Journal of Modern Applied Statistical Methods

Volume 22 | Issue 1 Article 4

The Influence of Number of Clusters and Level-2 Residual Distribution on Multilevel Parameter Estimates: A Latin Hypercube Design Perspective

Hao Jia
National Council of State Boards of Nursing, United States,
hjia@ncsbn.org

Yadira Peralta
Centro de Investigación y Docencia Económicas: Aguascalientes, Mexico,
Stanford Center for Biomedical Informatics research, United States,
yadira.peralta@cide.edu

Michael Harwell
University of Minnesota, Minneapolis, United States,
harwe001@umn.edu

Recommended Citation

Hao Jia, Yadira Peralta, Michael Harwell (2023). The Influence of Number of Clusters and Level-2 Residual Distribution on Multilevel Parameter Estimates: A Latin Hypercube Design Perspective. Journal of Modern Applied Statistical Methods, 22(1), https://doi.org/10.56801/Jmasm.V22.i1.4

Journal of Modern Applied Statistical Methods
2023, Vol. 22, No. 1,
Doi: 10.56801/Jmasm.V22.i1.4

ISSN 1538 – 9472

The Influence of Number of Clusters and Level-2 Residual Distribution on Multilevel Parameter Estimates: A Latin Hypercube Design Perspective

Hao Jia
National Council of State Boards of Nursing, United States

Yadira Peralta
Centro de Investigación y Docencia Económicas: Aguascalientes, Mexico, Stanford Center for Biomedical Informatics research, United States

Michael Harwell
University of Minnesota, Minneapolis, United States

Monte Carlo methods with Latin Hypercube Design (LHD) were used to explore the impact of increasing numbers of clusters considering four cluster residual distributions in a two-level model. The implementation of LHD is illustrated and the results update existing guidelines for the number of clusters needed to obtain accurate estimates.

Keywords: Multilevel modeling, Non-normally distributed residuals, Number of clusters, Latin Hypercube Design.

1. Introduction

The presence of hierarchical data structures resulting from multi-stage sampling, such as students nested within classrooms (clusters), has prompted the development of multilevel models which are widely used in disciplines such as education (e.g., DiPrete & Forristal, 1994), psychology (e.g., Rose et al., 2014), and sociology (e.g., Lee, 2000). Statistical details of these models can be found in Raudenbush and Bryk (2002).

1.1 Overview of Latin Hypercube Sampling (LHS) and Latin Hypercube Design (LHD)

Previous multilevel model simulation studies have adopted a full factorial design. Although informative, full factorial designs are not the ideal choice when prediction accuracy over the complete experimental region (the region where manipulated factors under study take values) is of interest, (Santner et al., 2018). A simulation sampling method with substantial potential for increasing the generalizability of simulation results, called Latin Hypercube Sampling (LHS), has been shown to have

desirable properties such as including dense coverage of a user-specified range of values of each manipulated factor in the simulation study which enhances generalizability. Thus, LHS is an attractive option when generalizing across a pool of simulation conditions is central to the rationale for a simulation study, compared to other sampling methods such as simple random sampling or stratified sampling of simulation conditions (Authors, 2017; Iman & Conover, 1980; McKay et al., 2000; Stein, 1987).

The key feature of LHS is spreading design points evenly over the range of each manipulated factor independently, which enhances generalizability, accommodates a variety of statistical simulation settings, handles both small- and high-dimensional problems in terms of manipulated factors, and, thanks to advancements in computing, can be implemented with relative ease (Santner et al. , 2018; Viana, 2016; Wang, 2003). The simulation design resulting from an LHS is called a Latin Hypercube Design (LHD).

LHDs have apparently not been used in Monte Carlo simulation studies of multilevel modeling. Instead, full factorial designs have traditionally been used with narrow ranges for the manipulated factors, such as J=10~30 and J=5~30 (Stegmueller, 2013), and few values, such as J=30, 50, 100 (Maas & Hox, 2004a) and J=50, 100, 200, 500 (Bell et al., 2008). Using a sampling mechanism that populates the entire experimental region to produce an experimental design that enhances generalizability by having evenly spread points that represent all segments of the experimental region can deepen our understanding of the manipulated factors' effects.

A feature of most simulation studies of multilevel models is varying the number of clusters required to warrant accurate parameter estimates. The aim of the current study is to revise and expand the current advice on the number of clusters by using an efficient experimental design, namely the LHD, that offers greater generalizability.

1.2 LHD vs. Full Factorial Design

Santner, Williams and Notz (2018) defined inputs in a simulation as numerical values of simulation factors that collectively define the experimental region, which in turn defines the design. Traditionally, the experimental design of two-level models' simulation studies was a specification that included J, n_j and ICC values in the experimental region at which the authors wished to compute an outcome. We further define design points as combinations of values of all manipulated factors, and those design points are sampled from the experimental region using one of several sampling methods. The number of design points and the way design points are sampled is directly related to the generalizability of simulation results.

The full factorial design appears to be the only experimental design that has been reported in existing multilevel model simulation studies. However, researchers usually must make trade-offs between generalizability and computational efficiency. Suppose previous literature suggests that the impact of J = 5, 10, 15, …, 200 and n_j= 5, 10, 15, …, 50 on estimation and hypothesis-testing in two-level models should be studied. A full factorial would consist of $40 \times 10 = 400$ design

points, which is unwieldy. Adding a third simulation factor such as ICC levels (e.g., 0.05, 0.1, 0.15, …, 0.5) makes the computational load unmanageable (4,000 design points in total). To reduce the load, the number of clusters might be specified as J = 10, 30, 50, 70, 200 and cluster sample sizes as n_j = 5, 10, 30, 50. These full factorial design points appear in Figure 1 as circles and reflect a predictable pattern because they fall at the intersection of specified horizontal and vertical lines. This pattern can limit the generalizability since it only considers a small portion of the possible design points in the experimental region, and there are several intervals that are not represented in the design.

A different sampling method, LHS, can distribute design points randomly throughout the experimental region and thus enhance the generalizability of findings. An example using LHS is given next following (Authors, 2017). Suppose J = 5, 10, 15, …, 200 and n_j = 5, 10, 15, …, 50 and a total of 20 design points (paired values of J and n_j) were randomly sampled. The choice of 20 is arbitrary and for illustrative purposes. The more design points, the greater the coverage of the experimental region. Contrasting the solid triangles in Figure 1 – representing design points under LHS – against the circles – representing design points for a full factorial – reveals that the design points of LHD are spread more evenly throughout the experimental region in a random pattern. Thus, simulation findings obtained from a LHD would be generalizable to the complete range of values specified by the researcher for the manipulated factors. The greater generalizability offered by LHD prompts us to use this design when studying the impact of numbers of clusters.

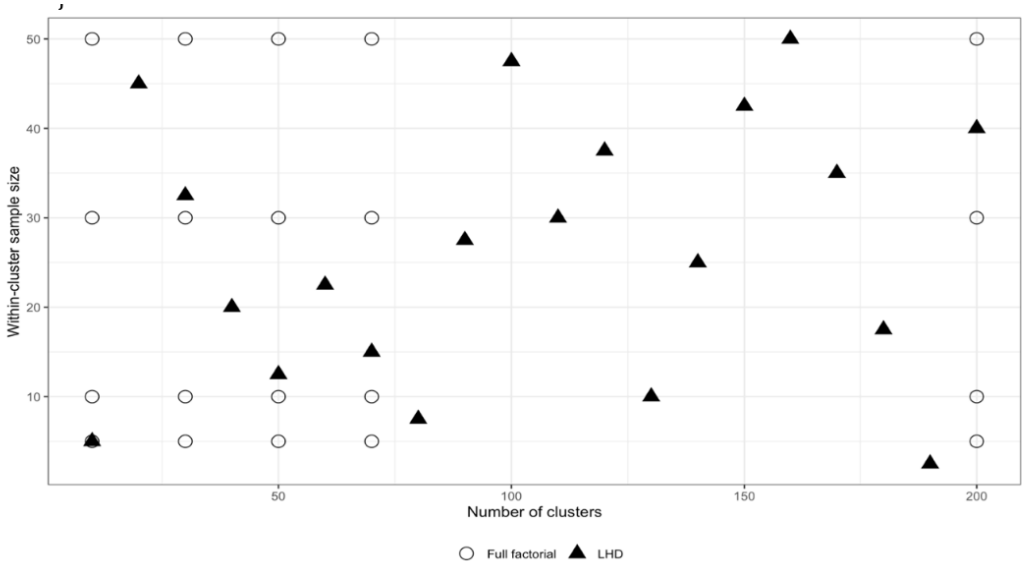

Figure 1. Contrasting the experimental region of a full factorial versus Latin Hypercube design for J and n_j (q = 20)

1.3 Number of Clusters and Level-2 Residual Distribution in Multilevel Models

The number of clusters (*J*) in multilevel analyses is particularly important because maximum likelihood, which is typically used to estimate some or all model parameters, requires a sufficiently large *J* to ensure properties of these estimators hold (Hox, 2002). For example, the accuracy of estimating the between-cluster variance component for intercepts via maximum likelihood generally improves with increasing *J* (Maas & Hox, 2005). Simulation studies have shown that the impact of within-cluster sample size n_j affects the accuracy of within-cluster estimates but its impact on the accuracy of level-2 effcts estimation is typically modest (Kraemer, 2012; Maas & Hox, 2004a; McNeish & Stapleton, 2016), leaving the sample size focus on *J*.

