The Effects of Including Observed Means or Latent Means as Covariates in Multilevel Models for Cluster Randomized Trials

Educational and Psychological

Measurement

2016, Vol. 76(5) 803–823

© The Author(s) 2015

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164415618705

epm.sagepub.com

Burak Aydin¹, Walter L. Leite², and James Algina²

Abstract

We investigated methods of including covariates in two-level models for cluster randomized trials to increase power to detect the treatment effect. We compared multilevel models that included either an observed cluster mean or a latent cluster mean as a covariate, as well as the effect of including Level 1 deviation scores in the model. A Monte Carlo simulation study was performed manipulating effect sizes, cluster sizes, number of clusters, intraclass correlation of the outcome, patterns of missing data, and the squared correlations between Level 1 and Level 2 covariates and the outcome. We found no substantial difference between models with observed means or latent means with respect to convergence, Type I error rates, coverage, and bias. However, coverage could fall outside of acceptable limits if a latent mean is included as a covariate when cluster sizes are small. In terms of statistical power, models with observed means performed similarly to models with latent means, but better when cluster sizes were small. A demonstration is provided using data from a study of the Tools for Getting Along intervention.

Keywords

cluster randomized trials, power, multilevel modeling, latent variables, aggregation

¹RTE University, Rize, Turkey

²University of Florida, Gainesville, FL, USA

Corresponding Author:

Walter L. Leite, Research and Evaluation Methodology Program, University of Florida, 1215 Norman Hall, Gainesville, FL 32611, USA.

Email: walter.leite@coe.ufl.edu

Research designs in which clusters of participants, for example, classrooms, schools, hospitals, communities, or geographically defined units, are randomly assigned to interventions are known as cluster randomized trials (CRTs), group randomized trials, or community trials (Donner & Klar, 2004; Moerbeek, 2006a; Murray, 1998). They have become very common in educational research since the U.S. congress passed the Education Sciences Reform Act of 2002, which created the Institute of Education Sciences and emphasized random assignment as the strongest and preferred method to establish causal relationships and perform scientifically valid educational evaluation (U.S. Congress, 2002). Since its creation, the Institute of Education Sciences has funded more than 100 grants using CRT designs.

Compared with randomizing individuals, CRT designs can reduce cost, reduce contamination, and increase the feasibility and administrative efficiency (Donner & Klar, 2004). In educational research, randomizing clusters can be more feasible than randomizing individuals because education systems are organized into clusters such as classrooms, schools, and school districts, and using these existing clusters can assist cost reduction. Also, compared with randomizing individuals, randomizing clusters reduces the risk of treatment contamination, which occurs when information transfers between the study's conditions, producing treatment diffusion and threatening the internal validity of the study (Cook & Campbell, 1979). Even though CRTs have considerable advantages, they can introduce design implementation difficulties, require complex statistical methods of analysis, and may produce low statistical power to detect treatment effects. The effect of clustering on statistical power can be traced back to earlier studies. Kish (1965) introduced the effect of clustering on the sample variance within a survey sampling framework. Cornfield (1978) studied the effect of clustering on creating equally distributed idiosyncratic characteristics for two condition groups and stated that randomizing clusters and then analyzing the data as if randomization occurred at the individual level to have larger statistical power is an exercise of self-deception. Barcikowski (1981) also examined the problem of using individuals as the unit of analyses and then making inferences at cluster level. He provided tables and equations to facilitate power estimates as a function of the intraclass correlation coefficient when using group means as the unit of analyses.

As the use of CRTs has become common in educational research, it has become particularly important to investigate how to increase power for testing for treatment effects. Increasing the number of clusters is an effective way to increase power in CRT, but given the possible high costs of increasing the number of clusters, including covariates in these designs is a generally less expensive strategy to increase statistical power. Covariates increase power by reducing the error variance if the correlation between the outcome and the covariate is substantial. In a CRT design, the researcher can measure covariates at different levels. In their presentation of power analysis for two-level CRTs, Spybrook et al. (2011) indicated that power can be enhanced by using a Level 2 covariate, and Level 1 covariates can be aggregated to create Level 2 covariates. Konstantopoulos (2012) showed that Level 2 predictors are generally more effective to increase statistical power compared with group-mean centered

Level 1 predictors.¹ This may not be the case with grand-mean centering or no centering of the Level 1 predictors. In such cases the Level 1 predictors are expected to explain variance at both levels. For a comprehensive discussion of the effects of centering in multilevel models, see Kreft, de Leeuw, and Aiken (1995); Enders and Tofighi (2007); and Algina and Swaminathan (2011).

Lüdtke et al. (2008) discussed two different types of aggregation: formative and reflective. In formative aggregation, an arithmetic average of a lower level variable within an upper level can be used as an upper level variable. In reflective aggregation, the expected value of a lower variable is used as an upper level variable. Formative aggregation results in an observed variable, whereas reflective aggregation results in a latent variable. The literature on power to test treatment effects in CRT data analysis with multilevel models has addressed the use of observed means as a Level 2 covariate (Hedges & Hedberg, 2007; Konstantopoulos, 2012). However, the literature has not addressed whether including a Level 2 covariate obtained with reflective aggregation in the multilevel model for CRT data results in different power than using a covariate obtained with formative aggregation. Furthermore, research on power for CRT analysis has not demonstrated conclusively whether including a group-mean centered Level 1 covariate in addition to the Level 2 covariate results in increased statistical power. Considering that a Level 2 covariate was created from a Level 1 covariate based on either formative or reflective aggregation, a Level 1 covariate could be included in the multilevel model as an individual deviation from either type of Level 2 aggregate. The goal of this study is to investigate the relative merits of formative or reflective aggregation to create Level 2 covariates and also of including the individual deviation from the covariate in multilevel models for estimating and testing treatment effects in CRTs. Thus, we expect to answer two different questions.

Research Question 1: How does including Level 2 covariates as latent means or observed means affect the power to detect the treatment effect in a CRT?

Research Question 2: How does including group-mean centered Level 1 covariate into the model, together with latent or observed means, affects statistical power to detect a treatment effect in CRT?

