Power to Detect Moderated Effects with Random Slopes in Partially Nested Designs

Kyle Cox
University of North Carolina at Charlotte

Ben Kelcey
University of Cincinnati

Hannah Luce
Oak Ridge Institute for Science and Education

Investigations of differential treatment effects through the inclusion of moderating variables advance individualized treatment decisions through improved knowledge of treatment effectiveness across individuals and contexts. Literature has explored planning strategies for partially nested designs with heterogeneous treatment effects but statistical power formulas for detecting moderated effects when the moderator-outcome relationship varies across clusters (i.e., random slopes) are currently unavailable. We derive these formulas and probe the roles of the governing parameters and likely scale required for detecting these moderation effects. Simulation study results suggest that substantial moderation effect heterogeneity warrants much larger sample sizes to consistently detect moderation effects. Power formulas are implemented in `R` (see Supplemental Materials).

For a range of treatments frequently seen in educational research settings collaboration among participants in a cluster (e.g., small-group reading instruction) or a shared facilitator (e.g., patients with the same therapist) can produce dependencies among participant outcomes (Lohr et al. 2014; Roberts & Roberts, 2005; Sterba, 2017). If ignored, standard modelling typically underestimates the standard error of treatment effect estimates and leads to inflated Type I error (Baldwin et al., 2011; Bauer et al., 2008; Candlish et al., 2018; Pals et al., 2008). Despite these analytic complications, administering cluster or shared facilitator interventions is often more effective because participant collaborations frequently bolster treatment effects. The advantages of these cluster and shared facilitator interventions have led to widespread adoption and growing implementation (Raudenbush et al., 2007; Spybrook, Shi et al., 2016; Sterba, 2017).

A common complication in these types of experimental studies arises when randomization splits individuals into either a clustered treatment group or an unclustered control group (e.g., waitlist). This type of data structure is often referred to as partial nesting because it results in cluster-based dependence among individuals in the treatment group but independence among individuals in the control group. Partially nested data is common in the fields of education, psychotherapy, and counselling (Bauer et al., 2008; Lohr et al., 2014; Sterba, 2017). For example, studies have compared a school-based intervention for autism and an individualized home-based control group (Roberts et al., 2011) and a group-therapy for high-risk adolescents against an individual therapy control group (Dishion et al., 2001). Partially nested study designs have also been utilized to study interventions involving cognitive behavioral therapy (Johnson et al., 2007) and other behavioral therapies (Powell et al., 2010; Rothschild et al., 2012).

A comprehensive collection of literature has been developed to address planning and analysis with a broad range of partially nested structures (Bauer et al., 2008; Cox & Kelcey, 2022; Cox et al., 2022; Lohr et al., 2014; Sterba, 2017; Roberts & Roberts, 2005; Lee & Thompson, 2005; Moerbeek & Wong, 2008; Roberts et al., 2016; Kelcey et al., 2020). For example, prior literature has established that many partially nested designs can be analyzed through extensions of multiple-arm multilevel models (Roberts et al., 2011; Sterba et al., Lachowicz et al., 2015). Literature has also outlined several study planning considerations including sample size and power formulas for various partially nested designs (Moerbeek & Wong, 2008; Roberts & Roberts, 2005; Li & Hedeker, 2017), main, moderation and mediation effects (Cox & Kelcey, 2022; Cox et al., 2022; Kelcey et al., 2020), continuous and binary outcomes (Roberts et al., 2016), and optimized sample allocation strategies (Moerbeek & Wong, 2008). Recent literature has also considered a broad range of partially nested data structures including three-level hierarchical structures and multisite structures (Cox et al., 2022; Kelcey et al., 2020; Li & Hedeker, 2017).

Of particular interest in the field of education are techniques that better elucidate for whom and under what conditions an intervention, policy or, program (i.e., treatment) is effective to inform more targeted intervention decisions. Literature has detailed design and analysis strategies including benefiting subgroup identification using the credible subgroups approach (Lazar et al., 2016) and the use of subpopulation

treatment effect pattern plots to identify treatment effect heterogeneity (Schnell et al., 2018). We focus on the inclusion of individual-level moderator variables to investigate heterogeneous treatment effects. Power formulas and sample size requirements for detecting moderation effects have been developed for two- and three-level cluster randomized trials with binary and continuous moderators and multiple moderators (Dong et al., 2018; Dong et al., 2021; Spybrook et al., 2016; Yang et al., 2020). More recently, power formulas for detecting moderation effects have been extended to partially nested designs (Cox & Kelcey, 2022; Cox et al., 2022).

An important additional consideration in the design of partially nested studies targeting the effects of individual-level moderators is the extent to which the relationship between the individual-level moderator and the outcome varies across clusters or groups (i.e., random slope coefficient for the moderator variable). Consider the potential moderating role of gender on the effectiveness of a small-group depression therapy. Under a fixed effect approach (i.e., nonrandom slope) the relationship between gender and the depression outcome varies across clusters only as a function of treatment status whereas under a random effect approach (i.e., random slope) the relationship between gender and the depression outcome potentially varies across clusters (e.g., slope or coefficient for gender varies across therapy groups after accounting for treatment effect). Strategies for planning studies that include moderation effects (fixed or random moderator slope) have been widely developed in fully nested or cluster randomized trials (Dong et al., 2018; Dong et al., 2021; Mathieu, 2012) but only for moderators with fixed slopes in partially nested designs (Cox & Kelcey, 2022; Cox et al., 2022).

Omitting heterogeneity or assuming a fixed slope for the moderator may lead to spurious inferences about moderation because it will underestimate the standard errors (Heisig & Schaeffer, 2019; LaHuis et al., 2020). That is, ignoring variability in the relationship between the moderator and outcome across groups makes detecting the moderation effect more likely (i.e., Type I error) when, in fact no moderation effect is present. From a design perspective, assuming a fixed slope for the moderator will generally underestimate the adequate sample size for detecting the moderation effect. Additionally, previous literature involving fully nested designs (Dong et al., 2021) found the relationship between design parameters (e.g., individual- and cluster sample size) and power to detect the moderation effect can depend on moderation effect heterogeneity. Put differently, guidance and recommendations for detecting moderators with fixed slopes in partially nested designs (e.g., Cox & Kelcey, 2022) may not be applicable when the moderator has a random slope.

The primary purpose of this study is to derive power formulas for detecting moderation effects when the moderator-outcome relationship varies across clusters (i.e., random slopes) in partially nested designs and assess their accuracy using Monte Carlo simulations. Within this context, we developed statistical power formulas that accommodate binary or continuous moderators assessed at the individual-level, are applicable with unbalanced sample sizes across study arms, and can incorporate other covariates. A secondary purpose of this study is to outline the roles of the parameters governing power to detect these moderation effects. Using examples, we highlight key results related to power and adequate sample sizes while varying several influential factors including cluster and individual per cluster sample size, moderation effect heterogeneity, variance structure of the outcome (i.e., ICC), distribution of the moderator, and variance explained in the outcome by covariates (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2018; Dong et al., 2021; Spybrook et al., 2016; Yang et al., 2020).

The article is divided into sections to map out statistical power formulas based on analytic models with and without covariates. Formula accommodations for binary moderators and study arms with unbalanced sample sizes are also noted. A simulation study follows the presentation of the analytic models to assess the accuracy of the power formulas. Following the second simulation, we present derivatives of the moderation effect error variance for key parameters. The final section extends this work for partially nested data with a three-level structure. The paper concludes with a discussion of results and their implications for study design and planning.

Analytical Method

We begin with a two/one partially nested design in which the treatment group has a two-level data structure and the control group has a single-level data structure. Our analyses consider designs that randomly assign individuals to cluster-administered treatments (e.g., small-group intervention) or treatments employing a shared facilitator (e.g., teacher). For two/one partial nesting the treatment induces

a two-level data structure consisting of individuals nested within clusters while individuals in the control group avoid any nesting or clustering. These formulas are also applicable when treatment eliminates nesting or clustering such that the control group has a two-level structure and the treatment group is a single-level (e.g., home-based therapy versus typical group therapy control condition) but we avoid further discussion of this scenario for clarity.

