Behavior Research Methods (2019) 51:243–257
https://doi.org/10.3758/s13428-018-1080-1

Comparison of model- and design-based approaches to detect the treatment effect and covariate by treatment interactions in three-level models for multisite cluster-randomized trials

Burak Aydin[1] · James Algina[2] · Walter L. Leite[2]

Published online: 31 July 2018

Abstract
In this study, we evaluated the estimation of three important parameters for data collected in a multisite cluster-randomized trial (MS-CRT): the treatment effect, and the treatment by covariate interactions at Levels 1 and 2. The Level 1 and Level 2 interaction parameters are the coefficients for the products of the treatment indicator, with the covariate centered on its Level 2 expected value and with the Level 2 expected value centered on its Level 3 expected value, respectively. A comparison of a model-based approach to design-based approaches was performed using simulation studies. The results showed that both approaches produced similar treatment effect estimates and interaction estimates at Level 1, as well as similar Type I error rates and statistical power. However, the estimate of the Level 2 interaction coefficient for the product of the treatment indicator and an arithmetic mean of the Level 1 covariate was severely biased in most conditions. Therefore, applied researchers should be cautious when using arithmetic means to form a treatment by covariate interaction at Level 2 in MS-CRT data.

Keywords Three-level models · Covariate by treatment interaction · Design-based · Model-based · Multisite cluster-randomized trials

The random assignment of study conditions to individuals or groups allows for the equivalence of potential outcomes in treated and control groups and provides a strong basis for causal inference of the treatment effect (Hong, 2015). In the social sciences, assigning clusters to conditions is generally more feasible than assigning individuals. Furthermore, randomizing individuals might be inappropriate when assessing the impact of interventions that naturally occur in clusters (Barbui & Cipriani, 2011; Donner & Klar, 2004). The frequency of utilizing cluster-randomized trials (CRTs), and especially multisite cluster-randomized trials (MS-CRTs), has been increasing (Bloom & Spybrook, 2017). In a CRT, clusters of participants are assigned to a condition. An MS-CRT is a type of CRT in which lower-level clusters are assigned from within levels of a higher-level cluster to at least two different levels of a condition. For example, an MS-CRT could have teachers within schools assigned to either treatment or control groups, and all students of a teacher would be in the same condition. By contrast, when the highest-level clusters are assigned to conditions, the design is not an MS-CRT and is typically referred to simply as a CRT. For example, a CRT could have several schools randomly assigned to treatment or control, resulting in a study in which all teachers in a school, and therefore all children in the school, are assigned to the same condition.

In the simplest MS-CRT there are three levels. In educational research, within-school random assignment can offer more efficient studies than random assignment of schools. For a fixed sample size, assigning teachers within a school to different study conditions offers increased statistical power to detect a treatment effect, when compared to random assignment of entire schools. For a fixed target power, random assignment of schools might require up to twice as many schools. (Bloom & Spybrook, 2017; Wijekumar, Hitchcock, Turner, Lei, & Peck, 2009). For an extensive discussion of MS-CRTs, readers are referred to Kelcey, Spybrook, Phelps, Jones, and Zhang (2017), Kraemer (2000), Raudenbush and Liu (2000), and Ruud et al. (2013). In this study we focused

✉ Burak Aydin
burak.aydin@erdogan.edu.tr

[1] School of Education, Recep Tayyip Erdogan University, Rize, Turkey

[2] School of Human Development and Organizational Studies in Education, University of Florida, Gainesville, FL, USA

on analyses of the simplest MS-CRT, in which the second-level units are randomly assigned to treatment conditions from within a third level.

In any variation of cluster-randomized designs, participants' scores within a cluster are dependent. For example, in an MS-CRT with children nested in classes nested in schools, scores for participants in a class are dependent, as are scores for participants in a school. Three main frameworks address this dependency when investigating a treatment effect: model-based, design-based, and permutation (Feng, Diehr, Peterson, & McLerran, 2001; Gardiner, Luo, & Roman, 2009; Ghisletta & Spini, 2004; Huang, 2016; Hubbard et al., 2010; Murray et al., 2006; Nevalainen, Oja, & Datta, 2017). However, for an MS-CRT targeting on detecting a covariate by treatment interaction, the choices are limited. Permutation tests, a partially model-free approach, do not allow for examining moderation effects and are used infrequently in educational research. A relatively recent approach, cluster bootstrapping, has been shown to produce results similar to those of multilevel modeling (MLM), but the results are computationally demanding, especially within a Monte Carlo simulation study (Huang, 2016). In brief, MLM, as a completely model-based approach, and generalized estimating equations (GEE) and cluster-robust standard errors (CRSE), as design-based approaches, are possible alternatives for analyzing MS-CRT data (McNeish, 2014; McNeish & Harring, 2017; McNeish & Wentzel, 2017). Among these three methods, MLM is currently the predominant method in the field of the social sciences (Bauer & Sterba, 2011; McNeish, Stapleton, & Silverman, 2017). However, CRSE offers a convenient approach, especially for the applied researchers familiar with single-level models that make fewer assumptions about random effects than MLM does. Furthermore, comparison studies have supported the use of design-based methods over model-based methods (Gardiner et al., 2009; McNeish et al., 2017; Sterba, 2009; Wu, Wang, & Pei, 2012). GEE has fewer advantages for MS-CRT data with a continuous outcome than does CRSE; furthermore, if GEE uses an independent working correlation matrix, its results are expected to be identical to those of CRSE (McNeish et al., 2017). In brief, in this study we investigated the performance of MLM and CRSE using M*plus* 7.4 (Asparouhov & Muthén, 2006; Muthén & Muthén, 2015) to detect both a treatment effect and Covariate × Treatment interactions in an MSCRT setup with a continuous outcome.

Investigating moderation effects in an MLM setting is of relatively recent interest, as compared to single-level models. The importance of the subject was emphasized by Bauer and Curran (2005) and by Preacher, Curran, and Bauer (2006). Mathieu, Aguinis, Culpepper, and Chen (2012) and Aguinis, Gottfredson, and Culpepper (2013) investigated estimation procedures for cross-level interactions. Preacher, Zhang, and Zyphur (2016) advised the examination of level-specific moderation, and explained the problems with commonly applied procedures in which moderation tests are completed without separating the lower- and higher-level effects into their orthogonal components. Studies addressing the level-specific moderation in a multilevel setting have been limited. Ryu (2015), focusing on Level 1 (L1) variables, investigated the effect of centering in a multilevel structural equation framework based on an orthogonal partitioning. One of her two simulation studies is relevant to an MS-CRT design in which the interaction between a Level 2 (L2) variable and the between-level component of an L1 variable is investigated. Ryu's approach of orthogonal partitioning does not allow for interaction between the within-level component of the L1 variable and the L2 variable, due to a homogeneous L1 covariance structure across clusters; hence, she studied the moderation effect at L2, and reported biased estimates due to the use of observed rather than latent means (Lüdtke et al., 2008) with cluster mean centering. A similar insight was provided by Preacher, Zhang, and Zyphur, who suggested using latent decomposition when investigating moderation effects. *Latent decomposition* refers to decomposing an L1 independent variable around its expected values at higher levels. For example, in a two-level design, the independent variable X_{ij} can be decomposed as $X_{ij} = \mu + (\mu_j - \mu) + (X_{ij} - \mu_j)$, where μ is the grand expected value and μ_j is the expected value of X_{ij} for the jth cluster. The mean μ_j is referred to as a *latent mean* (Lüdtke et al., 2008). However, the authors did not examine in detail the bias due to the use of observed means as covariates, and they reported that latent decomposition for a three-level model is not easily feasible. A recent work by Brincks et al. (2017), in which the authors investigated the effect of centering in three-level models, also mentioned the infeasibility of latent decomposition in three-level models. The necessity of using latent decomposition to study the main effect of an L1 reflective variable at L2 was also shown by earlier studies (Croon & van Veldhoven, 2007; Lüdtke et al., 2008; Shin & Raudenbush, 2010). However, an investigation comparing the bias of the treatment effect estimator in a two-level CRT revealed no bias when the L2 covariate comprised either observed or latent means; furthermore, the statistical power to detect the treatment effect was slightly lower when the latent means were used as the covariate and cluster sizes were small (Aydin, Leite, & Algina, 2016). Given that testing treatment effects is a primary purpose of CRTs, we addressed two questions:

> *Research Question 1:* In an MS-CRT, what are the effects of using the observed L2 means as a covariate on (a) the bias of estimation of the treatment effect and the level-specific treatment by covariate (T×C) interactions, and (b) the Type I error rate and power of the test of the treatment effect and level-specific T×C interactions?

Research Question 2: Do the effects of using L2 observed means as a covariate differ between design-based and model-based approaches?

To answer these questions, we considered the simplest MS-CRT, in which $k = 1 \ldots K$ Level 3 (L3) units are randomly selected from a population, $j = 1 \ldots J$ L2 units are randomly selected within each L3 unit and randomly assigned to a treatment or control group with equal probabilities, and $i = 1 \ldots n$ L1 units are selected in each L2 unit. At L1, an outcome (Y_{ijk}) and two covariates (X_{1ijk}, X_{2ijk}), all continuous, are assessed. The treatment indicator ($Z_{.jk}$) is an L2 binary variable. We used following terms in this three-level structure: A *site mean* refers to the arithmetic mean of all observations within the *k*th site, and a *cluster mean* refers to the arithmetic mean of all observations within the *j*th L2 cluster within the *k*th site. The article is structured as follows: We (a) briefly discuss decomposing interactions in an MS-CRT design, (b) introduce the competing approaches to analyze MS-CRT data, (c) explain our simulation study design, (d) report the results, and (e) provide an empirical example and a Discussion section.

