Hierarchical Modeling with Large-Scale Assessment Data: Influence of Intra-Class Correlation on Sampling Precision

Sabine Meinck[1] and Caroline Vandenplas[2]

Abstract

Most data collected in educational large scale assessments (LSA) is very well suited for multilevel modeling because sampled individuals are usually nested within clusters (e.g., students nested within schools). Hierarchical models allow for the effect of explanatory variables at different clustering levels; parameters can be determined as fixed or random effects depending on the research question. All model parameters are however estimated based on random samples and are therefore subject to sampling error. A Monte Carlo simulation was utilized to explore the connections between the sample sizes at different levels and intra-class correlation coefficients (ICCs) in settings that mimic scenarios typical for LSA. It was observed that varying levels of ICC influence the margins of sampling variance of the estimated model parameters in different amounts and even different directions. Assuming fixed sample sizes, the coefficients of variation (CVs) of the model parameters *mean of random intercepts* (γ_{00}) and *slope of random intercepts* (γ_{01}) increased with increasing ICC levels, as expected. However, the inverted relationship was observed for parameters U_0 – *variance of random intercepts,* γ_{10} – *mean of random slopes, and* β_1 – *fixed slope*: with increasing ICC, the CVs decreased. The findings can help us to determine sample sizes in LSA when particular hierarchical models are to be investigated.

Introduction: Purpose, Framework and Academic Significance

More and more often, researchers apply hierarchical linear models (HLM) to analyze data collected in LSA. The clustered structure of the population and the data collected in such surveys often fits this type of analyses very well. As for any analyses using sample data, it is important to keep in mind that the obtained results are only estimates of the population parameters, hence it is essential to indicate the precision of these estimates when presenting the results. Moreover, sampling error levels determine whether it is reasonable to test specific hypotheses or examine group differences. A question arises then naturally: what is the required sample size to obtain a certain precision of the estimated parameters of the hierarchical model? Or the other way around, researchers may wonder what precision they can expect for the estimates they are interested in, given a certain sample size at the hierarchical levels. Various authors reported on different aspects of the subject (Snijders & Bosker, 1993; Raudenbush, 1997; Snijders, 2005; Cohen, 1998; Moerbeek, Van Breukelen and Berger, 2000 & 2001), inferring from real data (Afshartous, 1995) or simulation studies (Mok, 1995, Bell, Morgan, Schoeneberger, Loudermilk, Kromrey, & Ferron, 2010; Maas & Hox, 2005). The specific properties of educational LSA were, however, only partly covered.

[1] IEA Data Processing and Research Center, Sampling Section, Hamburg, Germany
[2] University of Lausanne, Faculty of Social and Political Sciences, Lausanne, Switzerland

This paper is based on results published by Meinck & Vandenplas (2012), exploring one of the topics addressed in their book in more depth. We examine the relationship between the sample sizes at the different levels of hierarchical linear models and the precision of the outcome of these models, focusing on the influence of the intra-class correlation on this relationship.

ICCs depend on the variable of interest (e.g., achievement score, background variable), population characteristics and features of the clusters that determine the hierarchical levels. Moreover, it is a well-known fact that intra-class correlation is directly related with the efficiency of sampling designs (Kish, 1965). It is therefore of prime importance to have a good understanding of the effect of the ICC on the relationship between sample size and the precision of the outcome of the hierarchical modeling.

Data and Methods

Conducting Monte Carlo simulations, two-stage samples (mimicking a setting with students nested in schools) were drawn from an infinite population. This sampling design reflects, for example, the applied design in ICILS (IEA's International Computer and Information Literacy Study) or PISA (Program for International Student Assessment, OECD). The population parameters were chosen such that they reflect a realistic population assessed in educational LSA. An achievement variable and a social-economic status index were created, and sample sizes at both levels were varied. The ICC values of 0.2, 0.3, and 0.4 were examined. Sampling weights were applied at level 2, reflecting school samples selected with probabilities proportional to their size. To mimic sampling weights that typically arise from this sampling method, the weight values were created following a Poisson distribution. For further details on the study design, refer to Meinck & Vandenplas (2012), note however that we limit the considered scenarios to the ones mentioned above. Four increasingly complex hierarchical linear models were examined for each scenario, going from the null or unconditional model to models conditional on level 1 variables, models with random slopes and models with level 2 covariates. Detailed model specifications can be found in the appendix. The sampling error for each of the model parameters, displayed as coefficient of variation (CV), was the main outcome of the simulation study. We chose to report CVs rather than sampling errors because the latter are only meaningful with respect to the corresponding parameter. Instead, the CV is a standardized measure of sampling variation.