One of the earliest recommendations for *J* came from Kreft (1996) who advocated the 30/30 rule (30 clusters and 30 observations per cluster), although $J < 30$ often appears in empirical studies (e.g., Dedrick et al., 2009; Epstein et al., 2011; Jitendra et al., 2017; Marks & Printy, 2003; Shernoff et al., 2017). It is important to mention that simulation-based power analysis can be used to determine *apriori* the value of *J* needed to ensure a desired statistical power in multilevel models to detect an effect of interest (see Kumle et al. 2020, Lane and Hennes, 2018, Scherbaum, Ferreter, 2009). However, an investigation of how parameter estimates are impacted by varying sample sizes and other simulation factors might reveal details that the power analysis approach cannot. Since 1988, around 30 Monte Carlo simulation studies have included varying values of *J* as a simulation factor, but a consensus on minimum values of *J* to ensure accurate estimates and valid statistical inferences from hypothesis-testing for two-level models under realistic data conditions has not appeared.

For example, Shieh, Fouladi, and Pullum (2001) studied the impact of normally-distributed model residuals with $J = 5$, 20, and 80 for a two-level model. These authors reported that fixed effects estimates showed negligible bias regardless of *J* but the variance component for intercepts (τ_{00}), estimated using restricted maximum likelihood (REML), tended to be underestimated for smaller *J*. Maeda (2007) simulated normally-distributed model residuals and reported inflated Type I error rates for tests of τ_{00} for models with predictors at both levels for $J = 10$, 15, 20, and 25. McNeish and Stapleton (2016) used a Monte Carlo study to generate normally-distributed model residuals with predictors at both levels for $J = 4$, 8, 10, 14 and $n_j =$ 4, 8, 10, and 14 and found substantially biased estimates of τ_{00} under full maximum likelihood, and less (but still) biased estimates under REML.

Maas and Hox (2005) reported coverage rates for a 95% confidence interval for fixed effects and variance components (the latter based on REML estimates) of a model with predictors at both levels. These authors studied the impact of J = 30, 50, 100, nj = 5, 30, 50, and ICC = .10, .20, and .30. Fixed effects and their standard errors showed little bias regardless of *J,* whereas coverage rates for variance components were underestimated for J=30 (e.g., coverage rate of 91.1% rather than 95%). Seco, Garcia, Garcia, and Rojas (2013) reported that estimates of fixed effects and their

standard errors were essentially unbiased for normally-distributed model residuals but coverage rates for intervals for 95% confidence intervals for variance components estimated using REML were biased even for J = 100. Kraemer (2012) examined intervals for τ_{00} using several methods for constructing confidence intervals in an unconditional model for J = 5, 10, 20, 40, 80, and n_j = 10, 20, 40, 80, 160. Kraemer (2012) reported coverage rates based on REML estimates were satisfactory for a procedure due to Satterthwaite (Satterthwaite, 1946) for normally-distributed model residuals for $J \geq 10$ (see Kraemer, 2012 for details of the Satterthwaite procedure).

McNeish and Stapleton (2014) reviewed 20 Monte Carlo studies of two-level models, 14 of which assumed continuous and normally-distributed cross-sectional outcomes. These authors qualitatively synthesized these results and recommended a minimum of J = 15 for fixed effects estimates, and J = 30 for their standard errors and for level-2 variance component estimates. Authors (2018) used meta-analytic methods to summarize the results of 14 Monte Carlo studies of multilevel models (7 of these studies were used in McNeish and Stapleton 2014). Estimation bias and Type I error rates served as effect sizes of the meta-analysis, and the number of effect sizes ranged from K = 8 to 186 per study. Estimates of fixed effects and their statistical tests were generally insensitive to J whereas variance component estimates and their tests showed some bias even for J > 100 (K was typically > 50). Similar findings for normally-distributed residuals have been reported by Clarke and Wheaton (2008a), Korendijk, et al. (2008), Vallejo et al. (2015), and McNeish and Stapleton (2016). On the other hand, Browne and Draper (2006) found little evidence of bias in variance components estimated using REML for J = 6, but reported unsatisfactory coverage rates for τ_{00} for small J.

Guidelines for minimum values of J are also important when the assumption of normality of cluster residuals is violated because the standard errors of variance components are often badly biased, resulting in inaccurate estimates and unacceptable Type I and Type II error rates (Maas & Hox, 2004a, 2004c, 2005; Seco et al., 2013). Based on the work of Hill and Dixon (1982), Micceri (1989), and Dyer et al. (1999), non-normally distributed variables are common in empirical studies in several domains, suggesting non-normally distributed cluster residuals may also be common. However, the frequency and nature of non-normal distributions of cluster residuals in empirical studies has not been rigorously documented, in part because journal articles reporting multilevel model results rarely describe the distribution of cluster residuals (Dedrick et al., 2009).

A few Monte Carlo studies have examined the impact of non-normal cluster residuals. Maas and Hox (2004a, 2004b) reported that three non-normal cluster residual distributions ($\chi^2_{df=1}$, uniform, Laplace) had little effect on estimates of fixed effects and variance components regardless of J (30, 50, 100) or n_j, but reported poor coverage rates for REML-estimated variance components (e.g., 66% for a 95% confidence interval). Kraemer (2012) reported poor coverage rates for gamma, beta,

and t-distributions of cluster residuals for the estimated variance of the intercept in an unconditional model using REML estimation (e.g., 79.6% for a 95% interval and $J = 80$) that worsened as J increased.

Burch (2011) studied the impact of normal, uniform, and chi-square (df = 1) distributions for $J = 5$, 10, 50, and 100, and small, varying n_j on the ICC for the unconditional model. For $J = 5$ and normally- distributed cluster residuals, interval coverage tended to be overestimated (e.g., 98% for a 95% interval), and underestimated for $J = 100$. For the chi-square distribution this pattern was exacerbated. Similarly, Auda et al. (2019) used a random intercepts model with a single level-1 predictor whose slope variance was set to zero to study the impact on the ICC. These authors compared REML estimates against a nonparametric estimator for $J = 10$, 20, 30 and $n_j = 5$, 10, 30 for normal and "corrupted" normal distributions where the latter 5% of the values from N(0,1) were replaced with values from N(75, 900), which generated outliers and a heavy-tailed distribution. The results showed minimal bias of estimated fixed effects regardless of J, n_j , or cluster residual distribution, whereas estimates of the ICC for the corrupted normal distribution showed substantial bias regardless of J. Verbeke and Kesaffre (1996) showed that random effects may be poorly estimated if normality is assumed when the distribution of random effect is a mixture of normal distribution. On the other hand, McCulloch and Neuhaus (2011a) examined the symmetrical Tukey lambda distribution and suggested that estimation of the random effects variance appears relatively robust to misspecification of the random intercept distributional shape, and McCulloch and Neuhaus (2011b) added that the prediction accuracy of slopes variance is only minimally affected for mild-to-moderate violations of the normality assumption. By investigating the influence of skewed, bimodal and heteroscedastic residuals on the simple random intercept model, Scheilzeth, Dingemanse and Allegue (2020) found that estimates were usually robust to violations of assumptions, but estimates for random effects became less precise. These authors also pointed out that when the model became more complicated, the presence of non-normally-distributed random effects for random slopes can produce biased and imprecise estimates.

In sum, varying Monte Carlo results based on normally-distributed model residuals have produced a variety of recommendations for the minimum J needed to ensure accurate cluster-level parameter estimates including $J = 6$ (Browne & Draper, 2006), $J = 10$ (Kraemer, 2012), $J = 30$ (McNeish & Stapleton, 2014), $J = 50$ ((Hox, 1995; Maas & Hox, 2004a, 2005; Van der Leeden & Busing, 1994), and $J = 100$ (Hox & Maas, 2001; Seco et al., 2013). The lack of a consensus for minimum values of J is related to variation in the model features that are studied including the parameters that are estimated, number of level-1 and level-2 predictors, presence of level-2 slope models, the criteria used to assess the quality of estimation (bias, standard errors, coverage rates for 95% confidence intervals), and the possibility that cluster residuals are not normally-distributed. The net effect is that data analysts will find mixed guidance for the number of clusters needed to ensure accurate estimates assuming normally-distributed model residuals, and little guidance for J when non-normal cluster residual distributions are present.