Decomposing interactions for an MS-CRT setup

For empirical studies in which the main interest is in the treatment effect itself, investigating the interaction between a treatment indicator and a relevant variable is generally stated as an additional research question. One way to include an interaction term in a multilevel model is to multiply the treatment indicator by the relevant variable and, if necessary, decompose it into different levels by centering the product. This approach was considered by Josephy, Vansteelandt, Vanderhasselt, and Loeys (2015), but its use was criticized by Preacher, Zhang, and Zyphur (2016) because the results are uninterpretable. In this study, we investigated the treatment main effect and Covariate × Treatment interaction due to an L1 moderator (1 × (2→1) design) by decomposing the L1 predictor into between- and within-factor components and then multiplying the components by the treatment indicator. The decomposition of the covariate is $X_{1ijk} = \left(X_{1ijk} - \overline{X}_{1.ijk}\right) + \left(\overline{X}_{1.ijk} - \overline{\overline{X}}_{1.k}\right) + \overline{\overline{X}}_{1.k}$, and the product terms are $\left(X_{1ijk} - \bar{X}_{1.ijk}\right)Z_{.jk}$ at L1 and $\left(\bar{X}_{1.ijk} - \bar{\bar{X}}_{1.k}\right)Z_{.jk}$ at L2, where $\overline{X}_{1.ijk}$ and $\overline{\overline{X}}_{1.k}$ represent a cluster mean and a school mean, respectively.

Three approaches to analyze MS-CRT data

In this article, three approaches to address dependency due to the nested structure of an MS-CRT are compared. We utilized MLM, CRSE, and a combination of these two approaches. These approaches were implemented with the M*plus* 7.4 software (Muthén & Muthén, 2015). As we noted earlier, multilevel modeling is currently the predominant method among social scientists for analyzing clustered data. For extensive details on three-level models, readers are referred to Moerbeek and Teerenstra (2015), Raudenbush and Bryk (2002), and Snijders and Bosker (2012). These models can be estimated using several different software programs (e.g., SAS, HLM, and the R package nlme or lme4). Both regression coefficients and variance components can be estimated with a multilevel model. A three-level model without covariates for an MS-CRT can be written as

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + e_{ijk} \\ \beta_{0jk} &= \gamma_{00k} + \gamma_{01k} Z_{.jk} + u_{0jk} \\ \gamma_{00k} &= \pi_{000} + u_{00k} \\ \gamma_{01k} &= \pi_{010} + u_{01k} \end{aligned} \tag{1}$$

where Y_{ijk} is a continuous L1 outcome and $Z_{.jk}$ is the binary treatment indicator at L2. Fixed effects are represented by π; specifically, π_{000} is the intercept and π_{010} is the treatment main effect. The L1 random effect (e_{ijk}), L2 random effect (u_{0jk}), and L3 random effects for the intercept (u_{00k}) and the treatment (u_{01k}) are assumed to be normally distributed with a mean of 0 and a covariance matrix *T*:

$$T = \begin{bmatrix} \sigma^2 & & & \\ 0 & \tau_{\beta_{0jk}} & & \\ 0 & 0 & \tau_{\gamma_{00k}} & \\ 0 & 0 & \tau_{\gamma_{00k},\gamma_{01k}} & \tau_{\gamma_{01k}} \end{bmatrix} \tag{2}$$

where σ^2 is the within-cluster variance component, $\tau_{\beta_{0jk}}$ is the variance component due to clusters within sites, $\tau_{\gamma_{00k}}$ represents the variance due to sites, $\tau_{\gamma_{01k}}$ is the treatment effect variance between sites, and $\tau_{\gamma_{00k},\gamma_{01k}}$ is the covariance between the site-specific means and the site-specific treatment effects. The statistical power to detect the treatment effect in Eq. 1 varies as a function of the magnitude of π_{010}, the sample size, and the variance at each level including $\tau_{\gamma_{01k}}$ (Bloom & Spybrook, 2017; Spybrook et al., 2011, p. 86). When covariates are added to Eq. 1, the statistical power changes due to the adjustment in π_{010} and the variance components. The effect of the covariates on the conditional variance depends on the strength of the correlation between the covariates and the outcome, as well as on the correlation between the covariates.

CRSEs can account for dependency due to clusters. As an example of calculating CRSEs, consider a residual-based estimator for standard errors of a single-level model estimated from two-level data. The standard errors can be computed using a sandwich estimator (see, e.g., Raudenbush & Bryk, 2002, p. 277) to produce robust standard errors with large samples. According to Raudenbush and Bryk, the procedure allows for approximately correct tests and confidence

intervals, even when the residual for the single-level model is not normally distributed. For relatively small sample sizes, further modifications might be needed (McNeish & Stapleton, 2016).

Several variations of CRSEs are available, at least for a single-level model (McNeish, 2014; McNeish & Harring, 2017; Raudenbush & Bryk, 2002). M*plus* provides a CRSE procedure referred to as type = complex, which entails fitting a single-level model to data and correcting the standard errors for clustering at a higher level. Asparouhov (2005) describes the procedure for complex sampling with stratification, clustering, and sampling weights. In the following discussion, we adapt Asparouhov's description of a design without stratification or sampling weights, the type of design we investigated. Estimates are obtained by maximizing a likelihood defined assuming independence of the observations. Thus, when there are no sampling weights and no stratification, the estimates are ML estimates of a single-level model. Let l be the likelihood based on the independence assumption, l_{ij} the contribution to the likelihood by the ith individual in cluster j, $z_j = \sum_i \partial\left(\log\left(l_{ij}\right)\right) / \partial\theta$, z the average of the z_i, and L" the matrix of the second derivatives of log$l$. The asymptotic covariance matrix of the estimates is given by $(J/(J-1))L^{''-1}\sum_j \left(z-z_j\right)\left(z-z_j\right)^T L^{''-1}$, where J is the number of clusters and T is the transpose operator.

A combination of CRSE and MLM can be employed to analyze a three-level data structure by accommodating the first two levels with MLM and the third level with CRSE (McNeish & Wentzel, 2017; Rabe-Hesketh & Skrondal, 2006). Hence, instead of extensive model building at L3, the researchers can focus on a less complex model while accounting for the third-level clustering. M*plus* provides a combination procedure that is referred to as type = complex twolevel. Asparouhov and Muthén (2006) have described the procedure to compute standard errors in a multilevel setup that allows for stratification and sampling weights. Assuming no stratification and sampling weights equal to 1 at both L1 and L2, as in this study, the procedure can be described as follows, which we have adapted from Asparouhov and Muthén. Let l_{jk} be the likelihood of the observed data for the jth L2 unit nested in the kth L3 unit, $l = \prod_{j,k} l_{jk}$, $L = \log(l)$, and $L_{jk} = \log(l_{jk})$, CRSEs are computed using $(L^{''})^{-1}\mathrm{Var}(L^{'})(L^{''})^{-1}$, where the $'$ and $''$ refer to the first and second derivatives, respectively, of the log likelihood, and the second term Var($L^{'}$) is equal to $\mathrm{Var}\left(\sum_{k,j} L_{jk}\right)$. According to Asparouhov and Muthén, the last term "is computed according to the formulas for the variance of the weighted estimate of the total described in Cochran, Chapter 11 (1977) taking the appropriate design into account" (p. 2719).

Monte Carlo simulation study

The data generation process for the MS-CRT data was completed using R (R Core Team, 2016). The simulated datasets were analyzed using M*plus* 7.4 (Muthén & Muthén, 2015). The results from M*plus* outputs were investigated in a mixed analysis of variance (ANOVA) framework. These three main steps of the Monte Carlo simulation study are presented in this section. The data generation model was a three-level model, presented in Eq. 3:

$$
\begin{aligned}
Y_{ijk} &= \beta_{0jk} + \beta_{1jk}\left(X_{1ijk}-\mu_{1jk}\right) + \beta_{2jk}\left(X_{1ijk}-\mu_{1jk}\right)Z_{.jk} \\
&\quad + \beta_{3jk}\left(X_{2ijk}-\mu_{2jk}\right) + e_{ijk} \\
\beta_{0jk} &= \gamma_{00k} + \gamma_{01k}Z_{.jk} + \gamma_{02k}\left(\mu_{1jk}-\mu_{1k}\right) + \gamma_{03k}\left(\mu_{1jk}-\mu_{1k}\right)Z_{.jk} \\
&\quad + \gamma_{04k}\left(\mu_{2jk}-\mu_{2k}\right) + u_{0jk} \\
\beta_{1jk} &= \gamma_{10k} \\
\beta_{2jk} &= \gamma_{20k} \\
\beta_{3jk} &= \gamma_{30k} \\
\gamma_{00k} &= \pi_{000} + \pi_{001}(\mu_{1k}-\mu_1) + \pi_{002}(\mu_{2k}-\mu_2) + u_{00k} \\
\gamma_{10k} &= \pi_{100} \\
\gamma_{20k} &= \pi_{200} \\
\gamma_{30k} &= \pi_{300} \\
\gamma_{01k} &= \pi_{010} + u_{01k} \\
\gamma_{02k} &= \pi_{020} \\
\gamma_{03k} &= \pi_{030} \\
\gamma_{04k} &= \pi_{040}
\end{aligned}
\tag{3}
$$

where for each of the two L1 covariates s = 1, 2, μ_s is the grand mean, $\omega_{sk} = \mu_{sk} - \mu_s$ is the random effect for school k with variance component τ_{ω_s}, $\xi_{sjk} = \mu_{sjk} - \mu_{sk}$ is the random effect for class j in school k with variance component $\tau_{\xi s}$ and $R_{sijk} = X_{sijk} - \mu_{sjk}$ is the random effect for individual i in class j in school k with variance component $\sigma^2_{R_S}$. Each of the random effects has a mean of 0. The grand means for the covariates were set equal to 0. The decomposition of the variance of a covariate is $\sigma^2_{X_{sijk}} = \tau_{\omega_s} + \tau_{\xi s} + \sigma^2_{R_S}$. The variance $\sigma^2_{X_{sijk}}$ was set to 1, so that $\tau_{\omega_s} = ICC_{X_S-L3}$, $\tau_{\xi_s} = ICC_{X_S-L2}$, and $\sigma^2_{R_S} = 1-(ICC_{X_S-L2} + ICC_{X_S-L3})$, where ICC_{X_S-L3} and ICC_{X_S-L2} are the intraclass correlation coefficients for X_s. The variance components were equal for the two covariates, and the correlation coefficient for each component of the covariates is provided in Table 1.

