Experimental Methodology

American Journal of Evaluation
2023, Vol. 44(1) 133-152

DOI: 10.1177/1098214020977692
journals.sagepub.com/home/aje
SAGE

Statistical Power for Detecting Moderation in Partially Nested Designs

Kyle Cox[1] and Benjamin Kelcey[2]

Abstract

Analysis of the differential treatment effects across targeted subgroups and contexts is a critical objective in many evaluations because it delineates for whom and under what conditions particular programs, therapies or treatments are effective. Unfortunately, it is unclear how to plan efficient and effective evaluations that include these moderated effects when the design includes partial nesting (i.e., disparate grouping structures across treatment conditions). In this study, we develop statistical power formulas to identify requisite sample sizes and guide the planning of evaluations probing moderation under two-level partially nested designs. The results suggest that the power to detect moderation effects in partially nested designs is substantially influenced by sample size, moderation effect size, and moderator variance structure (i.e., varies within groups only or within and between groups). We implement the power formulas in the R-Shiny application PowerUpRShiny and demonstrate their use to plan evaluations.

Keywords

experimental design, partially nested, moderation, statistical power

A common goal of experimental evaluations is determining the average effectiveness of a program, intervention, or policy (i.e., treatment). However, treatment effectiveness can depend on the individual (for whom) and contextual factors (under what conditions). Inclusion of moderator variables that capture the factors by which effects vary is a common technique to investigate treatment effect heterogeneity. Planning evaluations that consider treatment effect moderation is aided by the availability of power formulas for detecting moderation effects in various randomized designs, but these formulas are unavailable for partially nested designs in which the grouping structure of the treatment condition is different from that of the control condition. In this study, we develop, describe, and

[1] Educational Research, Measurement, and Evaluation, Cato College of Education, University of North Carolina at Charlotte, NC, USA
[2] Quantitative and Mixed Methods Research Methodologies, University of Cincinnati, OH, USA

Corresponding Author:
Kyle Cox, Educational Research, Measurement, and Evaluation, Cato College of Education, University of North Carolina at Charlotte, Room 266, Charlotte, NC 28223, USA.
Email: kyle.cox@uncc.edu

investigate power formulas for moderated effects in evaluations with partial nesting to ensure adequate sample sizes and improve evaluation planning.

Our analyses focus on experimental evaluations because of their ability to produce strong causal evidence regarding the effectiveness of a treatment but recognize the value of supplementary questions. Understanding the variability of treatment effects across differing individual characteristics (e.g., race, gender, pretest score) or group characteristics (e.g., organization size or location) reveals for whom and under what conditions a treatment is effective. It also provides an avenue to investigate opposing treatment effects across groups that produce a null main effect (MacKinnon et al., 2011). For example, a near-zero treatment effect would be produced by an educational intervention that increases academic achievement among male students but decreases achievement among female students.

The value of capturing a more complete understanding of treatment effects is often reflected in evaluation literature through the use of moderator variables that detail variance in the effects of an intervention across different groups. For example, recent literature has used this approach to study an intervention aimed at reducing violent video game consumption across different lifestyles (Rivera et al., 2016); a treatment program for mental illness across race, gender, and age (Kenny et al., 2004); an online intervention for depression across education levels, attitudes toward online instruction, and willingness to change (Lüdtke et al., 2018); and an implementation intentions method to increase physical activity across levels of executive function (Hall et al., 2014). Funders and professional organizations also emphasize investigations that capture a more complete understanding of treatment effects (e.g., Institute of Education Sciences, 2016; Society for Research on Educational Effectiveness, 2012).

In conjunction with the increasing emphasis on moderated treatment effects is a growing literature detailing design and analysis techniques that support the inclusion of moderator variables. For example, existing research details the inclusion and analysis of different types of moderator variables (e.g., categorical or continuous) in various experimental study designs (Bloom, 2005; Dong et al., 2018; Jaciw et al., 2016; Spybrook et al., 2016). Many of these advancements have been implemented in software (e.g., Dong et al., 2016), expanding the capacity of evaluators to plan for and capture treatment effect moderation. This capacity does not, however, extend to designs with partial nesting which occur when there are disparate grouping or nesting structures across treatment conditions.