Specifically, there are two main shortcomings in the simulation literature: (1) The results often do not represent realistic data conditions (e.g., non-normality of cluster residuals); (2) Results are conditional on the simulation factor values modeled (e.g., J = 30, 50, 100) and do not support inferences about the impact of other values (e.g., J = 10, 40, 200), individually or in combination with other simulation factors (e.g., non-normal cluster residual distributions), which limits generalizability. An examination of the impact of systematically increasing *J* across a range of realistic conditions for normal and non-normal cluster residual distributions can update and extend existing recommendations for the minimum *J* needed to obtain accurate parameter estimates.

2. Methodology

A LHD in a simulation study was used to investigate the accuracy of estimates, along with the validity of statistical inferences from hypothesis-testing, for a two-level model with random slopes. This model was fitted to continuous cross-sectional data under a series of manipulated conditions to explore the impact of systematically increasing *J* considering four cluster residual distributions. Following Hoaglin and Andrews (1975), the simulation was treated as a statistical sampling experiment subject to established principles of research design and data analysis.

2.1 Model

The random slopes model was chosen because it is commonly used in many research domains (Dedrick et al., 2009). This model includes parameter estimation of main effects on both levels as well as cross-level interactions available by adding predictors to the model. Adding random slopes to a multilevel model would generally increase the minimum *J* required to obtain accurate parameter estimates and valid statistical inferences.

Consider a two-level model for continuous cross-sectional data. Following Raudenbush and Bryk (2002, pp. 100-101), the multilevel (mixed) model can be written as:

$$Y_{ij} = \gamma_{00} + \sum_{p=1}^{S_0} \gamma_{0p} W_{pj} + u_{0j} + \sum_{q=1}^{Q} \left(\gamma_{q0} + \sum_{p=1}^{S_q} \gamma_{qp} W_{pj} + u_{qj} \right) X_{qij} + r_{ij} \,. \qquad (1)$$

In equation (1), Y_{ij} is the outcome of the *i*(th) level-1 unit (i = 1, 2, …, n_j) nested within the *j*(th) level-2 unit or cluster (*j* = 1, 2, …, *J*), γ_{q0} and γ_{qp} are the intercept and linear slope(s) for the *j*(th) cluster of X_{qij}, X_{qij} is a predictor score of the i(th) level-1 unit on the q(th) level-1 predictor nested within the j(th) cluster, r_{ij} is a level-1 residual assumed to follow N(0,σ^2), γ_{00} is a weighted average intercept across the level-2 units, W_{pj} is the value of the p(th) level-2 predictor for the j(th) cluster, γ_{0p} is

the average regression slope across the level-2 units, and u_{0j} and u_{qj} are cluster-level random effects. A key assumption in equation (1) is that $\begin{bmatrix} u_{0j} \\ \vdots \\ u_{qj} \end{bmatrix} \sim MVN(0, \boldsymbol{T})$, where T is a Q x Q covariance matrix of level-2 random effects (Raudenbush & Bryk, 2002, p. 255).

The model examined in this simulation study included one level-1 predictor and one level-2 predictor. Following the naming convention in Equations (1), the model can be defined as:

$$Y_{ij} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j} + \gamma_{10}X_{1ij} + \gamma_{11}W_{1j}X_{1ij} + u_{1j}X_{1ij} + r_{ij} \quad (2)$$

Or for level-1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + r_{ij} \quad (3)$$

Level-2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_{1j} + u_{0j} \quad (4)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}W_{1j} + u_{1j} \quad (5)$$

In equation (4) and (5) γ_{00} ,γ_{10}, γ_{01} and γ_{11} are level-2 coefficients or fixed effects, u_{0j} and u_{1j} are level-2 residuals with variance τ_{00} and τ_{11}, and r_{ij} represents level-1 residuals that are assumed to be normally- distributed variance σ^2. The models we study employ non-Bayesian methods; (fully) Bayesian methods -- like those illustrated in Sorensen, Hohenstein, and Vasishth (2015) -- were not considered.

2.2 Simulation Study Design

2.2.1 Fixed Factors

The γ_{00} parameter was fixed at 1 throughout all simulation conditions, and the values of all other fixed effects were 0.3 because these values were associated with a medium effect size at level-2 (Maas & Hox, 2004a). Level-1 residuals were generated from a normal distribution ($r_{ij} \sim N(0,0.5)$), and level-1 and -2 predictors were generated from a standard normal distribution (e.g., $X_{1ij} \sim N(0,1)$ and $w_{1j} \sim N(0,1)$).

2.2.2 Manipulated Factors

Four factors were manipulated for the simulation using values that were informed by conditions typically present in empirical studies and in previous simulation studies of multilevel models (e.g., Clarke, 2008b; Jitendra et al., 2017; Maas & Hox, 2004b, 2005; Maeda, 2007; McNeish & Stapleton, 2016; Seco et al., 2013; ZOPLUOĞLU, 2012): (1) number of clusters (*J*), (2) cluster sample size (n_j); (3) intra-class correlation (ICC) and (4) distribution of cluster (level-2) residuals. *J*, n_j and ICC values used in the simulation were randomly sampled from a pool of values specified

by the authors using LHS, whereas the level-2 residual distributions were treated as fixed factors, meaning the findings are conditioned upon each specific distribution.

2.2.3 Data Generation

2.2.3.1 Number of clusters, within-cluster sample sizes and Intra-class correlations

To examine the impact of systematically increasing *J*, the current study selected a wide range of values ranging from 5 to 200 with an increment of 5. The range of within-cluster sample size, n_j, was chosen from 5 to 50 with an increment of 5, and the range of ICC values was 0.05 to 0.5 with an increment of 0.05. These three factors formed a pool from which a random sample of values was drawn using LHS and used in the simulation.

LHS initially constructs a design matrix based on the random factors and their specified inputs. The inputs here are *J* = 5, 10, 15, …, 200, n_j = 5, 10, 15, …, 50, and *ICC*=0.05, 0.1, …, 0.5 meaning there are $40 \times 10 \times 10$ = 4,000 design points (design points represent rows of the design matrix, columns are simulation factors). After partitioning the experimental region into 4,000 equal-sized segments, a sample of q = 100 design points was obtained at random from those segments such that one point was sampled at each level of each random factor, where the distribution of values of a random factor is uniform. The subset of 100 design points constitutes the LHD of the study. The choice of q is arbitrary but, as noted above, should be large enough to provide adequate coverage of the experimental region.

2.2.3.2 Distribution of cluster residuals

The accuracy of parameter estimates, along with the validity of statistical inferences about a null hypothesis, were studied for both normal and nonnormal cluster residual distributions. As Blanca et al., Bono and Bendayan (2013) and Bono et al. (2017) found, the most widely reported continuous non-normal distributions in empirical studies belong to the gamma, lognormal and exponential families, with ranges for skewness and kurtosis of -2.94 to 2.33 and -1.92 to 7.41. Based on these findings, and those of Hill and Dixon (1982), Micceri (1989), and Dyer et al. (1999), the current study investigated three non-normal distributions: symmetric but heavy-tailed ($t_{df} = 5$, skewness = 0 and kurtosis = 9), asymmetric and heavy-tailed (exponential, skewness= 2 and kurtosis = 9), and asymmetric and moderately-tailed (gamma(2,6), skewness = 1.41 and kurtosis = 4). Both level-2 residuals (i.e., u_{0j} and u_{1j}) followed the same distribution and the variances (τ_{00} and τ_{11}) were equal. Following Maas and Hox (2005) the ICC was based on the variance of the intercepts with the covariance of u_{0j} and u_{1j} set to zero.

Because cluster residual distributions represent fixed conditions, the LHD is conditioned upon each fixed distribution. Thus, three random factors and one fixed condition produced $100 \times 4 = 400$ cells in the simulation design used to assess the impact of J and level-2 residual distribution. For each cell we simulated

10,000 data sets. Reproducible streams of random numbers were specified in the simulations and all multilevel models were fitted with the R package nlme using REML (Pinheiro et al., 2017).

2.3 Monte Carlo Outcomes

Three measures of the impact of Monte Carlo conditions on parameter estimates were used. One was average relative bias (ARB) defined as

$$ARB = \frac{(E(\hat{\theta}) - \theta))}{\theta} \times 100\% = \frac{\left[\sum_{i=1}^{R} \frac{\hat{\theta_i}}{R} - \theta\right]}{\theta} \times 100\% \quad (5)$$

where $\hat{\theta}$ = estimated parameter and R = number of replications that converged.

The second outcome was the Root Mean Square Error (RMSE) defined as

$$RMSE = \sqrt{\frac{\sum_{i=1}^{R}(\hat{\theta}_i - \theta)^2}{R}} \quad (6)$$

The RMSE was transformed to ln(RMSE) as recommended by Raudenbush (1988). Smaller ARB values reflected more accurate (less biased) estimates, and smaller RMSEs (or more negative ln(RMSEs) reflected more accurate estimates.

