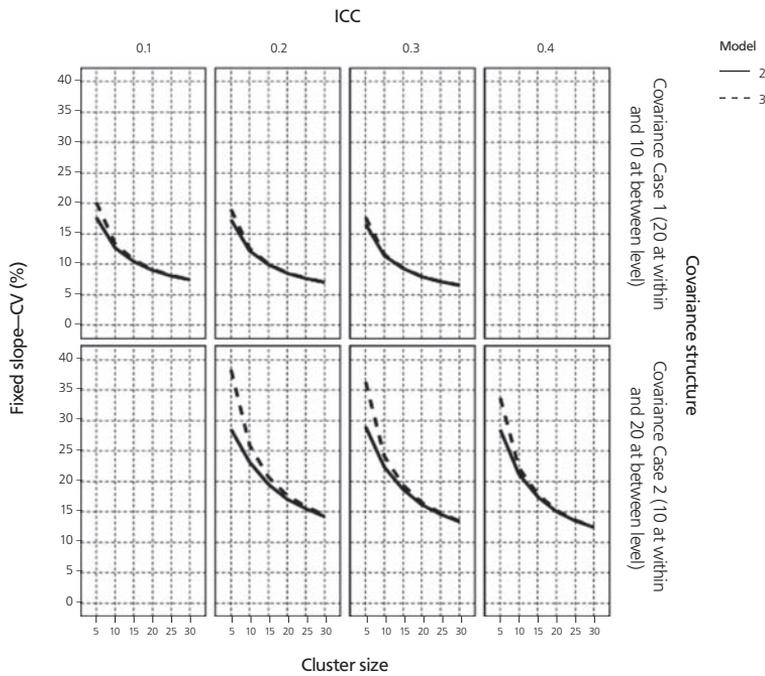
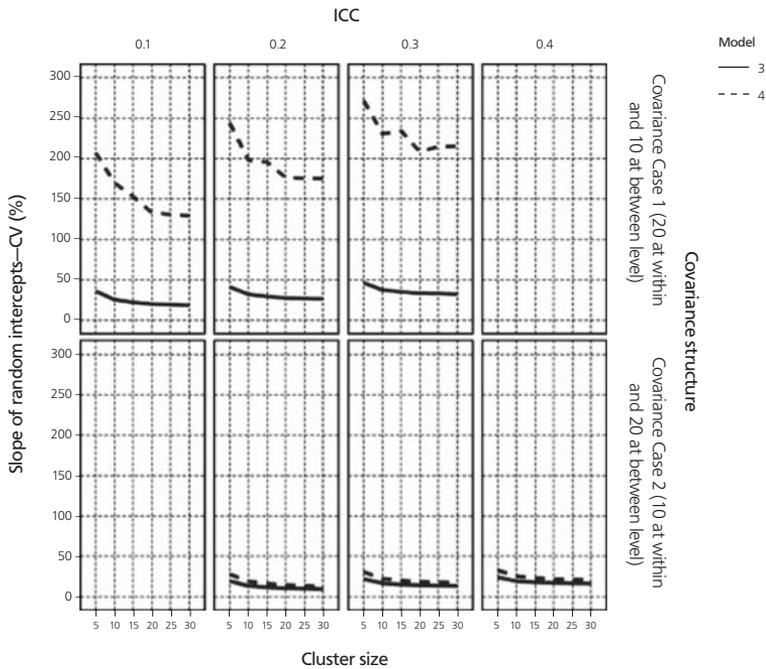


Figure 5.17: CV (%) of the slope of random intercepts by ICC, model, and cluster size: average over weight status and numbers of sampled clusters



5.3 Practical Use of Outcome Equations

Our intention in conducting this research was to provide outcomes that would be of practical use for researchers wanting to determine required sample sizes for a particular research project during its planning phase. We also anticipated that these outcomes would enable researchers doing secondary analysis of available LSA datasets to evaluate, in advance, the precision range that they could expect from the available data, depending on their research questions.