A wide variety of partial nesting structures occur in practice (e.g., Sterba et al., 2014). Such structures can occur when, for example, assignment to a treatment condition eliminates some nested structure (e.g., homeschooling treatment vs. typical schooling control condition) or when assignment induces or utilizes some nesting structure in the treatment condition that does not exist in the control condition. For example, a common design in the field of counseling involves randomly assigning individuals to therapy led by a counselor or to a waitlist control condition (e.g., Roberts & Roberts, 2005). In this setting, treatment individuals are nested within counselors while control individuals are not nested. Similar examples of partial nesting occur in a wide range of fields (e.g., Bauer et al., 2008; Lohr et al., 2014; Sterba et al., 2014). In the medical field, patients can receive novel treatments in a clinic setting or be assigned to receive the typical home-based care in the control condition (e.g., Morrell et al., 1998); in education settings, a treatment may consist of a school-based intervention and be compared to an individualized home-based intervention control condition (e.g., Roberts et al., 2011); and in psychotherapy, the treatment condition may involve a group therapy while the control condition utilizes individual therapy (e.g., Dishion et al., 2001).

Partial nesting occurs across these examples because assignment to the treatment condition creates a grouping structure that is dissimilar to the structure of the control condition. Those in the treatment condition of the provided examples experience the treatment as a group (e.g., classroom, organization, neighborhood) or share a common agent of implementation (e.g., patients with the same counselor or students with the same tutor). The result is a treatment condition with a two-level data structure brought on by treatment delivery but a control condition without any grouping

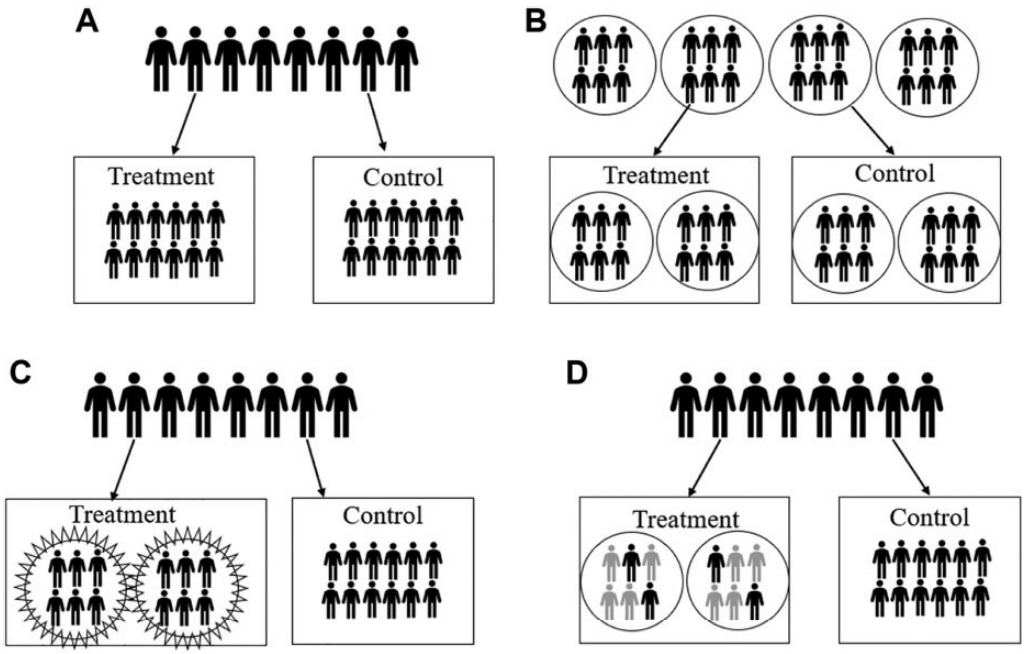

Figure 1. Individual, cluster, and partially nested randomized designs. (A) . . . .(D)

structure. Our examples focus on these "two/one" partially nested designs that combine the grouping structure of a two-level cluster-randomized trial in the treatment condition with the single-level structure of an individual-randomized trial in the control condition. The two/one partially nested nomenclature stems from the design having two levels in the first treatment condition and one level in the other condition. These designs are common in education, public health, and other social science settings (e.g., Bauer et al., 2008; Lohr et al., 2014; Sanders, 2011).

