COMPARING THREE APPROACHES FOR HANDLING A FOURTH LEVEL OF NESTING STRUCTURE IN CLUSTER-RANDOMIZED TRIALS

Ryan Glaman

Dissertation Prepared for the Degree of:

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

August 2017

APPROVED:

Robin K. Henson, Major Professor
Miriam C. Boesch, Committee Member
Colleen M. Eddy, Committee Member
Darrell M. Hull, Committee Member
Abbas Tashakkori, Chair of the Department of Educational Psychology
Randy Bomer, Dean of the College of Education
Victor R. Prybutok, Dean of the Toulouse Graduate School

Glaman, Ryan. *Comparing Three Approaches for Handling a Fourth Level of Nesting Structure in Cluster-Randomized Trials*. Doctor of Philosophy (Educational Psychology), August 2017, 41 pp., 5 tables, references. 45 titles.

This study compared 3 approaches for handling a fourth level of nesting structure when analyzing data from a cluster-randomized trial (CRT). CRTs can include 3 levels of nesting: repeated measures, individual, and cluster levels. However, above the cluster level, there may sometimes be an additional potentially important fourth level of nesting (e.g., schools, districts, etc., depending on the design) that is typically ignored in CRT data analysis. The current study examined the impact of ignoring this fourth level, accounting for it using a model-based approach, and accounting it using a design-based approach on parameter and standard error (SE) estimates. Several fixed effect and random effect variance parameters and SEs were biased across all 3 models. In the 4-level model, most SE biases decreased as the number of level 3 clusters increased and as the number of level 4 clusters decreased. Also, random effect variance biases decreased as the number of level 3 clusters increased. In the 3-level and complex models, SEs became more biased as the weight level 4 carried increased (i.e., larger intraclass correlation, more clusters at that level). The current results suggest that if a meaningful fourth level of nesting exists, future researchers should account for it using design-based approach; the model-based approach is not recommended. If the fourth level is not practically important, researchers may ignore it altogether.

ACKNOWLEDGEMENTS

I want to thank the members of my dissertation committee, my family, friends, and everybody else who has helped and supported me along the way. You have my sincerest gratitude.

TABLE OF CONTENTS

LIST OF TABLES

Page

COMPARING THREE APPROACHES FOR HANDLING A FOURTH LEVEL OF NESTING STRUCTURE IN CLUSTER-RANDOMIZED TRIALS

Introduction

Cluster-randomized trials (CRTs) are a category of experimental designs that represent the gold standard for research investigating large-scale interventions (Campbell, Mollison, Steen, Grimshaw, & Eccles, 2000). They can be used when it may be inappropriate or impossible to employ a traditional experiment, such as in the cases of classroom-based education interventions or studies of group counseling techniques (Sim & Dawson, 2012). Whereas in a traditional experiment, individuals are randomized into experimental conditions, in a CRT, clusters of individuals are randomized into conditions. For example, in an education intervention, entire classrooms of students may be randomized into the treatment or control groups, as opposed to the individual students being randomly assigned to different groups within the same classroom. Researchers administer treatment at the cluster level for a variety of reasons, such as increased administrative efficiency, decreased risk of experimental contamination, and enhanced participant compliance (Donner & Klar, 2004). CRTs are also useful when an intervention is designed to be administered to and affect entire clusters of participants (Edwards, Braunholtz, Lilford, & Stevens, 1999), as is often the case with education interventions.

One implication of CRT studies is that participants within a given cluster are more likely to respond similarly to the intervention, and can no longer be assumed to be independent of one another (Campbell et al., 2000; Jo, Asparouhov, Muthén, Ialongo, & Brown, 2008). This implication violates the independence of observations assumption underlying many single-level data analysis techniques such as multiple regression (Hox, 2010). Therefore, researchers should use statistical methods designed to account for the nested data structure and lack of independence of observations, such as hierarchical linear modeling (HLM) or latent growth curve modeling (LGCM).

Although CRT researchers frequently employ HLM to correctly reflect the multilevel nature of the data, they also often ignore potentially important higher levels of nesting above the level at which randomization occurs. For example, Russell, O'Dwyer, and Miranda (2009) investigated the impact of a diagnostic assessment system on students' misconceptions toward algebra. In their study, students were nested within teachers, and teachers were randomized into one of four intervention groups; the researchers treated the analysis as a 2-level HLM. However, teachers were also nested within schools, a potential third level above the level of randomization. When ignoring this level of nesting, the results of a study may have been adversely impacted in some way, such as by producing biased standard errors (*SE*s) or inappropriately distributing variance across levels (Hox, 2010; Moerbeek, 2004). Whereas some researchers disregarded higher levels of nesting structure for justifiable reasons, such as having a small number of clusters at that level (e.g., Al Otaiba, Connor, Folsom, Greulich, & Meadows, 2011; Hegedus, Dalton, & Tapper, 2015), other researchers did not address why they may have ignored higher levels of nesting (e.g., Russell et al., 2009). The purpose of the current study was to investigate the impact of ignoring a higher level of nesting structure on the analysis of CRT data as well as compare various methods of accounting for that nested data structure.

Hierarchical Data Structure

Hierarchical or nested data structure is common in social research (Raudenbush & Bryk, 2002), particularly in education where students are nested in classrooms, which are nested in schools, and so forth. Generally speaking, there are two overall approaches to handling multilevel data. The first is a model-based approach in which researchers specify separate models for each level of the multilevel data, thus modeling non-independence due to cluster sampling; this includes analytic techniques such as HLM and LGCM (Muthén & Muthén, 2012; Wu & Kwok, 2012).

The second method is a design-based approach that models non-independence by adjusting parameter estimate *SE*s based on the cluster sampling design. Specifically, the design-based approach uses sampling weights to account for clustering; these sampling weights are used to compute adjusted *SE*s for a single-level model (Muthén & Muthén, 2012; Muthén & Satorra, 1995). In the case of a 2-level analysis (e.g., students nested within classrooms), the result would be a single-level model that features adjusted *SE*s to reflect the non-independence of observations instead of a traditional 2-level model (Wu & Kwok, 2012). This design-based procedure represents a compromise between fully considering additional levels using multilevel modeling and ignoring those levels.

One important area of multilevel modeling research involves investigating the effects of ignoring levels of nesting. Broadly, ignoring nesting altogether tends to result in underestimated *SE*s (Hox, 2010), but it can also impact other facets of data analysis because many analytic techniques can involve multilevel data.

Research on the Impact of Ignoring Nested Data Structure using a Model-Based Approach

Ignoring a level of nesting in 3-level models causes estimated variance attributed to the ignored level to be distributed to adjacent levels (Moerbeek, 2004). Furthermore, when the ignored level features a predictor variable, the *SE*s of the predictors at the ignored level and the level below it may become biased (Van den Noortgate, Opdenakker, & Onghena, 2005); specifically, *SE*s of predictors at the ignored level may become negatively biased and *SE*s of predictors at the below level may become positively biased. Wampold and Serlin (2000) observed that treatment effect sizes were greatly overestimated when multilevel data structure was ignored in multilevel analysis of variance (ANOVA). In the context of multilevel confirmatory factor analysis, failing to model nesting structure can decrease overall model fit and the accuracy of estimated standardized parameters and *SE*s (Pornprasertmanit, Lee, & Preacher, 2014). In growth mixture modeling, ignoring a higher level of nesting structure can cause lower classification accuracy, overestimated lower-level variance estimates, and biased *SE*s (Chen, Kwok, Luo, & Willson, 2010).

Research has also investigated the effects of incorrectly modeling cross-classified data structures. Cross-classified data structures occur when individual units are not purely nested within two or more cluster levels (Beretvas, 2011). For example, students (a Level 1 unit) may be nested within both schools and neighborhoods (both Level 2 units). Findings from Luo and Kwok (2009; 2012) and Meyers and Beretvas (2006) mirrored those of ignoring nested data structures altogether: incorrectly modeling cross-classified data structures yielded underestimated *SE* estimates for the predictor variables.

Research on Design-Based Approaches

There are also design-based approaches that use statistical adjustments to account for nested data structure. Asparouhov (2005) found that design-based methods of accounting for multilevel data are effective for reducing parameter bias in multilevel confirmatory factor analysis. Additional simulation research on multilevel confirmatory factor analysis suggested that design-based approaches are capable of producing *SE* estimates that are as accurate as the model-based approaches' estimates when Level 1 and Level 2 feature the same underlying factor structure (Muthén & Satorra, 1995; Wu & Kwok, 2012). Together, these findings suggest that design-based methods of handling nesting structure can model data as accurately as model-based approaches, at least in multilevel confirmatory factor analysis.

In sum, the above findings suggest that failing to properly model nested data structure using either a model-based or design-based approach can negatively impact one's results in a variety of ways. The general conclusion that *SE*s become underestimated when levels of nesting are ignored has substantial practical implications. When *SE*s are underestimated, the probability of making a Type-I error increases. In the context of educational intervention studies such as CRTs, this may cause researchers to conclude that a particular intervention is effective when it really is not. This, in turn, could cause practitioners or administrators to adopt programs and policies that are ineffective.

In methodological research, CRTs have received attention regarding various issues such as power (Spybrook, Kelcey, & Dong, 2016) and effect sizes (Ames, 2013), but the potential impact of a higher level of nesting structure has not yet been explored. The following section discusses the importance of CRTs in empirical studies as well as some concerns regarding including the appropriate number of levels in CRT data analysis. Given the research findings on both the model-based and design-based approaches to account for nested data structure, it may be relevant to examine both types of approaches in the context of CRT data.