We operationalize our analysis using the multiple-arm multilevel framework for partially nested data (MA-PN; Lachowicz et al., 2015; Lohr et al., 2014; Sterba et al., 2014). The MA-PN framework easily accommodates moderation effects (Sterba, 2017; Sterba et al., 2014) and the heteroscedasticity that is possible across study arms with partially nested data (Sterba, 2017). The treatment group analytic model has a continuous outcome (y_{ij}), individual-level continuous moderator ($m_{ij}^{(t)}$) with variance $\sigma^2_{m^{(t)}}$, and a slope that may vary across clusters such that

$$\text{Level 1:} \qquad y_{ij}^{(t)} = \beta_{0j}^{(t)} + \beta_{1j}^{(t)} m_{ij} + \varepsilon_{ij}^{(t)}, \qquad \varepsilon_{ij}^{(t)} \sim N(0, \sigma^2_{y|^{(t)}}) \tag{1}$$

$$\text{Level 2:} \qquad \beta_{0j}^{(t)} = \delta_{00}^{(t)} + u_{0j}^{(t)}, \qquad \begin{pmatrix} u_{0j}^{(t)} \\ u_{1j}^{(t)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{00|^{(t)}} & \tau_{01|^{(t)}} \\ \tau_{01|^{(t)}} & \tau^2_{11|^{(t)}} \end{pmatrix} \right). \tag{2}$$

$$\beta_{1j}^{(t)} = \gamma_{10}^{(t)} + u_{1j}^{(t)}$$

The treatment group is identified using the superscript *t* with individuals per cluster and clusters identified using subscripts *i* and *j* respectively. The variance term $\sigma^2_{y^{(t)}}$ indicates variance in $y_{ij}^{(t)}$ at the individual-level. At the cluster-level, $\beta_{0j}^{(t)}$ is decomposed into $\delta_{00}^{(t)}$, the grand mean for the treatment group, and $u_{0j}^{(t)}$, the residuals with a mean of zero and variance of $\tau^2_{00^{(t)}}$. To capture moderated effects when the moderator has a random slope, the cluster-specific main effect $\beta_{1j}^{(t)}$ of the moderator variable (m_{ij}) is decomposed into the main effect of the moderator variable under exposure to the treatment, $\delta_{00}^{(t)}$, and an unexplained cluster specific deflection, $u_{1j}^{(t)}$, that has a mean of zero, a variance of $\tau^2_{11^{(t)}}$, and a covariance with the random intercept of $\tau_{01^{(t)}}$. Within the context of partially nested designs, estimates of the main effect draw on the contrast between the overall intercept in the treatment group, $\delta_{00}^{(t)}$, and the overall intercept in the control group (see $\delta^{(c)}$ below). Similarly, investigations of treatment effect moderation draw on the contrast between the overall effect of the moderator in the treatment group ($\gamma_{10}^{(t)}$) and the overall effect of the moderator in the control group (see $\beta_1^{(c)}$ below). Randomization of individuals to conditions ensures that the compositions of individuals across conditions and clusters will not systematically differ. As a result, we assume that aggregates of the moderating variable do not further modulate the treatment effect.

The outcome model for the single-level control group with a moderator ($m_i^{(c)}$) is

$$y_i^{(c)} = \delta^{(c)} + \beta_1^{(c)} m_i + \varepsilon_i^{(c)} \qquad \varepsilon_i^{(c)} \sim N(0, \sigma^2_{y|^{(c)}}). \tag{3}$$

Variables and parameters (e.g., *y*, *m*, $\delta^{(c)}$, and $\beta_i^{(c)}$) retain similar meanings from the outcome model in the treatment group. The superscript (*c*) now indicates the control group setting and terms are identified using a single subscript *i* because no clustering is present. Variance components are also simplified, with $\sigma^2_{y^{(c)}}$ representing outcome variance in the control group and $\varepsilon_i^{(c)}$ the associated error term with mean zero. Our analytic interest in this investigation is on the coefficient capturing the relationship between the moderator and outcome in the control group, $\beta_1^{(c)}$.

Moderation Effect and Sampling Variance

The moderation effect in this partially nested structure can be defined as the difference between the overall effect of the moderator in the treatment condition and the effect of the moderator in the control condition. Under the MA-PN framework, we utilize the coefficients capturing the moderator-outcome relationship in the treatment and control arms to estimate the moderation effect (*ME*) such that

$$ME = \gamma_{10}^{(t)} - \beta_1^{(c)}. \tag{4}$$

In terms of our example, the moderation effect is the difference between the overall effect of gender in the treatment condition and the effect of gender in the control condition (for more detailed discussions on multilevel moderation effects see Preacher et al., 2016; Bauer & Curran, 2005; Preacher et al., 2006). While

moderation effect estimates are not operationalized as a product term under the multiple-arm multilevel framework for partially nested data, follow-up procedures delineating moderation effects (e.g., plots and probes) remain similar.

In addition to the definition and size of a moderation effect, a concomitant consideration when designing a study is the statistical power with which a given sample size can be used to detect an effect if it exists. In order to predict statistical power, we must track the sampling variability of the moderation effect. Under the multiple-arm multilevel framework for partially nested data the control and treatment groups are independent such that the sampling variability of the moderation effect in Equation 4 is

$$\sigma^2_{ME} = \sigma^2_{(\gamma^{(t)}_{10} - \beta^{(c)}_1)} = \sigma^2_{\gamma^{(t)}_{10}} + \sigma^2_{\beta^{(c)}_1} . \tag{5}$$

To be precise, we have the variance of the moderation effect as the variance of the difference in the estimator of $\gamma^{(t)}_{10}$ (i.e., $\hat{\gamma}^{(t)}_{10}$) and the estimator of $\beta^{(c)}_1$ (i.e., $\hat{\beta}^{(c)}_1$). The uncertainty about a moderation effect is then the sum of the variance of the regression coefficients associated with the moderator in the treatment and control condition. Below, we unpack this error variance by connecting it to conventional random slope models (Dong et al., 2021; Snijders, 2001; Snijders, 2005). We then re-express the error variance as a function of common sample statistics that can be used to predict requisite sample sizes when designing a study.

The sampling variability (i.e., $\sigma^2_{\gamma^{(t)}_{10}}$) of the relationship between the moderating variable and the outcome in the treatment group can be tracked using (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021; Snijders, 2001; Snijders, 2005),

$$\sigma^2_{\gamma^{(t)}_{10}} = \frac{n^{(t)}_1 \tau^2_{11|^{(t)}} \sigma^2_{m^{(t)}} + \sigma^2_{y|^{(t)}}}{(n^{(t)}_2 - 2) n^{(t)}_1 \sigma^2_{m^{(t)}}} \tag{6}$$

The formula for $\sigma^2_{\gamma^{(t)}_{10}}$ includes $\sigma^2_{y^{(t)}}$, the individual-level outcome variance and $\sigma^2_{m^{(t)}}$, the variance of the moderator with $n^{(t)}_2$ as the number of clusters and $n^{(t)}_1$ as the number of individuals within each cluster. The $\tau^2_{11^{(t)}}$ term represents variance of the random slopes (i.e., variance between clusters on the effect of $m^{(t)}_{ij}$). Similarly, the sampling variability (i.e., $\sigma^2_{\beta^{(c)}_1}$) of the relationship between the moderating variable and the outcome in the control group is (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021; Snijders, 2001; Snijders, 2005),

$$\sigma^2_{\beta^{(c)}_1} = \frac{\sigma^2_{y|^{(c)}}}{(n^{(c)} - 2)\sigma^2_{m^{(c)}}} . \tag{7}$$

The terms in the formulation of $\sigma^2_{\gamma^{(t)}_{10}}$ and $\sigma^2_{\beta^{(c)}_1}$ are similar with total sample size in the control group represented with $n^{(c)}$.

Substituting the expanded formulations of $\sigma^2_{\gamma^{(t)}_{10}}$ and $\sigma^2_{\beta^{(c)}_1}$ into the σ^2_{ME} formula, the predicted sampling variability of the moderation effect is

$$\sigma^2_{ME} = \frac{n^{(t)}_1 \tau^2_{11|^{(t)}} \sigma^2_{m^{(t)}} + \sigma^2_{y|^{(t)}}}{(n^{(t)}_2 - 2) n^{(t)}_1 \sigma^2_{m^{(t)}}} + \frac{\sigma^2_{y|^{(c)}}}{(n^{(c)} - 2)\sigma^2_{m^{(c)}}} . \tag{8}$$

We can also present a standardized version using unit variance so that $\tau^2_{00^{(t)}} + \sigma^2_{y^{(t)}} = 1$ and $\sigma^2_{m^{(t)}} = 1$. Under this standardization $\tau^2_{00^{(t)}}$ is the unconditional variance of the outcome in the treatment group at the cluster level and $\sigma^2_{y^{(t)}}$ is the unconditional variance of the outcome in the treatment group at the individual-level. The $\tau^2_{00^{(t)}}$ term is equivalent to the unconditional intraclass correlation coefficient (ρ) of the outcome in the treatment group such that

$$\rho = \frac{\tau^2_{00^{(t)}}}{\tau^2_{00^{(t)}} + \sigma^2_{y^{(t)}}} \tag{9}$$

A statistical test of the null hypothesis that the *ME* is zero (H_0) versus an alternative hypothesis that it is not zero (H_α) can then be formed using the ratio of the effect to its variance. The resulting ratio forms *t* statistic such that if the null hypothesis (H_0) is true the statistic follows a central *t*-distribution and if the null hypothesis is false the statistic will follow a noncentral *t*-distribution with a non-centrality parameter of (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021; Snijders, 2001; Snijders, 2005)

$$t_{ME} = ME/\sigma_{ME} \tag{10}$$

and $n_2^{(t)}-2$ degrees of freedom (*df*, Raudenbush & Bryk; 2002). This formula to determine *df* is an approximation with more precise formulas possible using the number of explanatory variables in the model or a Satterthwaite approximation. However, any discrepancy with the reference *t*-distribution because of inaccuracies in the *df* would be minimal (i.e., slight inaccuracies in type I error rate) and limited to small cluster sample sizes. Statistical power for the two-sided test is formulated as

$$P(\textit{reject } H_0|H_a \textit{ is True}) = P(|t| > t_{critical}) \quad (11)$$

with *t* the observed value of t_{ME} in Equation 10, assumed to follow a *t*-distribution with $n_2^{(t)}-2$ degrees of freedom under H_0 and $t_{critical}$ as the corresponding critical value for a selected type one error rate (e.g., 1.96 in large samples).