Another outcome measure was confidence interval coverage rates of fixed effects and level-2 variance components. Several methods for constructing confidence intervals for variance components are currently available in different software, such as SAS ProcMixed (SAS Institute Inc, 2013) and confint in R (Venables & Ripley, 2002) (for a detailed illustration, see West et al., 2014). As argued by Browne and Draper (2006) and Kraemer (2012), the Satterthwaite method (Satterthwaite, 1946) appears to have the "best possible" performance for variances and was used in the current study. The 95% coverage rate was calculated as the percentage of replications whose 95% Satterthwaite confidence interval contained the specified true value. Comparing coverage rates to 95% provides evidence of relative error, and values above or below 95% were treated as reflecting biased standard errors. Bradley (1978) suggested that values within one-half of the nominal Type-I error rate be considered acceptable. Following this standard, we treated coverage rates less than 92.5 % or greater than 97.5 % as unacceptable.

2.4 Implementation

All simulation procedures described above were conducted in R that is comprised with three parts (annotated accordingly in R syntax in Appendix):

In Part I: *LHD Sampling*. The maximinLHS function in the lhs package (Carnell, 2020) was used to sample 100 design points randomly from an experimental region constructed with three standard uniform distributions. Once sampled, these values were ordered and categorized into groups, and then rescaled to the range of interest mentioned above. For example, in terms of *J*, 100 randomly sampled design points from a standardized uniform distribution were rounded

up to the nearest multiple of 0.025 (i.e., 0.025, 0.05, …, 1), and then rescaled to have values such as 5, 10, …, 200. This process was also performed for the n_j and ICC factors and resulted in each factor having the desired range, i.e., $5 \leq J \leq 200$; $5 \leq n_j \leq 50$ and $0.05 \leq ICC \leq 0.5$. Presenting sampled design points in blocks (e.g., for $5 \leq J \leq 200$ J is a multiple of 5) can enhance generalizability by further decreasing the appearance of adjacent values (e.g., J = 7 and 8). Unlike the full factorial design, the simulated factors are not fully crossed in the LHD, which means that values of n_j and *ICC* given each value of *J* are different. For example, n_j= 5 and ICC = .40 when *J*= 170 but n_j= 40 and ICC = .10 when *J*=65 (a full list of *J*, n_j and *ICC* values is provided in the Appendix).

In Part II: *Data Generation*. For each simulation replication, the level-1 and -2 residuals and predictors were generated first using the sample size generated from LHS. Then, those values were plugged into Equation (2) along with the pre-specified fixed effect coefficients, to calculate the outcome variable, Y_{ij}.

In Part III: *Model Analysis*. The generated data from Part II was fitted with the R package nlme (Pinheiro et al., 2017). An indicator of model convergence was dichotomously coded as converged/non-converged. The Monte Carlo outcomes were calculated only when the model converged. ARB and RMSE were calculated using Equation (5) and (6). The confidence interval coverage rate was calculated as the percentage of replications whose confidence interval contained the true value.

3. Results

Model convergence rates were investigated first because in empirical research, a converged model is always a prerequisite for the subsequent analysis. In the current study, nonconvergence was only found when $J \leq 30$ and was not viewed as problematic as 92.51% of all replications studied converged. Busing (1993) and Shieh et al. (2001) discuss common reasons for nonconvergence. When the model does not converge to stable estimates, the fixed and random effect estimates are not available and the R code produces the error message “Model failed to converge” and reports no results. The conditions that produced nonconvergence and the percentage of nonconvergence are displayed in Table 1. The replications that failed to converge were discarded and the simulation outcomes were calculated with converged replications only.

Table 1. Percentage of nonconvergence

J	n_j	ICC	Normal	t	Exponential	Gamma
5	10	0.05	27.54%	27.07%	26.43%	26.26%
10	5	0.15	15.23%	16.84%	17.07%	16.36%
10	30	0.05	0.41%	0.76%	0.95%	0.73%
5	45	0.20	0.17%	0.13%	0.34%	0.15%
10	15	0.40	0.02%	0.12%	0.15%	0.07%
30	10	0.15	0.02%	0.01%	0.04%	0.05%

3.1 Graphical illustration

Simulation outcomes were plotted against each random simulation factor while averaging outcomes across the other two random factors. For example, to plot the ARBs against *J*, all ARBs at the same value of *J* but with different values of n_j and ICC, were averaged and appear as one point on that plot. Thus, a plot of an outcome variable against a simulation factor may be influenced by the other two factors to different degrees, and produce a fluctuating pattern graphically rather than a consistent pattern. For example, a plot of ARB versus *J* may fluctuate as *J* increases because of the influence of n_j and the ICC. Put another way, the influence of the factor plotted on the x-axis on an outcome variable may be moderated by other factors. All figures appear in the appendix.

Figure 2 suggests that fixed effects estimates were generally accurate with average relative biases close to 0, although the intercept fixed effects (γ_{00}) had smaller ARBs than the slope fixed effects. When the number of clusters increased, fixed effects' ARBs approached zero, but no clear trend was observed between fixed effect's ARBs and n_j or ICC and both presented obvious fluctuations. Given the same value of any manipulated factor, the difference across level-2 residual distributions was negligible.

In terms of the variance components, Figure 3 shows that when $J<15$, the variance components were significantly overestimated (i.e., larger than 0.30 for τ_{00} and 0.40 for τ_{11} when $J=5$) for all distribution types, but the ARBs quickly approached 0 as *J* increased. Similar patterns were also seen in Darandari (2004) and Shih (2008). When n_j was plotted on the x-axis, positive ARBs were observed when $n_j = 5$ and 10, and the jump at $n_j = 10$ suggests that the estimates of variance components were also influenced by *J*, given that the average *J* was 27.5 when $n_j = 5$ but 16.7 when $n_j = 10$. The right column of Figure 3 shows that the largest positive bias was observed when ICC=0.1 and moved back to around 0, except when ICC=0.15. Besides the jumps, both Figures 2 and 3 showed that variance components' ARBs decreased as cluster sizes and ICC increased but did not show any notable differences between level-2 residual distributions.

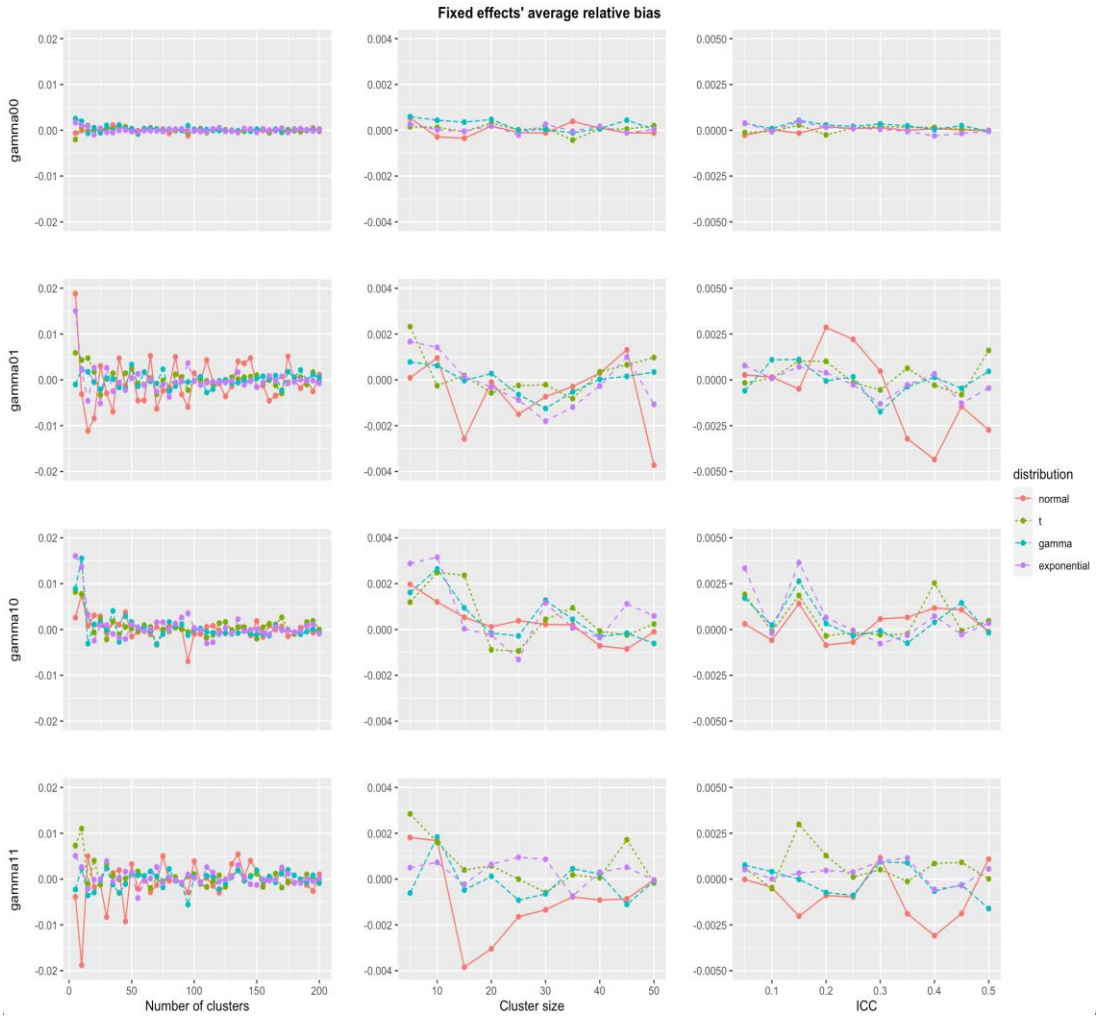

Figure 2. Impact of number of clusters, Cluster sizes and ICC on fixed effects estimates' ARB