Cluster-Randomized Trials

CRTs are experimental studies in which clusters of individuals, rather than individuals themselves, are randomized into experimental conditions (Donner & Klar, 2004). Examples of clusters include classrooms, schools, hospitals, families, neighborhoods, and so forth. CRTs are useful research designs because they retain the random nature of randomized controlled trials, but can be used in cases in which randomized controlled trials may be impractical due to the inability to randomize individuals directly (Sim & Dawson, 2012). As previously mentioned, researchers may also choose to administer treatment at the cluster level for several reasons such as increased administrative efficiency (Donner & Klar, 2004) or when interventions are intended to be administered at the cluster level (Edwards et al., 1999), as is the case for classroom-based interventions in education. For example, Sarama, Clements, Wolfe, and Spitler (2012) investigated the effects of a technology-based mathematics program in elementary mathematics education. In their study, students were nested within schools, and the schools were randomly assigned to one of three experimental groups. Because students typically receive instruction in a group format, it is more practical and efficient to test education interventions using an experiment that randomizes participants at the cluster or group level. For this reason, the CRT is a desirable approach to test interventions on a large scale (e.g., Abe & Gee, 2011; Kim et al., 2011; Rose, Hawes, & Hunt, 2014).

In CRTs, researchers typically employ HLM to account for the lack of independence of observations due to the clustering effect, but they also often ignore potentially important higher levels of clustering above the level at which randomization occurred (e.g., Al Otaiba et al., 2011; Hegedus et al., 2015; Karimi-Shahanjarini, Rashidian, Omidvar, & Majdzadeh, 2013; McDonald et al., 2006). One reason some authors cited for ignoring the higher level of nesting was due to a small number of clusters at that higher level (e.g., Al Otaiba et al., 2011; Hegedus et al., 2015). Hox (2010) suggested including a level of nesting if there are at least 30 clusters at that level. If there are too few clusters at the highest level of nesting (i.e., less than 30) and that level of nesting is accounted for in the analysis, *SE* estimates for fixed effect coefficients may become negatively biased, which can inflate researchers' probability of committing a Type-I error in hypothesis testing (Hox, 2010; Lai & Kwok, 2015). In sum, accounting for a level of nesting structure when it is inappropriate to do so (i.e., when there are too few clusters at that level) can negatively impact results in the same way as failing to account for a level of nesting structure when it is appropriate to do so. This, in turn, could cause researchers to make inaccurate inferences regarding the effectiveness of the intervention being studied.

However, Van den Noortgate and colleagues (2005) suggested that if researchers are interested in a predictor at a specific level, such as experimental group membership at the cluster level, then they should account for the level above and the level below that particular level, as *SE*s may otherwise become biased. Practical considerations for wanting to include additional levels of nesting also exist. In education, if students are clustered within classrooms and HLM is used to account for nesting structure because the students are no longer purely independent of one another, it would follow that classrooms nested within different schools are also not independent of one another (i.e., classrooms within the same school are more similar to each other compared to classrooms from different schools) and their nesting within schools should also be addressed. The existence of this additional level of nesting would have statistical implications as classrooms within the same school would likely have correlated error terms, and failing to properly account for the school level in the model may result in biased error terms (Luke, 2004).

Modeling Data from CRTs

Because CRTs feature nested data structure, individuals are no longer independent of one another (Campbell et al., 2000; Jo et al., 2008) and analytic techniques such as HLM or LGCM should be used to account for individuals' non-independence. Although HLM and LGCM come from two different analytic traditions (i.e., HLM is based on hierarchical regression and LGCM is based on structural equation modeling), when the two techniques are used to examine the same model, they produce identical parameter estimates and differ only in terms of model representation (Hox & Stoel, 2005; Stoel, van Den Wittenboer, & Hox, 2003). Therefore, one can model HLMs as multilevel LGCMs (MLGCMs), and vice-versa; either approach is appropriate for modeling data from CRTs.

Purpose of the Current Study

The purpose of this study was to compare three methods of handling a fourth level of nested data structure in simulated CRT data. Additional levels of nesting may be of practical or statistical importance in CRT designs, and different ways of handling them has received little attention in the current methodological literature.

The present study looked at three ways of handling a higher level of nesting across a variety of common conditions concerning CRT designs (e.g., number of clusters, intraclass correlations), including both ideal (e.g., meeting the minimum recommended 30 clusters at the highest level) and not ideal (e.g., having fewer than 30 clusters at the highest level) conditions to reflect the conditions described in empirical research (e.g., Al Otaiba et al., 2011; Hegedus et al., 2015). The current study addressed the following research questions:

1. How do model-based, level-ignoring, and design-based approaches to handling a fourth level of nesting impact fixed and random effect parameter estimate and *SE* biases in simulated CRT data? That is, compared to a specified threshold, does a 4-level model, a 3-level model that ignores Level 4, or a model that accounts for the fourth level using a design-based method introduce substantial parameter or *SE* bias?
2. How do design factors impact parameter or *SE* bias (if any)? That is, are the number individuals at Level 2, the number of Level 3 clusters, the number of Level 4 clusters, the

intraclass correlation (ICC) at Level 4, or any interactions among them related to parameter or *SE* bias?

The current paper examined these questions using a series of two studies. The first study compared the three methods of handling the fourth level of nesting using a model that included variables only at the first, second, and third levels. However, because *SE* estimates may become biased when a predictor variable is included at a level that is ignored (Van den Noortgate et al., 2005), the second study used a model that included a covariate at Level 4 to examine the impact of different ways of handling that level when a covariate was present.

Study 1

Method

Data Generation

CRTs can be comprised of at least three levels: repeated measures (Level 1), individuals (Level 2), and clusters (Level 3); randomization into treatment conditions occurs at the cluster level. However, in the current study, data for a 4-level model (e.g., repeated measures nested within students nested within classrooms nested within schools) was generated to reflect the real-world nesting structure found in education research, and therefore allow investigation of methods that can address this additional level. The 4-level model for data generation is shown below:

Level 1:

$$Y_{tijk} = \psi_{0ijk} + \psi_{1ijk}(Time_{tijk}) + e_{tijk} \quad (1)$$

$$\text{With } e_{tijk} \sim N(0, \sigma^2)$$

Level 2:

$$\psi_{0ijk} = \pi_{00jk} + \pi_{01jk}(Covariate_{ijk}) + r_{0ijk} \quad (2)$$

$$\psi_{1ijk} = \pi_{10jk} + r_{1ijk} \quad (3)$$

$$\text{With } \begin{bmatrix} r_{0ijk} \\ r_{1ijk} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\psi 00} & \tau_{\psi 01} \\ \tau_{\psi 10} & \tau_{\psi 11} \end{bmatrix}\right) \quad (4)$$

Level 3: $\pi_{00jk} = \beta_{000k} + \beta_{001k}(Group_{jk}) + u_{00jk}$ (5)

$$\pi_{01jk} = \beta_{010k} + \beta_{011k}(Group_{jk}) + u_{01jk} \quad (6)$$

$$\pi_{10jk} = \beta_{100k} + \beta_{101k}(Group_{jk}) + u_{10jk} \quad (7)$$

$$\text{With} \begin{bmatrix} u_{00jk} \\ u_{01jk} \\ u_{10jk} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} & \tau_{\pi 02} \\ \tau_{\pi 10} & \tau_{\pi 11} & \tau_{\pi 12} \\ \tau_{\pi 20} & \tau_{\pi 21} & \tau_{\pi 22} \end{bmatrix}\right) \quad (8)$$

Level 4: $\beta_{000k} = \gamma_{0000} + v_{000k}$ (9)

$$\beta_{001k} = \gamma_{0010} + v_{001k} \quad (10)$$

$$\beta_{010k} = \gamma_{0100} \quad (11)$$

$$\beta_{100k} = \gamma_{1000} \quad (12)$$

$$\beta_{011k} = \gamma_{0110} \quad (13)$$

$$\beta_{101k} = \gamma_{1010} + v_{101k} \quad (14)$$

$$\text{With} \begin{bmatrix} v_{000k} \\ v_{001k} \\ v_{101k} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\beta 00} & \tau_{\beta 01} & \tau_{\beta 02} \\ \tau_{\beta 10} & \tau_{\beta 11} & \tau_{\beta 12} \\ \tau_{\beta 20} & \tau_{\beta 21} & \tau_{\beta 22} \end{bmatrix}\right) \quad (15)$$

where $Time_{tijk}$ is a variable with four assessment waves (Time = 0, 1, 2, 3), $Covariate_{ijk}$ is a dichotomous variable with 0 and 1 representing two different groups at the individual level (e.g., male and female), and $Group_{jk}$ is a dichotomous variable with 0 and 1 representing two different intervention groups in the CRT design at the cluster level (e.g., a control group and a treatment group). The Level 2 covariate was included to reflect models from empirical research that include covariates at the individual level (e.g., Apthorp et al., 2012, Sarama et al., 2012). The simulated data featured a balanced design in which individuals were evenly distributed across all groups within the covariate and group variables.