For clarity, Figure 1 illustrates the two/one partially nested designs included in the scope of this investigation (see Panels C and D) along with the more common individual- (Panel A) and cluster-randomized design (Panel B). The purpose of Figure 1 is to show the differences and similarities in treatment and control condition grouping across these designs. Panel (A) in Figure 1 presents a typical individual-randomized design that does not have any grouping in the treatment or control condition. Panel (B) in Figure 1 displays a typical cluster-randomized design with intact groups (represented using circles) randomly assigned to the treatment or control condition. Panel (C) in Figure 1 displays a partially nested design that leverages individual assignment but whose treatment induces a two-level nesting structure in the treatment condition. We emphasize that this grouping is new (i.e., treatment-induced) using spikes around the group circles in the figure. Notice that the partially nested design combines an unclustered structure for the control condition with a two-level nesting structure for the treatment condition that was brought about by the nature of the treatment (e.g., sharing a therapist or teacher). Panel (D) in Figure 1 displays a similar type of partially nested design as in Panel (C) because it yields a two-level nesting structure in the treatment condition and an unclustered structure in the control condition. However, the clustering structure in the treatment condition of Panel (D) arises differently. In Panel (D), grouping is an artifact of a preexisting social structure (e.g., extant groups). To illustrate extant grouping, we have study participants (black figures) join others (gray figures) in existing groups (solid circles).

Evident from the panels in Figure 1, the structures of the control conditions are identical across the partially nested designs (Panels C and D) and the individual-randomized design (Panel A) because all three designs leverage individual assignment that results in an ungrouped or unclustered control condition. Similarly, the structures of the treatment conditions in Panels B (cluster-randomized trial), C (two/one treatment-induced partial nesting), and D (two/one partial nesting with extant groups) are the same as they each result in a two-level structure for the treatment condition. The distinguishing feature of the partially nested designs (C and D) is their use of different grouping structures in the treatment and control conditions. The specific partially nested designs are distinguished by the mechanism that produces the grouping or clustering. Recall in (C), the clustering is generated by the nature of the treatment (e.g., individuals assigned to treatment are grouped to form a new therapy group) whereas in (D), the clustering exists prior to the study (e.g., individuals are assigned to an existing therapy group).

The unique data structure of partial nesting is often disregarded, and past research has widely documented the various problems this can introduce in terms of efficiency, bias in standard errors of the treatment effect, and bias in estimates of variance components. These problems lead to inaccurate results and incorrect inferences (Baldwin et al., 2011; Bauer et al., 2008; Candlish et al., 2018; Hedges & Citkowicz, 2015; Korendijk et al., 2012; Lee & Thompson, 2005; Sanders, 2011; Schweig & Pane, 2016). Increasing attention has focused on development of analytic approaches and design strategies to address these issues in partially nested studies. For example, multiple-arm multilevel models have been extended for partially nested data (e.g., Lachowicz et al., 2015; Lohr et al., 2014; Sterba et al., 2014), and several studies have investigated sample size considerations and the use of covariates in partially nested designs (e.g., Moerbeek & Wong, 2008; Roberts & Roberts, 2005).

The purpose of this study is to develop statistical power formulas for moderated effects in common two/one partially nested designs and investigate these formulas to provide guidance and recommendations for evaluation planning. We structure our analyses to address two/one partially nested designs in which (a) treatment assignment induces a nesting structure in the treatment condition such that individual-level moderators plausibly vary only within groups and (b) treatment assignment inserts individuals into an extant nesting structure in the treatment condition such that individual-level moderators plausibly vary within and between groups.