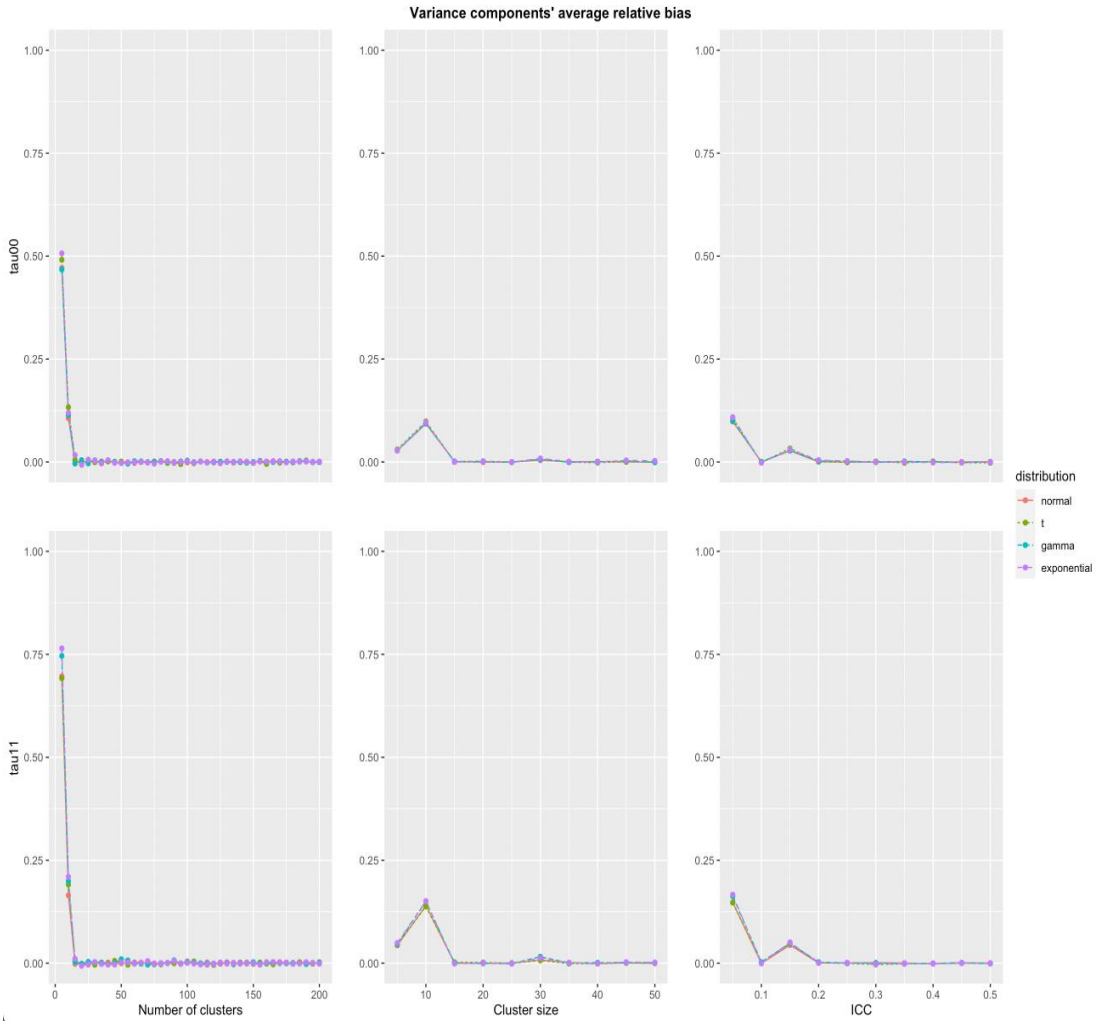

Figure 3. Impact of number of clusters, Cluster sizes and ICC on variance components estimates' ARB

Figure 4 shows that the ln(RMSE)s of fixed effect estimates decreased as *J* increased, with a decreasing pattern when cluster size was plotted on the x-axis. These results suggest that estimation accuracy is influenced by *J* to a larger extent than n_j, which is consistent with the results of Kraemer (2012), Maas and Hox (2004a), and McNeish and Stapleton (2016) On the contrary, the ln(RMSE)s became less negative when ICC increased, meaning the variance components estimation became less accurate. Similar results were seen in Clarke and Wheaton (2007). In addition, fixed effect estimates' ln(RMSE)s were quite similar across different level-2 residuals distributions.

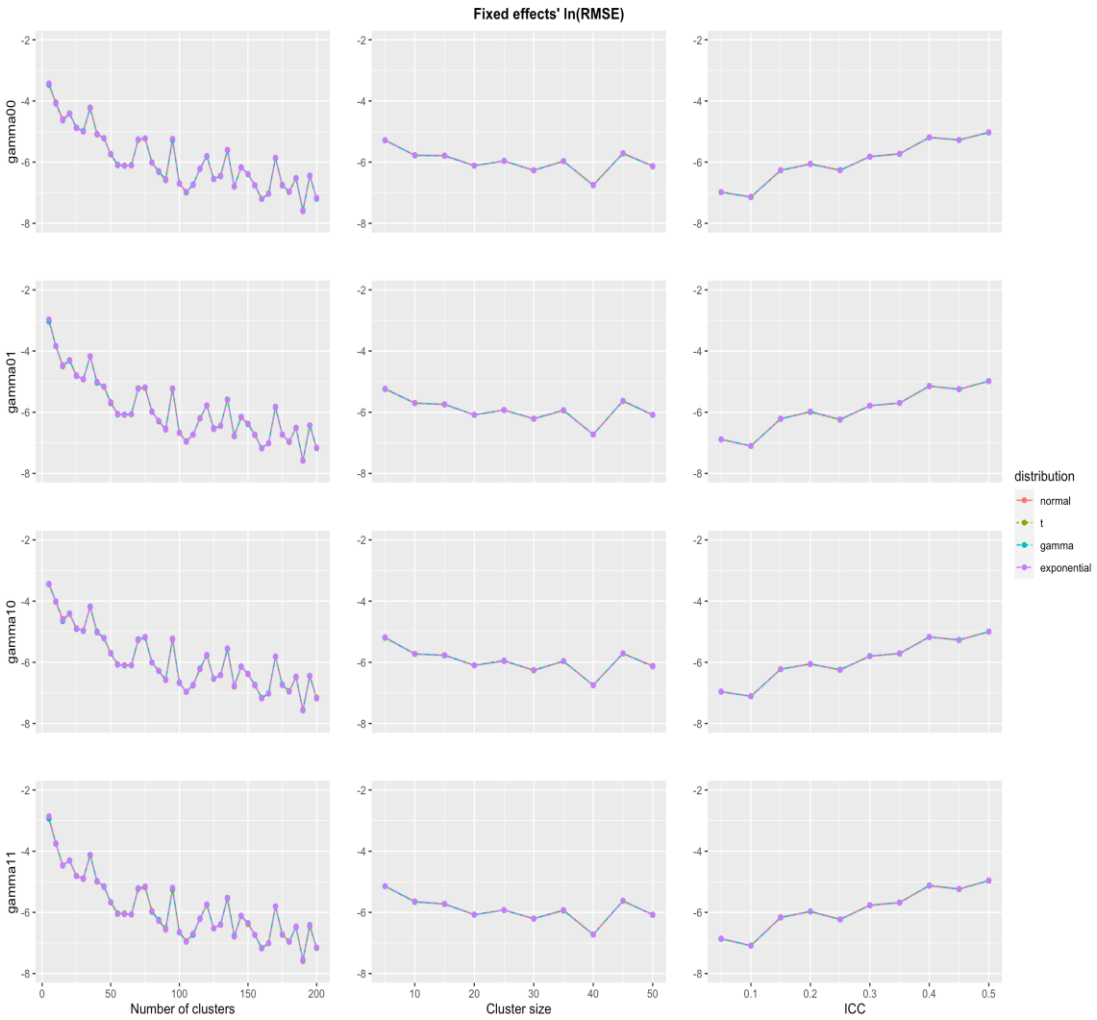

Figure 4. Impact of number of clusters, Cluster sizes and ICC on fixed effects estimates' ln(RMSE)

Similar to the fixed effects, the variance components' ln(RMSE)s decreased as J or n_j increased, but increased as ICC became bigger (see Figures 5). Different from the fixed effects, however, ln(RMSE)s followed a descending order in that normal < gamma < t < exponential.