In this 4-level model, six fixed effect coefficients (i.e., γ_{0000}, γ_{0010}, γ_{0100}, γ_{1000}, γ_{0110}, γ_{1010}) and nine variances of the random effects (σ^2, $\tau_{\psi 00}$, $\tau_{\psi 11}$, $\tau_{\pi 00}$, $\tau_{\pi 11}$, $\tau_{\pi 22}$, $\tau_{\beta 00}$, $\tau_{\beta 11}$, $\tau_{\beta 22}$) were estimated; no covariances among the random effects were estimated to reduce model complexity. To simulate the effect of an efficacious intervention, γ_{0000}, γ_{0010}, γ_{0100}, γ_{1000}, γ_{0110}, and γ_{1010} were fixed to 1.00, 0.10, 0.10, 1.00, 0.10, and 0.1044 respectively. In this simulated scenario, both the control group ($Group_{jk} = 0$) and the treatment group ($Group_{jk} = 1$) experienced growth over time, but the treatment group experienced growth at a faster rate than the control group, indicating the intervention was successful. More specifically, holding all other variables constant, the difference in growth trajectories between the control group and the treatment group was .1044 units, representing an effect size of about $\delta = .33$ (δ can be interpreted as a Cohen's *d* for growth trajectories). Values for the γ_{0000}, γ_{0010}, γ_{0100}, γ_{1000}, and γ_{0110} fixed effect parameters were determined following similar methodological research on multilevel modeling (e.g., Chen et al., 2010; Maas & Hox, 2005). The value for γ_{1010} was determined based on mean effect sizes observed in empirical education intervention studies (Hill, Bloom, Beck, Black, & Lipsey, 2008).

Following Raudenbush and Liu's (2000) criteria, the variances and covariances of the random effects were specified as follows:

$$\sigma^2 = 1.00,$$

$$T_\psi = \begin{bmatrix} \tau_{\psi 00} & \tau_{\psi 01} \\ \tau_{\psi 10} & \tau_{\psi 11} \end{bmatrix} = \begin{bmatrix} .20 & .00 \\ .00 & .10 \end{bmatrix},$$

$$T_\pi = \begin{bmatrix} \tau_{\pi 00} & \tau_{\pi 01} & \tau_{\pi 02} \\ \tau_{\pi 10} & \tau_{\pi 11} & \tau_{\pi 12} \\ \tau_{\pi 20} & \tau_{\pi 21} & \tau_{\pi 22} \end{bmatrix} = \begin{bmatrix} .15 & .00 & .00 \\ .00 & .10 & .00 \\ .00 & .00 & .10 \end{bmatrix},$$

$$T_\beta = \begin{bmatrix} \tau_{\beta 00} & \tau_{\beta 01} & \tau_{\beta 02} \\ \tau_{\beta 10} & \tau_{\beta 11} & \tau_{\beta 12} \\ \tau_{\beta 20} & \tau_{\beta 21} & \tau_{\beta 22} \end{bmatrix} = \begin{bmatrix} .0711 & .00 & .00 \\ .00 & .10 & .00 \\ .00 & .00 & .10 \end{bmatrix} \text{ when ICC = .05, and}$$

$$T_\beta = \begin{bmatrix} \tau_{\beta 00} & \tau_{\beta 01} & \tau_{\beta 02} \\ \tau_{\beta 10} & \tau_{\beta 11} & \tau_{\beta 12} \\ \tau_{\beta 20} & \tau_{\beta 21} & \tau_{\beta 22} \end{bmatrix} = \begin{bmatrix} .15 & .00 & .00 \\ .00 & .10 & .00 \\ .00 & .00 & .10 \end{bmatrix} \text{ when ICC = .10.}$$

As previously stated, the covariances among the random effects were fixed to zero to reduce overall model complexity.

The current study manipulated four design factors. These design factors were selected because they play a role in the accuracy of fixed and random effect parameter and *SE* estimates as well as statistical power in multilevel modeling (Hox, 2010; Spybrook, 2008).

Number of individuals per cluster. First, the number of individuals per cluster had two conditions, 20 and 40 individuals. These values were based on a literature review by Graves and Frohwerk (2009) on the state of multilevel modeling in school psychology and are representative of various cluster sizes found in empirical education research. Note that, whereas 20 individuals per cluster is a common cluster size in education, 40 individuals may represent the higher end of potential cluster sizes.

Number of clusters at Level 3. The number of clusters at Level 3 had three conditions: 30, 60, and 90 clusters. Again, values for the 30- and 60-cluster conditions were based on Graves and Frohwerk's (2009) literature review; the 90-cluster condition was created to represent cases that may feature an especially large number of clusters.

Number of clusters at Level 4. The number of clusters at the highest level of nesting (Level 4) had two conditions: 10 and 30 clusters. The 10-cluster condition was designed to reflect the small number of clusters observed at the highest level of nesting in empirical research (e.g., Al Otaiba et al., 2011; Hegedus et al., 2015), whereas the 30 cluster condition reflected a scenario in which the number of clusters at the highest level was the minimum ideal value for empirical research (Hox, 2010).

Intraclass correlation at Level 4. Lastly, the ICC at Level 4 had two conditions: a small ICC of .05 and a medium ICC of .10. Generally, an ICC of .05 would be considered small for educational research, whereas an ICC of .10 is more reasonable (Hox, 2010). To specify these different ICC values, the $\tau_{\beta 00}$ random effect variance estimate was manipulated. The variance ($\tau_{\beta 00}$) was set to .0711 to obtain an ICC of .05, and it was set to .15 to obtain an ICC of .10.

Simulation Design

The simulation used a 2 (number of individuals per Level 3 cluster: 20 or 40) x 3 (number of clusters at Level 3: 30, 60, or 90) x 2 (number of clusters at Level 4: 10 or 30) x 2 (ICC: .05 or .10) factorial design to generate the data. One thousand replications were generated for each condition using SAS 9.4, yielding a total of 24,000 datasets (1000 datasets * 24 conditions).

For each dataset, three different models were fitted, resulting in a total of 72,000 models (24,000 datasets * three models). The first was a 4-level HLM that accounted for all four levels of nesting structure using the model-based approach. The second was a 3-level HLM that ignored the highest level of nesting. These two models were fitted using maximum likelihood (ML) estimation in SAS version 9.4. The third was a 3-level model fitted as a MLGCM that accounted for the fourth level of nesting using the design-based approach. This model was estimated using the TYPE=COMPLEX TWOLEVEL routine and the maximum likelihood with robust *SE*s (MLR) estimator in Mplus version 7 (Muthén & Muthén, 2012), which produced a 3-level model featuring adjusted *SE* estimates that correct for the fourth level of nesting. These three models will be referred to as the 4-level model, 3-level model, and complex model respectively.

The current study examined the accuracy of fixed effect estimates, random effect variances, and all associated *SE*s. Note that although the current study analyzed the models as either HLMs or MLGCMs, the two types of models are interchangeable and their results produce the same estimates when used to analyze the same data (Hox & Stoel, 2005). Therefore, it was reasonable to compare parameter estimates and *SE*s across the two types of models. Furthermore, the current findings are all presented as HLM results to ease interpretation and be more consistent with the analyses more commonly used in empirical CRT research.

Analyses

For each set of conditions, only replications with estimates for all parameters under all three models were considered valid and used for further analysis. Replications that failed to compute parameter estimates for all fixed and random effects were considered invalid. Some models failed to compute random effect variance parameter estimates, resulting in non-positive definite G-matrices (for more information, see Kiernan, Tao, & Gibbs, 2012). All models with a non-positive definite G-matrix were excluded from further analyses; 7,735 replications were considered invalid based on this criterion, resulting in 16,265 valid replications being used for final analyses. Note that of these 7,735 invalid replications, 7,719 of them were considered invalid because the 4-level model had non-positive definite G-matrices. Parameter estimates from the 4-level model, 3-level model, and complex model were summarized across the valid replications for each of the 24 sets of conditions. The relative bias for each fixed effect parameter and random effect variance was computed using the following equation:

$$B\left(\bar{\hat{\theta}}\right) = \frac{\bar{\hat{\theta}}_{est} - \theta_{pop}}{\theta_{pop}}$$

where $\bar{\hat{\theta}}_{est}$ is the mean of a parameter estimate across the valid replications and θ_{pop} is the true parameter value under each design condition. A relative bias of zero indicates an unbiased parameter estimate whereas a negative relative bias indicates an underestimation of the parameter and positive relative bias indicates an overestimation of the parameter. The current study used a cutoff value of ±.05 for the acceptable relative parameter bias (Hoogland & Boomsma, 1998).

The relative bias of the estimated *SE*s was calculated using the following equation:

$$B(\hat{S}_{\theta}) = \frac{\bar{\hat{S}}_{\hat{\theta}\_est} - S_{\hat{\theta}\_true}}{S_{\hat{\theta}\_true}}$$

where $\bar{\hat{S}}_{\hat{\theta}\_est}$ is the mean of the estimated *SE*s of the parameter estimate across the valid replications in the four-level, three-level, and complex models, and $S_{\hat{\theta}\_True}$ is the standard deviation of the parameter estimate across the valid replications of the 4-level model within a particular design condition. The current study used a cutoff value of ±.10 for the acceptable relative *SE* bias (Hoogland & Boomsma, 1998).

Following other simulation procedures (e.g., Chen et al., 2010; Chung & Beretvas, 2012), a series of ANOVAs were used to examine the impact of the design factors on relative parameter and *SE* biases across the three types of models. Given the large number of replications to be included in the analysis, even small effects could be detected as statistically significant using the *F*-test. Therefore, the effect size indicator eta squared (η^2) was used to determine which design factors had a practically significant impact on relative bias; $\eta^2 \geq .01$ was used as the criterion to identify which factors and interactions had a meaningful effect. In the interest of space, only those effects with the largest effect sizes or that are most relevant in terms of interpretation will be discussed below. Also in the interest of space, no *F*-test results will be shown; all results discussed below were statistically significant at the $\alpha = .05$ level.