In the first design (see Panel C of Figure 1), random assignment introduces a nesting structure such that individual-level moderator variables vary within but not across groups. It may seem tenuous to assume that the variability of an individual-level moderator arises solely from differences among individuals and not from differences among groups. However, because the formation of treatment conditions in partially nested designs is frequently the specific feature that induces nesting that would not otherwise exist, the values of an individual-level pretreatment moderator variable are typically established before assignment and exposure to the cluster-inducing treatment. As a result, the values of individual-level moderators will typically not be clustered and will not have variation at the group level.

Consider an evaluation of a counseling therapy in which individuals are randomly assigned to participate in therapy with a therapist (treatment) or to remain on a waitlist (control). This design creates a partially nested structure because treatment individuals are nested within therapists whereas control individuals remain ungrouped on a waitlist. Now consider an individual-level moderator such as pretreatment mental health. With individuals randomly assigned to therapists, there is no reason to suspect that the therapist-level averages of pretreatment individual-level mental health will systematically differ across therapists. Pretreatment mental health should be equally dispersed across therapists because pretreatment mental health took on values before individuals were assigned to therapists.

In the second design (see Figure 1 Panel D), we consider evaluations in which individuals assigned to the treatment condition are inserted into a preexisting nested structure such that individual-level moderator variables plausibly vary within and across groups. In this design, treated individuals participate in extant groups rather than forming new groups. Consider an evaluation of a group therapy in which individuals are randomly assigned to group therapy sessions (treatment) in preexisting groups or

remain on a control waitlist. Treatment condition structure in this design has individuals nested in therapy groups whereas the control condition continues to be unclustered.

For example, let us consider pretreatment assessment of individual mental health as a moderator of group therapy effectiveness. When groups are formed prior to treatment assignment, it is plausible that the groups differ in their average pretreatment levels of mental health. Pretreatment mental health may be clustered within groups because of, for example, the prior progress of the groups or the self-selection of individuals with similar mental health levels into a therapy group. These average differences in prior mental health may play important contextual roles that moderate the effectiveness of the therapy. It is possible that the therapy is highly effective for groups with high average pretreatment mental health but ineffective for groups with low average pretreatment mental health. As a result, the average pretreatment mental health of a group may play an important moderating role.

Below, we further detail two/one partially nested designs setting a foundation for subsequent power formula development. We outline the analytic models, describe the error variance of the moderation effect, then provide formulas for estimating statistical power in two/one partially nested designs with treatment conditions that induce nesting. We repeat this process for designs that assign treatment to extant groups. A probe of the newly developed power formulas investigates the feasibility of detecting moderator effects in evaluations with partial nesting. This is followed by an illustrative example to demonstrate the application of the formulas in evaluation planning. To conclude, we summarize results, discuss implications, note limitations, and provide recommendations.

Two/One Partial Nesting

In designs with two/one partial nesting, individuals are randomly assigned to a treatment or control condition one of which has a two-level data structure while the other has a single-level data structure. Most often, individuals in the control condition are unaffected by treatment grouping. This creates a treatment condition with a two-level data structure and a control condition with a single-level structure. Our derivations apply to partially nested designs with a two-level data structure in the control condition, but we focus on those with a two-level data structure in the treatment condition (as illustrated in Figure 1 Panels C and D).

We noted an example in counseling when individuals are randomly assigned to receive a treatment delivered by a therapist or placed on a waitlist. This two/one partially nested design has a treatment-induced nesting structure (patients within a therapist) with a waitlist control condition comprised of independent individuals (i.e., the control condition retains a single-level structure). Two/one partially nested designs can also arise when treatments use extant grouping structure. In the context of our counseling example, assigning individuals to group therapy sessions utilizes a pre-existing two-level treatment (i.e., individuals within therapy groups) with wait-listed control individuals retaining a single-level data structure. In either case, the treatment condition has two levels while the control condition has one level.

We take up these two complementary types of two/one partial nesting separately because moderator variability likely differs under each type. We refer to the two types of partial nesting as (a) treatment-induced nesting (moderator plausibly varies only within groups) and (b) treatment assignment to extant nesting (moderator plausibly varies within and between groups). Below, we examine each of the scenarios assuming that moderators are continuous variables but formulas are adaptable to binary moderators (see Binary Moderator in the Technical Supplemental Appendix).