The data generation process was completed through the following steps: Simulate (a) the L3 components of covariates $\begin{pmatrix}\omega_{1.k}\\ \omega_{2.k}\end{pmatrix}$ from $N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix} ICC_{X1-L3} & \\ Cov(\omega_{1.k},\omega_{2.k}) & ICC_{X2-L3}\end{bmatrix}\right)$; (b) the L2 components of covariates $\begin{pmatrix}\xi_{1jk}\\ \xi_{2jk}\end{pmatrix}$ from $N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{bmatrix} ICC_{X1-L2} & \\ Cov(\xi_{1jk},\xi_{2jk}) & ICC_{X2-L2}\end{bmatrix}\right)$; (c) L 1

Table 1 Simulation symbols and population parameters

Factor	Symbol	Main	Add-1	Add-2	Add-3
L1 sample size	n	6,10	20	10	10
L2 sample size	J	6,10	10, 20	10	10
L3 sample size	K	20,40	40, 60	40	40
L2 ICC_X	$ICC_{X1–L2}$	.08, .16	.08, .16	.08, .16	.16
L3 ICC_X	$ICC_{X1–L3}$	.09	.09	.06, .12	.09
L2 ICC_Y	ICC_{Y-L2}	.08, .16	.16	.08, .16	.08,.16
L3 ICC_Y	ICC_{Y-L3}	.06, .12	.12	.06, .12	.06, .12
L2 int. comp.	π_{30} *or* π_{TX-L2}	0, .20	0, .20, .30	0, .20	.20
L1 int. comp.	π_{200} *or* π_{TX-L1}	0, .20	0, .20, .30	.20	.20
Treatment	π_{010} *or* π_T	0, .15	.15	.15	.15
L3 Treatment variance	$\tau_{\pi 01k}$	.05	.05	.05, .20	0, .05, .20
Correlation X_1 and X_2	Cor(X_1, X_2)	.30, .70	.50	.30, .70	.50
X_1 and X_2 effect	π_{001}, π_{002}, π_{020}, π_{040}, π_{100}, π_{300}	.20	.20	.20 at L1–3 .10, .20 at L2	.20
L3 correlation	Cor(u_{00k}, u_{01k})	.20	.20	.20, .40	0,.30, .70
Competing models		1–3	1–3	2 only	1–3
Total		1,024	72	1,024	36

Add = Additional simulation study, int. comp. = interaction component. Also notice that the L1 interaction component corresponds to a cross-level interaction.

components of covariates $\begin{pmatrix} R_{1ijk} \\ R_{2ijk} \end{pmatrix}$ from $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} 1-ICC_{X1-L2}+ICC_{X1-L3} & \\ Cov(R_{1ijk}, R_{2ijk}) & 1-ICC_{X2-L2}+ICC_{X2-L3} \end{bmatrix}\right)$; (d) the binary treatment indicator to randomly assign half of the L2 units within an L3 unit to treatment and half to control; (e) the L3 random components of the outcome variable $\begin{pmatrix} u_{00k} \\ u_{01k} \end{pmatrix}$ from $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau_{\pi 00k} & \\ \tau_{\pi 00k,\pi 01k} & \tau_{\pi 01k} \end{bmatrix}\right)$, where $\tau_{\pi 00k} = ICC_{Y-L3}$ and $ICC_{Y–L3}$ is the L3 ICC for Y, conditional on the treatment indicator, the covariates, and the products of the treatment indicator and covariates; (f) for each L3 unit, the L2 random component u_{ojk} from $N(0, ICC_{Y-L2})$; and finally (g) for each L2 unit, the L1 random component e_{ijk} from $N(0, 1-(ICC_{Y-L3}+ICC_{Y-L2}))$.

The mixed model for generating the data is

$$\begin{aligned} Y_{ijk} = {} & \pi_{000} + \pi_{001}(\mu_{1k}-\mu_1) + \pi_{002}(\mu_{2k}-\mu_2) + \pi_{010}Z_{.jk} \\ & + \pi_{020}\left(\mu_{1jk}-\mu_{1k}\right) + \pi_{030}\left(\mu_{1jk}-\mu_{1k}\right)Z_{.jk} \\ & + \pi_{040}\left(\mu_{2jk}-\mu_{2k}\right) + \pi_{100}\left(X_{1ijk}-\mu_{1jk}\right) \\ & + \pi_{200}\left(X_{1ijk}-\mu_{1jk}\right)Z_{.jk} + \pi_{300}\left(X_{2ijk}-\mu_{2jk}\right) \\ & + u_{00k} + u_{0jk} + u_{01k}Z_{.jk} + e_{ijk} \end{aligned} \quad (4)$$

Conditions and population parameters

Estimating a three-level model for 1,000 replications of each combination of factors (i.e., a condition) was computationally demanding; hence, we conducted four separate simulation studies, one main and three additional simulations. The parameters for these simulations are summarized in Table 1. The first and second additional simulations aimed to explore if the main study findings were consistent with larger sample sizes and factors that were not manipulated in the main simulation. The third additional simulation aimed to provide additional insight on model comparisons under relatively complex covariance structures.

For the main simulation study, the sample size combinations were selected on the basis of reviewing 357 projects funded by Institute of Education Sciences (Aydin et al., 2016); the L1 and L2 sample sizes were six and ten; the L3 sample sizes were 20 and 40. The ICC values for L2 (.08 and .16) and L3 (.09) were chosen in light of a meta-analysis on variance decomposition of academic achievement data due to schools and districts (Hedges & Hedberg, 2013). The correlations between the components of the two L1 covariates were equal across levels and were set so that the total correlation was either .30 or .70, to represent relatively small and large relationship strengths, respectively. The magnitude of the fixed effects for the treatment (.15), its interactions with the covariates (.20 and .30 at both L1 and L2), and the covariate (.20 for both L1 and L2) were chosen on the basis of the representative application studies from the decade of the 2000s, summarized by Mathieu, Aguinis, Culpepper, and Chen (2012). Given that we fixed the variance of X to 1 in our simulation studies, these magnitudes can be considered standardized effect sizes. The effect size variability magnitude, $\tau_{\pi 01k} = .05$, was chosen

on the basis of the Optimal Design software (Spybrook et al., 2011, p. 98).

Competing models

We examined the performance of three competing models. The default estimator for each model was maximum likelihood estimation with robust standard errors (MLR[1]). Syntax for fitting the competing models using M*plus* 7.4 is provided in the Appendix. Our first model, referred as M3L and presented in Eq. 5, is a three-level model, as is Eq. 4, but an observed mean decomposition of the covariates is employed: group mean centering of the continuous independent variables represented by $\left(X_{ijk}-\overline{X}_{jk}\right)$ at L2, and $\overline{X}_{jk}-\overline{\overline{X}}_{..k}$ at L3.[2] Fixed effects are represented by π ; specifically, π_{001} and π_{002} are the L3 effects, π_{010} and π_{040} are the L2 effects, π_{100} and π_{300} are the L1 effects of the covariates X_1 and X_2, respectively. Given that the treatment effect and the covariate by treatment interaction are of greater interest in a CRT than are the coefficients for the covariates, the fixed effect of treatment (π_{020}, referred below as π_T), the L2 interaction component (π_{030} or $\pi_{TX\text{-}L2}$), and the L1 interaction component (π_{200} or $\pi_{TX\text{-}L1}$) are investigated in detail.

$$\begin{aligned} Y_{ijk} = {} & \pi_{000} + \pi_{001}\bar{\bar{X}}_{1..k} + \pi_{002}\bar{\bar{X}}_{2..k} + \pi_{010}Z_{.jk} + \pi_{020}\left(\bar{X}_{1.jk}-\bar{\bar{X}}_{1..k}\right) \\ & + \pi_{030}\left(\bar{X}_{1.jk}-\bar{\bar{X}}_{1..k}\right)Z_{.jk} + \pi_{040}\left(\bar{X}_{2.jk}-\bar{\bar{X}}_{2..k}\right) \\ & + \pi_{100}\left(X_{1.jk}-\bar{X}_{1.jk}\right) + \pi_{200}\left(X_{1ijk}-\bar{X}_{1.jk}\right)\left(Z_{.jk}\right) \\ & + \pi_{300}\left(X_{2ijk}-\bar{X}_{2.jk}\right) + u_{00k} + u_{0jk} + u_{01k}Z_{.jk} + e_{ijk} \end{aligned} \quad (5)$$

Model 2, referred as M2L-C, is a two-level model and thus includes variables at L1 and L2. The two-level complex procedure, which corrects the standard error for clustering at L3, was used to estimate the parameters and to carry out hypothesis testing. The M2L-C model can be obtained from Eq. 5 by deleting u_{00k} and u_{01k}. Model 3, referred as M1L-C, is a single-level model and can be obtained from Eq. 5 by deleting u_{00k}, u_{01k} and u_{0jk}. The complex procedure, which corrects the standard error for clustering at L3, was used to estimate the parameters and to carry out hypothesis testing. All three models included the same fixed effects. In particular, Model 3, as well as Models 1 and 2, includes the L1 interaction component $\pi_{TX\text{-}L1}$ and the L2 interaction component $\pi_{TX\text{-}L2}$. Model 1 includes variance components at L1, L2, and L3; Model 2 includes variance components at L1 and L2; and Model 3 includes only an L1 variance.