Results and Discussion

Relation between ICC, sample sizes and CV

As expected, the CV generally decreased with increasing sample sizes on one or both levels. One main outcome of the simulation study was that this dependency always followed a quadratic curve progression within the explored settings (see Meinck & Vandenplas, 2012). This means that the gain in precision is bigger when, for example, one increases the cluster size from 5 to 10 than from 15 to 20. The effect of the ICC on this relationship was (close to) linear for most explored parameters. The direction of the effect was however depending on the model parameter of interest. This fact deserves accentuation: the general statement that higher ICCs always lead to higher sampling errors does actually not hold true for all estimated parameters of a hierarchical model. For example, the CV of the *mean of the random intercepts* (parameter γ_{00}) increases with increasing ICC for fixed sample sizes, independently of the studied model (figure 1). The described effect of the ICC amplifies a little with increasing within-cluster

sample sizes (lines get steeper) while it reduces for increasing level 2 sample sizes (lines get flatter). This is independent of the model complexity. We illustrate the effect in figure 2, which zooms in a part of figure 1. Similar effects of the ICC on the relation between sample sizes and CVs were found for the model parameter *slope of random intercepts* (γ_{01}).

The impact of the ICC on the CVs of the parameters U_0 *– variance of random intercepts,* γ_{10} *– mean of random slopes, and* β_1 *– fixed slope* is however inverted: with increasing ICC, the CVs of these parameters decrease. Figure 3 shows the effect for parameter β_1 (*fixed slope*). As can be seen, the effect is linear and amplifies slightly the less clusters are sampled. It remains stable for different cluster sizes in model 2 (lines run almost parallel) and reduces a little with increasing cluster sizes in model 3. The smallest clusters (i.e., with 5 units) behave a bit inconsistently here. Biased parameter estimates due to small samples may be one reason for this observation, bearing in mind that for such estimates we may also not want to trust their corresponding CVs (Meinck & Vandenplas, 2012).

For parameter γ_{10} (*mean of random slopes*), the effect is again linear and reduces with increasing sample sizes at both levels (figure 4).

We see from figure 5 that the effect for the *mean of random slopes* (U_0) decreases too when increasing sample sizes at both levels. In contradiction to the two parameters discussed above, the effect is non-linear and diminishes with increasing ICC levels.

When developing a cluster sampling design, researchers are usually wary of high ICC, because they are known to increase the sampling error. Therefore some of the results in the previous paragraphs may be surprising: why would the precision of some estimated parameters increase with increasing ICC? This is somehow contra-intuitive but can be explained by taking a further look at the factors influencing the sampling error of the different estimates.

From sampling theory, researchers know that the sampling error of the mean of a variable of interest is proportional to the standard deviation of the same variable (Kish, 1965):

$$SE(\hat{\bar{y}}) \sim \frac{SD(y)}{\sqrt{n}}$$

where *n* is the sample size. In case of cluster sampling, the random intercept can be seen as the variable of interest and the sampling error of the *mean of random intercepts* (γ_{00}) is then proportional to the standard deviation between the clusters. Hence, if the variance between clusters (σ_B^2) increases, the sampling error of the *mean of random intercepts* (γ_{00}) increases, too. And this is exactly what happens when the ICC increases and the total variance (σ^2) is kept constant (Kish, 1965):

$$ICC = \frac{\sigma_B^2}{\sigma^2}.$$

A similar reasoning can be done about the *slope of random intercepts* (γ_{01}). The sampling error of the slope is proportional to the standard deviation of the variable of interest divided by the standard deviation of the explaining variable (Cochran, 1977):

$$SE(\gamma_{01}) \sim \frac{SD(y)}{SD(x)}.$$

In this case, as the standard deviation of the explaining variable at the cluster level remains constant, the sampling error of the slope of random intercepts increases with increasing variance between clusters, i.e. increasing ICC.

At the contrary, γ_{10} – *mean of random slopes, and* β_1 – *fixed slope* depend on the variance within clusters. With increasing ICC and constant total variance, the variance within clusters diminishes and so do the sampling errors of the mean of random slopes and the fixed slope.