Treatment Assignment–Induced Nesting Structure

Treatment-induced nesting structure results in a moderator that plausibly varies within groups only. We use a working example to help ground the analytic models, moderator effect variance formulas,

and subsequent power formulas. This hypothetical evaluation in an educational setting investigates the effectiveness of a spatial intervention program on secondary student mathematics performance (Lowrie et al., 2019). The spatial intervention is implemented as a summer school program with a sample of students selected from those performing below proficient levels in mathematics. Students randomly assigned to the control condition will be placed on a waitlist and will continue with their summer as usual. Students assigned to the treatment condition will complete the classroom-based intervention program over 3 weeks in the summer. This evaluation has a two/one partially nested design with the treatment condition containing two levels (i.e., students nested within classrooms) and the control condition (i.e., wait-listed students) representing a single level.

In addition to considering the main effect of the spatial intervention program on secondary student mathematics performance, we include math anxiety as a possible moderator. Student math anxiety represents a typical individual-level continuous moderator. Math anxiety has a deleterious relationship with math performance (e.g., Ashcraft & Krause, 2007; Ashcraft & Moore, 2009), and it is possible that the spatial intervention program has differentiated effects based on a student's math anxiety. Given random assignment of students to intervention groups, we can assume that math anxiety will not have variation at the group level. There is no reason with treatment-induced nesting to suspect that average math anxiety in the intervention groups will systematically differ across groups. Although math achievement (outcome) likely varies across individuals and groups, random assignment of individuals to groups ensures that in expectation pretreatment covariates such as math anxiety will be evenly distributed across groups.

Analytic Models

We use two analytic models to reflect the different treatment and control conditions and draw on the common multiple-arm multilevel framework for partially nested data (MA-PN). This approach makes power formulas more accessible (Spybrook et al., 2016). For the two-level treatment condition with an individual-level continuous moderator that only varies within groups, we have

$$\begin{aligned} Y_{\mathrm{ij}} &= \pi_{0j}^{(t)} + \Delta^{(t)} M_{\mathrm{ij}} + \pi_1^{(t)}(X_{\mathrm{ij}} - \bar{X}_j) + \pi_2^{(t)} V_{\mathrm{ij}} + \varepsilon_{\mathrm{ij}}^{(t)} \qquad \varepsilon_{\mathrm{ij}}^{(t)} \sim N\left(0, \sigma^2_{Y^{(t)}_{|}}\right) \\ \pi_{0j}^{(t)} &= \delta^{(t)} + \zeta_1^{(t)} \bar{X}_j + \zeta_2^{(t)} W_j + u_{0j}^{(t)} \qquad u_{0j}^{(t)} \sim N\left(0, \tau^2_{Y^{(t)}_{|}}\right) \end{aligned} \tag{1}$$

The superscript t indicates the treatment arm, and subscripts i and j follow common multilevel model notation indicating individual and group, respectively. The outcome is represented by Y_{ij} and interpreted in our example as the math performance score from student (i) in classroom (j) after completing the spatial intervention program (t). Covariates in the model (X, V, and W) explain extraneous variation in the outcome reflecting a well-established design strategy that improves study efficiency, increases power to detect a treatment effect, and reduces the sample size necessary to achieve adequate power (see Covariates in Technical Supplemental Appendix; e.g., Raudenbush et al., 2007). Investigations of main effects focus on $\delta^{(t)}$, the overall intercept that represents the conditional average of the outcome value in the treatment condition. In terms of our example, $\delta^{(t)}$ is the conditional average math score for students who completed the spatial intervention program. Investigations of treatment effect moderation focus on M_{ij}, which represents the individual-level continuous moderator for individual i in group j. Math anxiety for student i in classroom j in our example. The coefficient $\Delta^{(t)}$ captures the relationship between the moderator (math anxiety) and outcome (math performance). We have a group- or classroom-specific residual, $u_{0j}^{(t)}$ which represents the classroom-level random effects for the treatment condition (i.e., what variance in math performance is attributable to the classroom) with $\tau^2_{Y^{(t)}_{|}}$ representing variation in $u_{0j}^{(t)}$ across groups or

classrooms. The individual-level error term is $\varepsilon_{ij}^{(t)}$ (i.e., an indicator of the precision with which we are predicting a student's math performance) with $\sigma^2_{Y_|^{(t)}}$ as its conditional within-group variance.