Multilevel Models for Cluster Randomized Trials

Several techniques can be used to analyze CRT data. These include single-level models applied to measures created by collapsing the data in each cluster, generalized estimating equations, multilevel models, and permutation tests. The comparisons of these methods are beyond the scope of this study and can be found elsewhere (Hox, Maas, & Brinkhuis, 2010; Hubbard et al., 2010; Murray et al., 2006; Murray, Varnell, & Blitstein, 2004; Snijders & Bosker, 2012). In this study, the focus is on multilevel modeling because it can estimate both regression coefficients and variance components and is frequently used in educational research. As an example of a two-level CRT, consider a treatment-control study in which J schools will be randomly assigned to two treatments. Assigning entire villages, hospitals, or families might

constitute a two-level design as well (Donner, 2009). Data from a two-level CRT with assignment at Level 2 can be analyzed by using the following model:

$$Y_{ij} = \gamma_{00} + \gamma_{01}X_j + \gamma_{02}W_j + u_j + \varepsilon_{ij}, \quad (1)$$

where Y_{ij} is the dependent variable for the i th Level 1 unit in the j th cluster, γ s are coefficients, X_j is a Level 2 covariate, the W_j is the treatment indicator, u_j is a Level 2 residual assumed to have a normal distribution with a mean of zero and variance of τ^2 , and ε_{ij} is a Level 1 residual assumed to have a normal distribution with a mean of zero and variance of σ^2 . The covariate X_j could be a variable that is not aggregated from scores on Level 1 units, such as school budget, or it could be a variable that is aggregated from scores on Level -1 units, such as a pretest measure.

Parameters in Equation 1 can be estimated with maximum likelihood (ML) or least squares estimation procedures. Murray (1998) noted that quite often least squares and ML estimates converge to the same result in large samples with normal error distribution. Swaminathan and Rogers (2008) noted that if cluster sizes differ, fixed effects and variance components should be estimated together with ML procedures.

If some of the multilevel data are missing, estimation methods can be modified or paired with missing data techniques (Black, Harel, & McCoach, 2011; Graham, 2012; Longford, 2008). We reviewed abstracts of studies that were funded by the IES over the last decade that were designed to use a CRT and found that researchers generally plan for a perfectly balanced data structure, where each cluster has the same number of students and each treatment condition has the same number of clusters. However, it is not common to obtain perfectly balanced data due to nonresponse, attrition from the study, as well as failure to recruit the same number of participants within clusters and the same number of clusters within study conditions.

Covariates in Multilevel Models for Cluster Randomized Trials

The goal of measuring covariates in CRT is to increase power given that the alternative, increasing the sample size, is generally more expensive. Thus, a strong relationship between the covariates and the dependent variable is desirable. Moerbeek (2006b) studied the cost and benefit for measuring covariates versus increasing the sample in terms of statistical power to test treatment effects in a CRT. The author showed that including covariates is generally more cost-efficient if the covariates have a sufficiently large correlation with the outcome.

A common practice to create school-level covariates is to aggregate the student-level variables. These types of variables are referred as contextual variables, and their effects are referred to as context effects (Raudenbush & Bryk, 2002). Lüdtke et al. (2008) investigated two methods to include contextual variables in multilevel models. The first method is the observed mean model (OMM), consisting of using the observed means of covariate scores as a Level 2 covariate:

$$Y_{ij} = \gamma'_{00} + \gamma'_{10}(X_{ij} - \bar{X}_j) + \gamma'_{01}\bar{X}_j + u'_j + \varepsilon'_{ij}. \quad (2)$$

In this equation, $\bar{X}_j$ is the observed mean of j th cluster, and $(X_{ij} - \bar{X}_j)$ is the group-mean centered Level 1 covariate. The coefficient γ'_{10} is the within-cluster coefficient and γ'_{01} is the between-cluster coefficient.

The second method investigated by Lüdtke et al. (2008) is the latent mean model (LMM), which includes the expected value of the covariate within each cluster in the model, as shown below:

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \mu_{jX}) + \gamma_{01}\mu_{jX} + u_j + \varepsilon_{ij}. \quad (3)$$

In this equation, μ_{jX} is the expected value of the covariate within the j th cluster, and $(X_{ij} - \mu_{jX})$ is the individual-level deviation from this expected value.

The LMM can be estimated by calculating empirical Bayes estimates of the latent means and using them as Level 2 means in a multilevel analysis, which is a two-stage approach, or in a single stage using full information ML estimation. Below we provide an overview of the two-stage approach, and the single stage approach is detailed by Lüdtke et al. (2008).

In the first step of the two-stage approach (Croon & van Veldhoven, 2007), the following unconditional model is estimated for the Level 1 covariate:

$$X_{ij} = \mu_{jX} + X_{ij} - \mu_{jX} = \mu_{jX} + R_{xij}. \quad (4)$$

In this model, the variance of R_{xij} is assumed to be constant for each cluster and is denoted by σ_x^2 and the variance of μ_{jX} is denoted by τ_x^2 . The reliability of the cluster mean can be obtained with the variance component estimates for the covariate (Raudenbush & Bryk, 2002):

$$\lambda_j = \frac{\hat{\tau}_x^2}{\hat{\tau}_x^2 + \hat{\sigma}_x^2/n_j}. \quad (5)$$

An empirical Bayes estimate of the latent group mean (Shin & Raudenbush, 2010), also referred to as unbiased predictor of the latent group mean (Croon & van Veldhoven, 2007), can be obtained with a reliability-weighted composite of the grand mean of the covariate $\bar{X}$ and the group mean $\bar{X}_j$:

$$\tilde{X}_j = (1 - \lambda_j)\bar{X} + \lambda_j\bar{X}_j. \quad (6)$$

The second stage of the two-stage approach is to estimate the model in Equation 2, but with $\tilde{X}_j$ replacing $\bar{X}_j$. The comparison by Lüdtke et al. (2008) between the one-stage and two-stage approaches indicated that the results are similar except under the condition with small sample sizes and low ICC, in which case the one-stage approach outperformed the two-stage approach, and they concluded that the full information procedure should be preferred. Therefore, in this study we will proceed using the one-stage approach to estimate a multilevel model with latent cluster means as covariates.

The observed cluster mean $\bar{X}_j$ is a less than perfectly reliable measurement of μ_{jX} , and a substantial level of unreliability of cluster means can produce differences

between estimates with models in Equations 2 and 3. Lüdtke et al. (2008) concluded that Equation 3 is the most appropriate method to estimate a context effect, because they showed that γ'_{10} in the OMM model and γ_{10} in the LMM model are the same, but γ'_{01} and γ_{01} are not. Instead, γ'_{01} in the OMM is an attenuated version of γ_{01} in the LMM. Due to this difference, the maximum likelihood estimator of γ'_{01} in the OMM is a biased estimated of γ_{01} in the LMM.