The CV of the *variance of random intercepts* (U_0) also diminishes with increasing ICC. This is very surprising as the sampling error of the variance is proportional to the variance itself (Ahn and Fessle, 2003), which, in our case escalates (variance between clusters) with increasing ICC. This counterintuitive finding is probably due to the fact that we are not looking at the sampling error but at the CV, dividing the sampling error of the estimated parameter by the estimated parameter itself. We carried our analysis out with this approach in order to facilitate comparisons of the results by using a standardized measure of sampling variation. For all the other parameters estimated in the different models this had no or little impact, as the value of the parameters did not change with different ICC. However, the *variance of random intercepts* increases with increasing between-cluster variance and, hence, increasing ICC, making the interpretation of this outcome more difficult. To fully understand the impact of ICCs on the CV of the *variance of random intercepts*, a similar simulation should be done with constant between-clusters variance, rather than constant total variance.

Interactions with model complexity

The effects of the ICC explained above can be influenced further by the model complexity. Evidently, not all parameters are part of all models.[3] Therefore, in this section we can only consider parameters that are part of more than one model. For our considered models, we found such relationship for two parameters.

Let us look first at the CV of parameter *mean of random intercepts* (γ_{00}). Figure 6 illustrates the relationship between the increasing ICC levels and increasing cluster sample sizes by model. When comparing models that include an explanatory variable at the level 2 (model 3 and 4) with those that don't (models 1 and 2), we find the following effect: although the CV of the estimated parameter is smaller in model 3 and 4, the influence of the ICC on the CV is bigger. In other words, the precision of the estimated parameter depends more on the ICC when an explanatory variable at cluster level is introduced to the model, while it is independent on whether an explanatory variable is introduced at level 1 or not (compare models 1 and 2). Also, it makes no difference whether the slope is considered to be fixed or random (compare models 3 and 4). The relation becomes more pronounced with fewer sampled clusters but less pronounced with smaller clusters (see figure 7).

[3] Please refer for further information on considered models to the appendix.

The interaction between ICC, model complexity and the CV of the *variance of random intercepts* (U_0), are displayed in figure 8. We ignore here the moderating effects of the varying sample size scenarios since they are negligible. Apparently, the influence of the ICC on the precision of this parameter is rather small in models 1 and 2 but increases dramatically when an explanatory variable at level 2 is introduced to the model (models 3 and 4). In fact, the parameter may become insignificant in similar scenarios when less than 50 clusters contribute to the estimate (refer to figure 5). This is very likely when subgroups within countries are to be compared.

ICC effects on precision - Impact on data analysis and model interpretation

If one looks at the CVs of the different model parameters, it becomes apparent that some are generally "easier" to measure than others. In other words, the sampling precision varies a lot between parameters. For example, the *mean of random intercepts* (γ_{00}) has a very small CV for all considered scenarios. Even in the least optimal considered case (smallest sample sizes and highest ICC), it remains below 2.5%. Hence, even if lower ICCs induce significantly higher precision, it is not that relevant for the interpretation of this parameter.

Let us illustrate this with an example. If parameter γ_{00} constitutes the mean achievement score and is estimated as 500, then, for the considered worst case scenario, its sampling error is around 11 (which is equal to 2.2% of 500). The corresponding confidence interval can be estimated then as 500 ± 1.96[4] x 11. Consequently, we could state that we are 95% certain that the true value of this parameter in the population lies somewhere in the interval between 478 and 522. If the ICC is rather 0.2 then 0.4, the sampling error would be around 9 (which is equal to 1.8% of 500). Hence, the confidence interval shrinks accordingly (482 to 518). Of course, if the research question rather aims on comparing this parameter for different subgroups, the precision plays an important role, too, and the difference between the two discussed precision levels may often make the distinction between being able to detect a significant group difference or not.

However, all other parameters are affected with much larger sampling errors than the *mean of the random intercepts* (γ_{00}). The CVs of most estimated parameters increase easily to 20% or 30% or even higher, depending on the specific conditions. For example, to determine the relationship between socio-economic background and achievement after controlling for the school-level socioeconomic background, a hierarchical model equivalent to model 2 should be applied, measuring this relation by parameter β_1 and its significance. The CV of this parameter is in most considered sampling scenarios between 10% and 20%, hence, it should usually be easy to detect whether this parameter is significantly different from zero. However, if one wishes to compare this parameter for different groups, the group difference would need to be very large in order to be detected.