Results

Four-Level Model

Table 1 shows the means and standard deviations of the relative biases for all parameter estimates and *SE*s across the four-level, three-level, and complex models. The mean relative biases of the fixed and random effect parameter estimates were examined using the cutoff criterion of ±.05, and the mean relative biases of the fixed and random effect *SE*s were evaluated using the cutoff of ±.10 (Hoogland & Boomsma, 1998); bias statistics exceeding these cutoffs are shown in bold.

Table 1

Study 1 Means and Standard Deviations for Relative Parameter and SE Bias Estimates for the Four-Level, Three-Level, and Complex Models

		4-level Model		3-level Model		Complex Model	
		M	*SD*	*M*	*SD*	*M*	*SD*
Fixed Effect Parameters	Intercept	.0006	.1347	.0006	.1353	.0006	.1353
	Time	-.0005	.0651	-.0005	.0651	-.0005	.0651
	Covariate	.0042	.8306	.0042	.8306	.0044	.8306
	Group	-.0154	1.2917	-.0175	1.3361	-.0175	1.3360
	Time*Group	-.0061	.9818	-.0068	.9892	-.0068	.9893
	Covariate*Group	-.0036	.7728	-.0036	.7728	-.0036	.7728
Random Effect Variances	Level 1 Residual	.0001	.0239	.0001	.0239	<.0001	.0239
	Level 2 Intercept	-.0017	.1332	-.0018	.1333	-.0018	.1333
	Level 2 Time	-.0003	.0892	-.0003	.0892	-.0002	.0892
	Level 3 Intercept	**-.2048**	.4382	**1.0518**	.7458	**1.0516**	.7458
	Level 3 Time	-.0462	.2979	**.4689**	.4576	**.4688**	.4576
	Level 3 Covariate	-.0452	.4148	-.0453	.4149	-.0435	.4156
	Level 4 Intercept	**.1481**	.9408	-	-	-	-
	Level 4 Group	**.3547**	1.1622	-	-	-	-
	Level 4 Time*Group	.0320	.7338	-	-	-	-
Fixed Effect *SE*s	Intercept	**-.1034**	.1685	**-.1323**	.1911	-.0743	.1955
	Time	-.0363	.1385	**.1740**	.1675	-.0317	.1978
	Covariate	-.0047	.1149	-.0048	.1150	-.0055	.2006
	Group	**.3972**	.3253	**.4651**	.4138	**.4638**	.3800
	Time*Group	**.2974**	.3111	**.1478**	.3356	**.3459**	.3704
	Covariate*Group	**.9477**	.4875	**.9473**	.4868	**.9852**	.6687
	Level 1 Residual	-.0124	.0312	-.0124	.0312	-.0327	.1922

Random Effect Variance *SE*s	Level 2 Intercept	-.0572	.0492	-.0570	.0491	-.0798	.1870
	Level 2 Time	**-.1232**	.1491	**-.1231**	.1491	**-.1484**	.2293
	Level 3 Intercept	**-.2628**	.3816	**.5479**	1.2462	**.6685**	1.2888
	Level 3 Time	**-.1660**	.2633	.0213	.3108	.0871	.4814
	Level 3 Covariate	**-.1843**	.2087	**-.1771**	.2085	**-.2327**	.2653
	Level 4 Intercept	-.0336	.3932	-	-	-	-
	Level 4 Group	**.1585**	.4399	-	-	-	-
	Level 4 Time*Group	.0278	.4214	-	-	-	-

Note. Bias estimates indicating a substantial amount of bias are shown in boldface.

First, as the number of individuals per cluster increased, so did the covariate*group interaction fixed effect *SE* bias (η^2 = .102). Next, the number of Level 3 clusters also affected several parameter and *SE* estimates. As the number of Level 3 clusters increased, biases for the Level 3 intercept variance estimate (η^2 = .215) and the Level 4 group random effect variance estimate (η^2 = .084) decreased; however, as the number of Level 3 clusters increased, *SE* bias increased for both the Level 3 time random effect variance *SE* (η^2 = .035) and the Level 3 covariate random effect variance *SE* (η^2 = .062). Also, as the number of Level 4 clusters increased, the Level 3 time random effect variance *SE* bias increased (η^2 = .083). Broadly speaking, *SE*s became more biased as the number of individuals per cluster, Level 3 clusters, and Level 4 clusters increased, but random effect variances became less biased as the number of Level 3 clusters increased.

Next, the interaction between several design factors impacted parameter and *SE* biases. The interaction between the number of individuals per cluster and the number of Level 3 clusters impacted bias in the Level 2 random effect variance *SE* (η^2 = .106). *SE* bias increased as number of individuals per cluster increased, except when there were 30 Level 3 clusters; in this case, parameter bias decreased as number of individuals increased.

The interaction between the number of Level 3 clusters and Level 4 clusters also impacted several parameter and *SE* biases. This interaction was related to severe biases in the

Level 3 intercept random effect variance estimate ($\eta^2 = .076$), the Level 4 intercept random effect variance estimate ($\eta^2 = .071$), the group fixed effect *SE* ($\eta^2 = .340$), the time*group fixed effect *SE* ($\eta^2 = .363$), the covariate*group fixed effect *SE* ($\eta^2 = .221$), the Level 3 intercept random effect variance *SE* ($\eta^2 = .347$), the Level 3 covariate random effect variance *SE* ($\eta^2 = .036$), and the Level 4 group random effect variance *SE* ($\eta^2 = .021$). As these interactions differed dramatically across the parameter and *SE* bias estimates, bias means across the various cluster numbers are shown in Table 2 to help examine the nature of the interactions. As shown in Table 2, some common patterns emerged from these interactions. Generally, parameter estimates became more biased when there were fewer Level 3 clusters, and some estimates and *SE*s became more biased with more Level 4 clusters. Interesting effects also emerged when numbers of Level 3 clusters and Level 4 clusters were equal; biases tended to be either very high or very low with equal cluster numbers, but these effects were inconsistent across the bias estimates.

Next, the triple interaction between Level 4 ICC, number of Level 3 clusters, and number of Level 4 clusters substantially impacted the Level 2 time random effect variance *SE* bias ($\eta^2 = .101$). Bias remained relatively consistent across all groups, except when number of Level 3 and Level 4 clusters were the same; *SE* bias decreased when the ICC was .05 and cluster numbers were the same, and increased when the ICC was .10 and cluster numbers were the same.

Lastly, the triple interaction between number of individuals per cluster, Level 3 clusters, and Level 4 clusters impacted the intercept fixed effect *SE* bias ($\eta^2 = .017$). When the number of individuals per cluster was 20, the amount of *SE* bias decreased as the number of Level 4 clusters increased; this effect was stronger as the number of Level 3 clusters increased, but was not present when the number of individuals per cluster was 40. This triple interaction also affected the Level 2 time random effect variance *SE* bias ($\eta^2 = .100$). *SE* bias remained relatively consistent across all groups, except when number of Level 3 and Level 4 clusters were the same; *SE* bias increased when there were 20 individuals per cluster and cluster numbers were equal, and decreased when there were 40 individuals per cluster and cluster numbers were equal.

Table 2

Study 1 Relative Parameter and SE Bias Means across All Three Models for Number of Level 3 Clusters by Number of Level 4 Clusters Interaction Effects

Parameter/*SE*	Number of Level 4 Clusters	Number of Level 3 Clusters	4-level Model Bias	3-level Model Bias	Complex Model Bias
Level 3 Intercept Random Effect Variance Estimate	10	30	-.237	.997	.997
		60	-.080	.941	.940
		90	-.028	.949	.949
	30	30	-.850	1.457	1.457
		60	-.168	1.043	1.043
		90	-.067	1.020	1.020
Level 3 Time Random Effect Variance Estimate	10	30	-	.398	.398
		60	-	.422	.421
		90	-	.413	.413
	30	30	-	.704	.704
		60	-	.456	.456
		90	-	.459	.459
Level 4 Intercept Random Effect Variance Estimate	10	30	.019	-	-
		60	-.120	-	-
		90	-.139	-	-
	30	30	1.354	-	-
		60	.106	-	-
		90	-.030	-	-
Time Fixed Effect *SE*	10	30	-	.114	-
		60	-	.162	-
		90	-	.160	-
	30	30	-	.244	-
		60	-	.184	-
		90	-	.179	-
Group Fixed Effect *SE*	10	30	.556	.682	.692
		60	.377	.349	.436
		90	.258	.138	.324
	30	30	-.024	-.024	-.016
		60	.683	1.056	.714

		90	.559	.714	.657
Time*Group Interaction Fixed Effect *SE*	10	30	.405	.292	.495
		60	.203	-.039	.267
		90	.123	-.198	.183
	30	30	-.050	-.050	-.034
		60	.659	.659	.681
		90	.492	.378	.530

(table continues)

Table 2 (*cont.*).