For the single-level control arm, the outcome model is

$$Y_i = \delta^{(c)} + \Delta^{(c)} M_i + \pi_1^{(c)} X_i + \pi_2^{(c)} V_i + \varepsilon_i^{(c)} \qquad \varepsilon_i^{(c)} \sim N\left(0, \sigma^2_{Y_|^{(c)}}\right) \tag{2}$$

Most variables (e.g., Y, M, X, and V) and parameters (e.g., $\delta^{(c)}$,$\Delta^{(c)}$, and π) retain similar meaning from the treatment condition model (see Equation 1). Differences include a superscript c indicating the control arm and a single subscript i indicating individuals (i.e., students) are not nested or grouped. Variance is also simplified, with $\sigma^2_{Y_|^{(c)}}$ representing outcome variance in the control condition and $\varepsilon_i^{(c)}$ representing the associated error term that varies across individuals. We focus on the coefficient capturing the relationship between the moderator and outcome in the control condition, $\Delta^{(c)}$.

Moderator Effect and Error Variance

Estimation of the moderator effect (ME) is possible by contrasting the coefficients capturing the relationship between the moderator and outcome with

$$\text{ME} = \Delta^{(t)} - \Delta^{(c)}. \tag{3}$$

A difference between $\Delta^{(t)}$ and $\Delta^{(c)}$ suggests that the relationship between the treatment and outcome differs by moderator value. In our example, a difference between $\Delta^{(t)}$ and $\Delta^{(c)}$ suggests that the effect of the spatial intervention is dependent on math anxiety. In other words, math anxiety plausibly moderates the effects of the spatial intervention program on math performance. The statistical significance of the moderated effect can be determined using a t test (see Test Statistic and Power Formula in Technical Supplemental Appendix) with two key components: (a) the estimate of the ME in Equation 3 and (b) the error variance of ME ($\sigma^2_{\text{ME}_{\text{within}}}$).

The novel contribution of our power formulas are expressions to track the expected uncertainty of the moderator effect ($\sigma^2_{\text{ME}_{\text{within}}}$) using summary statistics that can be predicted a priori (i.e., before data have been collected). The $\sigma^2_{\text{ME}_{\text{within}}}$ expressions are suitable for power analyses with summary statistic components identified using historical data, pilot study results, and published catalogs (e.g., Aguinis et al., 2005; Hedges & Hedberg, 2007; Stone-Romero & Liakhovitski, 2002). We provide a brief development of $\sigma^2_{\text{ME}_{\text{within}}}$ in Equations 4 and 5 with a detailed presentation available in the Technical Supplemental Appendix. With independence across treatment conditions, the variance of the moderated effect is the sum of the variances of the moderator coefficients:

$$\sigma^2_{\text{ME}_{\text{within}}} = \sigma^2_{\Delta^{(t)}} + \sigma^2_{\Delta^{(c)}}. \tag{4}$$

(see Moderator Effect and Error Variance in Technical Supplemental Appendix for expanded formulations). We unpack $\sigma^2_{\text{ME}_{\text{within}}}$ as a function of parameters that are accessible in the design stage to facilitate prospective power analyses such that (see Technical Supplemental Appendix for details)

$$\sigma^2_{\text{ME}_{\text{within}}} = \frac{\sigma^2_{Y^{(t)}}(1 - R^2_{Y^{L1}_{(t)}})/n_1^{(t)}}{(n_2^{(t)} - C_{(t)} - 1)\sigma^2_{M^{(t)}}(1 - R^2_{M^{L1}_{(t)}})} + \frac{\sigma^2_{Y^{(c)}}(1 - R^2_{Y^{L1}_{(c)}})}{(n^{(c)} - C_{(c)} - 1)\sigma^2_{M^{(c)}}(1 - R^2_{M_{(c)}})} \tag{5}$$