Summary and Conclusions

The results show that the interaction between the ICC, the sample sizes and the precision levels depend heavily on the parameter of interest. For parameters measured at the cluster level, increasing ICCs induce larger sampling errors. In contradiction, the CVs of the model parameters estimated at the

[4] For simplicity, we assume here infinite degrees of freedom.

individual level increase with lower ICCs, assuming a constant total variance. Further similar analysis for which the variance between clusters or within clusters would be kept constant instead of the total variance could lead to a deeper understanding of the relationship between ICC and CV, especially regarding the variance parameters at the different levels.

Publicly available LSA datasets come along with restricted sample sizes. In most cases, the sample sizes on level 1 (e.g., students within schools) do not exceed 20 to 30 units on average. On level 2 (e.g., schools), the sample sizes range mostly around 150 units if the full sample is used but are much smaller if specific subgroups (e.g., regions within a country) are examined. These restrictions limit the possibilities to perform analyses with meaningful results; i.e., recognizing significant effects of a model parameter or differences between subgroups is very unlikely if sampling errors are high, or, in return, sample sizes are too small.

The objective of secondary analysis of LSA data utilizing HLM is very often to compare subsamples within or across countries (e.g. disadvantaged/immigrant students, boys versus girls etc.). Researchers who would like to employ HLM to analyze such publicly available datasets should make themselves aware of the restrictions set by the limited sample sizes and formulate their research questions accordingly. The influence of varying ICCs in populations (and hence, also in the data) on the expected precision of parameters in focus of the research questions should also be considered. When interpreting model outcomes, it is of utmost importance to closely monitor and report achieved precision levels of parameter estimates.

When designing a new study, sample sizes can be chosen with respect to the research questions and expected precision levels of parameters of interest, bearing the specific requirements for HLM analysis in mind.

References

Ahn, S. & Fessle, J. A. (2003). Standard errors of mean, variance, and standard deviation estimators. *Technical Report 413,* Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122.

Afshartous, D. (1995). Determination of sample size for multilevel model design. In: V. S. Williams, L. V. Jones, & I. Olkin (Eds.). Perspectives on statistics for educational research: Proceedings of the National Institute for Statistical Sciences (NISS), Technical Report #35.

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Loudermilk, B. L., Kromrey, J. D., & Ferron, J. M. (2010): Dancing the sample size limbo with mixed models: How low can you go?, SAS Global Forum 2010 Posters Paper 197-2010. Retrieved from http://support.sas.com/resources/papers/proceedings10/197-2010.pdf

Cochran, W.G. (1977). *Sampling Technics*, Third Edition, New York, Wiley.

Cohen, J. (1998). Determining sample sizes for surveys with data analyzed by hierarchical linear models. *Journal of Official Statistics, 14*(3), 267-275.

Kish, L. (1965). *Survey sampling*, New York, NY: Wiley.

Meinck, S. & C. Vandenplas (2012). Sample size requirements in HLM: An empirical study. IERI Monograph Series Issues and Methodologies in Large-Scale Assessments. IER Institute, Special Issue 1, Educational Testing Service and International Association for the Evaluation of Educational Achievement.

Mok, M. (1995). Sample size requirements for 2-level designs in educational research. Multilevel Modelling Newsletter, 7(2), 11-15.

Raudenbush, S.W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology 2005, 1*(3), 86-92.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design issues for experiments in multilevel populations. *Journal of Educational and Behavioral Statistics, 25*(3), 271-284.

Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2001). Optimal experimental design for multilevel logistic models. *Journal of the Royal Statistical Society, Series D (The Statistician), 50*(1), 17-30.

Snijders, T. (2005). Power and sample size in multilevel linear models. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (1570-1573). Volume 3, Chichester, U.K.: Wiley.

Snijders, T., & Bosker, R. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics, 18*(3), 237-259.

Appendix

Specifications of Examined Hierarchical Models

Model 1: The empty (or null) model: No explanatory variable in the model, the intercept is random.

$$\begin{cases} y = \beta_0 + \varepsilon \\ \beta_0 = \gamma_{00} + U_0 \end{cases}$$

Model 2: One explanatory variable at level 1, the intercept is random and the slope is fixed.

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

Model 3: explanatory variable at level 1 and level 2, the intercept is random and the slope is fixed

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

Model 4: One explanatory variable at level 1 and level 2, the intercept and the slope are random.