The graphs presented throughout Sections 5.1 and 5.2 above and the graphs accompanying the appendix tables offer a good starting point when determining required sample sizes under specific conditions. Ranges of sampling errors of specific model parameters to be expected for varying sample and population conditions can be easily read off these graphs. They also allow users unable to exactly specify population parameters, such as ICC or covariance distributions, in advance to draw conclusions relating to possible variations of sampling error.

The appendix contains the estimated quadratic equations for all sample and population conditions explored throughout this publication. Note that although some of the approximated quadratic equations appear to be insignificant, they still have very high R^2 values. This is the case, for example, for the equations describing the relationships between sample sizes and the coefficients of variation of the variance

of the random intercepts (Appendix Tables A8 to A15), and it is due to the fact that very few numerical values were used to build the curves (four or six respectively in this case). In fact, each of these numerical values represents an average of 6,000 replicates and therefore has a high reliability in its own right. This means that if the model goodness of fit is high, the equations are very reliable, even when their p values are not below 0.05.

In the remainder of this chapter, we use two examples to explain how the appendix equations can be of practical use. We emphasize, though, that the results are only valid under the specified model assumptions and cannot be generalized to different sample and population conditions (refer also to Section 6). For example, it is not certain whether the equations will hold for, say, within-cluster sample sizes of 50, or for different ranges or distributions of covariances. In particular, caution is advised for all model parameters where the coefficient of variation depends heavily on the covariance distribution. More research is required to explore the degree to which we can generalize the results to other settings and conditions.

In order to make use of the outcome equations in the appendix, the following steps need to be performed:

1. Formulate research question;
2. Specify the hierarchical model and the parameter of interest;
3. Determine fixed population and sample parameters;
4. Choose, apply, and solve equation;
5. Discuss limitations of observed result(s)

Example 1

A researcher is interested in reporting levels of association between family SES and academic achievement for various education systems. Model 2 is the appropriate model to use when endeavoring to answer the research question. The parameter of interest is β_1 (fixed slope). The researcher determines the parameter as being sufficiently precise if the SE takes no more than 10% of its value. For instance, if the estimate of β_1 is 18, its sampling error must be below 1.8. (Note that such a requirement can also be driven by the request to identify relevant differences between education systems within certain significance levels; see also Example 2 in this regard.)

The first step requires specification of each education system's sample and population conditions. (For the sake of simplicity, we use only one education system when describing the steps to be conducted.)

Let's assume the researcher wants to test one class per school; the average class size is 25. From previous research, it is known that the covariance between SES and achievement is stronger within than between groups and that the ICC is 0.3. The sample of schools will be selected with probabilities proportional to the

size of the schools, which means that the researchers will need to apply weights when they later carry out their analyses.

So, how many schools need to be sampled in order to answer the research question with the specified confidence levels? The correct equation can be found in Table A17 of the appendix, an excerpt from which is given here as Figure 5.18.