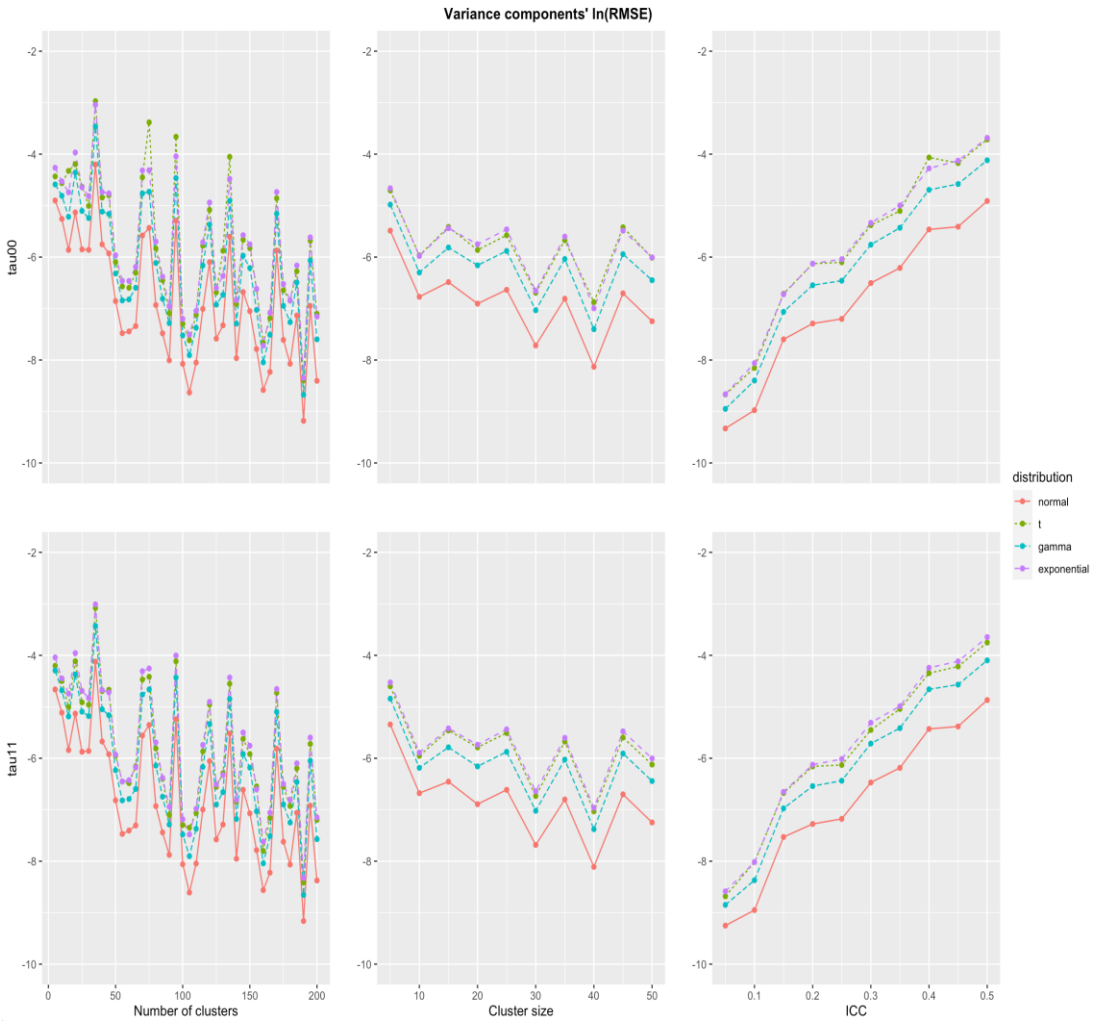

Figure 5. Impact of number of clusters, Cluster sizes and ICC on variance components estimates' ln(RMSE)

Regarding the 95% confidence interval coverage rates, all fixed effects except γ_{01} had coverage rates less than 95% when J was less than 50 (Figure 6). Beyond that, all coverage rates were within the acceptable range. Plotting coverage rates against n_j or ICC does not show a clear pattern, and fluctuations suggest that coverage rates were influenced by other factors. In addition, for the results of the intercept fixed effects, the symmetric distributions (i.e., normal and t distribution) produced coverage rates that were closer to 0.95 than the asymmetric distributions, with the exponential distribution associated with the poorest coverage.

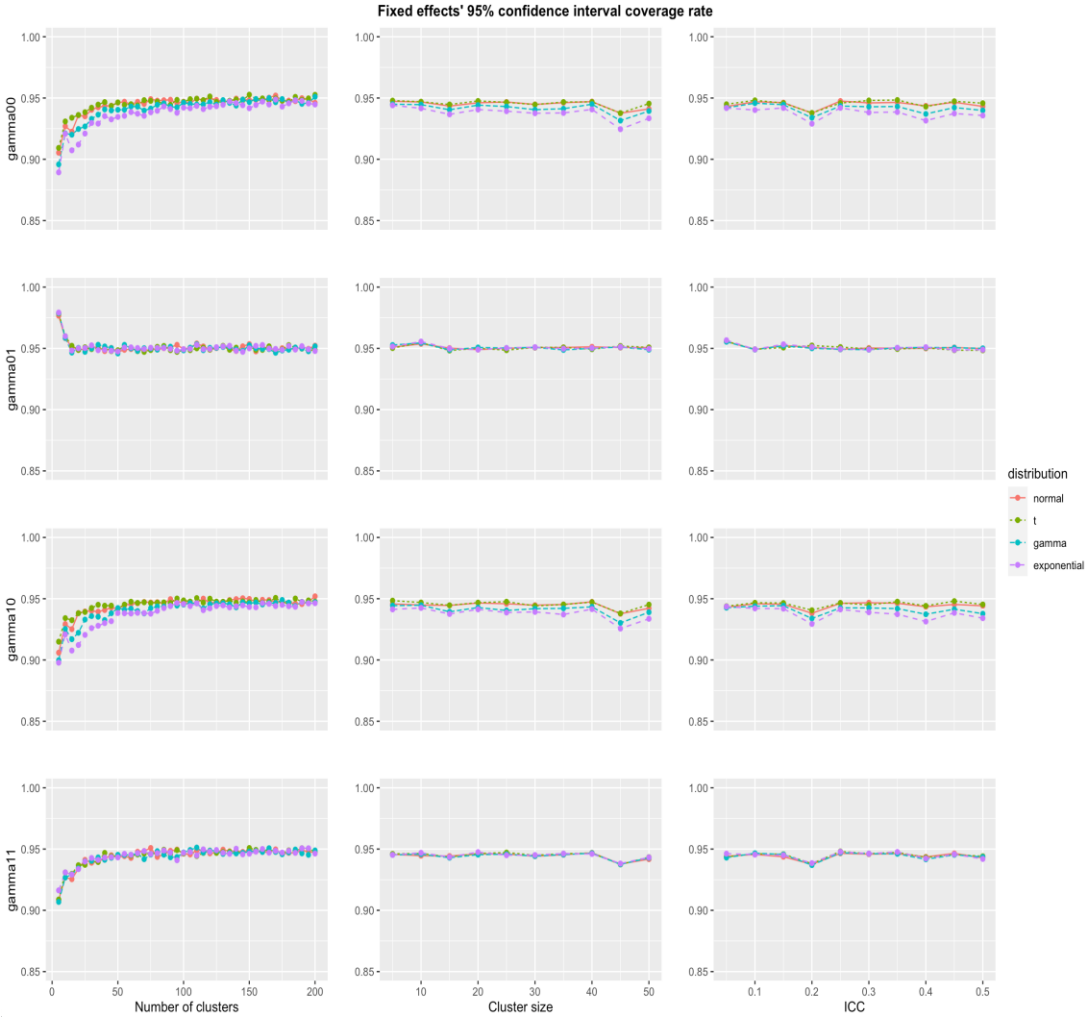

Figure 6. Impact of number of clusters, Cluster sizes and ICC on fixed effects estimates' 95% confidence interval coverage rate

---

Figure 7 shows that coverage rates of variance components for all distributions increased from J=5 to 10, but after that, only the normal distribution had rates near 0.95. The non-normal distributions were associated with coverage rates smaller than 0.90, which became even smaller as J increased, a finding that consistent with Kraemer (2012) and Burch (2011). When n_j was plotted on the x-axis, variance components' coverage rates were close to 0.95 if level-2 residuals were normally distributed. The t and gamma distributions were associated with similar coverage rates between 0.82 and 0.87, but the exponential distribution had coverage rates of less than 0.80 for most n_js. In addition, all non-normal distributions' coverage rates were noticeably influenced by other simulated conditions, compared to a normal

distribution. When level-2 residuals followed a normal distribution, coverage rates increased as ICC increased and then maintained around 0.95. However, when level-2 residuals were non-normally distributed, variance components' coverage rates decreased as ICC increased. Again, the exponential distributed residuals were associated with the poorest coverage rates.