with $\sigma^2_{Y^{(t)}}$ and $\sigma^2_{Y^{(c)}}$ capturing the sample variance of the outcome (e.g., variation in student math performance) in the treatment and control condition, respectively. Paired with each outcome variance term is a $1 - R^2$ term. The $R^2_{Y^{L1}_{(t)}}$ and $R^2_{Y^{L1}_{(c)}}$ terms represent variance explained in the outcome by

predictor variables (i.e., M, X, and V). The R^2 value can vary between zero and one such that increasing values of R^2 result in smaller and smaller $1 - R^2$ values. The product of the $1 - R^2$ and $\sigma^2_{Y(.)}$ produces a smaller value indicating variance in the outcome after conditioning on predictors. For instance, a pretest on math performance would often be an effective covariate for reducing outcome variance in our working example. Using a student's math performance score from a previous year, we might reasonably achieve an R^2 value around .75. With $\sigma^2_{Y(.)} = .8$, we have $0.8(1 - 0.75) = 0.2$ (Bloom et al., 2007). Inclusion of the covariate has reduced $\sigma^2_{Y(.)}$ from .8 to .2 leading to a reduction in the overall variance of the moderator effect and an increased ability to detect significant moderation effects.

Next, we have individual per group sample size ($n_1^{(t)}$) in the term representing the treatment condition moderator coefficient variance ($\sigma^2_{\Delta(t)}$). As a divisor of $\sigma^2_{Y(t)}$, $n_1^{(t)}$ reduces error variance of the moderated effect which typically leads to increased power to detect the moderated effect. A group sample size term, $n_2^{(t)}$, is also present in the $\sigma^2_{\Delta(t)}$ term. The $n_1^{(t)}$ and $n_2^{(t)}$ terms are absent from the control condition term because the single-level control condition only has a total sample size ($n^{(c)}$). Increases in students per classroom in our example reduces moderator effect variance through reductions in outcome variance. Increases to the sample of classrooms in the treatment condition and/or total sample in the control condition also reduce moderator effect variance. The remaining terms in the formula represent variance of the moderator ($\sigma^2_{M(.)}$) and variance explained in the moderator by predictors ($R^2_{M_{(.)}}$; see Moderator Model in Technical Supplemental Appendix). In terms of our example, they represent the dispersion or variation in math anxiety and variance in math anxiety explained by covariates (i.e., X and V). The relationship between these terms parallels that described for $\sigma^2_{Y(.)}$ and $R^2_{Y_{(.)}}$.

The variance of the moderated effect formulas indicate which parameters and design components are necessary for an a priori power analysis in a two/one partially nested design. Evaluators must predict the magnitude of the moderated effect, outcome and moderator variance structure (i.e., intraclass correlation coefficients [ICCs]), proportion of variance explained by predictors, and several sample sizes ($n_1^{(t)}$, $n_2^{(t)}$, and $n^{(c)}$). Predicted moderator effect and variance structure are typically identified through a pilot study or based on previous empirical research. Evaluators may test a range of sample sizes to identify those that provide adequate power while considering practical constraints (e.g., classroom size) or budgetary limitations.

To summarize, we developed a formulation for moderator effect variance in two/one partially nested designs with a treatment-induced nesting structure that plausibly limits moderator variance to within groups. The moderator effect variance formulation is suitable for power analysis in the planning stages of an evaluation with formula structure suggesting that greater variance of the outcome ($\sigma^2_{Y(t)}$ and $\sigma^2_{Y(c)}$) increases moderator effect variance while increasing sample sizes ($n_1^{(t)}$, $n_2^{(t)}$, and $n^{(c)}$) and moderator variance ($\sigma^2_{M(t)}$ and $\sigma^2_{M(c)}$) decreases the error variance of the moderator effect. Outcome and moderator variance are typically not malleable and often standardized for evaluation planning purposes. These terms are encoded in the effect size in a power analysis so we limit further discussion. Actionable implications include increasing the sample of groups and individuals per group to reduce $\sigma^2_{Y(t)}$ and $\sigma^2_{Y(c)}$ or including prognostic covariates that explain variance in $\sigma^2_{Y(t)}$ and $\sigma^2_{Y(c)}$. These actions increase the likelihood of finding a significant moderated effect.