To accommodate binary moderators, the proportion of the first moderator subgroup is designated *Q* with the remaining moderator subgroup (1–*Q*). The variance of the binary moderator follows a Bernoulli distribution such that $b_{ij} \sim Bernoulli(Q)$ where the variance of the moderator in the treatment and control group (i.e., $\sigma^2_{M^{(t)}}$ and $\sigma^2_{M^{(c)}}$) is $Q(1-Q)$. The separation of treatment and control groups in partially nested designs and the reflection of this separation in our formulation of σ^2_{ME} also allows easy accommodation of differing sample sizes in the treatment and control conditions. Possible treatment and control sample sizes (i.e., $n_1^{(t)}$, $n_2^{(t)}$, and $n^{(c)}$) can be selected assuming a balanced sample across study arms such that $n^{(c)} = n_1^{(t)}n_2^{(t)}$ or set individually for unbalanced samples.

Simulation Study I

A Monte Carlo simulation study was conducted in `R` (R Core Team, 2021) to establish the accuracy of the newly derived power formulas and probe several factors that influence power rates including cluster and individual per cluster sample size, magnitude of the moderated effect, moderation effect heterogeneity, and intraclass correlation coefficient of the outcome. We utilized the `lm` function in `R` (R Core Team, 2021) to analyze the single level models and the `lme4` package with the default REML estimator (Bates et al., 2015) when estimating the multilevel models.

Specifically, we generated data sets with sample sizes of $n_1^{(t)}$ = 10, 25, 50, and 100 and $n_2^{(t)}$ = 25, 50, and 100 with $n^{(c)} = n_1^{(t)}n_2^{(t)}$. Under a limited set of $n_1^{(t)}$ and $n_{12}^{(t)}$ values we included $n^{(c)} = 2n_1^{(t)}n_2^{(t)}$ and $n^{(c)} = n_1^{(t)}n_2^{(t)}/2$ to investigate unbalanced sample sizes across study arms. The magnitude of the moderation effect was set using $\beta_1^{(c)}$ = 0 and $\gamma_{10}^{(t)}$ = 0.1, 0,2 and 0, which respectively produces *ME* = 0.1, 0.2, and 0 (to evaluate formula Type I error rates at α=0.05). The core development in power formulas presented here is their ability to handle an individual-level moderator with a random slope so we varied $\tau^2_{11^{(t)}}$ and $\tau^2_{00^{(t)}}$ such that moderation effect heterogeneity ($\omega = \tau^2_{11^{(t)}}/\tau^2_{00^{(t)}}$) was 0.2, 0.4, 0.6, and 0.8 with a 0.0 condition to reflect a moderator with a non-random slope. The range of values for ω allowed us to consider varying degrees of moderation effect heterogeneity and aligned with previous simulation studies (e.g., Dong et al., 2021).

Finally, we examined these scenarios when the unconditional intraclass correlation coefficient of the outcome was ρ = 0.1 and 0.2. This set of simulation conditions was examined with a continuous and binary moderator. Simulation conditions were guided by previous simulation literature examining partially nested designs (Cox & Kelcey, 2022; Esserman et al., 2013; Heo et al., 2017; Roberts, 2021; Roberts et al., 2016) and moderation in cluster-randomized trials (Cox & Kelcey, 2022; Dong et al., 2018; Dong et al., 2021; Spybrook et al., 2016; Yang et al., 2020) as well as empirical analyses of values for multilevel structures and variables in education (e.g., Bai et al., 2024; Hedges & Hedberg, 2006; Kelcey et al., 2016). The data for the simulation were generated using the previously presented analytic models with $\tau_{10^{(t)}} = \tau_{01^{(t)}} = 0$, $\sigma^2_{m^{(t)}} = 1$, $\sigma^2_{y^{(c)}} = 1$, $\sigma^2_{m^{(c)}} = 1$, $\delta^{(t)}_{00} = 0.3$, and $\delta^{(c)} = 0$. Using these parameter values , we predicted the moderated effect (*ME*) based on Equation (4) and the respective expected sampling variance of the pertinent coefficients ($\sigma^2_{\gamma_{10}^{(t)}}$ and $\sigma^2_{\beta_1^{(c)}}$) as well as the sampling variance of the moderation effect (σ^2_{ME}).

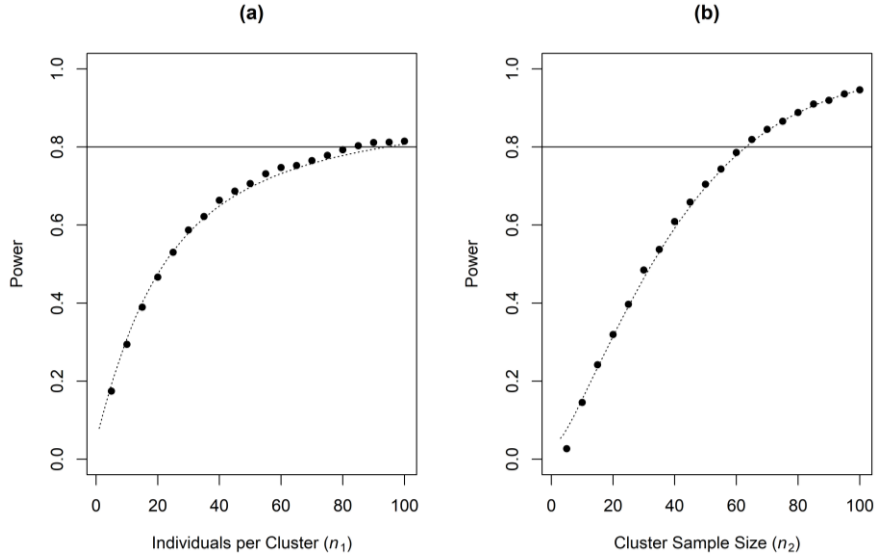
This process was replicated across 10,000 data sets. To estimate the observed power rate across simulation draws, we conducted hypothesis testing for each draw by checking whether the observed *t* exceeded $t_{critical}$. To predict the power rate using the expressions developed in this study, we substituted the parameter values for each condition outlined above into the applicable expressions (4-8). To assess the accuracy of our expressions, we compared the power and Type I error rate between the observed simulated power and the predicted power using the formulas under various conditions. All analyses were also completed using `R` (R Core Team, 2021).

Table1. Selected Comparisons of Formula-Based Statistical Power and Type I Error Rate and Monte Carlo Simulation Results for a Continuous Moderator without Covariates in a Design with Two/One Partially Nested Data

Scenario	1	2	3	4	5	6	7	8	9	10	11	12
ρ	0.2	0.2	0.2	0.2	0.2	0.2	0.1	0.1	0.1	0.1	0.1	0.1
ω	0.2	0.4	0.8	0.2	0.4	0.8	0.2	0.4	0.8	0.2	0.4	0.8
n_2	25	25	25	100	100	100	25	25	25	100	100	100
n_1	100	100	100	25	25	25	100	100	100	25	25	25
Simulation rejection rate	0.50	0.33	0.21	0.83	0.72	0.54	0.69	0.50	0.40	0.89	0.82	0.75
Formula power rate	0.50	0.33	0.20	0.84	0.72	0.55	0.67	0.50	0.39	0.89	0.83	0.76
Absolute Difference	0.00	0.00	0.01	0.01	0.00	0.01	0.02	0.00	0.01	0.00	0.01	0.01
Empirical Type I error rate	0.032	0.032	0.032	0.044	0.044	0.044	0.035	0.035	0.035	0.043	0.043	0.043
Formula Type I error rate	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
Absolute Difference	0.018	0.018	0.018	0.006	0.006	0.006	0.015	0.015	0.015	0.007	0.007	0.007

Note: Results were based on 10,000 replications with balanced sample sizes in the treatment and control condition, ρ representing intraclass correlation, ω representing moderation effect heterogeneity, and *ME*=0.1 for power and *ME*=0.0 for Type I error.