Analysis of the simulation results

We focused on three coefficients and their standard errors: $\hat{\pi}_T$, $\hat{\pi}_{TX-L1}$ and $\hat{\pi}_{TX-L2}$. The convergence rate, the ratio of normally terminated estimations to the total number of replications was 100% for each condition. We examined coverage rates, coefficient bias, relative bias of the standard error, power and Type I error rates. We used a mixed-design ANOVA model; factors of the simulation design were treated as between-subjects factors and the analysis method was treated as the within-subjects factor. We conducted 15 separate analyses of variance (ANOVAs), one for each combination of the coefficients, on the one hand, and for coverage rate, coefficient bias, relative standard error bias, Type I error rate, and power, on the other. The dependent variables in these analyses were:

- Coefficient bias—The dependent variable was $\hat{\theta}-\theta$ coefficient bias was calculated as the average of $\hat{\theta}-\theta$ over replications of a condition.
- Coverage rate—The dependent variable was an indicator variable for the 95% confidence interval (CI) in a replication of a condition: $\hat{\theta} \pm (z_{.975})S\left(\hat{\theta}\right)$, where θ was π_T, π_{TX-L1}, *or* π_{TX-L2}. The coverage rate was calculated as the percentage of intervals that contained θ, and rates within .925 and .975 were considered acceptable (Bradley, 1978).
- Relative standard error bias—The dependent variable was $SE\left(\hat{\theta}\right) = \left[SE\left(\hat{\theta}\right)-SD\left(\hat{\theta}\right)\right]/SD\left(\hat{\theta}\right)$, where $SD\left(\hat{\theta}\right)$ is the standard deviation of the parameter estimate across all replications of a condition (Bandalos & Leite, 2013). The relative standard error bias was calculated as the average of $\left[SE\left(\hat{\theta}\right)-SD\left(\hat{\theta}\right)\right]/SD\left(\hat{\theta}\right)$ over replications of a condition. Following Hoogland and Boomsma (1998), we considered the relative bias of the standard errors acceptable if the average over replications were between – 0.1 and 0.1.
- Type I error rate—For conditions in which $\theta = 0$, the dependent variable was an indicator variable for whether $z = \hat{\theta}/SE\left(\hat{\theta}\right)$ did not result in rejection of $H_0 : \theta = 0$, with $\pm z_{.975}$ as the critical value. The Type I error rate was calculated as the proportion of replications in which $H_0 : \theta = 0$ was not rejected, and rates within .025 and .075 were considered acceptable (Bradley, 1978).
- Power—For conditions in which $\theta \neq 0$, the dependent variable was an indicator variable for whether $z =$

[1] Conventional maximum likelihood (ML), one of the main estimation methods for multilevel modeling, is not available with the type = complex option in M*plus*, whereas restricted maximum likelihood (REML) currently is not an option in M*plus* at all (see McNeish, 2017).

[2] This approach corresponds to CWC1/CWC2 as described in Brincks et al. (2017).

$\hat{\theta}/SE\left(\hat{\theta}\right)$ did result in rejection of $H_0 : \theta = 0$, with $\pm z_{.975}$ as the critical value. Power was calculated as the proportion of replications in which $H_0 : \theta = 0$ was rejected.

Generalized η^2 (Olejnik & Algina, 2003) was calculated as the effect size measure, and effects with generalized $\eta^2 < .001$ were not interpreted.

Results

Main simulation study

The main simulation study was completed in approximately 2,088 hours, divided across six computers, each of which had 16 GB RAM and a 3.70-GHz central processing unit. As is reported in Table 1, a total of ten between-subjects factors were manipulated, and each factor had only two levels, resulting in 2^{10} = 1,024 conditions. Estimation converged for all 1,024 × 1,000 replications.

Coefficient bias The average coefficient bias across iterations for $\hat{\pi}_T$ ranged between – .010 and .009, with a mean of 0; these values did not change across different models. The η^2 values were all smaller than .001. Similarly, the coefficient bias for $\hat{\pi}_{TX-L1}$ was acceptable and ranged between – .006 and .006

The coefficient bias for $\hat{\pi}_{TX-L2}$ included large values and ranged between – .14 and .14, with a mean of 0; these values did not change across different models. The mixed-design ANOVA revealed substantial effects of π_{TX-L1} (η^2 = .060) and π_{TX-L2} (η^2 = .059), and relatively weak effects for (a) the π_{TX-L1} by $\mathrm{ICC}_{\mathrm{X1–L2}}$ interaction (η^2 = .003), (b) the π_{TX-L2} by $\mathrm{ICC}_{\mathrm{X1–L2}}$ interaction (η^2 = .003), (c) the π_{TX-L1} by n interaction (η^2 = .001), and (d) the π_{TX-L2} by n interaction (η^2 = .001). The results in Table 2 indicate the sources of these effects. Bias occurred when $\pi_{TX-L1} \neq \pi_{TX-L2}$. The magnitude of the bias differed after the third decimal place across the different models, and it decreased as n and $\mathrm{ICC}_{\mathrm{X1–L2}}$ increased. The direction of the bias was positive when $\pi_{TX-L2} > \pi_{TX-L1}$, and negative otherwise.

Coverage rates The mean coverage rates across iterations for M2L-C and M1L-C were similar to each other and differed generally only after the third decimal place, but M3L had slightly different rates. The ranges of coverage rates for π_T were [.906, 960] for the M3L and [.910, .963] for the M2L-C and M1L-C. The mixed-design ANOVA did not result in any values of $\eta^2 > .001$. For M3L, 21% of all 1,024 simulated conditions resulted in mean coverage rates lower than .925; for the other two models, low coverage rates were observed in approximately 9% of the conditions. Among all low-coverage conditions, 91% occurred with K = 20.

Table 2 Average biases of $\hat{\pi}_{TX-L2}$ for the main study

π_{TX-L2}	π_{TX-L1}	n	$\mathrm{ICC}_{\mathrm{X1–L2}}$	M3L	M2L-C	M1L-C
0	0	6	.08	.000	.000	.000
0	0	6	.16	.000	.000	.000
0	0	10	.08	.000	.000	.000
0	0	10	.16	.000	.001	.001
0	.2	6	.08	– .124	– .126	– .126
0	.2	6	.16	– .087	– .088	– .088
0	.2	10	.08	– .101	– .102	– .102
0	.2	10	.16	– .063	– .064	– .064
.2	0	6	.08	.127	.128	.128
.2	0	6	.16	.087	.088	.088
.2	0	10	.08	.102	.102	.102
.2	0	10	.16	.064	.064	.064
.2	.2	6	.08	.001	.002	.002
.2	.2	6	.16	.000	.000	.000
.2	.2	10	.08	.001	.002	.002
.2	.2	10	.16	.001	.001	.001

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

A similar pattern was observed for π_{TX-L1}: The ranges of coverage rates were [.898, .965] for M3L, and [.904, .966] for the other two models. For M3L, 21% of all 1,024 simulated conditions had mean coverage rates lower than .925, and for the other two models this percentage was 8%. Again, low-coverage conditions occurred largely (91%) when K = 20.

The coverage rates were more problematic for π_{TX-L2}, with ranges in coverage rates equal to [.751, .953], [.782, .957], and [.782, .958] for M3L, M2L-C, and M1L-C, respectively. Only 23% of the simulated conditions resulted in acceptable mean coverage rates. The mixed-design ANOVA resulted in η^2 = .006 for the π_{TX-L1} by π_{TX-L2} interaction. Table 3 shows the ranges in coverage rates for π_{TX-L2} as a function of model. The coverage ranges when π_{TX-L1} and $\hat{\pi}_{TX-L2}$ were equal were [.893, .953], [.900, .957], and [.900, .958]. These rates are similar to those reported for π_T and π_{TX-L1}, and rates lower than .925 occurred largely (81%) when K = 20.

Relative bias of the standard errors The results for 16 replications out of 1,024,000 had standard errors larger than 10. All of these outlying standard error estimates were for M3L.[3] For $\hat{\pi}_T$, the median relative bias of standard errors across 1,000 replications of the 1,024 conditions were in the range [– .113, .047] for M3L, and the range [– .094, .059] for M2L-C and M1L-C. The mixed-design ANOVA did not reveal any

[3] Outlying standard errors occurred for $\hat{\pi}_{TX-L2}$, and four of these 16 replications also had outlying standard errors for $\hat{\pi}_T$.