Figure 5.18: Excerpt from Appendix Table 17

Case	ICC	Cluster size	Weight status									
			Unweighted					Weighted				
			Model summary		Parameter estimates			Model summary		Parameter estimates		
R square	Sig.	Constant	b1	b2	R square	Sig.	Constant	b1	b2			
Covariance Case 1 (20 at within and 10 at between level)	.1	5	.993	.084	33.1	-.206	5.08E-04	.995	.068	36.2	-.222	5.35E-04
		10	.997	.055	23.7	-.146	3.54E-04	.997	.056	26.1	-.161	3.91E-04
		15	.995	.073	19.8	-.126	3.18E-04	.995	.067	21.6	-.136	3.36E-04
		20	.996	.067	16.8	-.106	2.63E-04	.995	.069	18.5	-.115	2.86E-04
		25	.995	.069	14.9	-.092	2.26E-04	.994	.076	16.7	-.107	2.75E-04
		30	.995	.068	14.0	-.088	2.20E-04	.995	.072	15.3	-.096	2.38E-04
	.2	5	.993	.083	32.6	-.204	5.03E-04	.994	.076	35.3	-.216	5.25E-04
		10	.997	.055	22.7	-.140	3.39E-04	.997	.058	24.9	-.151	3.61E-04
		15	.995	.073	18.8	-.120	3.02E-04	.995	.072	20.7	-.132	3.30E-04
		20	.996	.067	15.9	-.100	2.49E-04	.996	.066	17.6	-.110	2.74E-04
		25	.995	.068	14.1	-.086	2.13E-04	.996	.064	15.9	-.100	2.48E-04
		30	.995	.069	13.2	-.083	2.08E-04	.996	.066	14.6	-.093	2.32E-04
.3	5	.993	.083	31.2	-.195	4.81E-04	.995	.069	34.3	-.218	5.46E-04	
	10	.997	.055	21.3	-.132	3.19E-04	.996	.062	23.3	-.144	3.54E-04	
	15	.995	.073	17.6	-.112	2.82E-04	.995	.072	19.7	-.128	3.27E-04	
	20	.996	.066	14.9	-.093	2.31E-04	.996	.062	16.6	-.105	2.64E-04	
	25	.995	.068	13.1	-.080	1.98E-04	.997	.059	14.5	-.089	2.20E-04	
	30	.995	.069	12.3	-.078	1.94E-04	.996	.067	13.6	-.086	2.16E-04	
		.994	.077	53.6	-.336	1.12E-03	.995	.072	58.0	-.355	8.65E-04	

When the numbers are inserted into equation (6),

$$y = b_0 + b_1z + b_2z^2$$

becomes

$$10 = 14.5 - .089z + 2.2E^{-4}z^2,$$

with z being the number of clusters to be sampled. Solving the equation for z by applying the binomial theorem⁵⁰ leads to two results (rounded):

$$z = \{345; 59\}$$

Because the formula is only valid in a range from 50 to 200 clusters, the larger value, 345, should be dismissed. Therefore, 59 would be the estimated appropriate Level 2 sample size for that education system.

50 For convenience, users can refer to various programs, available on the internet, that solve quadratic equations (e.g., <http://www.math.com/students/calculators/source/quadratic.htm>).

When determining the expected coefficient of variation (in this example, 10%), the researcher would need to consider the valid ranges of this ratio under the given sample and population conditions. If he considers the values outside the valid range, the equation is either not solvable or the solution(s) of the equation take(s) values outside the explored ranges. If, for instance, a value of 5% rather than 10% is considered, the equation cannot be solved. The graphics in the results sections of this section of the monograph provide good reference points for valid ranges of coefficients of variation.

Example 2

A researcher is interested in comparing levels of association between family SES and academic achievement for two countries. Because she first wants to determine if a dataset originating from an LSA survey has data that will allow her to answer her research question, she needs to know how precisely the parameter was measured in that survey.

Again, Model 2 is the appropriate model to be applied and β_1 is the relevant parameter. Let's assume that both education systems fulfill the same preconditions as in Example 1. Let's also assume that a sample of 150 schools has been selected in both countries. The researcher can apply the same equation as the one in the previous example, but this time the sampled number of schools would need to be inserted:

$$y = 14.5 - .089 \times 150 + 2.2E^{-4} \times 150^2$$

$$y = 6.1\%.$$

The researcher can thus expect the 95% confidence interval of the parameter β_1 to be roughly within $\pm 12\%$ of β_1 , which implies that she will be able to identify, with respect to this parameter, only rather large differences between countries as significant under the given sample and population conditions.

Note that many further considerations other than those discussed above and in Section 5.2.1 drive the decision on sample sizes. For example, expected non-response rates, booklet rotation schemes, research interest in subgroups, and the like also need to be considered. However, these aspects are beyond the scope of this research and so are not addressed here.