Parameter/*SE*	Number of Level 4 Clusters	Number of Level 3 Clusters	4-level Model Bias	3-level Model Bias	Complex Model Bias
Covariate*Group Interaction Fixed Effect *SE*	10	30	1.062	1.059	1.138
		60	1.105	1.104	1.144
		90	1.133	1.132	1.167
	30	30	-.029	-.028	-.018
		60	1.087	1.089	1.137
		90	1.119	1.118	1.145
Level 3 Intercept Random Effect Variance *SE*	10	30	-.152	.195	.222
		60	-.220	.203	.412
		90	-.222	.227	.602
	30	30	-.973	3.11	2.957
		60	.120	-.018	-.004
		90	-.182	.033	.127
Level 3 Covariate Random Effect Variance *SE*	10	30	-.196	-.192	-.277
		60	-.231	-.220	-.288
		90	-.217	-.206	-.274
	30	30	-.002	-.002	-.045
		60	-.196	-.191	-.226
		90	-.218	-.211	-.246
Level 4 Group Random Effect Variance *SE*	10	30	.147	-	-
		60	.082	-	-
		90	.037	-	-
	30	30	.235	-	-
		60	.442	-	-
		90	.117	-	-

Three-Level Model

Substantially biased parameter and *SE* estimates are shown in bold in Table 1. Several design factors contributed to the severe parameter and *SE* biases. As the Level 4 ICC increased (i.e., more variability in the outcome was attributed to Level 4), bias in the Level 3 intercept random effect variance estimate also increased (η^2 = .113). As the number of individuals per cluster increased, the covariate*group fixed effect *SE* bias also increased (η^2 = .102). As the number of Level 3 clusters increased, the intercept fixed effect *SE* bias also increased (η^2 = .129). As the number of Level 4 clusters increased, the intercept fixed effect *SE* bias decreased (η^2 = .335). Generally, bias estimates increased as the Level 4 ICC, number of individuals per cluster, and number of Level 3 clusters all increased.

The interaction between number of individuals per cluster and number of Level 3 clusters also impacted the Level 2 time random effect variance *SE* bias (η^2 = .105). *SE* bias increased as the number of individuals per cluster increased, except when there were 30 Level 3 clusters; in this case, parameter bias decreased as number of individuals increased.

Next, the interaction between number of Level 3 clusters and Level 4 clusters impacted several parameter and *SE* biases including: the Level 3 intercept random effect variance estimate (η^2 = .014), the Level 3 time random effect variance estimate (η^2 = .016), the time fixed effect *SE* (η^2 = .020), the group fixed effect *SE* (η^2 = .488), the time*group interaction fixed effect *SE* (η^2 = .399), the covariate*group interaction fixed effect *SE* (η^2 = .220), the Level 3 intercept random effect variance *SE* (η^2 = .298), and the Level 3 covariate random effect variance *SE* (η^2 = .038). To help examine the nature of these interactions, bias means across the various groups are shown in Table 2. Generally, random effect variance and *SE* biases increased when there were more Level 4 clusters, though there were some exceptions. Furthermore, the relationship between bias and number of Level 3 clusters was generally stronger when there were more Level 4 clusters. As with the 4-level model, interesting and inconsistent effects emerged when cluster numbers were equal.

Lastly, two triple interaction effects impacted the Level 2 time random effect variance *SE* bias. The triple interaction between Level 4 ICC, number of Level 3 clusters, and number of Level 4 clusters impacted this *SE* bias (η^2 = .101). Bias remained relatively consistent across all groups, except when number of Level 3 and Level 4 clusters were the same. *SE* bias decreased when the ICC was .05 and cluster numbers were the same, and increased when the ICC was .10 and cluster numbers were the same. Next, the triple interaction between the number of individuals per cluster, Level 3 clusters, and Level 4 clusters impacted the Level 2 time random effect variance *SE* bias (η^2 = .100). Similar to the previous triple interaction, bias remained consistent across all groups except when number of Level 3 and Level 4 clusters were the same. *SE* bias increased when there were 20 individuals per cluster and cluster numbers were the same, and decreased when there were 40 individuals per cluster and cluster numbers were the same.

Complex Model

Severely biased parameter and *SE* estimates are shown in bold in Table 1. Four-way ANOVA results showed that bias statistics were impacted by several factors. Similar to the 3-level model, as the Level 4 ICC increased, the Level 3 intercept random effect variance estimate bias also increased (η^2 = .113). As the number of individuals per cluster increased, the covariate*group interaction fixed effect *SE* bias also increased (η^2 = .056).

Next, the interaction between the number of individuals per cluster and Level 3 clusters impacted the Level 2 time random effect variance *SE* bias (η^2 = .044). *SE* bias increased as number of individuals per cluster increased, except when there were 30 Level 3 clusters; in this case, *SE* bias decreased as number of individuals increased.

The interaction between the number of Level 3 clusters and Level 4 clusters impacted several parameter and *SE* biases including: the Level 3 intercept random effect variance estimate

(η^2 = .014), the Level 3 time random effect variance estimate (η^2 = .016), the group fixed effect *SE* (η^2 = .331), the time*group fixed effect *SE* (η^2 = .281), the covariate*group fixed effect *SE* (η^2 = .133), the Level 3 intercept random effect variance *SE* (η^2 = .290), and the Level 3 covariate random effect variance *SE* (η^2 = .024). To help examine the nature of these interactions, bias means across the various groups are shown in Table 2. Similar to the 3-level model, random effect variance and *SE* biases increased when there were more Level 4 clusters, though there were some exceptions. Also similar to the previous models, interesting effects were present when clusters numbers were equal.

Lastly, a triple interaction between the Level 4 ICC, number of Level 3 clusters, and number of Level 4 clusters substantially impacted the Level 2 time random effect variance *SE* bias (η^2 = .045). *SE* bias remained relatively consistent across all groups, except when the number of Level 3 and Level 4 clusters were the same. *SE* bias decreased when the ICC was .05 and cluster numbers were the same, and bias increased when the ICC was .10 and cluster numbers were the same.

Study 2

Method

Data Generation

Study 2 served as an extension to Study 1 and investigated the impact of different methods of handling a fourth level of nesting structure when a covariate was present at Level 4. For example, in a 4-level analysis from a CRT (e.g., repeated measures, students, classrooms, schools) in which the treatment is administered at Level 3, a researcher may wish to include a relevant Level 4 covariate in the model, such as school-level socioeconomic status (SES), school urbanicity, or public-versus-private school status. Variables such as these occurring at the higher cluster level may be important to consider as they would add additional context to the outcome of testing the impact of an educational intervention. In this situation, ignoring the fourth level of nesting may cause the *SE*s of both the Level 4 covariate(s) and the Level 3 treatment predictors to become biased. Therefore, exploring the outcome of this potential scenario would be of interest to CRT researchers.

Data generation for Study 2 was identical to that of Study 1, except for a Level 4 covariate was included in the model. The model equations for Levels 1, 2, and 3 (i.e., Equations 1 through 8) remained the same as those in Study 1. The Level 4 equations for Study 2 were:

Level 4:

$$\beta_{000k} = \gamma_{0000} + \gamma_{0001}(Covariate2_k) + v_{000k} \quad (16)$$

$$\beta_{001k} = \gamma_{0010} + v_{001k} \quad (17)$$

$$\beta_{010k} = \gamma_{0100} \quad (18)$$

$$\beta_{100k} = \gamma_{1000} \quad (19)$$

$$\beta_{011k} = \gamma_{0110} \quad (20)$$

$$\beta_{101k} = \gamma_{1010} + v_{101k} \quad (21)$$

$$\text{With} \begin{bmatrix} v_{000k} \\ v_{001k} \\ v_{101k} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\beta 00} & \tau_{\beta 01} & \tau_{\beta 02} \\ \tau_{\beta 10} & \tau_{\beta 11} & \tau_{\beta 12} \\ \tau_{\beta 20} & \tau_{\beta 21} & \tau_{\beta 22} \end{bmatrix} \right) \quad (22)$$

where $Covariate2_k$ was a dichotomous variable with 0 and 1 representing two different groups at the fourth level (e.g., public versus private school at the school level).

In this 4-level model, seven fixed effect coefficients (i.e., γ_{0000}, γ_{0010}, γ_{0100}, γ_{1000}, γ_{0110}, γ_{1010}, γ_{0001}) and nine variances of the random effects (σ^2, $\tau_{\psi 00}$, $\tau_{\psi 11}$, $\tau_{\pi 00}$, $\tau_{\pi 11}$, $\tau_{\pi 22}$, $\tau_{\beta 00}$, $\tau_{\beta 11}$, $\tau_{\beta 22}$) were estimated; as with Study 1, no covariances among the random effects were estimated to reduce model complexity. Values for all fixed effect coefficients present in Study 1 remained the same for Study 2 in order to simulate the outcome of an effective intervention. The value of γ_{0001}

was fixed at 0.10 to indicate that there was a small difference between Level 4 covariate groups at the start of the intervention (i.e., Time = 0). The variances and covariances of the random effects were fixed at the same values as those used in Study 1, following Raudenbush and Liu's (2000) criteria.

The same four design factors manipulated in Study 1 were manipulated in the present study as well: number of individuals per Level 3 cluster, number of Level 3 clusters, number of Level 4 clusters, and conditional ICC. The current study also used the same values for the design factors as Study 1.

Analyses

The same analytic procedures from Study 1 were used in Study 2. That is, for each set of conditions, only replications with estimates for all parameters under all three models were considered valid and used for further analysis. Replications that failed to compute any parameter estimate were considered invalid; 8,215 replications were considered invalid based on this criterion, resulting in 15,785 valid replications being used for final analyses. Note that, of these 8,215 invalid replications, 8,200 of them were invalid because the 4-level model failed to compute all parameter estimates. Study 2 examined the relative bias estimates of the fixed effect parameters, random effect variances, and corresponding *SE*s using the equations and acceptable bias thresholds described in Study 1 (Hoogland & Boomsma, 1998). Furthermore, the current simulation used a series of ANOVAs and the $\eta^2 \geq .01$ criterion to identify which design factors and interactions had a meaningful impact on relative parameter and *SE* bias. As with Study 1, only those effects with the largest effect sizes or that were most relevant in terms of interpretation will be described below and no *F*-test results will be shown in the interest of space. All results discussed below were statistically significant at the $\alpha = .05$ level.