In a two-level CRT with Level 2 treatment assignment, the multilevel models in Equations 2 and 3 can be extended as shown in Equations 7 and 8, respectively, to estimate treatment effects:

$$Y_{ij} = \gamma'_{00} + \gamma'_{10}(X_{ij} - \bar{X}_j) + \gamma'_{01}\bar{X}_j + \gamma'_{02}W_j + u'_j + \varepsilon'_{ij} \quad (7)$$

and

$$Y_{ij} = \gamma_{00} + \gamma_{10}(X_{ij} - \mu_{jX}) + \gamma_{01}\mu_{jX} + \gamma_{02}W_j + u_j + \varepsilon_{ij}. \quad (8)$$

In this study, we will refer to the model in Equation 7 as OMM with Level 1 and Level 2 covariates (OMML1L2), because the observed mean is a Level 2 covariate and the individual deviation from this mean is a Level 1 covariate. Similarly, we will refer to the model in Equation 8 as LMM with Level 1 and Level 2 covariates (LMML1L2), because the latent mean is included as a Level 2 covariate and the individual deviation from the latent mean is included in Level 1.

Alternatively, the multilevel models in Equations 7 and 8 for estimating the effect of a Level 2 treatment effect in CRT can be specified without the individual-level deviation from the mean as a covariate, as shown in Equations 9 and 10, which we refer as OMM with Level 2 covariate (OMML2) and LMM with Level 2 covariate (LMML2), respectively:

$$Y_{ij} = \gamma'_{00} + \gamma'_{01}\bar{X}_j + \gamma'_{02}W_j + u'_j + \varepsilon'_{ij} \quad (9)$$

and

$$Y_{ij} = \gamma_{00} + \gamma_{01}\mu_{jX} + \gamma_{02}W_j + u_j + \varepsilon_{ij}. \quad (10)$$

Power to Detect the Treatment Effect

CRT designs generally require a larger total sample size to successfully detect a treatment effect when compared with random assignment of individuals. Bloom, Richburg-Hayes, and Black (2007) showed that the required number of clusters to detect a treatment effect in a CRT can be reduced by using covariates and including Level 2 covariate in the model is more effective than including Level 1 covariates. In Optimal Design (Spybrook et al., 2011), a popular software program to perform power analysis for CRT, power analysis for a two-level CRT with assignment at Level 2 is conducted assuming that there is only a Level 2 covariate. If a cluster-

level predictor and an individual-level predictor are included in the multilevel model for CRT treatment effect estimation, assuming there is no interaction between the treatment and the covariate, the noncentrality parameter for the test of the treatment effect in a two-level balanced CRT is (Hedges & Hedberg, 2007; Konstantopoulos, 2009):

$$\lambda = \delta \sqrt{\frac{Jn}{4}} \sqrt{\frac{1}{(1-\rho_W^2) + (n(1-\rho_B^2) - (1-\rho_W^2))\rho}}, \quad (11)$$

where δ is the effect size, ρ is the unconditional intraclass correlation coefficient, ρ_B^2 is the proportion of between cluster variance explained by the Level 2 covariate, ρ_W^2 is the proportion of within cluster variance explained by the Level 1 covariate, J is the number of clusters, and n is the cluster size. When there is no Level 1 covariate, Equation 11 becomes (Raudenbush, 1997):

$$\lambda = \sqrt{\frac{J\delta^2}{4[(1-\rho_B^2)\rho + (1-\rho)/n]}}. \quad (12)$$

The noncentrality parameter indicates that, when OML1L2 is used, power increases as ρ_B^2 and ρ_W^2 increase, providing a justification for including Level 2 and Level 1 covariates in the model. Although an expression for a noncentrality parameter for a multilevel model with latent means as Level 2 covariates has not been provided in the literature, it would be expected to be similar to Equation 12. Although Level 1 covariates may be helpful to increase power, Konstantopoulos (2012) noted that the noncentrality parameter in Equation 11 indicates that Level 2 variables are expected to have a larger impact on power in most cases (i.e., as long as cluster size is greater than 10 and ICC is greater than .10).

The noncentrality parameter indicates that when OML1L2 is used, power is a function of both ρ_B^2 and ρ_W^2 , therefore suggests that one might expect OMML1L2 to have higher power than OMML2 because of the inclusion of the group-mean centered Level 1 covariate in addition to the group-mean covariate as the Level 2 covariate. However, application of results presented in Snijders and Bosker (1994) indicates that when OMML1L2 is used in place of OMML2, the variance component at Level 1 goes down but the variance component at Level 2 goes up. Because the estimated standard error of the treatment effect estimator is a function of both variance components, it does not necessarily decline when the Level 1 covariate is added to the Level 2 covariate. Based on this result, it would be reasonable to only use Level 2 covariates to increase power in multilevel models for CRT.

The objective of the present study is to investigate whether convergence rates, Type I error rates, power of the test of the treatment effect, coverage of confidence intervals (CIs), and treatment effect bias differ when multilevel models with observed or latent Level 2 covariates and with or without a Level 1 covariate are used to analyze CRT data. More specifically, we extend the existing literature on statistical analysis of CRT designs by comparing the four models shown in Equations 7 to 10 through a Monte Carlo simulation study, which is presented below. Then, we provide

an illustration of the models of interest using data from the *Tools for Getting Along* project (Daunic, Smith, Brank, & Penfield, 2006; Smith, Lochman, & Daunic, 2005).

Monte Carlo Simulation Study

Method

CRT data were simulated using the R software (R Core Team, 2013) and analyzed using *Mplus* 7 (Muthén & Muthén, 2012). Data were simulated by manipulating seven factors: Level 1 sample size, Level 2 sample size, intraclass correlation of the outcome (ICC_Y), squared correlation between the Level 2 covariate and the outcome (ρ_B^2), squared correlation between the Level 1 covariate and the outcome (ρ_W^2), missing data pattern, and effect size. These manipulated factors are detailed in the next section. Simulated data sets were analyzed with four multilevel models (i.e., OMML1L2, LMML1L2, OMML2, and LMML2). The intraclass correlation of the covariate (ICC_X) was held constant across all condition. The simulation design consisted of a total of 3,456 conditions, with 1,000 simulated datasets per condition.