Figure 1. Statistical Power and Empirical Rejection Rates for A Moderated Effect with Two/One Partially Nested Data



Note: This figure plots formula-based power curves to detect a moderation effect with simulation rejection rates for a continuous moderator with balanced sample sizes across study arms, without covariates across increasing (a) individuals per cluster and (b) cluster sample sizes. The dotted line represents the formula-based power curve with black points marking the simulation-based rejection rate at matching sample sizes. A horizontal line marks 80% power. For (a) n_2=50 and for (b) n_1=50 with remaining conditions set at ω=0.2, *ME*=0.1, and ρ=0.2

Results

Power rate accuracy was considered for a continuous and binary moderator across 192 conditions with an additional 24 conditions to assess accuracy of the Type I error rate (see Table 1 for selected results and Supplemental Materials for complete results). Across all conditions, formula predicted power rates for the moderated effect closely approximated simulation rejection rates. Figure 1 illustrates formula accuracy with (a) a cluster sample size of $n_2^{(t)}$ = 50 across various $n_1^{(t)}$ sample sizes and (b) an individual per cluster sample size of $n_1^{(t)}$ = 50 across various $n_2^{(t)}$ sample sizes. Simulation rejection rates (dots) track closely with formula-based power curves across the sample size conditions with ω = 0.2, *ME* = 0.1, and ρ = 0.2. Our simulations also evaluated the Type I error rate when *ME* = 0. The results suggested deflated Type I error rates when cluster level sample sizes were small. For example, when cluster-level sample sizes were $n_2^{(t)}$ = 25, rejection

rates were below the designated 0.05 level. We found empirical rejection rates quickly approached the 0.05 level as cluster-level sample sizes increased. This is likely a result of slight discrepancies in the reference distribution for the *t*-statistic based on the $n_2^{(t)}-2$ *df* formula. To ensure the accuracy of our power formulas, we compared simulation and formula-based moderation effect error variance values and found no systematic or substantial differences. These results suggest discrepancies between formula predicted Type I error rate and empirical rejection rate stems from difficulties with the empirical rejection rates with small cluster-level sample sizes and minor discrepancies in the reference *t* distribution rather than inaccuracies in our variance or power formulas. When $\sigma^2_{M^{(t)}} = \sigma^2_{M^{(c)}} = Q(1\text{-}Q)$ for binary moderators, formula predicted power to detect the moderation effect again closely approximated simulation rejection rates (see Supplemental Materials for complete results).

We also examined power formula accuracy in partially nested designs with unbalanced sample sizes across study arms. While balanced sample sizes in the treatment and control condition typically maximize power, they are rarely achieved in practice and sometimes explicitly avoided. For example, researchers may plan to assign many more clusters to the treatment group to encourage study participation or limit treatment participation due to logistical constraints (e.g., limited resources or budget) that inflate the control group sample size. Following previous literature, we consider unbalanced study arm sample sizes with a larger sample in the control group ($n^{(c)} = 2n_1^{(t)}n_2^{(t)}$) and a smaller sample in the control group ($n^{(c)} = n_1^{(t)}n_2^{(t)}/2$); Dong et al., 2018; Dong et al., 2021; Spybrook et al., 2016). Formula predicted power closely approximated simulation rejection rates with unbalanced sample sizes across study arms (see Supplemental Materials for complete results).

Analytic Model with a Covariate

It is a common and beneficial practice in experimental studies to include prognostic covariates in the analytic model. Covariates that explain variation in the outcome increase the power to detect both main and moderation effects (Spybrook et al., 2016). Therefore, it is important for power formulas to accommodate the inclusion of prognostic covariates. We replicate the work presented thus far with a revised analytic model that includes a covariate. We again draw on the common multiple-arm multilevel framework for partially nested data (MA-PN). In the treatment group we again have a continuous outcome (y_{ij}) and individual-level moderator ($m_{ij}^{(t)}$) with variance $\sigma^2_{m^{(t)}}$ but the analytic model includes a covariate at the individual-level ($x_{ij}^{(t)}$) with variance of $\sigma^2_{x^{(t)}}$ such that

Level 1:
$$y_{ij}^{(t)} = \beta_{0j}^{(t)} + \beta_{1j}^{(t)} m_{ij} + \beta_{2j}^{(t)} x_{ij} + \varepsilon_{ij}^{(t)}, \qquad \varepsilon_{ij}^{(t)} \sim N(0, \sigma^2_{y|^{(t)}}) \tag{12}$$

$$\beta_{0j}^{(t)} = \delta_{00}^{(t)} + u_{0j}^{(t)}, \qquad \begin{pmatrix} u_{0j}^{(t)} \\ u_{1j}^{(t)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{00|^{(t)}} & \tau_{01|^{(t)}} \\ \tau_{10|^{(t)}} & \tau^2_{11|^{(t)}} \end{pmatrix} \right)$$

Level 2:
$$\beta_{1j}^{(t)} = \gamma_{10}^{(t)} + u_{1j}^{(t)} \tag{13}$$
$$\beta_{2j}^{(t)} = \gamma_{20}^{(t)}$$

A revised error variance formulation can be expressed as (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021)

$$\sigma^2_{\gamma_{10}^{(t)}} = \frac{n_1^{(t)}\tau^2_{11|^{(t)}}\sigma^2_{m^{(t)}} + \sigma^2_{y|^{(t)}}(1 - R^2_{y^{(t)}})}{(n_2^{(t)} - C_{(t)} - 1)n_1^{(t)}\sigma^2_{m^{(t)}}} \tag{14}$$

where $C_{(t)}$ is the number of additional covariates in the analytic model and variance explained in the outcome at the individual-level is $R^2_{y^{(t)}} = 1 - [\sigma^2_{y|^{(t)}}/\sigma^2_{y^{(t)}}]$ again assuming standardization using unit variance so that $\tau^2_{00^{(t)}} + \sigma^2_{y^{(t)}} = 1$ and $\sigma^2_{m^{(t)}} = 1$ (see also Equation 9).

A covariate in the outcome model for the control group, $x_i^{(c)} \sim N(0, \sigma^2_{x^{(c)}})$ requires similar adjustments

$$y_i^{(c)} = \delta^{(c)} + \beta_1^{(c)} m_i + \beta_2^{(c)} x_i + \varepsilon_i^{(c)} \qquad \varepsilon_i^{(c)} \sim N(0, \sigma^2_{y|^{(c)}}) \tag{15}$$

and an error variance formulation of (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021)

$$\sigma^2_{\beta_1^{(c)}} = \frac{\sigma^2_{y|^{(c)}}}{(n^{(c)} - C_{(c)} - 1)\sigma^2_{m^{(c)}}} \tag{16}$$

The subsequent expression for moderator coefficient variance in the control group is (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021)

$$\sigma^2_{\beta_1^{(c)}} = \frac{\sigma^2_{y|^{(c)}}(1-R^2_{y^{(c)}})}{(n^{(c)} - C_{(c)} - 1)\sigma^2_{m^{(c)}}} \tag{17}$$

and the variance explained in the outcome of the control group is $R^2_{y^{(c)}} = 1 - [\sigma^2_{y|^{(c)}}/\sigma^2_{y^{(c)}}]$ with standardized unit variance such that $\sigma^2_{y^{(c)}} = 1$. The variance of the moderation effect when the model includes covariates is then

$$\sigma^2_{ME} = \frac{n_1^{(t)}\tau^2_{11|^{(t)}}\sigma^2_{m^{(t)}} + \sigma^2_{y|^{(t)}}(1-R^2_{y^{(t)}})}{(n_2^{(t)} - C_{(t)} - 1)n_1^{(t)}\sigma^2_{m^{(t)}}} + \frac{\sigma^2_{y|^{(c)}}(1-R^2_{y^{(c)}})}{(n^{(c)} - C_{(c)} - 1)\sigma^2_{m^{(c)}}} \tag{18}$$

As with the previous model, we can standardize the variance terms, employ a *t*-statistic to conduct a hypothesis test for the *ME*, and determine power using Equation 11. An approximation of the *df* can be found using $n_1^{(t)} - C_{(t)} - 1$ degrees of freedom where $C_{(t)}$ is the number of predictor variables in the treatment arm outcome model. Extensions of these formulas to accommodate binary moderators and unbalanced sample sizes in the treatment and control condition parallel those based on the analytic model without covariates.

Simulation Study II

We repeat the first simulation study but with a covariate (x_{ij}) added to the analytic model explaining variance in the outcome (y). Specifically, we set $R^2_{y^{(t)}} = R^2_{y^{(c)}} = 0.4$ and 0.7 reducing $\sigma^2_{y^{(t)}}$ and $\sigma^2_{y^{(c)}}$ and terms in the variance of the moderation effect formulation. We again focus on power formula accuracy while varying several influential factors. Simulation conditions are retained from the first study but we only consider a moderation effect of $ME = 0.1$.

We found predicted power rates from the formulas that accommodated covariates closely aligned to empirical rejection rates across a multitude of conditions including continuous and binary moderators and in unbalanced sample sizes across study arms. Type I error rates when *ME* = 0were also closely approximated by the formulas, though typical discrepancies arose when cluster-level sample sizes were limited (e.g., $n_2^{(t)} \leq 25$). As with our first simulation, we found power and error variance formulas to be accurate. The problem again stemmed from difficulties with empirical rejection rates when $n_2^{(t)}$ values are small. Table 2 presents selected results demonstrating power formula accuracy including those with $R^2_{y^{(t)}} = R^2_{y^{(c)}} = 0$ for comparative purposes (see Supplemental Materials for complete results).

Table 2. Selected Comparisons of Formula-Based Statistical Power and Type I Error Rate and Monte Carlo Simulation Results for a Continuous Moderator with Covariates in a Design with Two/One Partially Nested Data.