Results

Four-Level Model

Table 3 shows the means and standard deviations of the relative biases for all parameter estimates and *SE*s across the four-level, three-level, and complex models. As with Study 1, the mean relative biases of the fixed and random effect parameter estimates were examined using the cutoff criterion of ±.05, and the mean relative biases of the fixed and random effect *SE*s were evaluated using the cutoff of ±.10 (Hoogland & Boomsma, 1998); bias statistics exceeding these cutoffs are shown in bold.

Four-way ANOVA results revealed that several factors impacted parameter and *SE* biases. First, as the Level 4 ICC increased, the Level 2 time random effect variance *SE* bias also increased (η^2 = .059). Next, as the number of individuals per cluster increased, the covariate*group interaction fixed effect *SE* bias (η^2 = .306) and the Level 3 covariate random effect variance *SE* bias (η^2 = .015) increased, and the Level 2 Level 2 time random effect variance *SE* bias decreased (η^2 = .052). Broadly, as the Level 4 ICC and the number of individuals per cluster increased, *SE*s became more biased, with one exception.

The number of Level 3 clusters impacted biases for the Level 3 intercept random effect variance estimate (η^2 = .227), the Level 4 group random effect variance estimate (η^2 = .040), the time*group fixed effect *SE* (η^2 = .119), the covariate2 fixed effect *SE* (η^2 = .024), the Level 3 time random effect variance *SE* (η^2 = .029), the Level 4 group random effect variance *SE* (η^2 = .061), and the Level 4 time*group random effect variance *SE* (η^2 = .053). Because these effects took on different patterns, Table 4 shows the mean parameter and *SE* biases across numbers of Level 3 clusters. Most parameter and *SE* biases decreased as the number of Level 3 clusters increased, though there were some exceptions.

Table 3

Study 2 Means and Standard Deviations for Relative Parameter and SE Bias Estimates for the 4-Level, 3-Level, and Complex Models

	4-level Model		3-level Model		Complex Model	
	M	*SD*	*M*	*SD*	*M*	*SD*
Fixed Effect Parameters						
Intercept	.0006	.1872	.0007	.1901	.0007	.1901
Time	.0006	.0656	.0006	.0656	.0006	.0656
Covariate	.0035	.8450	.0035	.8451	.0038	.8451
Group	-.0137	1.1263	-.0125	1.1703	-.0126	1.1703
Time*Group	-.0047	.8481	-.0061	.8525	-.0060	.8525
Covariate*Group	-.0072	.5935	-.0072	.5934	-.0071	.5935
Covariate2	-.0051	2.4826	-.0082	2.5450	-.0082	2.5450
Random Effect Variances						
Level 1 Residual	-.0003	.0236	-.0003	.0236	-.0004	.0236
Level 2 Intercept	-.0022	.1358	-.0021	.1359	-.0021	.1359
Level 2 Time	.0001	.0923	.0001	.0923	.0001	.0923
Level 3 Intercept	**-.2234**	.4452	**.8801**	.6086	**.8798**	.6087
Level 3 Time	-.0409	.3008	**.4301**	.4141	**.4300**	.4141
Level 3 Covariate	-.0483	.4339	-.0499	.4348	-.0483	.4354
Level 4 Intercept	-.0220	.7428	-	-	-	-
Level 4 Group	**.2743**	.9893	-	-	-	-
Level 4 Time*Group	**-.0555**	.5465	-	-	-	-
Fixed Effect SEs						
Intercept	**-.1895**	.1520	**-.2682**	.1563	**-.1548**	.2167
Time	-.0399	.1406	**.1535**	.1596	-.0350	.1977
Covariate	-.0234	.1173	-.0238	.1174	-.0229	.1969
Group	**.4267**	.2685	**.4666**	.3417	**.4883**	.3230
Time*Group	**.3855**	.2927	**.2375**	.3577	**.4336**	.3454
Covariate*Group	**1.0152**	.3010	**1.0144**	.3012	**1.0524**	.5174
Covariate2	**-.2021**	.1609	**-.3753**	.1381	**-.1355**	.1751
Random Effect Variance SEs						
Level 1 Residual	.0025	.0342	.0026	.0342	-.0166	.1926
Level 2 Intercept	-.0675	.0422	-.0672	.0422	-.0905	.1835
Level 2 Time	**-.1059**	.0422	**-.1058**	.0422	**-.1320**	.1823
Level 3 Intercept	**-.2758**	.3929	**.5309**	1.1688	**.6165**	1.1885
Level 3 Time	**-.1646**	.2637	-.0115	.2689	.0428	.4208
Level 3 Covariate	**-.2126**	.1960	**-.2061**	.1965	**-.2642**	.2487
Level 4 Intercept	-.0036	.4021	-	-	-	-
Level 4 Group	**.1926**	.4414	-	-	-	-
Level 4 Time*Group	**.1196**	.4448	-	-	-	-

Note. Bias estimates indicating a substantial amount of bias are shown in boldface.

Table 4

Study 2 Parameter and SE Bias Means across Numbers of Level 3 Clusters in the 4-level Model

	Number of Level 3 Clusters		
Parameter/*SE* bias	30	60	90
Level 3 Intercept Random Effect Variance Estimate	-.625	-.117	-.049
Level 4 Group Random Effect Variance Estimate	.605	.202	.119
Time*Group Interaction Fixed Effect *SE*	.531	.405	.279
Covariate2 Fixed Effect *SE*	-.246	-.181	-.190
Level 3 Time Random Effect Variance *SE*	-.103	-.148	-.216
Level 4 Group Random Effect Variance *SE*	.337	.235	.069
Level 4 Time*Group Random Effect Variance *SE*	.285	.126	.010

Next, as the number of Level 4 clusters increased, the time*group interaction fixed effect *SE* bias (η^2 = .213), the Level 4 group random effect variance *SE* bias (η^2 = .058), and the Level 4 time*group random effect variance *SE* bias (η^2 = .044) also increased. However, the Level 3 time random effect variance *SE* bias decreased as number of Level 4 clusters increased (η^2 = .076).

The interaction between the number of Level 3 clusters and Level 4 clusters impacted several parameter and *SE* bias estimates including: the Level 3 intercept random effect variance estimate (η^2 = .071), the intercept fixed effect *SE* (η^2 = .012), the group fixed effect *SE* (η^2 = .159), and the Level 3 intercept random effect variance *SE* (η^2 = .340). To help examine the nature of these interactions, bias means across the various cluster numbers are shown in Table 5. These interaction effects took on a variety of patterns, but random effect variances and *SE*s generally became less biased as the number of Level 3 clusters increased, but were impacted inconsistently by the number of Level 4 clusters.

Table 5

Study 2 Relative Parameter and SE Bias Means across All Three Models for Number of Level 3 Clusters by Number of Level 4 Clusters Interaction Effects

Parameter/*SE*	Number of Level 4 Clusters	Number of Level 3 Clusters	4-level Model Bias	3-level Model Bias	Complex Model Bias
Level 3 Intercept Random Effect Variance Estimate	10	30	-.254	-	-
		60	-.073	-	-
		90	-.024	-	-
	30	30	-.877	-	-
		60	-.177	-	-
		90	-.074	-	-
Intercept Fixed Effect *SE*	10	30	-.229	-	-
		60	-.201	-	-
		90	-.199	-	-
	30	30	-.233	-	-
		60	-.117	-	-
		90	-.164	-	-
Group Fixed Effect *SE*	10	30	.544	.635	.654
		60	.369	.307	.426
		90	.313	.159	.380
	30	30	.267	.265	.284
		60	.616	.955	.644
		90	.528	.661	.619
Level 3 Intercept Random Effect Variance *SE*	10	30	-.124	.142	.134
		60	-.205	.157	.310
		90	-.238	.152	.440
	30	30	-.973	2.909	2.765
		60	.121	-.033	-.019
		90	-.184	.004	.096

Lastly, although the Level 4 time*group random effect variance estimate was severely overestimated (see Table 3), none of the design factors were substantial contributors to this bias according to the four-way ANOVA results and the $\eta^2 \geq .01$ criterion.

Three-Level Model

Severely biased parameter and *SE* estimates are shown in bold in Table 3. ANOVA results indicated that bias statistics were impacted by several factors. First, as the Level 4 ICC increased, several biases also increased, including: the Level 3 intercept random effect variance estimate (η^2 = .118), the time fixed effect *SE* (η^2 = .010), the group fixed effect *SE* (η^2 = .025), and the Level 2 time random effect variance *SE* (η^2 = .059). Next, as the number of individuals per cluster increased, covariate*group fixed effect *SE* bias (η^2 = .306) and the Level 3 covariate random effect variance *SE* bias (η^2 = .019) both increased and the Level 2 time random effect variance bias decreased (η^2 = .051). In general, *SE* biases increased as the Level 4 ICC and the number of individuals per cluster increased, with one exception.