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

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 slope (SES indicator)
γ_{10}	Mean of random slopes
U_1	Variance of random slopes

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

Figure 1: Coefficient of variation of the parameter γ_{00} (mean of random intercepts) in dependency of the ICC, the number of sampled clusters and cluster size by model

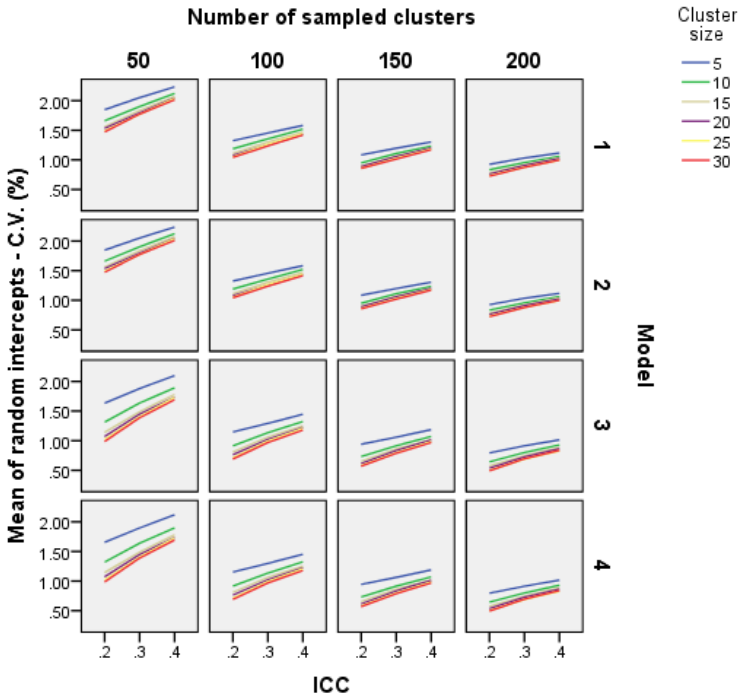

Figure 2: Coefficient of variation of the parameter γ_{00} (mean of random intercepts) in dependency of the ICC, the number of sampled clusters and cluster size, model 1

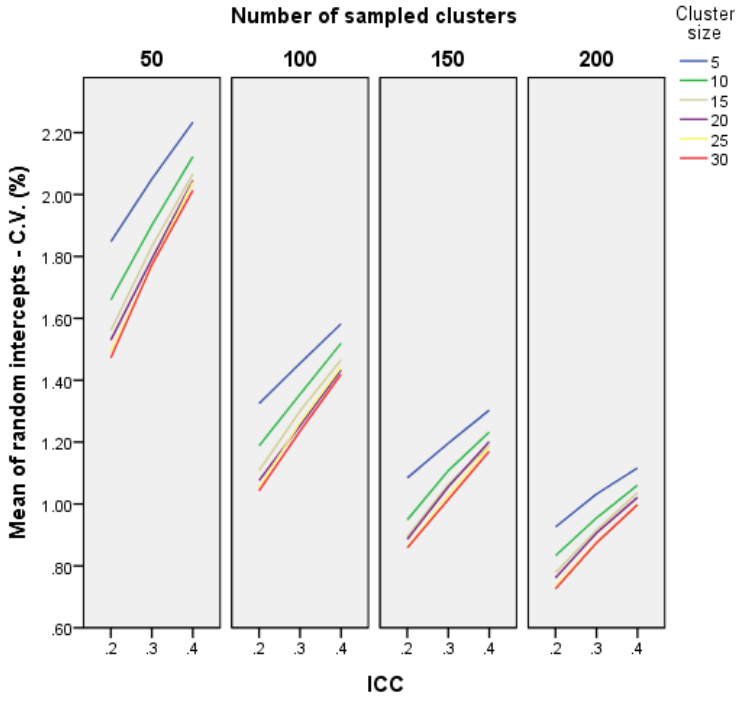

Figure 3: Coefficient of variation of the parameter β_1 (fixed slope) in dependency of the ICC, the number of sampled clusters and cluster size by model