Scenario	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
$R^2_{y^{(t)}} = R^2_{y^{(c)}}$	0	0	0	0	0.4	0.4	0.4	0.4	0.7	0.7	0.7	0.7
ω	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8	0.2	0.8
n_2	25	25	100	100	25	25	100	100	25	25	100	100
n_1	100	100	25	25	100	100	25	25	100	100	25	25
Simulation rejection rate	0.50	0.21	0.83	0.54	0.57	0.21	0.93	0.58	0.61	0.22	0.98	0.64
Formula power rate	0.50	0.20	0.84	0.55	0.56	0.21	0.93	0.59	0.60	0.22	0.98	0.64
Absolute Difference	0.00	0.01	0.01	0.01	0.01	0.00	0.00	0.01	0.01	0.00	0.00	0.00
Empirical Type I error rate	0.032	0.032	0.044	0.044	0.036	0.034	0.041	0.043	0.034	0.031	0.045	0.043
Formula Type I error rate	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050	0.050
Absolute Difference	0.018	0.018	0.006	0.006	0.014	0.016	0.009	0.007	0.016	0.019	0.005	0.007

Note: Results were based on 10,000 replications with balanced sample sizes in the treatment and control condition, $\rho = 0.2$, ω representing moderation effect heterogeneity, and *ME*=0.1 for power and *ME*=0.0 for Type I error.

Implications for Design

Prior literature has delineated the role of parameters for a broad range of designs and effects (e.g., main, mediation, moderation) and used those roles to develop a number of strategies that guide efficient design (Raudenbush et al., 2007). To outline the roles of the parameters governing design and develop design principles and strategies in the current moderation context, we examined how the error variance of the moderation effect shrinks or grows with increments in a parameter value while holding other parameters constant. Using derivatives of the error variance, we examined two broad categories of parameters: (a) potentially mutable parameters such as sample sizes ($n_2^{(t)}$, $n_1^{(t)}$, $n^{(c)}$) and outcome variance explained through the introduction of covariates ($R^2_{y^{(t)}}$, $R^2_{y^{(c)}}$) and (b) typically immutable parameters such as the variance decomposition of the outcome and moderator ($\tau_{11^{(t)}}$, $\sigma^2_{y^{(t)}}$, $\sigma^2_{y^{(c)}}$, $\sigma^2_{m^{(t)}}$, $\sigma^2_{m^{(c)}}$).

The first derivatives of the error variance in terms of each sample size were

$$\frac{\partial \sigma^2_{ME}}{\partial n_2^{(t)}} = -\frac{\sigma^2_{y^{(t)}}\left(1-R^2_{y^{(t)}}\right)+n_1^{(t)}\tau^2_{11^{(t)}}\sigma^2_{m^{(t)}}}{n_1^{(t)}\sigma^2_{m^{(t)}}\left(n_2^{(t)}-C_{(t)}-1\right)^2} < 0 \tag{19}$$

$$\frac{\partial \sigma^2_{ME}}{\partial n_1^{(t)}} = \frac{\tau^2_{11^{(t)}}}{n_1^{(t)}\left(n_2^{(t)}-C_{(t)}-1\right)} - \frac{\sigma^2_{y^{(t)}}\left(1-R^2_{y^{(t)}}\right)+n_1^{(t)}\tau^2_{11^{(t)}}\sigma^2_{m^{(t)}}}{\left(n_1^{(t)}\right)^2\sigma^2_{m^{(t)}}\left(n_2^{(t)}-C_{(t)}-1\right)} < 0 \tag{20}$$

$$\frac{\partial \sigma^2_{ME}}{\partial n^{(c)}} = -\frac{\sigma^2_{y^{(c)}}\left(1-R^2_{y^{(c)}}\right)}{\sigma^2_{m^{(c)}}\left(n^{(c)}-C_{(c)}-1\right)^2} < 0 \cdot \tag{21}$$

These derivatives were uniformly negative replicating the well-established pattern of increases in sample size at any level decreasing error variance (and increasing power). The reduction of error variance by the level- and condition-specific sample size ($n_2^{(t)}$, $n_1^{(t)}$) was fairly dependent on other parameters. This result parallels other examinations of power to detect moderation effects with random slopes in fully nested designs (Dong et al., 2021). These results also highlight different sampling variances for *ME* estimates when comparing models with fixed versus random moderator slopes in partially nested designs (e.g., Cox & Kelcey, 2022).

Analysis of the parameters describing the reduction in outcome variance associated with conditioning on covariates ($R^2_{y^{(t)}}$, $R^2_{y^{(c)}}$) also consistently reduced error variance. Put differently, covariates that were correlated with the outcome steadily reduced the error variance of the moderated effect as explained variance in the outcome increased thus increasing power. Specifically, each derivative was uniformly negative such that

$$\frac{\partial \sigma^2_{ME}}{\partial R^2_{y^{(t)}}} = -\frac{\sigma^2_{y^{(t)}}}{n_1^{(t)}\sigma^2_{m^{(t)}}\left(n_2^{(t)}-C_{(t)}-1\right)} < 0 \tag{22}$$

$$\frac{\partial \sigma^2_{ME}}{\partial R^2_{y^{(c)}}} = -\frac{\sigma^2_{y^{(c)}}}{\sigma^2_{m^{(c)}}\left(n^{(c)}-C_{(c)}-1\right)} < 0 \tag{23}$$

Similar to the disparate roles of the sample sizes, the relative contribution of these parameters to power rates depended heavily on the values of other parameters including variance of the moderator and the variance of the outcome (see Supplemental Materials for complete results). This again paralleled results for power to detect moderation effects with random slopes in fully nested designs (Dong et al., 2021).

Analysis of the variance components ($\tau_{11^{(t)}}$, $\sigma^2_{y^{(t)}}$, $\sigma^2_{y^{(c)}}$, $\sigma^2_{m^{(t)}}$, $\sigma^2_{m^{(c)}}$) demonstrated the diverse roles of these parameters in regards to power to detect the moderation effect. For moderation effect heterogeneity and the variance of the outcome in both the treatment and control group, the analysis indicated that the first derivatives were uniformly positive

$$\frac{\partial \sigma^2_{ME}}{\partial \tau^2_{11^{(t)}}} = \frac{1}{\left(n_2^{(t)}-C_{(t)}-1\right)} > 0 \tag{24}$$

$$\frac{\partial \sigma^2_{ME}}{\partial \sigma^2_{y^{(t)}}} = \frac{1 - R^2_{y^{(t)}}}{n_1^{(t)} \sigma^2_{m^{(t)}} \left(n_2^{(t)} - C_{(t)} - 1 \right)} > 0 \tag{25}$$

$$\frac{\partial \sigma^2_{ME}}{\partial \sigma^2_{y^{(c)}}} = \frac{1 - R^2_{y^{(c)}}}{\sigma^2_{m^{(c)}} \left(n^{(c)} - C_{(c)} - 1 \right)} > 0 \tag{26}$$

As a result, increases in each of these parameters was associated with increases in the error variance (and decreasing power). In contrast, results suggested that increases in the moderator variance (holding other parameters constant) in either the treatment or the control group was associated with decreases in the error variance such that

$$\frac{\partial \sigma^2_{ME}}{\partial \sigma^2_{m^{(t)}}} = \frac{\tau^2_{11^{(t)}}}{\sigma^2_{m^{(t)}} \left(n_2^{(t)} - C_{(t)} - 1 \right)} - \frac{\sigma^2_{y^{(t)}} \left(1 - R^2_{y^{(t)}} \right) + n_1^{(t)} \tau^2_{11^{(t)}} \sigma^2_{m^{(t)}}}{n_1^{(t)} (\sigma^2_{m^{(t)}})^2 \left(n_2^{(t)} - C_{(t)} - 1 \right)} < 0 \tag{27}$$

$$\frac{\partial \sigma^2_{ME}}{\partial \sigma^2_{m^{(c)}}} = -1 \frac{\sigma^2_{y^{(c)}} \left(1 - R^2_{y^{(c)}} \right)}{\left(\sigma^2_{m^{(c)}} \right)^2 \left(n^{(c)} - C_{(c)} - 1 \right)} < 0 \cdot \tag{28}$$

Essentially, more variance in the moderator increases the likelihood of detecting a moderated effect.

Collectively, our analyses suggest the following intuitive roles for design parameters:

- Increases in sample size ($n_2^{(t)}$, $n_1^{(t)}$, $n^{(c)}$) increase the statistical power to detect the moderation effects but the magnitude of these increases is dependent on other factors.
- Increases in moderation effect heterogeneity ($\tau_{11^{(t)}}$) and outcome variance components ($\sigma^2_{y^{(t)}}$ and $\sigma^2_{y^{(c)}}$) decrease the statistical power to detect moderation effects.
- Increases in moderator variance ($\sigma^2_{m^{(t)}}$ and $\sigma^2_{m^{(c)}}$) increase the statistical power to detect moderation effects.

Based on these roles and the totality of our results, we suggest the following strategies for the mutable parameters a) incorporating covariates that explain variance in the outcome, b) prioritizing cluster sample size ($n_2^{(t)}$), and c) planning for an even sampling of treatment and control conditions when statistical power is paramount (see Supplemental Materials for an Illustrative Example).