Data Generation

The first step in the data generation was to generate the data on the covariate. The covariate variable was decomposed into two uncorrelated components, μ_{jX} and R_{Xij} , as shown in Equation 4. The corresponding decomposition of the variance of the covariate is $\sigma_X^2 = \tau_X^2 + \sigma_{R_X}^2$. Without loss of generality, the covariate variance was set equal to one. Then τ_X^2 is equal to ICC_X and $\sigma_{R_X}^2$ is equal to $1 - ICC_X$. Therefore, scores on the variable μ_{jX} were generated by multiplying a standard normal variable by τ_X and scores on R_{ij} were generated by multiplying a standard normal variable by σ_{R_X} .

The population model for the outcome was defined according to Equation 8. The outcome variable and its variance can be decomposed as $Y_{ij} = \mu_{jY} + R_{Yij}$ and $\sigma_Y^2 = \tau_Y^2 + \sigma_{R_Y}^2$. The outcome variance was also set equal to one without loss of generality. Then τ_Y^2 is equal to ICC_Y and $\sigma_{R_Y}^2$ is equal to $1 - ICC_Y$.

Within each treatment group, the relationship between the cluster-level means for the outcome and predictor is $\mu_{jY} = \gamma_{00} + \gamma_{01}\mu_{jX} + u_{0j}$ (Snijders & Bosker, 2012). Using standard results in regression theory, the squared correlation for the equation is $\rho_B^2 = 1 - \sigma_u^2/\tau_Y^2$. Thus, once ICC_Y and the Level 2 correlation coefficient was set, the variance of the Level 2 residual was determined; the Level 2 residual was generated by multiplying a standard normal variable by σ_u . Within each treatment group, the relationship between the Level 1 deviations from the cluster mean for the predictor and outcome is $R_{Yij} = \gamma_{10}R_{Xij} + r_{ij}$ (Snijders & Bosker, 2012). The squared correlation for the equation is $\rho_W^2 = 1 - \sigma_r^2/(1 - \tau_Y^2)$. Once ICC_Y and the Level 1 correlation coefficient were set, the variance of the individual-level residual was determined; the Level 1 residual was generated by multiplying a standard normal variable by σ_r . In addition, the coefficients γ_{01} and γ_{10} were $\gamma_{01} = \tau_Y\rho_B/\tau_X$ and

Table 1. Factor Levels and Population Parameters.

Manipulated factor		Levels			
Level 1 sample size	3	6	10	20	
Level 2 sample size	40	50	80	140	
ICC _Y	.10	.30			
ρ _B ² and ρ _W ²	.10	.25	.5		
Missing data	10% at Level 1	10% at Level 2	5% at each		None

Note. ρ_B² = between association; ρ_W² = within association.

$\gamma_{10} = \sqrt{(1 - \tau_Y^2)\rho_w^2 / (1 - \tau_X^2)}$. Finally γ_{02} was determined using the noncentrality parameter from Equation 12 with the squared Level 2 correlation equal to ρ_B^2 .

Population Parameters

The levels of the factors and population parameters are presented in Table 1. The population treatment effect γ_{02} was set to zero or selected so that target power would be .70 or .90. We manipulated target power rather than effect size to investigate how well the four models perform in terms of achieving target power. For each condition we set γ_{02} equal to the minimum detectable effect size (MDES), that is, to the treatment effect that can be detected with a specified target power and a directional test of γ_{02} in the LMML2 model:

$$Y_{ij} = \gamma_{00} + \gamma_{01}\mu_{jX} + \gamma_{02}W_j + u_{0j} + \varepsilon_{ij}. \tag{13}$$

Due to the fact that both the outcome and covariate had standard deviation of one, the MDESs in the study are Cohen’s effect size with the denominator equal to the total standard deviation for a model excluding covariates. The Level 2 sample sizes in conjunction with the values for target power resulted in MDESs in the ranges .14 to .59 and .19 to .80 when target power is .70 and .90, respectively. Thus, the conditions used to investigate power were originated based on MDESs that are common in CRT designs. We also included a condition with the treatment effect equal to zero to investigate the effect of manipulated factors and models on Type I error rates. The data were generated by using the LMML1L2 model.

Maas and Hox (2005) investigated the effect of Level 1 (i.e., 5, 30, and 50) and Level 2 (i.e., 30, 50, and 100) sample sizes on the results of a model that included a Level 1 covariate, a Level 2 covariate, and a cross-level interaction. They concluded that estimators of fixed effects, variance components, and standard errors of fixed effects are unbiased regardless of the sample-size combination, but the standard errors of the Level 2 variance components are too small when the Level 2 sample size is substantially lower than 100. Bell, Ferron, and Kromrey (2008) and Bell, Morgan, Kromrey, and Ferron (2010) found similar results. To inform the selection of sample sizes for the study, we searched research grants and contracts funded by

Institute of Education Sciences. A total of 357 abstracts emerged when the term *cluster randomized trial* was searched. Among these 357 abstracts, 131 reported the planned Level 2 sample size, with sample sizes equal to 32, 48, 80, and 140 corresponding to the 20th, 40th, 60th, and 80th percentiles, respectively, among identified studies. Planned Level 1 sample size was reported in 85 abstracts. We found that 30% of the reported cluster sizes were smaller than 6. The median cluster size was 10 and the 3rd quartile was 20. Given this review, we simulated Level 2 sample sizes of 40, 60, 80, and 140 and Level 1 sample sizes of 3, 6, 10, and 20.

Hedges and Hedberg (2007) reported that in the Early Childhood Longitudinal Study–Kindergarten Cohort, the ICCs for reading and mathematics in Grade K were .23 and .24, respectively. Results from data collected by Snyder, Hemmeter, Sandall, and McLean (2007) indicated that the average ICC across nine constructs was .22. In light of these results, we simulated $ICC_X = .20$ and $ICC_Y = .20$ or $.30$.

The values ρ_B^2 and ρ_W^2 in the current study were chosen to represent small, moderate, and strong relationships. Bloom et al. (2007) reported R_B^2 and R_W^2 values for pretest–posttest relations for a sample including approximately 28,000 students from 450 elementary schools in five different districts. District-level aggregated R_B^2 and R_W^2 values were ranged between .12 and .78. We simulated values equal to .10, .25, and .50, which are realistic given these results.