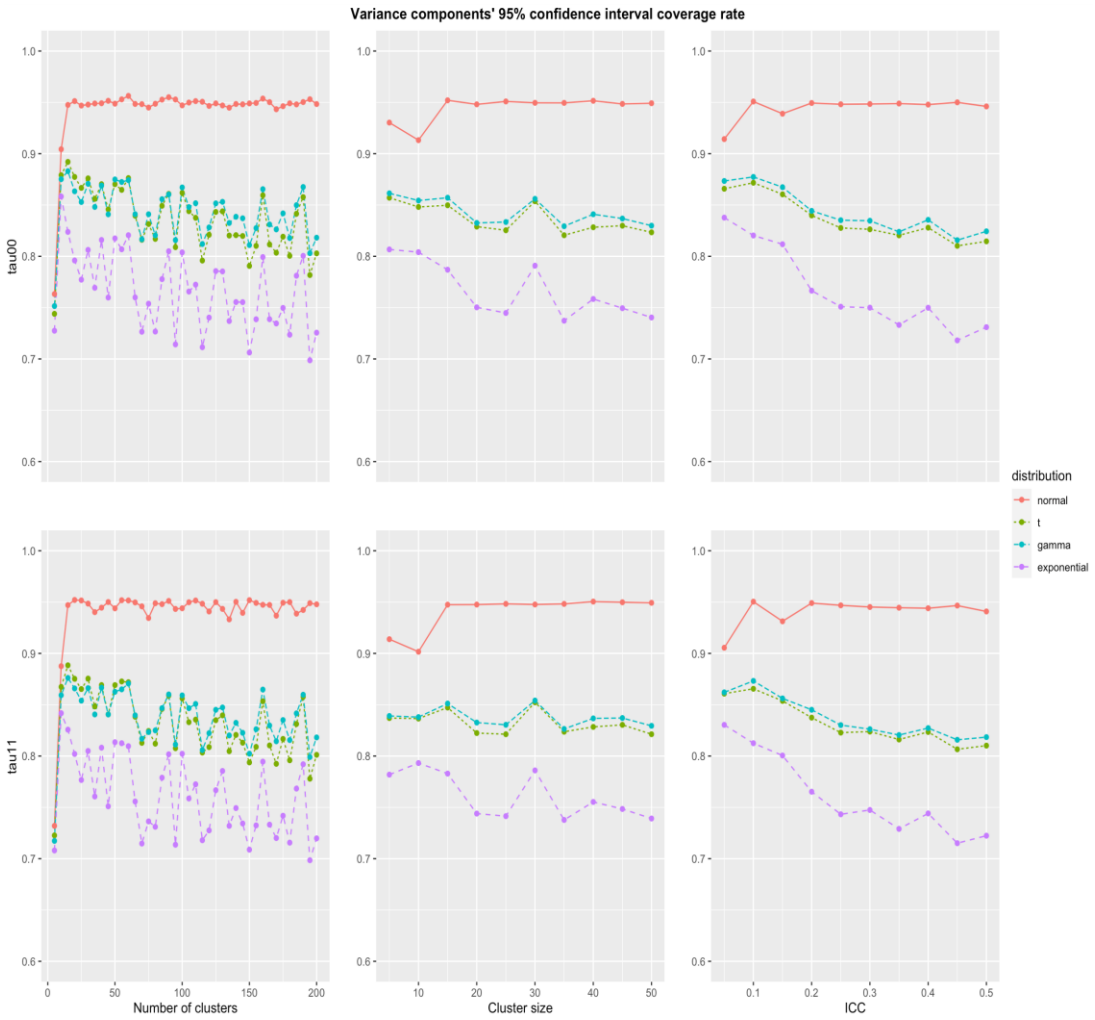

Figure 7. Impact of number of clusters, Cluster sizes and ICC on variance components estimates' 95% confidence interval coverage rate

4. Conclusion

A Latin Hypercube design (LHD) was used in a Monte Carlo simulation study to investigate the impact of number of clusters and non-normally distributed cluster residuals in multilevel modeling. The choice of a LHD responds to the need to

increase the generalizability of simulation results, and is accomplished by uniformly populating the range of values of each manipulated factor independently, resulting in greater generalizability. The ability of LHDs to enhance generalizability of simulation results, and the ease with which they can be implemented with existing software, make these designs attractive for simulation studies.

The current study used an LHD approach with inputs for J, n_j, and ICC that produced 4,000 design points. After partitioning the experimental region into 4,000 equal-sized segments, a sample of q = 100 design points was obtained at random from those segments such that one point was sampled at each level of each random factor, where the distribution of values of a random factor was uniform. This sampling strategy can enhance the generalizability of simulation results (e.g., across a wide range of J, n_j, and ICC values) compared to traditional fixed factorial designs typically used in simulation studies. Of course, LHDs should not necessarily be used for all simulation studies because no single experimental design is going to be suitable for all simulations (Santner et al., 2018; Viana, 2015). However, methodological researchers conducting simulation studies who place a premium on generalizing their findings to enhance their impact on statistical practice should find LHDs attractive.

An important limitation of LHDs is related to the number of random factors and the number of values of these factors specified by a researcher. Even though advancements in computing and optimization support the implementation of LHS for studies with multiple random factors, as the number of random factors in the simulation design increases, the optimization of the LHD can become cumbersome (Viana, 2015). This occurs because LHS solves a combinatorial problem to select the points that conform to the LHD and assure uniformity such that all portions of each manipulated factor´s range of values is represented in the design (McKay et al., 2000; Viana, 2015). Increasing the number of manipulated factors typically makes it harder to solve the combinatorial problem, and even if the combinatorial problem can be solved, the property of having design points evenly spread over the experimental region is harder to achieve as the number of random factors involved increases (Viana, 2015). In addition, no a priori dependency between factors is assumed, helping to ensure each factor is well-represented in the experimental design and that simulation results can be analyzed independently for each factor. Methods that account for dependencies in the simulation factors have been developed (Stein, 1987; Owen, 1994), but this topic is beyond the scope of the current study.

The results of the current simulation study produced two key findings: (1) Bias is not a concern when estimating fixed effects regardless of J and cluster residual distribution, and is not a concern when estimating variance components for $J \geq 20$ regardless of level-2 residual distribution, (2) Coverage rates for fixed effects were substantially different from 0.95 for J<30 especially for the exponential level-2 residual distribution, but converged quickly to 95% for $J \geq 30$; coverage rates for 95% confidence intervals for variance components were significantly underestimated

when J<15 for normally-distributed cluster residuals and converged to 95% once J≥15. But for non-normal level-2 residual distributions coverage rates were uniformly underestimated and increasing J increased the underestimation. These results are summarized in Table 2 and suggest that for a random slopes model with normally-distributed residuals the minimum J is 30, and the presence of any of the non-normal level-2 distributions in the current study signal that variance components are unlikely to be accurately estimated even for large J. Importantly, the summary of findings and recommendations in Table 2 apply to the entire pool of J values specified earlier.

Table 2. Summary of Findings and Recommendations for Minimum J for the Random Slope Model

Distribution of Level-2 Residuals	Parameter	Minimum Recommended J
Normal	Fixed effects ($\gamma_{00}, \gamma_{01}, \gamma_{10}$, and γ_{10})	15
	Variance components (τ_{00} and τ_{11})	30
$t_{df} = 5$	Fixed effects	15
	Variance components	---
gamma(2,6)	Fixed effects	25
	Variance components	---
exponential	Fixed effects	30
	Variance components	---

It is important to note that our results provide partial support for the 30/30 rule of thumb advocated by Kreft (1996) for normally-distributed model residuals. For the non-normal cluster residual distributions we studied, our findings reveal that parameter estimation accuracy can decrease with larger sample sizes, which contradicts the commonly held belief that increasing sample size will alleviate the effects of violated assumptions. Our findings speak to the importance of model checking to ensure normality at level-2 is plausible (along with other model assumptions).

References

Auda, H. A., McKean, J. W., Kloke, J. D., & Sadek, M. (2019). A Monte Carlo study of REML and robust rank-based analyses for the random intercept mixed model. Communications in Statistics-Simulation and Computation, 48(3), Article 3.

Bell, B. A., Ferron, J. M., & Kromrey, J. D. (2008). Cluster Size in Multilevel Models: The Impact of Sparse Data Structures on Point and Interval Estimates in Two-Level Models. 8.

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. Methodology.

Bono, R., Blanca, M. J., Arnau, J., & Gómez-Benito, J. (2017). Non-normal distributions commonly used in health, education, and social sciences: A systematic review. Frontiers in Psychology, 8, 1602.

Bradley, J. V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31(2), Article 2. https://doi.org/10.1111/j.2044-8317.1978.tb00581.x

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. Bayesian Analysis, 1(3), 473–514. https://doi.org/10.1214/06-BA117

Burch, B. D. (2011). Assessing the performance of normal-based and REML-based confidence intervals for the intraclass correlation coefficient. Computational Statistics & Data Analysis, 55(2), 1018–1028.