Three-Level Partially Nested Data

We now extend our power formulas for moderation effects from individual-level moderators with a random slope to studies with three-level partially nested data structures. In three/one partially nested designs the analytic model includes an additional level of nesting in the intervention arm compared to previously described models. The model with a continuous outcome ($y_{ijk}^{(t)}$), a continuous individual-level moderator, $m_{ijk}^{(t)} \sim N(0, \sigma^2_{m^{(t)}})$ with random slopes (at levels two and three), and individual-level covariate, $x_{ijk}^{(t)} \sim N(0, \sigma^2_{x^{(t)}})$ is:

$$\begin{aligned}
&\text{Level 1:} \quad y_{ijk}^{(t)} = \beta_{0jk}^{(t)} + \beta_{1jk}^{(t)} m_{ijk} + \beta_{2jk}^{(t)} x_{ijk} + \varepsilon_{ijk}^{(t)}, \qquad \varepsilon_{ijk}^{(t)} \sim N(0, \sigma^2_{y|^{(t)}}) \\
&\qquad \beta_{0jk}^{(t)} = \gamma_{00k}^{(t)} + u_{0jk}^{(t)}, \qquad \begin{pmatrix} u_{0jk}^{(t)} \\ u_{1jk}^{(t)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau^2_{00|^{(t)}} & \tau_{01|^{(t)}} \\ \tau_{10|^{(t)}} & \tau^2_{11|^{(t)}} \end{pmatrix} \right) \\
&\text{Level 2:} \quad \beta_{1jk}^{(t)} = \gamma_{10k}^{(t)} + u_{1jk}^{(t)} \\
&\qquad \beta_{2jk}^{(t)} = \gamma_{20k}^{(t)} \\
&\qquad \gamma_{00k}^{(t)} = \pi_{000}^{(t)} + v_{00k}^{(t)} \qquad \begin{pmatrix} v_{00k}^{(t)} \\ v_{10k}^{(t)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \phi^2_{000|^{(t)}} & \phi_{010|^{(t)}} \\ \phi_{100|^{(t)}} & \phi^2_{110|^{(t)}} \end{pmatrix} \right) \\
&\text{Level 3:} \quad \gamma_{10k}^{(t)} = \pi_{100}^{(t)} + v_{10k}^{(t)} \\
&\qquad \gamma_{20k}^{(t)} = \pi_{200}^{(t)}
\end{aligned} \tag{29}$$

.

Model notation and interpretation parallel those for the two/one partially nested design with a covariate with the notable addition of the *k* subscript to identify the third level of nesting.

Contrasting the coefficients capturing the relationship between the moderator and outcome again provides an estimate of the moderation effect (*ME*) such that

$$ME = \pi_{100}^{(t)} - \beta_1^{(c)}. \tag{30}$$

The error variance formula is also similar but the variance term for the moderator coefficient in the intervention arm now includes sample size and variance terms for three-levels (e.g., $n_1^{(t)}$, $n_2^{(t)}$, $n_3^{(t)}$, $\phi^2_{110^{(t)}}$) such that (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021)

$$\sigma^2_{\pi_{100}^{(t)}} = \frac{n_1^{(t)}\phi^2_{110|^{(t)}}\tau^2_{11|^{(t)}}\sigma^2_{m^{(t)}} + \sigma^2_{y|^{(t)}}(1-R^2_{y^{(t)}})}{(n_2^{(t)}n_3^{(t)} - C_{(t)} - 1)n_1^{(t)}\sigma^2_{m^{(t)}}} \tag{31}$$

As with the models for two/one partially nested data, we can standardize the variance terms such that $\phi^2_{000^{(t)}} + \tau^2_{00^{(t)}} + \sigma^2_{y^{(t)}} = 1$ and $\sigma^2_{m^{(t)}} = 1$. The single-level analytic model for the control arm and variance formula for $\beta_1^{(c)}$ remains the same.

Three/Two Partially Nested Designs

We can further extend our power formulas to accommodate three-level partially nested designs with two-level data structures in the control arm. The intervention arm analytic models (see Equations 29) and moderation effect error variance terms (see Equation 31) remain the same but a two-level analytic model for the control arm is now required. We again limit consideration to models with a continuous outcome ($y_{ij}^{(c)}$), individual-level continuous moderator ($m_{ij}^{(c)}$) with variance $\sigma^2_{m_{ij}^{(c)}}$, and whose slope may vary such that

$$\text{Level 1:} \quad y_{i0k}^{(c)} = \beta_{00k}^{(c)} + \beta_{10k}^{(c)}m_{i0k} + \beta_{20k}^{(c)}x_{i0k} + \varepsilon_{i0k}^{(c)}, \qquad \varepsilon_{i0k}^{(c)} \sim N(0, \sigma^2_{y|^{(c)}}) \tag{32}$$

$$\beta_{00k}^{(c)} = \pi_{000}^{(c)} + v_{00k}^{(c)} \qquad \begin{pmatrix} v_{00k}^{(c)} \\ v_{10k}^{(c)} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \phi^2_{000|^{(c)}} & \phi_{010|^{(c)}} \\ \phi_{100|^{(c)}} & \phi^2_{110|^{(c)}} \end{pmatrix}\right)$$

$$\text{Level 2:} \quad \beta_{10k}^{(c)} = \pi_{100}^{(c)} + v_{10k}^{(c)} \tag{33}$$

$$\beta_{20k}^{(c)} = \pi_{200}^{(c)}$$

Model structure and interpretation parallel those for the analytic model for the two-level treatment arm of a two/one partially nested design although we utilize notation to parallel the corresponding three-level intervention arm (e.g., $v_{00k}^{(c)}$ and $\phi^2_{110^{(c)}}$ for the upper-level). The contrast of coefficients to estimate the moderation effect is now

$$ME = \pi_{100}^{(t)} - \pi_{100}^{(c)} \tag{34}$$

with the error variance term from the intervention arm ($\sigma^2_{\pi_{100}^{(t)}}$) found in Equation 31. The error variance term in the control arm is formulated similar to the error variance term in the analytic model for the two-level treatment arm of a two/one partially nested design such that (Cox & Kelcey, 2022; Cox et al., 2022; Dong et al., 2021)

$$\sigma^2_{\pi_{100}^{(c)}} = \frac{n_1^{(c)}\phi^2_{110|^{(c)}}\sigma^2_{m^{(c)}} + \sigma^2_{y|^{(c)}}(1-R^2_{y^{(c)}})}{(n_2^{(t)} - C_{(c)} - 1)n_1^{(t)}\sigma^2_{m^{(c)}}}. \tag{35}$$

Variance components remain the same in the treatment arm as described for three/one partial nesting. As for the control arm, we can standardize the variance terms such that, $\phi^2_{000^{(c)}} + \sigma^2_{y^{(c)}} = 1$ and $\sigma^2_{m^{(c)}} = 1$. We again employ a *t*-statistic to conduct a hypothesis test for the *ME*, and determine power using Equation 11. Under the assumption that the null hypothesis is false, the *t*-statistic follows a noncentral *t*-distribution with the non-centrality parameter presented in Equation 10. An approximation of the *df* can be found using $n_3^{(t)} - C_{(t)} - 1$ degrees of freedom where $C_{(t)}$ is the number of predictor variables in the treatment arm outcome model (Raudenbush & Bryk; 2002). Extensions of these formulas to accommodate binary moderators and unbalanced sampling across study arms parallel those based on previous analytic models.

Simulation Study III

We conducted a third simulation of limited scope to establish the accuracy of the power formulas for detecting moderation effects in three-level partially nested designs. Specifically, we consider power to detect moderation effects in three/one partially nested designs when the moderator-outcome relationship may vary across groups (i.e., random slope) and in three/two partially nested designs when the moderator-outcome relationship may vary across groups (i.e., random slope). We utilized conditions similar to those in the second simulation study. Specifically, we generated data sets with sample sizes of $n_3^{(t)}$, $n_2^{(t)}$, and $n_1^{(t)}$ of 10 and 25 with $n^{(c)} = n_3^{(t)} n_2^{(t)} n_1^{(t)}$ for three/one partial nesting and $n_1^{(c)} = n_2^{(t)} n_1^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$ for three/two partial nesting. We varied τ_{11}^2, $\tau_{00^{(t)}}^2$, $\phi_{110^{(t)}}^2$, and $\phi_{000^{(t)}}^2$ such that moderation effect heterogeneity at each level ($\omega_{L2}^{(t)} = \tau_{11}^2/\tau_{00^{(t)}}^2$ and $\omega_{L3}^{(t)} = \phi_{110^{(t)}}^2/\phi_{000^{(t)}}^2$) was 0.2, 0.4, and 0.8 with a 0.0 condition to reflect moderation effects when the moderator had a non-random slope. For three/two partially nested data we set $\phi_{110^{(c)}}^2$ and $\phi_{000^{(c)}}^2$ such that moderation effect heterogeneity ($\omega_{L2}^{(c)} = \phi_{110^{(c)}}^2/\phi_{000^{(c)}}^2$) was 0.0, 0.2, 0.4, and 0.8. As for other variance components, $\sigma_{m^{(t)}}^2 = 1$, $\sigma_{y^{(t)}}^2 = 0.8$, $\tau_{00^{(t)}}^2 = 0.1$, $\phi_{000^{(t)}}^2 = 0.1$, $\sigma_{y^{(c)}}^2 = 1$, and $\sigma_{m^{(c)}}^2 = 1$ for three/one partial nesting and, $\sigma_{y^{(c)}}^2 = 0.9$ and $\phi_{000^{(c)}}^2 = 0.1$ when applicable with three/two partial nesting.