Missing data conditions were included to allow evaluation of the effect of lack of balance on power, Type I error, and coverage rates. Missing data were generated in the outcome with a missing completely at random mechanism to generate unbalanced designs. Ten percent of the cases were randomly selected to have missing data at (a) both levels (5% in each level), (b) Level 1 only, (c) Level 2 only, and (d) at neither level. ML estimation was used for all models of interest, which estimates parameters based on available data. Black et al. (2011) conducted a simulation study examining the performance of ML estimation and multiple imputation with multilevel data when values are missing at random on the dependent variable at rates 10%, 30%, and 50%, and concluded that both perform well with 10% missing data.

Analysis of Simulated Data

The four models of interest were fit to each simulated data set using ML estimation in *Mplus* 7. We used the ML estimator rather than the MLR estimator because Hox et al. (2010) showed that when the data simulation procedure does not include departures from normality and variance homogeneity in a multilevel structural equation framework, ML, compared with MLR results in more accurate standard errors unless there are at least 200 clusters.

The outcomes of interest were convergence rates, power, Type I error rates, coverage rates, and treatment effect bias. *Mplus* 7 provides information on whether the estimation terminated normally for every replication. This information is used for calculating the convergence rate. To investigate the effects of the manipulated factors on power for the 2,304 nonzero treatment effect conditions, $H_0 : \gamma_{02} = 0$ was tested

Table 2. Summary of Type I Rates by Analysis Method.

Model	Percentile				
	Minimum	25th	50th	75th	Maximum
LMML1L2	.033	.049	.054	.059	.078
LMML2	.019	.052	.059	.066	.094
OMML1L2	.033	.049	.053	.059	.080
OMML2	.033	.048	.053	.058	.080

by using $t = \hat{\gamma}_{02} / S_{\hat{\gamma}_{02}}$ with $t_{.95, N-3}$ as the one-tail critical value. The same strategy was followed for Type I error rates. We also examined accuracy of interval estimation by using the coverage of the 95% CI: $\hat{\gamma}_{02} \pm (t_{.975, N-3})S_{\hat{\theta}}$.

We attempted to use a repeated measures logistic regression model to quantify the effects of manipulated factors and analysis method on power and Type I error rates. However, estimation would not converge. Instead, we used an ANOVA model where the factors of the simulation design were the between-subjects factors and the analysis method was the within-subjects factor. The generalized eta squared (η^2_G) (Olejnik & Algina, 2003) and proportion of effect variance ($PEV_i = SS_{effect_i} / \sum SS_{effects_i}$) for each factor and interaction were calculated as effect size measures.

Results

Convergence Rates. Convergence problems were negligible. In 91% of the simulated conditions, estimation converged for all replications. In the remaining, the lowest convergence rate was 99.5%.

Type I Error Rates. Table 2 shows the summary of Type I error rates. Average Type I error rates were .054, .059, .053, and .053 for LMML1L2, LMML2, OMML1L2, and OMML2 models, respectively. All manipulated factors and interactions had a negligible effect on Type I error rates, except for one effect with a PEV larger than .05 (i.e., a five-way interaction between N , n , ICC_Y , ρ^2_b , and ρ^2_w explained 8% of the total effect variance), but Type I error rates were between .025 and .075 in all conditions and therefore can be considered acceptable.

Power to Detect the Treatment Effect. We manipulated the size of the treatment effect to achieve a target power of .7 or .9. An examination of the PEV of manipulated conditions and interactions revealed that 86% of the total effect sums of squares variance was explained by the target power factor. Table 3 shows a summary of statistical power for all conditions by methods and target power. For conditions with target power of .7, the mean statistical power was .722, .707, .742, and .740 for LMML1L2, LMML2, OMML1L2, and OMML2, respectively. With target power of .9, the mean statistical power was .902, .893, .920, and .919 for LMML1L2,

Table 3. Summary of Statistical Power by Target Power and Analysis Method.

Target power	Model	Percentiles				
		Minimum	25th	50th	75th	Maximum
.7	LMML1L2	.577	.696	.722	.748	.870
	LMML2	.581	.684	.711	.733	.796
	OMML1L2	.646	.715	.737	.764	.882
	OMML2	.646	.714	.736	.764	.869
.9	LMML1L2	.763	.886	.906	.925	.978
	LMML2	.781	.878	.898	.913	.958
	OMML1L2	.848	.903	.919	.935	.990
	OMML2	.848	.903	.919	.935	.987

Table 4. Summary of Selected Effects on Power.

Source	PEV	F value	p Value	Eta squared
Target = .7				
ρ^2_W	.36	2,065	<.001	.003
Method	.12	10,139	<.001	.001
Missing	.12	441	<.001	.001
Method $\times$ <i>n</i>	.09	2,476	<.001	.001
Target = .9				
ρ^2_W	.31	2,287	<.001	.003
Method	.15	11,950	<.001	.002
Missing	.10	467	<.001	.001
Method $\times$ <i>n</i>	.11	2,947	<.001	.001

Note. ρ^2_W within association.

LMML2, OMML1L2, and OMML2, respectively. Overall, in terms of power, OMML1L2 and OMML2 performed very similarly and slightly better than the LMMs. Furthermore, excluding the Level 1 covariate from the LMM caused only a slight decrease in power.

Because PEV for target power was so large, it may have obscured smaller effects. To address this possibility, a mixed ANOVA was conducted separately for conditions with target power = .7 and target power = .9. Results are presented in Table 4. The ANOVA models produced consistent findings for the two levels of target power indicating that ρ^2_W , method, missing data pattern, and method by cluster size (*n*) effects produced PEV larger than .05. Specifically, more than 30% of the total effect variance was explained by the effect of ρ^2_W , and each of the remaining three effects explained roughly 10%.

Table 5 shows the average power by method and cluster size. The results showed that power increased as *n* increased when LMM methods were used and power

Table 5. Average Power by Method and Cluster Size.