Busing, F. M. T. A. (1993). DISTRIBUTION CHARACTERISTICS OF VARIANCE ESTIMATES IN TWO-LEVEL MODELS: A Monte Carlo study. https://doi.org/10.13140/RG.2.2.18116.94088

Carnell, R. (2020). Basic Latin hypercube samples and designs with package lhs. https://cran.r-project.org/web/packages/lhs/vignettes/lhs_basics.html

Clarke, P. (2008a). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. Journal of Epidemiology & Community Health, 62(8), Article 8. https://doi.org/10.1136/jech.2007.060798

Clarke, P. (2008b). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. Journal of Epidemiology & Community Health, 62(8), Article 8. https://doi.org/10.1136/jech.2007.060798

Clarke, P., & Wheaton, B. (2007). Addressing Data Sparseness in Contextual Population Research: Using Cluster Analysis to Create Synthetic Neighborhoods. Sociological Methods & Research, 35(3), Article 3. https://doi.org/10.1177/0049124106292362

Darandari, E. Z. (2004). Robustness of Hierarchical Linear Model Parameter Estimates Under Violations of Second-Level Residual Homoskedasticity and Independence Assumptions. 232.

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., Niles, J. D., & Lee, R. S. (2009). Multilevel Modeling: A Review of Methodological Issues and Applications. Review of Educational Research, 79(1), Article 1. https://doi.org/10.3102/0034654308325581

DiPrete, T. A., & Forristal, J. D. (1994). Multilevel models: Methods and substance. Annual Review of Sociology, 331–357.

Dyer, J. R., Pilcher, C. D., Shepard, R., Schock, J., Eron, J. J., & Fiscus, S. A. (1999). Comparison of NucliSens and Roche Monitor assays for quantitation of levels of human immunodeficiency virus type 1 RNA in plasma. Journal of Clinical Microbiology, 37(2), Article 2.

Epstein, J. L., Galindo, C. L., & Sheldon, S. B. (2011). Levels of leadership: Effects of district and school leaders on the quality of school programs of family and community involvement. Educational Administration Quarterly, 47(3), Article 3.

Harwell, M., Kohli, N., & Peralta, Y. (2017). Experimental design and data analysis in computer simulation studies in the behavioral sciences. Journal of Modern Applied Statistical Methods, 16(2), 2.

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clinical laboratory data. Biometrics, 377–396.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. The American Statistician, 29(3), Article 3.

Hox. (2002). Quantitative methodology series. Multilevel Analysis Techniques and Applications. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Hox, J. J. (1995). Applied multilevel analysis. TT-publikaties.

Iman, R. L., & Conover, W. J. (1980). Small sample sensitivity analysis techniques for computer models. With an application to risk assessment. Communications in Statistics-Theory and Methods, 9(17), 1749–1842.

Jitendra, A. K., Harwell, M. R., Dupuis, D. N., & Karl, S. R. (2017). A randomized trial of the effects of schema-based instruction on proportional problem-solving for students with mathematics problem-solving difficulties. Journal of Learning Disabilities, 50(3), Article 3.

Korendijk, E. J., Maas, C. J., Moerbeek, M., & Van der Heijden, P. G. (2008). The influence of misspecification of the heteroscedasticity on multilevel regression

parameter and standard error estimates. Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 4(2), Article 2.

Kraemer, K. (2012). Confidence intervals for variance components and functions of variance components in the random effects model under non-normality.

Kreft, I. G. (1996). Are multilevel techniques necessary. An Overview, Including Simulation Studies.

Lee, V. E. (2000). Using hierarchical linear modeling to study social contexts: The case of school effects. Educational Psychologist, 35(2), 125–141.

Maas, C. J. M., & Hox, J. J. (2004a). Robustness issues in multilevel regression analysis. Statistica Neerlandica, 58(2), Article 2. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2004b). Robustness issues in multilevel regression analysis. Statistica Neerlandica, 58(2), Article 2. https://doi.org/10.1046/j.0039-0402.2003.00252.x

Maas, C. J. M., & Hox, J. J. (2004c). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Computational Statistics & Data Analysis, 46(3), Article 3. https://doi.org/10.1016/j.csda.2003.08.006

Maas, C. J. M., & Hox, J. J. (2005). Sufficient Sample Sizes for Multilevel Modeling. 1, 7.

Maeda, Y. (2007). Monte Carlo Evidence Regarding the Effects of Violationg Assumed Conditions of Two-Level Hierarchical Models for Cross-Sectional Data [Dissertation]. University of Minnesota.

Marks, H. M., & Printy, S. M. (2003). Principal leadership and school performance: An integration of transformational and instructional leadership. Educational Administration Quarterly, 39(3), Article 3.

McCulloch, C. E., & Neuhaus, J. M. (2011a). Misspecifying the shape of a random effects distribution: Why getting it wrong may not matter. Statistical Science, 26(3), 388–402.

McCulloch, C. E., & Neuhaus, J. M. (2011b). Prediction of random effects in linear and generalized linear models under model misspecification. Biometrics, 67(1), 270–279.

McKay, M. D., Beckman, R. J., & Conover, W. J. (2000). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics, 42(1), 55–61.

McNeish, D. M., & Stapleton, L. M. (2014). The effect of small sample size on two-level model estimates: A review and illustration. Educational Psychology Review, 28(2), Article 2.

McNeish, D., & Stapleton, L. M. (2016). Modeling Clustered Data with Very Few Clusters. Multivariate Behavioral Research, 51(4), 495–518. https://doi.org/10.1080/00273171.2016.1167008

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105(1), Article 1.

Owen, A. B. (1994). Controlling correlations in Latin hypercube samples. Journal of the American Statistical Association, 89(428), 1517–1522.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., Heisterkamp, S., Van Willigen, B., & Maintainer, R. (2017). Package 'nlme.' Linear and Nonlinear Mixed Effects Models, Version, 3(1).

R Core Team. (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Raudenbush, S. W. (1988). Estimating change in dispersion. Journal of Educational Statistics, 13(2), 148–171.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (Vol. 1). sage.

Rose, K., Hawes, D. J., & Hunt, C. J. (2014). Randomized controlled trial of a friendship skills intervention on adolescent depressive symptoms. Journal of Consulting and Clinical Psychology, 82(3), Article 3.

Santner, T. J., Williams, B. J., & Notz, W. I. (2018). Space-filling designs for computer experiments. In The design and analysis of computer experiments (pp. 145–200). Springer.

Santner, T. J., Williams, B. J., Notz, W. I., & Williams, B. J. (2003). The design and analysis of computer experiments (Vol. 1). Springer.

SAS Institute Inc. (2013). SAS® 9.4 Statements: Reference.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. Biometrics Bulletin, 2(6), 110–114.

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. Methods in Ecology and Evolution, 11(9), 1141–1152.

Seco, G. V., García, M. A., García, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. Psicothema, 25(4), Article 4.

Shernoff, D. J., Ruzek, E. A., & Sinha, S. (2017). The influence of the high school classroom environment on learning as mediated by student engagement. School Psychology International, 38(2), Article 2.

Shieh, Y. Y., Fouladi, R. T., & Pullum, T. W. (2001). The effect of error term non-normality on multilevel model parameter estimates and standard errors: A focus on estimation bias. Multiple Linear Regression Viewpoints, 27(1), Article 1.

Shih, T.-H. (2008). Adequate sample sizes for viable 2-level hierarchical linear modeling analysis: A study on sample size requirement in HLM in relation to different intraclass correlations [University of Virginia]. https://doi.org/10.18130/V3J63G

Sorensen, T., & Vasishth, S. (2015). Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. ArXiv Preprint ArXiv:1506.06201.

Stegmueller, D. (2013). How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches: HOW MANY COUNTRIES? American Journal of Political Science, 57(3), Article 3. https://doi.org/10.1111/ajps.12001

Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. Technometrics, 29(2), 143–151.

Vallejo, G., Fernández, P., Cuesta, M., & Livacic-Rojas, P. E. (2015). Effects of modeling the heterogeneity on inferences drawn from multilevel designs. Multivariate Behavioral Research, 50(1), Article 1.

Van der Leeden, R., & Busing, F. M. T. A. (1994). First iteration versus IGLS/RIGLS estimation in two-level models: A Monte Carlo study with ML3. Preprint PRM, 94(03), Article 03.

Venables, W. N., & Ripley, B. D. (2002). Random and mixed effects. In Modern applied statistics with S (pp. 271–300). Springer.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. Journal of the American Statistical Association, 91(433), 217–221.

Viana, F. A. (2016). A tutorial on Latin hypercube design of experiments. Quality and Reliability Engineering International, 32(5), 1975–1985.

Wang, G. G. (2003). Adaptive response surface method using inherited latin hypercube design points. J. Mech. Des., 125(2), 210–220.

West, B. T., Welch, K. B., & Galecki, A. T. (2014). Linear Mixed Models: A Practical Guide Using Statistical Software, Second Edition (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b17198

ZOPLUOĞLU, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 3(1), 242–278.