Table 3 Coverage rates for π_{TX-L2} for the main study

π_{TX-L2}	π_{TX-L1}	M3L	M2L-C	M1L-C
0	0	[.899, .953]	[.900, .957]	[.900, .957]
0.2	0	[.751, .922]	[.782, .934]	[.782, .934]
0	0.2	[.767, .922]	[.806, .928]	[.806, .928]
0.2	0.2	[.893, .951]	[.901, .955]	[.901, .958]

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

interpretable effects, and – .113 was observed when $K = 20$ and $J = 6$. A similar pattern was observed for $\hat{\pi}_{TX-L1}$: [– .115, .045] with M3L, and [– .092, .060] with the other two models; – .11 was observed when $K = 20$ and $J = 6$. All four models performed similarly when K was larger. Table 4 reports the mean relative biases of the standard errors after removing the 16 outliers.

For $\hat{\pi}_{TX-L2}$, the ranges for the median relative standard error bias were [– .141, .041], [– .151, .022], and [– .150, .028] for Models 1–3, respectively. The mixed-design ANOVA did not reveal interpretable effects. Out of the 1,024 manipulated conditions, median values lower than – .10 occurred in 98, 120, and 119 conditions for Models 1–3, respectively, mainly when $K = 20$ and $J = 6$. Table 5 reports the mean values after removing the 16 outliers. These results indicated that on average, standard errors were slightly underestimated when K was smaller and that J had a larger effect when K was smaller.

Power and Type I error rate The average empirical powers to detect the treatment effect were similar across the different models: .533 for M3L and .523 for M2L-C and M1L-C. The sample size at each of the three levels and the ICC_{Y-L2} factors had $\eta^2 > .001$. The largest effect size was for K, $\eta^2 = .061$. The mean power was .402 for $K = 20$, and .65 for $K = 40$. The effects of sample size at L1 and L2 were smaller: $\eta^2 = .004$ for n, with means of .49 and .56 for $n = 6$ and $n = 10$, respectively, and $\eta^2 = .017$ for J, with means of .46 and .59 for $J = 6$ and $J = 10$, respectively. The effect size for ICC_{Y-L2} was $\eta^2 = .006$, with means of .56 and .49 for $ICC_{Y-L2} = .08$ and .16, respectively. The larger value of ICC_{Y-L2} indicates a larger conditional variance for Y, and this accounts for the reduction in power.

Table 4 Mean relative biases of the standard errors for the main study

Coefficient	K	M3L	M2L-C	M1L-C
$\hat{\pi}_T$	20	– .041	– .023	– .023
	40	– .020	– .009	– .009
$\hat{\pi}_{T-L1}$	20	– .041	– .016	– .016
	40	– .019	– .006	– .006

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

Table 5 Mean relative biases of the standard errors for $\hat{\pi}_{TX-L2}$ for the main study

K	J	M3L	M2L-C	M1L-C
20	6	– .067	– .073	– .072
20	10	– .051	– .055	– .055
40	6	– .040	– .038	– .036
40	10	– .030	– .031	– .029

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

The average Type I error rates for the treatment effect were in the ranges [.040, .094], [.039, .090], and [.039, .090], with mean values of .067, .062, and .063 for Models 1–3, respectively. No effect had $\eta^2 > .001$; however, 20% of the conditions resulted in Type I error rates larger than .075 for M3L, and 8% for the other two models. Among the conditions with large Type I error rates, 90% occurred when $K = 20$.

The empirical powers to detect π_{TX-L1} were similar across models: .926 for M3L, and .920 for the other models. The mixed design ANOVA revealed six effects with $\eta^2 > .001$, each involving sample size: (a) $\eta^2 = .051$ for K, (b) $\eta^2 = .037$ for J, (c) $\eta^2 = .028$ for n, (d) $\eta^2 = .015$ for the K by n interaction, (e) $\eta^2 = .010$ for the K by J interaction, and (f) $\eta^2 = .006$ for the J by n interaction. Table 7 reports the empirical power to detect π_{TX-L1} as a function of sample size. In addition, $\eta^2 = .001$ for ICC_{Y-L2}, with means equal to .913 for .08 and .930 for .16, and also for ICC_{X1-L2}, with means equal to .930 for .08 and .914 for .16. Averaged across iterations, the Type I error rates when testing $\pi_{TX-L1} = 0$ were in the ranges

Table 6 Empirical power to detect $\pi_T = .15$ for the main study

K	J	n	ICC_{Y-L2}	M3L	M2L-C	M1L-C
20	6	6	.08	.357	.346	.346
20	6	6	.16	.310	.298	.298
20	6	10	.08	.425	.411	.411
20	6	10	.16	.345	.333	.333
20	10	6	.08	.467	.452	.452
20	10	6	.16	.410	.394	.394
20	10	10	.08	.535	.521	.521
20	10	10	.16	.460	.446	.446
40	6	6	.08	.583	.575	.575
40	6	6	.16	.505	.497	.497
40	6	10	.08	.672	.665	.665
40	6	10	.16	.565	.558	.558
40	10	6	.08	.733	.726	.726
40	10	6	.16	.658	.650	.650
40	10	10	.08	.801	.795	.795
40	10	10	.16	.708	.701	.701

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

Table 7 Empirical power to detect π_{TX-L1} for the main study

K	J	n	M3L	M2L-C	M1L-C
20	6	6	.703	.686	.686
20	6	10	.907	.899	.899
20	10	6	.886	.876	.876
20	10	10	.987	.985	.985
40	6	6	.933	.930	.930
40	6	10	.995	.995	.995
40	10	6	.993	.992	.992
40	10	10	1.000	1.000	1.000

The empirical power for $K = 40$, $J = 10$, and $n = 10$ was 1, and the other values of empirical power were near 1. Therefore, we conducted the same analyses on z values for all conditions combined, and separately for $K = 20$. The results were essentially the same, and the model effect was absent. Furthermore, we investigated Type I error rates separately for $K = 40$, and they were on average .061 for M3L, and .058 for the other models.

[.035, .102] and [.034,.096], and on average they were .067 for M3L and .062 for the other models. No effect had $\eta^2 >$ .001; however, 17% of the conditions resulted in Type I error rates larger than .075 for M3L, and 7% for the other two models. Among these conditions with large Type I error rates, 92% occurred when $K = 20$.

The L2 interaction effect, $\hat{\pi}_{TX-L2}$, was estimated without bias only when $\pi_{TX-L2} = \pi_{TX-L1}$, and therefore we studied empirical power for the 216 conditions in which $\pi_{TX-L2} = \pi_{TX-L1} = .20$. The effect size η^2 was at least .001 for sample size at L2 (η^2 = .017), L3 (η^2 = .015), and the interaction effect for the L2 and L3 sample sizes (η^2 = .001). The model had η^2 = .005. The empirical power rates in Table 8 show that power increases as either the L1 or L2 sample size increases, and the impact of the L1 sample size is larger when the L2 sample size is larger. In addition, M3L had a relatively larger power than the other models, which had very similar powers. An increase in a covariate's ICC at L2 resulted in a larger power (η^2 = .005), with mean empirical power rates equal to .27 for M3L and .22 for the other models when ICC_{X-L2} = .08, and .35 for M3L and .28 for the other models when ICC_{X-L2} = .16. An increase in conditional variance for Y at L2 resulted in lower empirical power (η^2 = .003), with mean empirical power rates equal to .34 for M3L and .27 for the other models when ICC_{Y-L2} = .08, and .28 for M3L and .23 for the other models when ICC_{Y-L2} = .16. Type I error rates for testing $\pi_{TX-L2} = 0$ were in the ranges [.047, .101] and [.043, .100], and on average they were .074 for M3L and .072 for the other models. No effect had $\eta^2 >$ .001; however, 45% of the conditions resulted in Type I error rates larger than .075 for M3L, and 33% for the other two models. Among these conditions with large Type I error rates, 80% occurred when $K = 20$.

Table 8 Empirical power to detect π_{TX-L2} for the main study when $\pi_{TX-L2} = \pi_{TX-L1}$

K	J	ICC_{Y-L2}	ICC_{X1-L2}	M3L	M2L-C	M1L-C
20	6	.08	.08	.190	.159	.159
20	6	.08	.16	.236	.192	.192
20	6	.16	.08	.160	.134	.134
20	6	.16	.16	.197	.170	.170
20	10	.08	.08	.288	.227	.226
20	10	.08	.16	.363	.281	.281
20	10	.16	.08	.231	.195	.195
20	10	.16	.16	.293	.240	.240
40	6	.08	.08	.278	.212	.211
40	6	.08	.16	.356	.278	.278
40	6	.16	.08	.224	.191	.190
40	6	.16	.16	.284	.234	.234
40	10	.08	.08	.451	.346	.345
40	10	.08	.16	.581	.437	.437
40	10	.16	.08	.360	.286	.285
40	10	.16	.16	.460	.373	.373

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

Additional simulation studies

The estimation for M3L was computationally demanding and slow, and this prevented manipulating more than two levels for each factor. On the basis of the main study findings, we conducted three additional small scale simulation studies to explore further (a) $\hat{\pi}_{TX-L2}$ bias, (b) the effect of the parameters that were not manipulated in the main study, and (c) model comparisons with relatively more complex variance structures.

Additional Simulation Study 1 Table 2 indicated that π_{TX-L2} was estimated with bias for some conditions. The amount of bias varied, mainly due to the interaction magnitude at both levels (i.e., π_{TX-L1} and π_{TX-L2}) and to the covariate's ICC at L2. Thus, in our first additional study we manipulated the interaction magnitude to be 0, .2, and .3 at both levels, and the ICC to be .08 and .16. We also increased the sample size at all levels: n = 20, J = 10, 20, and K = 40, 60. Consistent with the main study results, substantial bias was detected when $\pi_{TX-L1} \neq \pi_{TX-L2}$, even with larger sample sizes; the results are reported in Tables 9 and 10.