Model	Cluster size			
	3	6	10	20
Target = .7				
LMML1L2	.703	.728	.731	.725
LMML2	.678	.708	.721	.723
OMML1L2	.757	.744	.737	.728
OMML2	.753	.744	.737	.728
Target = .9				
LMML1L2	.877	.910	.912	.909
LMML2	.868	.893	.904	.907
OMML1L2	.927	.923	.918	.911
OMML2	.926	.923	.918	.911

decreased as n increased for OMM methods. Inspection of mean estimates of the treatment effect for conditions with no missing data indicated that holding levels of all factors in the simulation constant except n , the estimate of the treatment effect declined as n increased. The decline occurred because the simulation was designed to maintain power at either .70 or .90 for conditions with no missing data as J , n , and ρ_B^2 varied, and consequently, the effect size had to be reduced as n increased. Inspection of the mean standard errors for OMM and LMM methods shows that standard errors for both declined as n increased. However, the decline was fairly slow for OMM methods and slower than the rate of decline of treatment effect estimates. Therefore, power decreased as n increased. In contrast, the decline in the standard error for the LMM methods was faster than the decline of treatment effect estimates. Thus, when the LMM methods were used power increased as n increased. The more important result is that, for each simulated sample size, OMML1L2 and OMML2 performed similarly and provided more power than LMM methods, with a particular advantage when n was 3. LMML2 had slightly lower power than LMML1L2. At $n = 20$ power was very similar for all methods, suggesting that the choice between these methods is inconsequential with n equal to or greater than about 20.

Table 6 shows the average power by ρ_W^2 . The results indicate that power increases as ρ_W^2 increases. It is important to note that this result does not imply that adding a Level 1 covariate to a model with a Level 2 covariate will increase power. Rather, it implies that for a fixed value of ρ_B^2 power will be higher when ρ_W^2 is higher. It may seem counterintuitive that the increase in power occurs even with LMML2 and OMML2, which do not include a Level 1 covariate. Given that the simulation was designed to maintain power at either .70 or .90 for conditions with no missing data as ρ_W^2 increased and J , n , and ρ_B^2 were held constant, the treatment effect was constant for these conditions and the increase in power for LMML2 and OMML2 were due to a decline in the standard error as ρ_W^2 increased. In the case of OMML2, the

Table 6. Average Power by Method and ρ^2_W .

Model	ρ^2_W		
	.10	.25	.50
Target = .7	.700	.723	.761
Target = .9	.889	.909	.929

Note. ρ^2_W = within association.

Table 7. Average Power by Method and Missing Data Pattern.

Model	Missing data pattern			
	No missing	Missing only at Level 1	Missing only at Level 2	Missing at both
Target = .7	.749	.730	.710	.721
Target = .9	.922	.911	.897	.904

decline can be explained by noting that when the within-cluster sample size is the same for all clusters, the single-level model.

$$\bar{Y}_j = \gamma'_{00} + \gamma'_{01}\bar{X}_j + \gamma'_{02}W + r_j \tag{14}$$

yields the same estimate and standard error as OMML2 (Raudenbush & Bryk, 2002). Using results from Lüdtke et al. (2008), the residual variance for the model is $\tau^2_Y + \sigma^2_Y/n - \gamma'_{01}2(\tau^2_X + \sigma^2_X/n)$, where $\gamma'_{01} = (\tau_{XY} + \sigma_{XY}/n)/(\tau^2_X + \sigma^2_X/n)$, and σ_{XY} and τ_{XY} are the Level 1 and Level 2 covariances, respectively. The standard error of the treatment effect for OMML2 is affected by ρ^2_W even though a Level 1 covariate is not included in the OMML2 model. It is not clear why power for LMML2 increases as ρ^2_W increases, but results show that the standard error of the treatment effect for LMML2 also declines as ρ^2_W increases.

Table 7 shows the average power by missing data pattern. As expected, the highest power was achieved when there was no missing data. The greatest power loss occurred when data were missing due to only deleted clusters. The smallest power loss occurred when data were missing due to only deleted individuals. Power loss between the two extremes occurred when data were missing due to deleted cluster and individuals.

Coverage Probability. Table 8 shows the summary of coverage rates for all conditions by analysis method. Mean coverage rates were equal to the median coverage rates. Among all conditions the mean and median coverage rates were equal and .943, .938, .944, and .945 for LMML1L2, LMML2, OMML1L2, and OMML2, respectively. Coverage rates within .925 and .975 can be considered acceptable (Bradley, 1978).

Table 8. Coverage Rates for the Treatment Population Parameter.

Model	Minimum	Percentiles			Maximum
		25th	50th	75th	
LMML1L2	.909	.938	.943	.949	.970
LMML2	.886	.930	.938	.946	.990
OMML1L2	.912	.939	.944	.949	.974
OMML2	.912	.939	.945	.950	.972

The mixed ANOVA model revealed that the analysis method (PEV = .056) was the only effect with PEV larger than .05. Results indicate that the smallest coverage rates tended to occur when LMML2 was used. Approximately 14% of the conditions with LMML2 had a coverage rate below .925.

Treatment Effect Bias. Among all conditions, treatment effect bias ranged from $-.033$ to $.021$ with an approximately normal distribution and a mean of 0 for each method. Overall these results indicate that bias was small. Out of 3,456 conditions, 2,304 included a nonzero treatment effect. Among all nonzero treatment conditions and analysis method combinations, relative bias (i.e., bias divided by the population parameter) ranged from $-.048$ to $.049$ with a mean of 0. Following Hoogland and Boomsma (1998), we considered relative bias acceptable if the value was smaller than 0.05. Overall these results indicate that relative bias was acceptable in all conditions.

Illustration

We will demonstrate the multilevel models addressed in this study with data from a study of *Tools for Getting along*, which is a curriculum designed for upper elementary school students to prevent emotional and behavioral problems (Daunic et al., 2006; Smith et al., 2005; Smith, Graber, & Daunic, 2009). The sample included 2,000 students nested in 135 classrooms, with average class size of 14.81. The outcome for this example analysis is the Behavioral Regulation Index (BRI) score at the postintervention assessment. The outcome was obtained based on the BRI of Executive Function Teacher Form, a standardized instrument that consists of 86 Likert-type items (1-3) comprising 8 clinical scales. The BRI score is calculated by summing two scales as an indicator of ability to use inhibitory control to manage emotions and behavior (Gioia, Isquith, Guy, & Kenworthy, 2000). The pre-intervention BRI score was used as to generate Level 1 and Level 2 covariates. The BRI scores ranged from 20 to 60 in which a high score indicates higher risk. Details about the initial sample, setting, intervention, instruments, and study procedures can be found in Daunic et al. (2012).

Table 9. Results of the Illustration Comparing LMML1L2 to OMML1L2 and OMML2.