For designs with three/one partial nesting, we found predicted power rates for detecting moderation effects from the formulas that accommodated covariates closely aligned to empirical rejection rates across a multitude of conditions (e.g., different sample sizes and degrees of moderation effect heterogeneity). Table 3 presents selected results demonstrating power formula accuracy for detecting moderation effects in three/one partially nested designs (see Supplemental Materials for complete results). Results paralleled those when data had a two/one partially nested structure with increases to sample size increasing power rates and increases in moderation effect heterogeneity decreasing power rates. Power formulas also accurately reflected Type I error rate when *ME* = 0. As with previous simulation studies, we noted some discrepancies between formula predicted Type I error rate and empirical rejection rate when upper-level sample sizes were limited (e.g., $n_3^{(t)} = 10$) but found these quickly dissipated as upper-level sample sizes increased. Accuracy of the moderation effect error variance formulas was also confirmed using a comparison of empirical and formula-based values. These results suggest discrepancies between formula predicted Type I error rate and empirical rejection rate stems from difficulties with empirical rejection rates when $n_3^{(t)}$ values are small not inaccuracies with the variance and power formulas.

For designs with three/two partial nesting, we found predicted power rates for detecting moderation effects also closely aligned to empirical rejection rates across a multitude of conditions. Table 4 presents selected results demonstrating power formula accuracy for three/two partially nested designs (see Supplemental Materials for complete results). We again found increases to sample size increased power rates and increases in moderation effect heterogeneity decreased power rates. Upper-level sample sizes (e.g., $n_3^{(t)}$) were the primary driver of these increases. Power formulas also accurately reflected Type I error rate when $ME = 0.0$. Selected results in Table 4 do include some discrepancies between formula predicted Type I error rate and empirical rejection rate when upper-level sample sizes were limited (e.g., $n_3^{(t)} = 10$) but these again quickly dissipated as upper-level sample sizes increased.

Discussion

A more wholistic understanding of treatment effectiveness is possible through the inclusion of moderator variables to capture differences in treatment effects across groups and contexts (i.e., treatment effect heterogeneity). Unfortunately, power formulas were unavailable for detecting moderation effects when the relationship between the moderator and outcome varies randomly across clusters (i.e., random slopes) in studies with partially nested data. We derived these formulas and assessed their accuracy in three simulation studies that demonstrated their validity across a range of conditions including binary moderators, unbalanced sampling across study arms, disparate sample sizes, and covariates. The availability of these formulas improves study planning with partially nested designs that include important considerations regarding for treatment effect heterogeneity. We recommend researchers consider a range of values for moderation effect heterogeneity ($\tau_{11^{(t)}}^2$) to capture a range of adequate sample sizes. This approach will provide researchers who are uncertain about the slope of the moderator (i.e., fixed or random) a general idea of adequate sample sizes and the consequences of moderation effect heterogeneity. Our subsequent moderator effect error variance derivations provide researchers with an initial understanding of the factors that influence power to detect these effects.

Table 3. Comparisons of Formula-Based Statistical Power and Type I Error Rate and Monte Carlo Simulation Results for a Continuous Moderator with Covariates in a Design with Three/One Partially Nested Data

Scenario	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
ME	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0	0	0
ω	0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.4	0.8	0.8	0.8	0.8	0	0	0	0
n_3	25	10	10	10	25	10	10	10	25	10	10	10	25	10	10	10
n_2	25	25	10	10	25	25	10	10	25	25	10	10	25	25	10	10
n_1	25	25	25	10	25	25	25	10	25	25	25	10	25	25	25	10
Simulation rejection rate	0.89	0.45	0.39	0.32	0.64	0.26	0.24	0.21	0.36	0.15	0.15	0.13	0.04	0.02	0.02	0.02
Formula power rate	0.88	0.44	0.38	0.32	0.63	0.26	0.24	0.22	0.36	0.15	0.15	0.14	0.05	0.05	0.05	0.05
Absolute Difference	0.01	0.01	0.01	0.00	0.01	0.00	0.00	-0.01	0.00	0.00	0.00	-0.01	-0.01	-0.03	-0.03	-0.03

Note: Results were based on 10,000 replications with balanced sample sizes across study arms, ω representing moderation effect heterogeneity, $R^2_{y(t)} = R^2_{y(c)} = 0.4$ and *ME*=0.1 for power and *ME*=0 for Type I error.

Table 4. Comparisons of Formula-Based Statistical Power and Type I Error Rate and Monte Carlo Simulation Results for a Continuous Moderator with Covariates in a Design with Three/Two Partially Nested Data

Scenario	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
ME	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0	0	0
ω	0.2	0.2	0.2	0.2	0.4	0.4	0.4	0.4	0.8	0.8	0.8	0.8	0	0	0	0
n_3	25	10	10	10	25	10	10	10	25	10	10	10	25	10	10	10
n_2	25	25	10	10	25	25	10	10	25	25	10	10	25	25	10	10
n_1	25	25	25	10	25	25	25	10	25	25	25	10	25	25	25	10
Simulation rejection rate	0.48	0.18	0.16	0.15	0.27	0.11	0.10	0.10	0.15	0.06	0.07	0.06	0.03	0.01	0.01	0.01
Formula power rate	0.46	0.18	0.18	0.16	0.27	0.12	0.11	0.11	0.16	0.08	0.08	0.08	0.05	0.05	0.05	0.05
Absolute Difference	0.02	0.00	-0.02	-0.01	0.00	-0.01	-0.01	-0.01	-0.01	-0.02	-0.01	-0.02	-0.02	-0.03	-0.04	-0.04

Note: Results were based on 10,000 replications with balanced sample sizes across study arms, ω representing moderation effect heterogeneity, $R^2_{y(t)} = R^2_{y(c)} = 0.4$ and *ME*=0.1 for power and *ME*=0 for Type I error.

To conclude, we summarize six study design considerations for detecting moderation effects in partially nested designs:

- The presence and increasing magnitude of moderation effect heterogeneity ($\tau^2_{11^{(t)}}$) reduces the power to detect the moderation effect (e.g., compare across three scenarios in Table 1 and between odd and even scenarios in Table 2).
- Increasing intraclass correlation coefficient of the outcome (ρ) reduces the power to detect the moderation effect in two/one partially nested designs (e.g., compare scenarios 1-6 to 7-12 in Table 1).
- Cluster sample size ($n_2^{(.)}$) has a stronger influence on power to detect moderation effects than individual per cluster sample size ($n_1^{(t)}$) and this difference increases as moderation effect heterogeneity increases (e.g., compare scenario 1 with 3 and 2 with 4 in Table 2).
- Design considerations are applicable to continuous and binary moderators but studies with reduced moderator variance (e.g., binary moderators with a maximum variance of 0.25) require substantially larger sample sizes to detect moderation effects (e.g., see Table 5a in Supplemental Materials).
- Only minor decreases in power are likely when the control group is half to twice the size of the treatment group given equal total sample sizes (see Table 9a in Supplemental Materials).
- The use of prognostic covariates in the outcome model of treatment and control groups represents an effective strategy to increase the power to detect moderation effects but, like sample size, is dependent on the magnitude of moderation effect heterogeneity (e.g., compare scenarios 1-4, 5-8, and 9-12 in Table 2).

There are several important implications related to these design considerations. First, our newly derived formulas will improve the design of partially nested studies considering moderation effects as there are clear and substantial differences between the power to detect moderation effects with and without moderation effect heterogeneity. A study with partially nested data planned without regard to moderation effect heterogeneity will overestimate power if heterogeneity is present. Second, the influence of design parameters on power to detect moderation effects vary based on the magnitude of moderation effect heterogeneity. Study design recommendations for increasing power to detect moderation effects should consider the presence of moderation effect heterogeneity. Third, adequate power is still achievable with feasible sample sizes (e.g., $n_2^{(t)} < 100$ and $n_1^{(t)} < 100$) under a variety of conditions (i.e., various ω, ρ, and *ME* values) but larger sample sizes are needed for detecting moderation effects when moderation effect heterogeneity is present. We highly recommend the inclusion of covariates as a method to achieve adequate power with feasible sample sizes but note that the benefits of including covariates is dependent on moderation effect heterogeneity. Fourth, study planning should emphasize cluster sample size ($n_2^{(t)}$ or $n_3^{(t)}$). Previous literature detailing power to detect moderation effects in partially nested designs has noted increased power rates when increasing $n_1^{(t)}$ (Cox & Kelcey, 2022). However, our results indicate this relationship is negated by moderation effect heterogeneity which aligns with the findings of Dong et al., (2021).