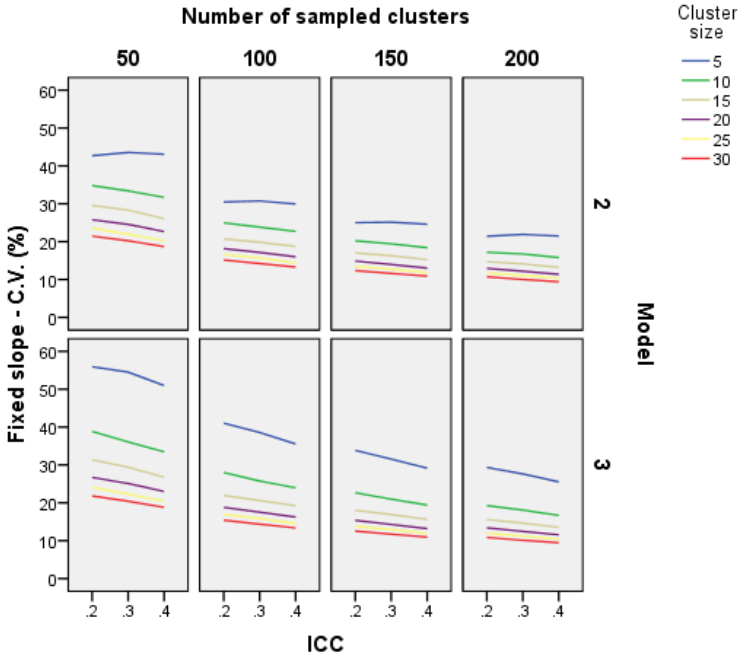

Figure 4: Coefficient of variation of the parameter γ_{10} (*mean of random slopes*) in dependency of the ICC, the number of sampled clusters and cluster size, model 4

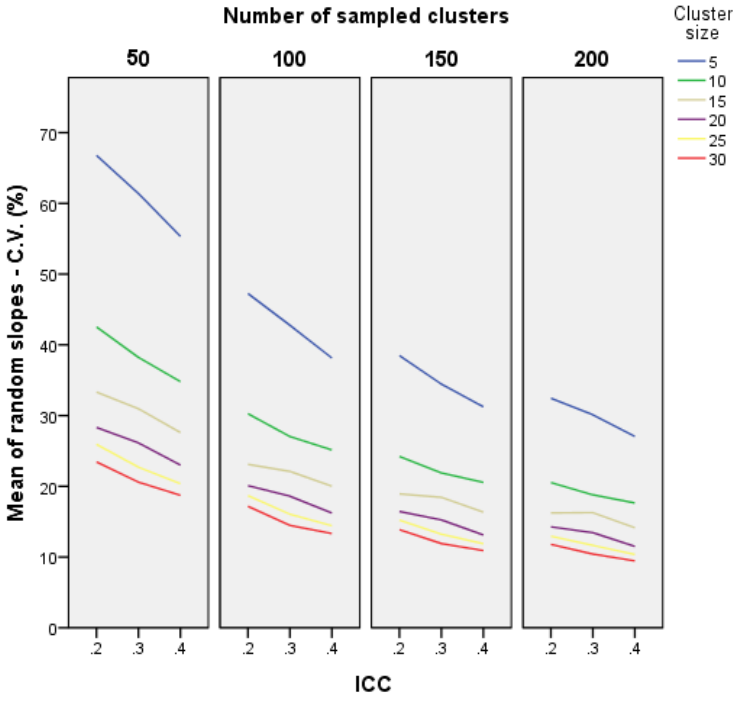

Figure 5: Coefficient of variation of the parameter U_0 (*variance of random intercepts*) in dependency of the ICC, the number of sampled clusters and cluster size, averaged over all models