Model	Fixed effects (SE)			Random effects		Fit	
	$\hat{\gamma}_{02}$	$\hat{\gamma}_{01}$	$\hat{\gamma}_{10}$	$\hat{\tau}^2$	$\hat{\sigma}^2$	AIC	BIC
OMML1L2	−.090(.109)	.684 (.106)	.692 (.030)	.125	.404	1,255	1,281
OMML2	−.090(.109)	.684 (.106)	NA	.085	.797	1,620	1,642
LMML1L2	−.089(.113)	.680 (.160)	.692 (.030)	.125	.404	2,905	2,931

Note. $\hat{\gamma}_{02}$ = treatment coef.; $\hat{\gamma}_{01}$ = between coef.; $\hat{\gamma}_{10}$ = within coef.; AIC = Akaike information criterion; BIC = Bayesian information criterion.

For illustration purposes, we randomly sampled 10 students from each class that had more than 10 students from the original data in order to have equal cluster sizes ($n = 10$). The subsample includes 60 clusters—26 receiving the control condition and 34 receiving the treatment. For this subsample, the ICC was .21 and .18 for pre- and postintervention BRI scores, respectively. The correlation between pretest and posttest scores was .69. The estimates of the between- and within-group squared correlations were .37 and .49, respectively. Table 9 reports the results for OMML1L2, OMML2, and LMML1L2 using standardized variables. The models were fit using ML estimation in *Mplus* 7. Consistent with the simulation findings, the results showed that OMML1L2 and OMML2 produced the exact same estimate and standard error for the treatment effect.

The inclusion of the deviation score in OMM method caused a decrease in the Level 1 variance but also results in an increase in the Level 2 variance component. When LMML1L2 was used, the estimated treatment effect was very similar to the estimates from the other methods but, consistent with the results of the simulation, the standard error of the treatment effect was slightly larger than the standard error under either OMM approach.

Discussion

We compared four methods to estimate and test the treatment effect in CRT designs with respect to convergence, Type I error rates, power, coverage, and bias of estimates. The results showed that OMM methods were slightly more powerful than LMM methods. Inspection on the standard errors of the treatment effect revealed that on average LMM methods produced approximately 1.06 times larger standard errors within our simulation conditions. The larger standard errors are the likely source of the difference in power. These results are consistent with Lüdtke et al. (2008) who stated that the estimates of the between coefficient in the LMM approach are more variable than those of the OMM approach. However, the method by n interaction was also notable: when $n = 20$, all four methods performed equally well in terms of statistical power, but OMM performed better than LMM methods with smaller cluster sizes.

Among the manipulated factors, our results indicate that ρ_{η}^2 has the largest effect on power. We found that changes in ρ_{η}^2 affect power by reducing standard errors, regardless of whether the Level 1 deviation score is included in the model. We also found that missing data results in a decrease in power, with the largest decrease occurring when all missing data are due to clusters dropping out of the study. This was expected given that the Level 2 sample size (N) is known to have a larger impact than the Level 2 sample size (n) on the sampling variance of the treatment effect estimator (Raudenbush, 1997).

In terms of Type I error rate, all four methods performed acceptably. For OMML1L2, OMML2, and LMML1L2, the proportion of conditions with unacceptable Type I error rate was smaller than .01%. For LMML2, only 6% of the 1,152 conditions produced unacceptable Type I error rates, with almost all occurring when $n = 3$ or 6. No difference due to the analysis method was detected.

LMML2 performed less well than the other methods in terms of coverage rate. Although all four methods produced some coverage rates below the lower limit of .925, for LMML2 coverage rates below .925 were found for 14% of the simulation conditions produced, indicating that either the standard errors were too small or the distribution of $(\gamma_{02} - \hat{\gamma}_{02})/\hat{S}_{\hat{\gamma}_{02}}$ did not follow a t distribution. Further inspection revealed that the 86% of the liberal coverage rates (rates below .925) occurred when $n = 3$ or 6. The remaining three methods also produced coverage rates below the lower limit in approximately 1% of the conditions. These generally occurred when $N = 40$.

Among the models examined, coverage rates above the upper limit only occurred with the LMML2 in 2% of the conditions and almost all occurred with $n = 3$. OMM methods and LMML1L2 produced accurate estimates for the fixed treatment effect even with small sample sizes. This finding agrees with the results reported by Bell et al. (2008) and Maas and Hox (2005).

Conclusion

Including covariates in CRT designs plays an important role to increase statistical power. We compared different methods of including covariates in multilevel models for CRT data and found evidence supporting that, in general, the difference between OMM and LMM is not substantial in terms of statistical power. Especially when the reliability of the aggregated variable is high, the difference between the LMML1L2 and OMM methods in terms of statistical power is inconsequential. Even though the difference is not substantial, compared with LMML1L2, statistical power might be slightly larger with OMM methods when the cluster sizes are smaller than 6.

We found that coverage rates might be problematic with LMM methods when the cluster size is small, especially if the individual-level deviation score from the latent mean is not included in the model. Our simulation study and illustrative analysis both showed that the effect of including the deviation score as a Level 1 covariate in the models on statistical power to detect a treatment effect is negligible. However, when using Equation 11 in power analyses, researchers should keep in mind that the

proportion of between cluster variance explained by the covariate might be different than when using Equation 12. This is due to the increase in the between-level variance estimate when the deviation scores are added. Snijders and Bosker (1994) explained this issue by stating that the decrease in the Level 1 variance component produced by the inclusion of the deviation scores in the model must be balanced by an increase in Level 2 variance given that the unexplained group variability remains the same. This behavior is seen for both ML and REML estimators.

Our results are restricted to two-level CRT designs, but readers are referred to Konstantopoulos (2008, 2009) and Schochet (2005) for a discussion of three-level CRT designs. Power in multisite randomized trials was not addressed in the current study, but it has been investigated by Raudenbush and Liu (2000), and a comparison with CRT designs was presented by Moerbeek (2005). Additional research is needed to study power to detect variance or cross-level interactions with the latent mean approach. The performance of multilevel models for CRT designs with data missing at random and not at random, and with different missing data techniques, such as multiple imputation, is another important area of future research.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Note

1. The author suggested that when $n \geq 10$ and $\rho \geq .10$, including just a Level 2 covariate will have a larger impact on power when compared with including just a Level 1 covariate (group-mean centered) with the same R^2 .