As the number of Level 3 clusters increased, the intercept fixed effect *SE* bias (η^2 = .129) and the covariate2 fixed effect *SE* bias (η^2 = .181) both increased and the time*group fixed effect *SE* bias (η^2 = .216) decreased. Next, as the number of Level 4 clusters increased, the intercept fixed effect *SE* bias (η^2 = .383) and the covariate2 fixed effect *SE* bias (η^2 = .341) both decreased whereas the time*group fixed effect *SE* bias increased (η^2 = .409).

Next, the interaction between the number of individuals per cluster and Level 3 clusters impacted the time fixed effect *SE* bias (η^2 = .020). *SE* bias increased as the number of individuals per cluster increased, except when there were 90 Level 3 clusters. In that case, *SE* bias decreased as the number of individuals per cluster increased.

The interaction between number of Level 3 clusters and Level 4 clusters impacted the group fixed effect *SE* bias (η^2 = .013) and the Level 3 intercept random effect variance *SE* bias (η^2 = .296). To help examine the nature of these interactions, bias means across the various cluster numbers are shown in Table 5. The group fixed effect *SE* bias decreased as the number of

Level 3 clusters increased, but only when there were 10 Level 4 clusters. The Level 3 intercept random effect variance *SE* bias was severely overestimated when the numbers of Level 3 clusters and Level 4 cluster were equal. Lastly, although the Level 3 time random effect variance estimate was severely overestimated (see Table 3), none of the design factors substantially contributed to this bias according to the four-way ANOVA results and the $\eta^2 \geq .01$ criterion.

Complex Model

Substantially biased parameter and *SE* estimates are displayed in bold in Table 3. Four-way ANOVA results showed that parameter and *SE* biases were impacted by several factors. As the ICC at Level 4 increased, the Level 3 intercept random effect variance estimate bias also increased ($\eta^2 = .118$). As the number of individuals per cluster increased, the covariate*group fixed effect *SE* bias increased as well ($\eta^2 = .117$). Overall, as ICC and the number of individuals per cluster increased, *SE*s became more biased.

Next, as the number of Level 3 clusters increased, biases for the intercept fixed effect *SE* ($\eta^2 = .016$), time*group fixed effect *SE* ($\eta^2 = .088$), and covariate2 fixed effect *SE* ($\eta^2 = .023$) all decreased. As the number of Level 4 clusters increased, biases for the time*group fixed effect *SE* ($\eta^2 = .106$) and the covariate2 fixed effect *SE* ($\eta^2 = .028$) both increased. Similar to previous models, as the number of Level 3 and Level 4 clusters increased, *SE* biases became more biased.

The interaction between the number of Level 3 clusters and Level 4 clusters impacted the group fixed effect *SE* bias ($\eta^2 = .153$) and the Level 3 intercept random effect variance *SE* bias ($\eta^2 = .293$). To help examine the nature of these interactions, bias means across the various cluster numbers are shown in Table 5. Similar to the 3-level model, the group fixed effect *SE* bias decreased as the number of Level 3 clusters increased, but only when there were 10 Level 4 clusters. The Level 3 intercept random effect variance *SE* bias was severely overestimated when the number of Level 3 clusters and Level 4 cluster were equal. Lastly, although the Level 2 time random effect variance *SE*, the Level 3 time random effect variance estimate, and the Level 3 covariate random effect variance *SE* were severely biased (see Table 3), ANOVA results indicated that none of the design factors were substantially related to these biases based on the $\eta^2 \geq .01$ criterion.

Discussion

The purpose of the current study was to examine the impact of different methods of accounting for a fourth level of nesting structure on parameter and *SE* estimates in the context of CRT designs. Previous research has suggested that ignoring potentially meaningful levels of nesting can result in the improper allocation of explained variance across different levels of the model (Moerbeek, 2004). Furthermore, Van den Noortgate and colleagues (2005) suggested that when a researcher is interested in a predictor at a specific level, such as treatment group membership at the cluster level, then they should account for both levels of nesting adjacent to that level. However, ignoring higher levels of nesting can be justified in some situations, such as when the number of clusters at that level is small (Hox, 2010). The current study observed several interesting findings regarding the impact of different methods of accounting for a fourth level of nesting structure.

None of the fixed effect parameter estimates were severely biased for the four-level, three-level, or complex models. However, several biased *SE*s and random effect variances were present in all three models. Although all three models had a great deal of overlap in their biases (see Tables 1 and 3), they also each featured unique biased parameters and *SE*s, suggesting that all models have both common and unique issues.

First, the 4-level model had several biased random effect parameters and *SE*s at Level 4. These biases were driven primarily by the number of Level 3 clusters, the number of Level 4 clusters, and their interaction. Generally speaking across both studies, most *SE* biases decreased as the number of Level 3 clusters increased (with some exceptions), and increased as the number of Level 4 clusters increased. Furthermore, random effect variance biases decreased as the number of Level 3 clusters increased.

The 4-level model likely had several biased *SE*s and random effect variances because multilevel models with three or more levels are more difficult to estimate than simpler models (Hox, 2010). Including additional levels of nesting simultaneously increases the number of parameters that need to be estimated and reduces the cluster size at the highest level, making it more difficult to compute robust parameter and *SE* estimates. Overall, the 4-level model had more severely biased parameter and *SE* estimates than the 3-level and complex models, several of which occurred at the fourth level of nesting.

In the 3-level and complex models, the Level 3 intercept random effect variance estimate was severely overestimated; generally, bias was greater when there were more Level 4 clusters and a larger Level 4 ICC. This finding is reasonable in the context of previous research (Moerbeek, 2004). Because the fourth level of nesting was not completely accounted for, intercept variance that should have been attributed to that level was instead reallocated to the level below it. The Level 4 ICC played a particularly large role in this; as the amount of variability attributed to Level 4 increased, variance biases increased dramatically.

Furthermore, numerous *SE*s were also biased in the 3-level and complex models. This was unsurprising given that prior research has suggested that failing to properly model nesting structure can negatively impact the accuracy of *SE* estimates (Hox, 2010; Pornprasertmanit et al., 2014). These biases were driven primarily by the Level 4 ICC and the interaction between the number of Level 3 and Level 4 clusters. Broadly, the more weight Level 4 carried (i.e., larger ICC, more clusters at that level), the more biased *SE* estimates became, though there were some exceptions.

It is worth noting that in both studies, the complex model, which accounted for the fourth level of nesting using the design-based approach, had fewer severely biased *SE*s than the 4-level and 3-level models. This is likely due to two reasons. First, because the TYPE=COMPLEX routine in Mplus uses adjusted *SE* estimates that account for the presence of a higher level of nesting (Muthén & Muthén, 2012), the complex model performed better and featured fewer severely biased *SE* estimates than the 3-level model, which ignored the fourth level altogether. Second, because the complex model accounted for the highest level of nesting, but did not need to compute any of the random effect variances or *SE*s that existed at Level 4, it featured fewer biased estimates and *SE*s than the 4-level model.

Study 2 included a covariate at Level 4 to evaluate the impact of the various methods of handling Level 4 on its parameter and *SE* bias. Although the Level 4 covariate parameter estimate was not biased in any of the models, its *SE* was substantially underestimated in all three models. This finding is consistent with previous research which suggested that parameter and *SE* estimates may become biased when a level featuring a predictor is ignored (Van den Noortgate et al., 2005). Current results suggested that if researchers plan to acknowledge the fourth level of nesting using either the model- or design-based approach, the Level 4 covariate *SE* is less biased when there is a larger number of Level 3 clusters in the analysis. However, if researchers ignore the highest level of nesting, then *SE* bias actually decreases when there are fewer Level 3 clusters and more Level 4 clusters.

Also of note, upon examination of the interaction between number of Level 3 and Level 4 clusters' effect on parameter and *SE* biases, several unique effects emerged when the number of Level 3 and Level 4 clusters were both 30. Several parameters and *SE*s became either highly accurate or highly biased relative to other bias means when cluster numbers were equal (e.g., see the covariate*group interaction fixed effect *SE* or the Level 3 intercept random effect variance *SE* shown in Table 2). It is unknown at this time why these effects occurred, but future research can explore the impact on parameter and *SE* bias when lower- and higher-level cluster numbers are equal.

Recommendations for CRT Researchers

The present findings carry implications for future researchers employing CRT designs. Based on the current results, if a meaningful level of nesting structure exists above the level at which randomization occurs (i.e., having a Level 4 ICC of about .10, having about 30 clusters at Level 4), then researchers should consider accounting for it in their analyses using a design-based approach. In the current study, the complex model employing a design-based approach to handling Level 4 had fewer biased *SE*s and performed better than the 3-level model. Therefore, researchers should account for that level using a design-based approach rather than ignoring it altogether. If the Level 4 ICC is very small and/or there are few Level 4 clusters, then it would be appropriate to ignore that level and analyze the data using a 3-level model.

Accounting for Level 4 using a model-based approach is not recommended based on the current findings. The present study encountered model estimation issues because some 4-level models failed to compute random effect variances. Furthermore, among the three models that were tested, the 4-level model featured the largest number of biased parameters and *SE*s. Several of these estimates remained severely biased even under the more optimal conditions (i.e., having a Level 4 ICC of .10, having 30 clusters at Level 4). Researchers would need more than 30 clusters at Level 4 to help reduce model estimation issues and severe parameter bias. Because one reason researchers ignore the higher level is due to having too few clusters (e.g., Al Otaiba et al., 2011; Hegedus et al., 2015), it is unlikely that having more than 30 Level 4 clusters would occur in empirical research. Therefore, because of the estimation issues and highly biased results associated with the 4-level model, it is recommended that future CRT researchers do not account for Level 4 using a model-based approach.