Additional Simulation Study 2 The main study did not reveal substantial differences due to model choice, except for the empirical power difference to detect L2 interactions. In our second additional simulation study, focused on the parameter estimates, we aimed to explore the effects of the factors that were not manipulated in the main study. A total of 1,024

Table 9 The average biases of $\hat{\pi}_{TX-L2}$ for Additional Study 1

π_{TX-L2}	π_{TX-L1}	ICC_{X1-L2}	M3L	M2L-C	M1L-C
0	0	.08	– .002	– .003	– .003
0	0	.16	.004	.005	.005
0	.2	.08	– .071	– .072	– .072
0	.2	.16	– .039	– .038	– .038
0	.3	.08	– .102	– .101	– .101
0	.3	.16	– .058	– .057	– .057
.2	0	.08	.068	.066	.066
.2	0	.16	.037	.036	.036
.2	.2	.08	– .003	– .003	– .003
.2	.2	.16	– .001	.001	.001
.2	.3	.08	– .037	– .034	– .034
.2	.3	.16	– .019	– .019	– .019
.3	0	.08	.101	.102	.102
.3	0	.16	.057	.057	.057
.3	.2	.08	.032	.034	.034
.3	.2	.16	.020	.020	.020
.3	.3	.08	.003	.003	.003
.3	.3	.16	.003	.003	.003

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

conditions were generated (see Table 1) and analyzed only with M2L-C. Consistent with the main study results, the coefficient estimate was unbiased for $\hat{\pi}_T$ and $\hat{\pi}_{TX-L1}$; the coefficient bias for the $\hat{\pi}_{TX-L2}$ varied as a function of π_{TX-L2} (η^2 = .075), ICC_{X1-L2} (η^2 = .004), and the π_{TX-L2} by ICC_{X1-L2} interaction (η^2 = .004), but not as a function of the newly added factors. The relative biases of the standard errors for $\hat{\pi}_T$, $\hat{\pi}_{TX-L1}$ and $\hat{\pi}_{TX-L2}$ were all acceptable: – .01, – .01, and – .03 on average, respectively.

Additional Simulation Study 3 We designed our final additional simulation study to investigate the model effect under a wider range of L3 variances of the treatment random effect ($\tau_{\pi 01k}$) and of the L3 covariances between the L3 random effect and the L3 treatment random effect ($\tau_{\pi 00k,\ \pi 01k}$). We expected these new conditions to affect the results for π_T, but we also report results for π_{TX-L1} and π_{TX-L2}. The coverage rates for π_T, π_{TX-L2}, and π_{TX-L2} were all 94%, and there was no bias for the estimates. The relative bias of the standard errors for these three parameter estimates were within the range [– .075, .054] and were acceptable for each of the 36 manipulated conditions.

The empirical power to detect the treatment effect was varied as a function of $\tau_{\pi 01k}$ (η^2 = .174), ICC_{Y-L2} (η^2 = .007), and the $\tau_{\pi 01k}$ by ICC_{Y-L2} interaction (η^2 = .001). Under the manipulated conditions, these results were expected (Bloom & Spybrook, 2017; Spybrook et al., 2011, p. 86). Table 11 reports the average power for these factors.

Table 10 The average biases of $\hat{\pi}_{TX-L2}$ by sample size when $\pi_{TX-L1} \neq \pi_{TX-L2}$ for Additional Study 1

K	J	π_{TX-L2}	π_{TX-L1}	M3L	M2L-C	M1L-C
40	10	0	.2	– .055	– .054	– .054
40	10	0	.3	– .078	– .078	– .078
40	10	.2	0	.055	.054	.054
40	10	.2	.3	– .029	– .03	– .029
40	10	.3	0	.078	.078	.078
40	10	.3	.2	.022	.022	.023
40	20	0	.2	– .056	– .055	– .056
40	20	0	.3	– .084	– .083	– .083
40	20	.2	0	.052	.051	.052
40	20	.2	.3	– .029	– .029	– .028
40	20	.3	0	.079	.081	.081
40	20	.3	.2	.023	.022	.023
60	10	0	.2	– .052	– .056	– .056
60	10	0	.3	– .077	– .076	– .077
60	10	.2	0	.05	.048	.049
60	10	.2	.3	– .025	– .024	– .023
60	10	.3	0	.079	.079	.079
60	10	.3	.2	.032	.033	.033
60	20	0	.2	– .057	– .055	– .055
60	20	0	.3	– .079	– .078	– .079
60	20	.2	0	.051	.051	.051
60	20	.2	.3	– .028	– .027	– .027
60	20	.3	0	.081	.079	.079
60	20	.3	.2	.028	.026	.027

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

The empirical power to detect π_{TX-L1} was high; hence, we examined the effects of conditions on z values. The method effect was small (η^2 = .008); the z values varied as a function of ICC_{Y-L2} (η^2 = .016) and ICC_{Y-L3} (η^2 = .009). Table 12 reports mean z values. M2L-C and M1L-C produced average z values that are equal and slightly smaller than those for M3L. An increase in conditional variance at L2 and L3, which

Table 11 Empirical power to detect π_T for Additional Study 3

$\tau_{\pi 01k}$	ICC_{Y-L2}	M3L	M2L-C	M1L-C
0	.08	.957	.955	.955
0	.16	.868	.863	.863
.05	.08	.798	.792	.792
.05	.16	.712	.706	.706
.20	.08	.471	.467	.467
.20	.16	.441	.435	.435

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

Table 12 Average of z values (estimate/*SE*) for $\hat{\pi}_{TX-L1}$ for Additional Study 3

$ICC_{Y\text{-}L2}$	$ICC_{Y\text{-}L3}$	M3L	M2L-C	M1L-C
.08	.06	5.784	5.711	5.711
.08	.12	6.003	5.927	5.927
.16	.06	6.078	6.001	6.001
.16	.12	6.343	6.263	6.263

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

resulted in a decreased L1 variance, was associated with larger z values.

The empirical power to detect $\hat{\pi}_{TX-L2}$ varied as a function of model choice (η^2 = .018), ICC_{Y-L2} (η^2 = .007), $\tau_{\pi 01k}$ (η^2 = .002), and ICC_{Y-L3} (η^2 = .001). Moreover, three different two-way interactions affected the results, of model by ICC_{Y-L3}, ICC_{Y-L2}, and $\tau_{\pi 01k}$ each with η^2 = .001. Table 13 reports average power for these factors. M2L-C and M1L-C performed similarly under the manipulated conditions, and all three models performed similarly when the variance components at L3 were smaller. The power difference between M3L and the other models reached its maximum with larger L3 variance but smaller L2 variance.

Illustration

We selected a subsample of 22 schools from the data for an ongoing early childhood education project. Each school had three control and three intervention classrooms, and each classroom had three children who met the criteria for risk of developing emotional/behavioral disorders. The total sample size was 396. The outcome measure was selected to be School Readiness Composite (SRC-post) scores at the postintervention. The preintervention SRC score, the Social Awareness Composite (SAC) score at L1, and the binary treatment indicator at L2 served as independent variables. We standardized SRC and SAC scores to have a grand mean of 0 and a standard deviation of 1. The classroom and school arithmetic means of SRC and SAC were added into the models, along with the SRC L1 and L2 deviation by treatment interactions, so that the models investigated in the simulations could be estimated.

The correlation between the treatment indicator and preintervention scores was 0, as we would expect from an MS-CRT design. The correlation between SRC-post and SRC-pre scores was .57; that between SRC-pre and SAC was .38. Three-level empty models revealed that the L1 variance components were .81, .87, and .82; the L2 variance components were .13, .10, and .12; and the L3 variance components were .06, .04, and .07, for SRC-post, SRC-pre, and SAC, respectively. Table 14 reports the results from the models addressed in this study; all were estimated using MLR in M*plus*. Consistent with the simulation studies, all three models provided similar results to detect the treatment effect and the L1 interaction component. The estimates for the L2 interaction coefficient and its standard error were the same for M2L-C and M1L-C, and slightly smaller than for M3L. We also tested a two-level model, M2L, in which we ignored L2, listed all covariates at L1, and declared L3 as the second level. As expected, with M2L the standard error estimates were different, due to the distribution of L2 variances over the bottom and top levels (Moerbeek, 2004).