Our power formulas for moderated effects when the moderator-outcome relationship varies across clusters (i.e., random slopes) in partially nested designs accommodate continuous and binary moderators, analytic models with and without covariates, and balanced and unbalanced sample sizes across study arms. They are, however, restricted to moderation effects at the individual-level from a moderator located at the individual-level. Future research should examine upper-level moderation effects from aggregated individual-level moderators with a random slope or moderators located at the cluster-level with random slopes. While these investigations are not applicable for two/one or three/one partially nested designs they can be considered with three/two partial nested designs.

Our derivations of moderator effect error variance and subsequent probe of factors that influence power only serves as an initial investigation. We also encourage future research to more comprehensively map out the factors that influence these power rates and adequate sample sizes under various conditions. This should include examining the most efficient sample size allocation across various conditions (e.g., optimal sample allocation), multiple arm partially nested designs, and the influence of unequal cluster sizes. We would expect power rates for detecting moderation effects to decrease as severe imbalance in cluster sample sizes are observed (Guittet et al., 2006; Moerbeek, 2018) and for multiple arm partially nested designs to provide

increased efficiency and power rates compared to equivalent separate trials (Li et al, 2017; Odutayo et al., 2020) but these relationships have not been examined. Additionally, research is needed to establish empirically based estimates for the design parameters necessary for planning partially nested designs that include moderation effects. Extensive work has been completed in this area for cluster randomized designs (e.g., Hedges & Hedberg, 2007) but has yet to be developed for partially nested designs. This work is important to better plan these types of studies and guide simulation parameters in future methodological research. Even considering these limitations, this work represents a significant contribution to study design literature. We have derived, established the accuracy of, and investigated power formulas for detecting individual-level moderation effects in partially nested designs. These advancements improve the planning of applicable studies and encourage more personalized treatment decisions by allowing researchers to elucidate populations or conditions in which a treatment is effective.

References

Ashcraft, M., & Krause, J. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review, 14*(2), 243-248.

Ashcraft, M., & Moore, A. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment, 27*(3), 197-205.

Bai, F., Xie, Y., Kelcey, B., Ataneka, A., McLean, L., & Phelps, G. (2024). Design parameter values for planning mediation studies with teacher and student mathematics Outcomes. *Journal of Research on Educational Effectiveness*, 1–35.

Baldwin, S., Bauer, D., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods, 16*, 149–165.

Bates D, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(*1*), 1–48.

Bauer, D., Sterba, S. K., & Hallfors, D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research, 43*(2), 210–236.

Cox, K., & Kelcey, B. (2023). Statistical power for detecting moderation in partially nested designs. *American Journal of Evaluation, 44*(*1*), 133–152.

Cox, K., Kelcey, B., & Luce. (2024). Power to detect moderated effects in studies with three-level partially nested data, *The Journal of Experimental Education*, *92*(*1*), 130–153.

Candlish, J., Teare, M., Dimairo, M., Flight, L., Mandefield, L., & Walters, S. (2018). Appropriate statistical methods for analyzing partially nested randomized controlled trials with continuous outcomes: a simulation study. *BMC Medical Research Methodology, 18*, 105.

Dishion, T., Poulin, F., & Burraston, B. (2001). Peer group dynamics associated with iatrogenic effects in group interventions with high-risk young adolescents. *New Directions for Child and Adolescent Development, 91*, 79–92.

Dong, N., Kelcey, B., & Spybrook, J. (2018). Power analyses for moderator effects in three-level cluster randomized trials. *The Journal of Experimental Education*, *86*(3), 489–514.

Dong, N., Spybrook, J., Kelcey, B., & Bulus, M. (2021). Power analyses for moderator effects with (non)randomly varying slopes in cluster randomized trials. *Methodology, 17*(2), 92–110.

Esserman, D., Zhao, Y., Tang, Y., & Cai, J. (2013). Sample size estimation in educational intervention trials with subgroup heterogeneity in only one arm. *Statistics in Medicine, 32*(12), 2140–2154.

Hedges, L., & Hedberg, E. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Heisig, J. P., & Schaeffer, M. (2019). Why you should always include a random slope for the lower-level variable involved in a cross-Level interaction. *European Sociological Review*, *35*(2), 258–279.

Heo, M., Litwin, A. H., Blackstock, O., Kim, N., & Arnsten, J. H. (2017). Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. *Statistical Methods in Medical Research*, *26*(1), 399–413.

Kelcey, Bai, F., & Xie, Y. (2020). Statistical power in partially nested designs probing multilevel mediation. *Psychotherapy Research*, *30*(8), 1061–1074.

Kelcey, B., Shen, Z., & Spybrook, J. (2016). Intraclass correlation coefficients for designing cluster-randomized trials in Sub-Saharan Africa education. *Evaluation Review, 40*(6), 500-525.

Lachowicz, M. J., Sterba, S. K., & Preacher, K. J. (2015). Investigating multilevel mediation with fully or partially nested data. *Group Processes & Intergroup Relations*, *18*(3), 274–289.

LaHuis, D., Jenkins, D. R., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2020). The effects of misspecifying the random part of multilevel models. *Methodology*, *16*(3), 224–240.

Lazar, A., Bonetti, M., Cole, B. F., Yip, W., & Gelber, R. D. (2016). Identifying treatment effect heterogeneity in clinical trials using subpopulations of events: STEPP. *Clinical Trials*, *13*(2), 169–179.

Li, H., & Hedeker, D. (2017). Statistical methods for continuous outcomes in partially clustered designs. *Communications in Statistics. Theory and Methods*, *46*(8), 3915–3933.

Lee, K. J., & Thompson, S. G. (2005). The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials*, *2*(2), 163–173.

Lohr, S., Schochet, P.Z., and Sanders, E. (2014). Partially nested randomized controlled trials in education research: A guide to design and analysis. (NCER 2014-2000) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the Institute website at http://ies.ed.gov/

Lowrie, Harris, D., Logan, T., & Hegarty, M. (2021). The Impact of a Spatial Intervention Program on Students' Spatial Reasoning and Mathematics Performance. *The Journal of Experimental Education*, *89*(2), 259–277.

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*(5), 951–966.

Moerbeek, M., & Wong, W. K. (2008). Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine*, *27*(15), 2850–2864.

Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., & Baker, W. L. (2008). Individually randomized group treatment trials: A critical appraisal of frequently used design and analytic approaches. *American Journal of Public Health*, *98*(8), 1418–1424.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*(1), 5–29.

Roberts, C. (2021). The implications of noncompliance for randomized trials with partial nesting due to group treatment. *Statistics in Medicine*, *40*(2), 349–368.

Roberts, C., Batistatou, E., & Roberts, S. A. (2016). Design and analysis of trials with a partially nested design and a binary outcome measure: Partially Nested Binary Data. *Statistics in Medicine*, *35*(10), 1616–1636.

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*(2), 152–162.

Roberts, C., Williams, K., Carter, M., Evans, D., Parmenter, T., Silove, N., Clark, T., & Warren, A. (2011). A randomized controlled trial of two early intervention programs for young children with autism: Centre-based with parent program and home-based. *Research in Autism Spectrum Disorders*, *5*(4), 1553–1566.

Rothschild, S. K., Martin, M. A., Swider, S. M., Lynas, C. T., Avery, E. F., Janssen, I., & Powell, L. H. (2011). The Mexican-American trial of community health workers (MATCH): Design and baseline characteristics of a randomized controlled trial testing a culturally tailored community diabetes self-management intervention. *Contemporary Clinical Trials*, *33*(2), 369–377.

Schnell, P., Müller, P., Tang, Q., & Carlin, B. P. (2018). Multiplicity-adjusted semiparametric benefiting subgroup identification in clinical trials. *Clinical Trials*, *15*(1), 75–86.

Shin, Y. (2012). Do Black children benefit more from small classes? Multivariate instrumental variable estimators with ignorable missing data. *Journal of Educational and Behavioral Statistics*, *37*(4), 543–574.

Snijders T. (2001). Sampling. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modeling of health statistics*. (pp. 159-173). Wiley.

Snijders T. (2005). Power and sample size in multilevel linear models. In: B.S. Everitt and D.C. Howell (eds.), *Encyclopedia of statistics in behavioral science.* (pp. 1570–1573). Wiley.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effects in two- and three-level cluster randomized trials. *Journal of Educational and Behavioral Statistics*, *41*(6), 605–627.

Spybrook, J., Shi,R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research and Method in Education*, *39* (3), 255-267.

Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research*, *27*(4), 425–436.

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research*, *49*(2), 93–118.

Yang, S., Li, F., Starks, M. A., Hernandez, A. F., Mentz, R. J., & Choudhury, K. R. (2020). Sample size requirements for detecting treatment effect heterogeneity in cluster randomized trials. *Statistics in Medicine*, *39*(28), 4218–4237.

Send correspondence to: Kyle Cox
University of North Carolina at Charlotte
Email: kyle.cox@charlotte.edu