References

- Algina, J., & Swaminathan, H. (2011). Centering in two-level nested designs. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 285-312). New York, NY: Taylor & Francis.
- Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 6, 267-285.
- Bell, B., Ferron, J., & Kromrey, J. (2008). Cluster size in multilevel models: The impact of sparse data structures on point and interval estimates in two-level models. *Section on Survey Research Methods*. Retrieved from <http://www.amstat.org/sections/srms/proceedings/y2008/Files/300933.pdf>
- Bell, B. A., Morgan, G. B., Kromrey, J., & Ferron, J. (2010). The impact of small cluster size on multilevel models: A Monte Carlo examination of two-level models with binary and

- continuous predictors. *Section on Survey Research Methods*. Retrieved from https://www.amstat.org/sections/srms/Proceedings/y2010/Files/308112_60089.pdf
- Black, A. C., Harel, O., & Betsy McCoach, D. (2011). Missing data techniques for multilevel data: Implications of model misspecification. *Journal of Applied Statistics*, 38, 1845-1865. doi:10.1080/02664763.2010.529882
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30-59. doi:10.3102/0162373707299550
- Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31, 144-152.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston, MA: Houghton Mifflin.
- Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology*, 108, 100-102.
- Croon, M. A., & van Veldhoven, M. J. P. M. (2007). Predicting group-level outcome variables from variables measured at the individual level: A latent variable multilevel model. *Psychological Methods*, 12, 45-57. doi:10.1037/1082-989X.12.1.45
- Daunic, A. P., Naranjo, A. H., Smith, S. W., Garvan, C. W., Barber, B. R., Becker, M. K., & Li, W. (2012). Reducing developmental risk for emotional/behavioral problems: A randomized controlled trial examining the tools for getting along curriculum. *Journal of School Psychology*, 50, 149-166. doi:10.1016/j.jsp.2011.09.003
- Daunic, A. P., Smith, S. W., Brank, E. M., & Penfield, R. D. (2006). Classroom based cognitive-behavioral intervention to prevent aggression: Efficacy and social validity. *Journal of School Psychology*, 44, 123-139.
- Donner, A. (2009). *Cluster unit randomized trials*. Retrieved from http://www.esourceresearch.org/portals/0/uploads/documents/public/donner_fullchapter.pdf
- Donner, A., & Klar, N. (2004). Pitfalls of and controversies in cluster randomization trials. *American Journal of Public Health*, 94, 416-422. Retrieved from <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1448267&tool=pmcentrez&rendertype=abstract>
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12, 121-138. doi:10.1037/1082-989X.12.2.121
- Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behavior Rating Inventory of Executive Function professional manual*. Lutz, FL: Psychological Assessment Resources.
- Graham, J. W. A. (2012). *Missing data: Analysis and design*. New York, NY: Springer.
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87. doi:10.3102/0162373707299706
- Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research*, 26, 329-367.
- Hox, J. J., Maas, C. J. M., & Brinkhuis, M. J. S. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, 64, 157-170. doi:10.1111/j.1467-9574.2009.00445.x
- Hubbard, A. E., Ahern, J., Fleischer, N. L., Van der Laan, M., Lippman, S. A., Jewell, N., & Satariano, W. A. (2010). To GEE or not to GEE: Comparing population average and mixed models for estimating the associations between neighborhood risk factors and health. *Epidemiology (Cambridge, Mass.)*, 21, 467-474. doi:10.1097/EDE.0b013e3181caeb90

- Kish, L. (1965). *Survey sampling*. New York, NY: Wiley.
- Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, 1, 265-288.
- Konstantopoulos, S. (2009). Using power tables to compute statistical power in multilevel experimental designs. *Practical Assessment, Research & Evaluation*, 14(10). Retrieved from <http://eric.ed.gov/?id=EJ933664>
- Konstantopoulos, S. (2012). The impact of covariates on statistical power in cluster randomized designs: Which level matters more? *Multivariate Behavioral Research*, 47, 392-420. doi:10.1080/00273171.2012.67389
- Kreft, I., de Leeuw, J., & Aiken, L. (1995). The effect of different forms of centering in hierarchical linear-models. *Multivariate Behavioral Research*, 30(1), 1-21.
- Longford, N. T. (2008). *Handbook of multilevel analysis*. New York, NY: Springer. doi:10.1007/978-0-387-73186-5
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203-229. doi:10.1037/a0012869
- Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86-92. doi:10.1027/1614-2241.1.3.86
- Moerbeek, M. (2005). Randomization of clusters versus randomization of persons within clusters: Which is preferable? *The American Statistician*, 59, 173-179. doi:10.1198/000313005X43542
- Moerbeek, M. (2006a). Cluster randomized trials. In H. Czichos, T. Saito, & L. Smith (Eds.), *Springer handbook of materials measurement methods* (pp. 705-718). Berlin, Germany: Springer.
- Moerbeek, M. (2006b). Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine*, 25, 2607-2617. doi:10.1002/sim.2297
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York, NY: Oxford University Press.
- Murray, D. M., Hannan, P. J., Pals, S. P., McCowen, R. G., Baker, W. L., & Blitstein, J. L. (2006). A comparison of permutation and mixed-model regression methods for the analysis of simulated data in the context of a group-randomized trial. *Statistics in Medicine*, 25, 375-388. doi:10.1002/sim.2233
- Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, 94, 423-432.
- Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.
- Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8, 434-447.
- R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173-185. doi:10.1037//1082-989X.2.2.173
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5, 199-213.

- Schochet, P. Z. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research.
- Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics*, 35(1), 26-53. doi:10.3102/1076998609345252
- Smith, S. W., Graber, J., & Daunic, A. P. (2009). Cognitive-behavioral interventions for anger/aggression: Review of research and research-to-practice issues. In M. Mayer, R. Van Acker, J. Lochman, & F. Gresham (Eds.), *Cognitive-behavioral interventions for emotional and behavioral disorders: School-based practice* (pp. 111-142). New York, NY: Guilford Press.
- Smith, S. W., Lochman, J. E., & Daunic, A. P. (2005). Managing aggression using cognitive-behavioral interventions: State of the practice and future directions. *Behavioral Disorders*, 30, 227-240.
- Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, 22, 342-363.
- Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.
- Snyder, P., Hemmeter, M. L., Sandall, S., & McLean, M. (2007). *Impact of professional development on preschool teachers' use of embedded-instruction practices*. Gainesville: University of Florida.
- Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design for longitudinal and multilevel research: Documentation for the optimal design software Version 3.0*. Retrieved from www.wtgrantfoundation.org
- Swaminathan, H., & Rogers, H. J. (2008). Estimation procedures for HLM. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 469-519). Charlotte, NC: Information Age Press.
- U.S. Congress. (2002). *Educational Sciences Reform Act*. Retrieved from <http://www2.ed.gov/policy/rschstat/leg/PL107-279.pdf>