Limitations

The present study had a few limitations. First, the current study examined a relatively specific set of models and design conditions; therefore, the results may be relevant only for the conditions examined here. Also, the models used in the current study only estimated random effects variances. Although random effects covariances are typically estimated in empirical research, they were excluded in this study to reduce model complexity and due to hardware limitations.

Future Directions

Although the current study found some interesting results regarding different methods of handling a level of nesting that features a predictor variable, these findings are not definitive. Future researchers can further explore the impact of handling a level of nesting structure that features a predictor in different types of multilevel models and contexts (e.g., cross-classified models, etc.). Also, the current findings suggest that, if researchers plan to account for the highest level of nesting using a model-based approach, having 30 clusters at that level does not produce unbiased parameter estimates and *SE*s. Future research can examine the impact of having more than 30 clusters at the highest level and examine how many clusters are necessary to get biases below the threshold values.

The current study also observed some interesting effects on parameter and *SE* biases when the number of Level 3 and Level 4 clusters were equal. Whereas some research has examined the impact of having a small number of individuals per cluster (Bell, Morgan, Kromney, & Ferron, 2010), research has not yet explored cases in which there may be a smaller number of lower-level clusters nested within higher-level clusters. Additional research is needed to examine why parameters and *SE*s may have become either highly biased or highly accurate when the cluster numbers are equal. Lastly, future simulation research on 4-level models could implement simpler models, such as by excluding covariates or some random effects. This would allow for the estimation of random effects covariances, and researchers could examine the impact of different methods of handling a higher level of nesting on these parameters as well.

Conclusion

In summary, in a CRT, an additional higher level of nesting may exist above the level at which group randomization occurred; using different methods of handling this higher level of nesting impacted parameter and *SE* biases in a variety of ways. The current findings suggest that, if a meaningful fourth level exists, it would be beneficial to account for it using a design-based approach in multilevel modeling. The model-based approach is not recommended due to having issues regarding model estimation and parameter bias. If the higher level is not meaningful or practically important, then future researchers may ignore it altogether.

REFERENCES

Abe, Y., & Gee. K. A. (2014). Sensitivity analyses for clustered data: An illustration from a large-scale clustered randomized controlled trial in education. *Evaluation and Program Planning, 47*, 26-34. doi: 10.1016/j.evalprogplan.2014.07.001

Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., & Meadows, J. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction. *The Elementary School Journal, 111*, 535-560. doi: 10.1086/659031

Ames, A. J. (2013). Accuracy and precision of an effect size and its variance from a multilevel model for cluster randomized trials: A simulation study. *Multivariate Behavioral Research, 48,* 592-618. doi: 10.1080/00273171.2013.802978

Apthorp, H., Randel, B., Cherasaro, T., Clark, T., McKeown, M., & Beck, I. (2012). Effects of a supplemental vocabulary program on word knowledge and passage comprehension. *Journal of Research on Educational Effectiveness, 5*, 160-188. doi: 10.1080/19345747.2012.660240

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling, 12*, 411-434. doi: 10.1207/s15328007sem1203_4

Bell, B. A., Morgan, G. B., Kromney, J. D., & Ferron, J. M. (2010, July-August). *The impact of small cluster size on multilevel models: A Monte Carlo examination of two-level models with binary and continuous predictors*. Paper session presented at the meeting of the Joint Statistical Meetings, Vancouver, Canada.

Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313-334). New York, NY: Routledge.

Campbell, M. K., Mollison, J., Steen, N., Grimshaw, J. M., & Eccles, M. (2000). Analysis of cluster randomized trials in primary care: A practical approach. *Family Practice, 17,* 192-196.

Chen, Q., Kwok, O., Luo, W., & Willson, V. L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling, 17,* 570-589. doi: 10.1080/10705511.2010.510046

Chung, H., & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology, 65,* 185-200. doi: 10.1111/j.2044-8317.2011.02023.x

Donner, A., & Klar, N. (2004). Pitfalls and controversies in cluster randomization trials. *American Journal of Public Health, 94,* 416-422.

Edwards, S. J. L., Braunholtz, D. A., Lilford, R. J., & Stevens, A. J. (1999). Ethical issues in the design and conduct of cluster randomised controlled trials. *British Medical Journal, 318*, 1407-1409.

Graves, S. L., & Frohwerk, A. (2009). Multilevel modeling and school psychology: A review and practical example. *School Psychology Quarterly, 24*(2), 84-94. doi: 10.1037/a0016160

Hegedus, S. J., Dalton, S., & Tapper, J. R. (2015). The impact of technology-enhanced curriculum on learning advanced algebra in US high school classrooms. *Education Technology Research and Development*, *63*, 203-228. doi: 10.1007/s11423-015-9371-z

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives, 2*, 172-177. doi: 10.1111/j.1750-8606.2008.00061.x

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research, 26*, 329–367.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Hox, J. J., & Stoel, R. D. (2005). Multilevel and SEM approaches to growth curve modeling. In B. S. Everitt & D. C. Howell (Series Eds.), *Encyclopedia of Statistics in Behavioral Science: Vol. 3* (pp. 1296-1305). Hoboken, NJ: John Wiley & Sons.

Jo, B., Asparouhov, T., Muthén, B. O., Ialongo, N. S., & Brown, C. H. (2008). Cluster randomized trials with treatment noncompliance. *Psychological Methods, 13*(1), 1-18. doi: 10.1037/1082-989X.13.1.1

Karimi-Shahanjarini, A., Rashidian, A., Omidvar, N., & Majdzadeh, R. (2013). Assessing and comparing the short-term effects of TPB only and TPB plus implementation intentions interventions on snacking behavior in Iranian adolescent girls: A cluster randomized trial. *American Journal of Health Promotion, 27*, 152-161. doi: 10.4278/ajhp.110311-QUAN-113

Kiernan, K., Tao, J., & Gibbs, P. (2012, April). *Tips and strategies for mixed modeling with SAS/STAT procedures*. Paper session presented at the meeting of the SAS Global Forum, Orlando, Florida.

Kim, J. S., Olson, C. B., Scarcella, R., Kramer, J., Pearson, M., van Dyk, D., … Land, R. E. (2011). A randomized experiment of a cognitive strategies approach to text-based analytical writing for mainstreamed Latino English language learners in grades 6 to 12. *Journal of Research on Educational Effectiveness, 4*, 431-463. doi: 10.1080/19345747.2010.523513

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

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

Luo, W. & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research, 44*, 182-212. doi: 10.1080/00273170902794214

Luo, W., & Kwok, O. (2012). The consequences of ignoring individuals’ mobility in multilevel growth models: A Monte Carlo study. *Journal of Educational and Behavioral Statistics, 37*(1), 31-56. doi: 10.3102/1076998610394366

Maas, C. J. M., & Hox. J. J. Sufficient sample sizes for multilevel modeling. *Methodology, 1*(3), 86-92. doi: 10.1027/1614-1881.1.3.86

McDonald, L., Moberg, D. P., Rodriguez-Espiricueta, I., Flores, N. I., Burke, M. P., & Coover, G. (2006). After-school multifamily groups: A randomized controlled trial involving low-income, urban, Latino children. *Children & Schools, 28*(1), 25-34. doi: 10.1093/cs/28.1.25

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41*, 473-497. doi: 10.1207/s15327906mbr4104_3

Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research, 39*(1), 129-149. doi: 10.1207/s15327906mbr3901_5

Muthén, L.K. and Muthén, B.O. (2012). *Mplus user’s guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology, 25*, 267-316.

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: Some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research, 49*, 518-543. doi: 10.1080/00273171.2014.933762

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

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199-213. doi: 10.1037///1100:82989X.5.2.199

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,* 510-520. doi: 10.1037/a0035827

Russell, M., O'Dwyer, L. M., & Miranda, H. (2009). Diagnosing students' misconceptions in algebra: Results from an experimental pilot study. *Behavior Research Methods, 41*, 414-424. doi: 10.3758/BRM.41.2.414

Sarama, J., Clements, D. H., Wolfe, C. B., & Spitler, M. E. (2012). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies. *Journal of Research on Educational Effectiveness, 5*(2), 105-135. doi: 10.1080/19345747.2011.627980

Sim, J., & Dawson, A. (2012). Informed consent and cluster-randomized trials. *American Journal of Public Health, 102,* 480-485. doi: 10.2105/AJPH.2011.300389

Spybrook, J. (2008). Power and sample size for classroom and school-level interventions. In A. O'Connell & B. McCoach (Eds.), *Multilevel analysis of educational data* (pp. 273-311). Greenwich, CT: Information Age Publishing.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two-and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41,* 605-627. doi: 10.3102/1076998616655442

Stoel, R. D., van Den Wittenboer, G., & Hox, J. J. (2003). Analyzing longitudinal data using multilevel regression and latent growth curve analysis. *Metodologia de las Ciencas del Comportamiento, 5*, 21-42.

Van den Noortgate, W., Opdenakker, M., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement, 16*, 281-303. doi: 10.1080/09243450500114850

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods, 5*, 425-433. doi: 10.I037//1082-989X.5.4.425

Wu, J., & Kwok, O. (2012). Using SEM to analyze complex survey data: A comparison between design-based single-level and model-based multilevel approaches. *Structural Equation Modeling, 19*(1), 16-35. doi: 10.1080/10705511.2012.634703