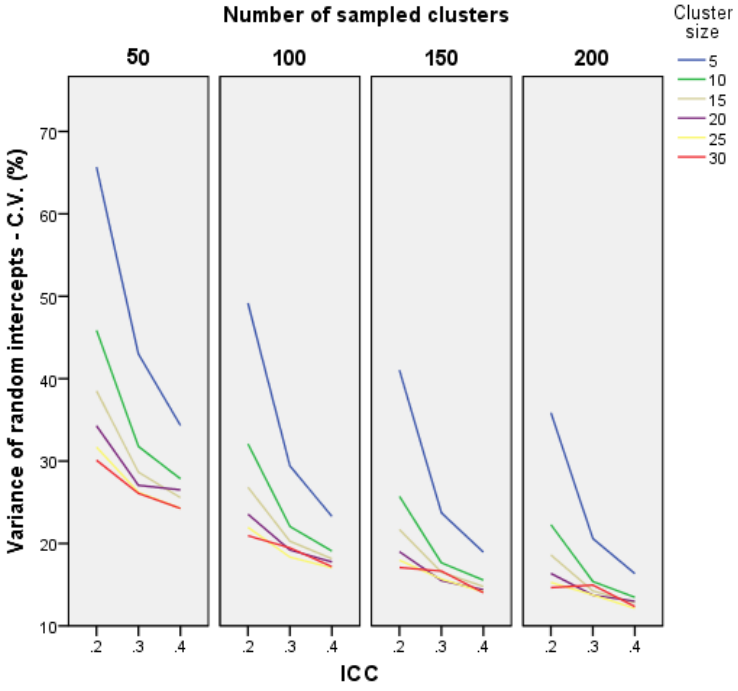

Figure 6: Coefficient of variation of the parameter γ_{00} (mean of random intercepts) in dependency of the ICC and the number of sampled clusters by model

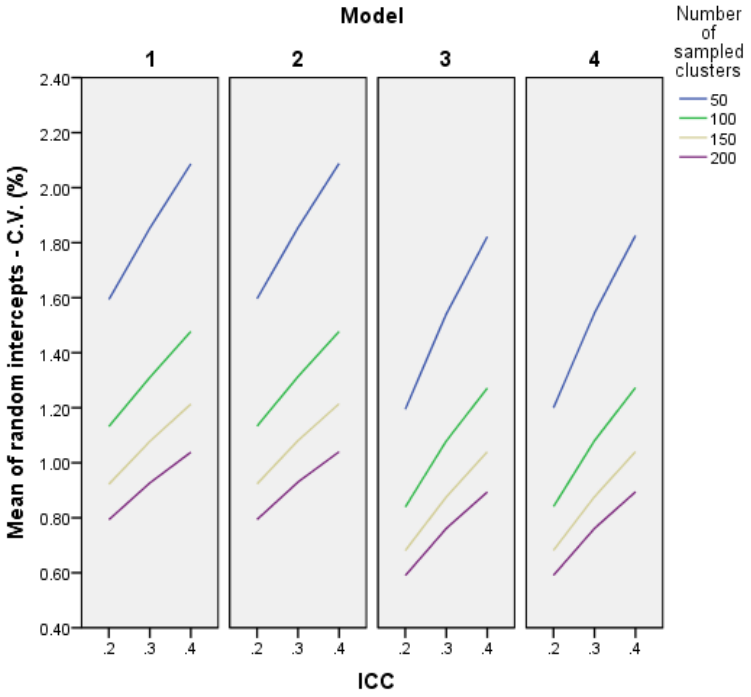

Figure 7: Coefficient of variation of the parameter γ_{00} (mean of random intercepts) in dependency of the ICC and the cluster size by model

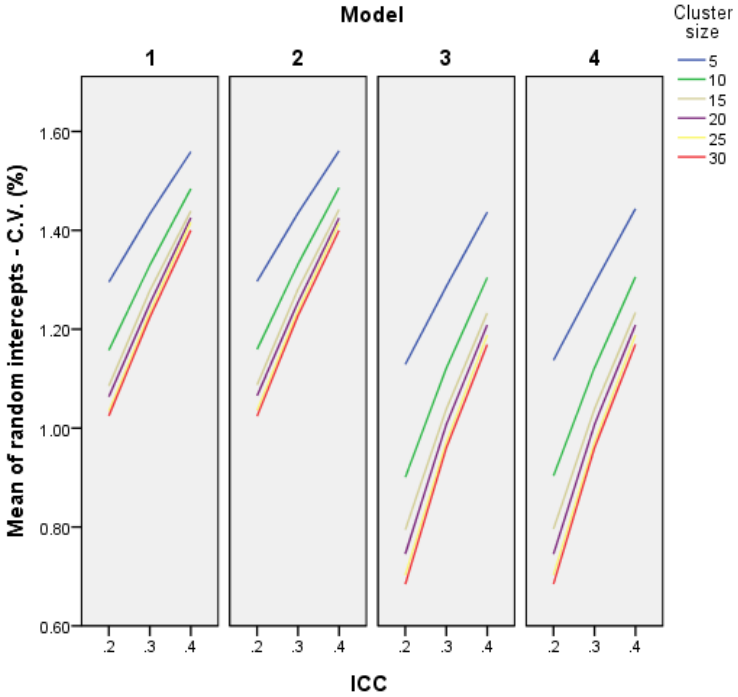

Figure 8: Coefficient of variation of the parameter U_0 (*variance of random intercepts*) in dependency of the ICC by model