Table 13 The empirical power to detect π_{TX-L2} for the Additional Study 3

$ICC_{Y\text{-}L2}$	$\tau_{\pi 01k}$	$ICC_{Y\text{-}L3}$	M3L	M2L-C	M1L-C
.08	0	.06	.595	.503	.503
.08	0	.12	.598	.440	.440
.08	.05	.06	.584	.457	.457
.08	.05	.12	.610	.383	.383
.08	.2	.06	.592	.402	.402
.08	.2	.12	.603	.352	.352
.16	0	.06	.446	.401	.401
.16	0	.12	.474	.367	.367
.16	.05	.06	.478	.394	.394
.16	.05	.12	.471	.344	.344
.16	.2	.06	.463	.353	.353
.16	.2	.12	.459	.311	.311

M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex model

Discussion and conclusion

Motivated by Bloom and Spybrook (2017) observation that the use of MS-CRT has been increasing, we compared

Table 14 Illustration results

Fixed Effects	M3L	M2L-C	M1L-C	M2L
$\hat{\pi}_T$	.051 (.063)	.051 (.064)	.051 (.064)	.051 (.062)
$\hat{\pi}_{TX-L1}$	.151 (.098)	.151 (.100)	.151 (.100)	.151 (.098)
$\hat{\pi}_{TX-L2}$	.274 (.161)	.242 (.156)	.242 (.156)	.294 (.171)
Random Effects				
$\hat{\sigma}^2$	.492	.492	.599	.579
$\hat{\tau}_{\beta_{0jk}}$	.093	.107	NA	Ignored
$\hat{\tau}_{\pi 00k}$	.023	NA	NA	.019
$\hat{\tau}_{\pi 01k}$	.01	NA	NA	NA
$\hat{\tau}_{\pi 00k,\pi 01k}$	– .01	NA	NA	NA

Standard errors are in parentheses. M3L = three-level model, M2L-C = two-level complex model, M1L-C = single-level complex, M2L = two-level model

three methods to estimate and test the treatment effect and covariate by treatment interaction due to L1 covariate moderation both at L1 and L2 with respect to convergence, Type I error rates, power, coverage and bias of estimates. In this section, we discuss our findings in threefold: (a) L2 interaction, (b) treatment main effect and L1 interaction component, and (c) comparison of competing models. We then present the limitations of the present study.

The results showed that, regardless of the model choice, L2 interaction estimates were biased unless the magnitude of the L1 and L2 interaction were equal to L1. The bias was upward when $\pi_{TX-L2} > \pi_{TX-L1}$ and downward otherwise and bias was larger when difference between π_{TX-L2} and π_{TX-L1} was larger. The magnitude of bias was as large as .13 for a population value of .20. This is clearly unacceptable, and the problem of bias persisted even with larger sample sizes studied in this article. For example, in our Additional Simulation Study 1 (see Table 10), the magnitude of bias was .051 for a population value of .20 even when $K = 60$ and $J = 20$; it was .050 when $K = 60$ and $J = 10$; .052 when $K = 40$ and $J = 20$; and .055 when $K = 40$ and $J = 10$. As was reported by Ryu (2015) and mentioned by Preacher, Zhang, and Zyphur (2016), this bias is due to unreliability (or sampling error) of the aggregated L1 covariate. Furthermore, the amount of bias parallels with the results calculated using bias derivation formula for two-level models given by Lüdtke et al. (2008). A possible solution is to use a latent decomposition; however, in a three-level model it is not easily computable (Brincks et al., 2017; Preacher et al., 2016), and to date it is not possible to compute with the M*plus* software.[4] We included conditions with $\pi_{TX-L2} = \pi_{TX-L1}$ in the simulations and found that the estimates of L2 interaction were not biased. The coverage rates, Type I error rates and relative bias of the standard errors were mainly acceptable when $\pi_{TX-L2} = \pi_{TX-L1}$ and $K = 40$ with all three competing models under the manipulated conditions in the main study and additional studies; however, when $K = 20$, slightly unacceptable rates and relative bias values occurred, especially with M3L. The empirical power to detect an unbiased L2 interaction increased with larger sample sizes and larger ICC_{X-L2} values. The results for M3L are consistent with Dong, Kelcey, and Spybrook (2017, Eq. 28). In terms of model comparison, M3L was slightly more powerful than the other three models when detecting an unbiased L2 interaction but these results are in alignment with slightly underestimated standard errors. The difference between M3L and the other models reached its maximum with larger L3 variance but smaller L2 variance. McNeish and Wentzel (2017) reported comparison between a M3L with small sample adjustment and M2L-C under a relatively simpler covariance matrix than was included in the present study and small L3 sample size (four, seven, and ten). The authors could not compare the power difference between these two models due to poor performance of M2L-C in terms of biased variance estimates, but they also noted that the poor performance was less severe with small L3 variance. Furthermore, our results also emphasize the importance of small sample adjustment given that slightly underestimated standard errors occurred more often for M3L when $K = 20$.

When $K = 40$, our results indicate that the coefficient bias for $\hat{\pi}_T$ and $\hat{\pi}_{TX-L1}$ was absent for all four models, Type I error rates were also acceptable. The coverage rates for population parameters of π_T and π_{TX-L1} were between .925 and .975 under 79% of the conditions in the main study; the remaining 21% had coverage rates ranged between .898 and .925. The poor performance in terms of coverage rates occurred mainly (91%) with $K = 20$ and was due to slight downward bias in standard errors. This finding is also consistent with McNeish and Wentzel's (2017) study in which they utilized ML estimation; in our study we were limited to MLR given that M*plus* does not offer ML with CRSE. The poor performance of MLR than of ML in a multilevel structural equation framework with no assumption violations was reported by Hox, Maas, and Brinkhuis (2010). The downward bias in standard errors was slightly larger for M3L than for the other models, but the difference among the competing models vanished with a larger L3 sample size. Using REML or REML with the Kenward–Roger correction (Kenward & Roger, 2009) could be a possible remedy for M3L's poor performance with small L3 sample sizes (McNeish, 2017; McNeish & Wentzel, 2017). These alternatives, however, are not available with M*plus*. The empirical power to detect a non-zero π_T or π_{TX-L1} did not change across competing models and generally increased as a function of sample size at L3 and then L2. These findings are expected for M3L as shown by Dong, Kelcey, and Spybrook (2017, Eq. 50), Spybrook et al. (2011, p. 86), and Bloom and Spybrook (2017).

In addition to the model comparisons above, one interesting outcome of this study is that M1L, at least with the M*plus* software, provided roughly the same results as M2L-C under the conditions of this study and the specifications of the two models. The specifications included correctly including the L1 and L2 components of interaction. If the model used in the program implementing M1L-C had only included a total product term—that is, $X_{1ijk}Z_{jk}$ and its coefficient—estimates of the coefficient would likely not be equal to either $\hat{\pi}_{TX-L1}$ or $\hat{\pi}_{TX-L2}$ obtained from the program we used to implement M1L-C in the simulations. If, alternatively, the model used in the program implementing M1L had only included only the product term $\left(X_{1ijk}-\overline{X}_{1jk}\right)\left(Z_{jk}\right)$ and its coefficient, estimates of the coefficient would likely be equal to $\hat{\pi}_{TX-L1}$ obtained from the program we used to implement M1L-C in the simulations. Similar to results reported by McNeish, Stapleton, and

[4] Confirmed by Bengt Muthén on the M*plus* discussion forum; see www.statmodel.com/discussion/messages/12/9389.html?1490056072#POST127856.

Silverman (2017), our results also support the use of design-based methods (M1L-C) or the combination of design-based and model-based methods (M2L-C) as an alternative to completely model-based methods even with a complex multilevel design, in our case an MS-CRT. Our illustration section agrees with this development. A note of caution is due here: Raudenbush and Bloom (2015) related the MS-CRT to "a fleet of experiments" that is particularly useful to study effect heterogeneity. According to Raudenbush and Bloom, investigating both mean program impact and impact heterogeneity is a necessary step when moving forward from a field study to public policy, program theory or professional practice. One dimension of the impact heterogeneity can be studied by examining the L3 random effects and design-based analyses removes the possibility of detecting this type of heterogeneity. Another point is that M2L in the illustration study ignores the intermediate level, instead treating all covariates at L1 and adjusting for L3 clustering only, and thus produced different results than M1L, since it is known that ignoring a level is problematic (Moerbeek, 2004). Therefore, M2L-C should be used rather than M2L.

As is true for all simulations and illustrations, there were some limitations to our study. Our results were restricted to a balanced MS-CRT without any missing data and with all assumptions satisfied. We focused on only three parameters, $\hat{\pi}_T$, $\hat{\pi}_{TX-L1}$, and $\hat{\pi}_{TX-L2}$. We were also limited to a single estimator implemented in M*plus*. Furthermore, in order to study model comparison with unbiased estimates of L2 interaction we set $\pi_{TX-L2} = \pi_{TX-L1}$ for a substantial portion of the conditions in each of the four simulations. Another limitation is that M3L demanded computational power, and therefore we refrained from combining the main study and the small simulation study conditions. Even though they occurred for only 0.0015% of the main study conditions, we observed outlying standard error estimates with M3L for the L2 interaction. To address this limitation, we reported both mean and median relative standard error biases.

Appendix

OSmean refers to observed school mean, OCmean refers to observed classroom mean centered around the school mean, Ocnt refers to L1 deviation scores centered around the observed classroom mean.

TITLE: M3L

```
DATA: dataname.csv;
VARIABLE:
   NAMES = cid2 y trt schid OSmeanx OSmeanx2 OCmeanx
OCmeanx2 Ocntx Ocntx2 IntL1 IntL2;
   CLUSTER = schid cid2;
   WITHIN = Ocntx Ocntx2 IntL1;
   BETWEEN = (cid2) OCmeanx OCmeanx2 trt IntL2
(schid) OSmeanx OSmeanx2;
DEFINE:
   IntL1 = trt*Ocntx;
   IntL2 = trt*OCmeanx;
ANALYSIS: TYPE = THREELEVEL RANDOM;
MODEL:
   %WITHIN%
   y ON Ocntx Ocntx2 intl1;
   %BETWEEN cid2%
   s2 | y ON trt;
   y ON OCmeanx OCmeanx2 intl2;
   %BETWEEN schid%
   y ON OSmeanx OSmeanx2;
   y with s2;
```

TITLE: M2L-C

```
DATA: dataname.csv;
VARIABLE:
   NAMES = cid2 y trt schid OSmeanx OSmeanx2 OCmeanx
OCmeanx2 Ocntx Ocntx2 IntL1 IntL2;
   CLUSTER = schid cid2;
   WITHIN = Ocntx Ocntx2 IntL1;
   BETWEEN = OCmeanx OCmeanx2 trt IntL2 OSmeanx
OSmeanx2;
DEFINE:
   IntL1 = trt*Ocntx;
   IntL2 = trt*OCmeanx;
ANALYSIS: TYPE = TWOLEVEL COMPLEX ;
MODEL:
   %WITHIN%
   y ON Ocntx Ocntx2 intL1;
   %BETWEEN%
   y ON trt Ocmeanx Ocmeanx2 intL2 Osmeanx Osmeanx2;
```

TITLE: M1L-C

```
DATA: dataname.csv;
VARIABLE:
   NAMES = cid2 y trt schid OSmeanx OSmeanx2 OCmeanx
OCmeanx2 Ocntx Ocntx2 IntL1 IntL2;
   CLUSTER = schid;
DEFINE:
   IntL1 = trt*Ocntx;
   IntL2 = trt*Ocmeanx;
ANALYSIS: TYPE = COMPLEX;
MODEL:
   y ON Ocntx Ocntx2 intL1 trt Ocmeanx Ocmeanx2 intL2
OSmeanx OSmeanx2;
```

References

Aguinis, H., Gottfredson, R. K., & Culpepper, S. A. (2013). Best-practice recommendations for estimating cross-level interaction effects using multilevel modeling. *Journal of Management*, *39*, 1490–1528.

Asparouhov, T. (2005). Sampling weights in latent variable modeling. *Structural Equation Modeling*, *12*, 411–434.

Asparouhov, T., & Muthén, B. O. (2006). Multilevel modeling of complex survey data. Los Angeles, CA: ASA Section on Survey Research Methods. Available from www.statmodel.com

Aydin, B., Leite, W. L., & Algina, J. (2016). The effects of including observed means or latent means as covariates in multilevel models for cluster randomized trials. *Educational and Psychological Measurement*, *76*, 803–823.

Bandalos, D. L., & Leite, W. L. (2013). Use of Monte Carlo studies in structural equation modeling research. In G. R. Hancock & R. O. Mueller (Eds.), Structural equation modeling: A second course (2nd ed.) (pp. 564–666). Greenwich, CT: Information Age.

Barbui, C., & Cipriani, A. (2011). Cluster randomised trials. *Epidemiology and Psychiatric Sciences*, *20*, 307–309.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: Performance of alternative specifications and methods of estimation. *Psychological Methods*, *16*, 373–390. doi: https://doi.org/10.1037/a0025813

Bauer, D., & Curran, P. (2005). Probing interactions in fixed and multilevel regression: Inferential and graphical techniques. *Multivariate Behavioral Research*, *40*, 373–400. https://doi.org/10.1207/s15327906mbr4003_5

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*, 877–902. https://doi.org/10.1080/19345747.2016.1271069

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Brincks, A. M., Enders, C. K., Llabre, M. M., Bulotsky-Shearer, R. J., Prado, G., & Feaster, D. J. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research*, *52*, 149–163. https://doi.org/10.1080/00273171.2016.1256753

Cochran, W. G. (1977). Sampling techniques. New York, NY: Wiley.

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. https://doi.org/10.1037/1082-989X.12.1.45

Dong, N., Kelcey, B., & Spybrook, J. (2017). Power analyses for moderator effects in three-level cluster randomized trials. *Journal of Experimental Education*, *86*, 489–514. https://doi.org/10.1080/00220973.2017.1315714

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

Feng, Z., Diehr, P., Peterson, A., & McLerran, D. (2001). Selected statistical issues in group randomized trials. *Annual Review of Public Health*, *22*, 167–187.

Gardiner, J., Luo, Z., & Roman, L. (2009). Fixed effects, random effects and gee: What are the differences? *Statistical Medicine*, *28*, 221–239. https://doi.org/10.1002/sim.3478

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics*, *29*, 421–437.

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*, 445–489.

Hong, G. (2015). Causality in a social world: Moderation, mediation, and spill-over. West Sussex, UK: Wiley-Blackwell.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods & Research*, *26*, 329–367. https://doi.org/10.1177/0049124198026003003

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.

Huang, F. L. (2016). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and Psychological Measurement*, *78*, 297–318. https://doi.org/10.1177/0013164416678980

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*, *21*, 467–474.

Josephy, H., Vansteelandt, S., Vanderhasselt, M.-A., & Loeys, T. (2015). Within-subject mediation analysis in ab/ba crossover designs. *International Journal of Biostatistics*, *11*, 1–22.

Kelcey, B., Spybrook, J., Phelps, G., Jones, N., & Zhang, J. (2017). Designing large-scale multisite and cluster-randomized studies of professional development. *Journal of Experimental Education*, *85*, 389–410.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics and Data Analysis*, *53*, 2583–2595.

Kraemer, H. C. (2000). Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin*, *26*, 533–541.

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. https://doi.org/10.1037/a0012869

Mathieu, J. E., Aguinis, H., Culpepper, S. A., & Chen, G. (2012). Understanding and estimating the power to detect cross-level interaction effects in multilevel modeling. *Journal of Applied Psychology*, *97*, 951–966. https://doi.org/10.1037/a0028380

McNeish, D. M. (2014). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological Methods*, *19*, 552–563. https://doi.org/10.1037/met0000024

McNeish, D. (2017). Multilevel mediation with small samples: A cautionary note on the multilevel structural equation modeling framework. *Structural Equation Modeling*, *24*, 609–625.

McNeish, D. M., & Harring, J. R. (2017). Clustered data with small sample sizes: Comparing the performance of model-based and design-based approaches. *Communications in Statistics: Simulation and Computation*, *46*, 855–869.

McNeish, D., & Stapleton, L. M. (2016). Modeling clustered data with very few clusters. *Multivariate Behavioral Research*, *51*, 495–518. https://doi.org/10.1080/00273171.2016.1167008

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods*, *22*, 114–140. https://doi.org/10.1037/met0000078

McNeish, D., & Wentzel, K. R. (2017). Accommodating small sample sizes in three-level models when the third level is incidental. *Multivariate Behavioral Research*, *52*, 200–215. https://doi.org/10.1080/00273171.2016.1262236

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

Moerbeek, M., & Teerenstra, S. (2015). Power analysis of trials with multilevel data. Boca Raton, FL: CRC 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.

Muthén, L.K., & Muthén, B.O. (1998–2015). Mplus user's guide (7th ed.). Los Angeles, CA: Muthén & Muthén.

Nevalainen, J., Oja, H., & Datta, S. (2017). Tests for informative cluster size using a novel balanced bootstrap scheme. *Statistics in Medicine*, *36*, 2630–2640. https://doi.org/10.1002/sim.7288

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. https://doi.org/10.1037/1082-989X.8.4.434

Preacher, K. J., Curran, P. J., & Bauer, D. J. (2006). Computational tools for probing interactions in multiple linear regression, multilevel modeling, and latent curve analysis. *Journal of Educational and Behavioral Statistics*, *31*, 437–448.

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, *21*, 189–205. https://doi.org/10.1037/met0000052

R Core Team. (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from www.R-project.org/

Rabe-Hesketh, S., & Skrondal, A. (2006). Multilevel modelling of complex survey data. *Journal of the Royal Statistical Society: Series A*, *169*, 805–827.

Raudenbush, S. W., & Bloom, H. S. (2015). Learning about and from a distribution of program impacts using multisite trials. *American Journal of Evaluation*, *36*, 475–499.

Raudenbush, S. W., & Bryk, A. S. (2002). Hierarchical linear models: Applications and data analysis methods (2nd ed., Vol. 1). Thousand Oaks, CA: Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199–213. https://doi.org/10.1037/1082-989X.5.2.199

Ruud, K. L., LeBlanc, A., Mullan, R. J., Pencille, L. J., Tiedje, K., Branda, M. E., . . . Montori, V. M. (2013). Lessons learned from the conduct of a multisite cluster randomized practical trial of decision aids in rural and suburban primary care practices. *Trials*, *14*, 267. https://doi.org/10.1186/1745-6215-14-267

Ryu, E. (2015). The role of centering for interaction of level 1 variables in multilevel structural equation models. *Structural Equation Modeling*, *22*, 617–630. https://doi.org/10.1080/10705511.2014.936491

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*, 26–53.

Snijders, T. A. B., & Bosker, R. J. (2012). Multilevel analysis: An introduction to basic and advanced multilevel modeling. Los Angeles, CA: Sage.

Spybrook, J., Bloom, H., Congdon, R., Hill, C., Martinez, A., & Raudenbush, S. (2011). *Optimal design plus empirical evidence: Documentation for the "optimal design" software* (Software manual). Retrieved from http://hlmsoft.net/od/od-manual-20111016-v300.pdf

Sterba, S. K. (2009). Alternative model-based and design-based frameworks for inference from samples to populations: From polarization to integration. *Multivariate Behavioral Research*, *44*, 711–740. https://doi.org/10.1080/00273170903333574

Wijekumar, K., Hitchcock, J., Turner, H., Lei, P., & Peck, K. (2009). A multisite cluster randomized trial of the effects of compass-learning odyssey [r] math on the math achievement of selected Grade 4 students in the mid-Atlantic region (Final report. NCEE 2009-4068). Washington, DC: National Center for Education Evaluation and Regional Assistance.

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