Treatment Assignment to Extant Nesting Structure

When the treatment assignment utilizes extant grouping, it is no longer tenable to assume that moderators vary within groups only. With preexisting groups, the average moderator values may systematically differ across groups. Consider a new working example evaluation that investigates the effectiveness of a group-based intensive lifestyle intervention on weight loss (e.g., Mayer-Davis et al., 2004). The intervention consists of weekly group sessions encouraging physical activity and

proper nutrition. An evaluation design could randomly assign a pool of volunteers to attend ongoing group-based lifestyle intervention sessions (i.e., two-level treatment) or continue with their current care (i.e., single-level control). That is, volunteers are randomly assigned to join groups formed independently and before the onset of the study. Group-based intensive lifestyle interventions have demonstrated an ability to increase weight loss among participants (Mayer-Davis et al., 2004), but these effects have been shown to be moderated by personal characteristics such as optimism (e.g., Scheier & Carver, 1992; Van Nguyen et al., 2018). Optimism represents an individual-level continuous moderator that plausibly varies across groups under this design because the intervention groups were formed prior to assignment. The use of extant groups makes it plausible that groups will differ in their average level of optimism. These average differences may influence the effectiveness of the intensive lifestyle intervention on weight loss.

To consider situations, like our working example, in which a variable's aggregate or average may moderate treatment effects, we adjust the treatment outcome model to allow the continuous moderator (accommodations for binary moderators remain unchanged) to vary within and between groups such that

$$\begin{aligned} Y_{\text{ij}} &= \pi_{0j}^{(t)} + \Delta_1^{(t)}(M_{\text{ij}} - \bar{M}_j) + \pi_1^{(t)}(X_{\text{ij}} - \bar{X}_j) + \pi_2^{(t)} V_{\text{ij}} + \varepsilon_{\text{ij}}^{(t)} \qquad \varepsilon_{\text{ij}}^{(t)} \sim N\left(0, \sigma^2_{Y_{|}^{(t)}}\right) \\ \pi_{0j}^{(t)} &= \delta^{(t)} + \Delta_2^{(t)} \bar{M}_j + \zeta_1^{(t)} \bar{X}_j + \zeta_2^{(t)} W_j + u_{0j}^{(t)} \qquad u_{0j}^{(t)} \sim N\left(0, \tau^2_{Y_{|}^{(t)}}\right) \end{aligned} \tag{6}$$

When considering moderators that vary within and between groups, a moderation effect can occur at Level 1 (ME_W) and a moderation effect can occur at Level 2 (ME_B). The Level 1 moderation effect (ME_W) describes how the individual-level component of the moderator influences the relationship between treatment and outcome (e.g., for whom style questions). The Level 2 moderation effect (ME_B) describes how context (e.g., average of moderator) moderates the relationship between the treatment and outcome (e.g., under what circumstances style questions). Equation 6 includes the aggregated moderator, $\bar{M}_j$, with accompanying coefficient $\Delta_2^{(t)}$ to consider a moderator that varies between groups. By utilizing group mean centering on the individual level $(M_{\text{ij}} - \bar{M}_j)$, $\Delta_2^{(t)}$ captures the total relationship between the moderator and outcome at both levels (i.e., ME_{W+B}). Separate investigations of moderation effects at each level are possible with minor adaptations to the analytic model and moderated effect estimates, but for simplicity, we only discuss the total moderation effect ($\Delta_2^{(t)}$; see Between Group Only Moderation Effects in the Technical Supplemental Appendix). The outcome model for the control condition is unaffected by the additional considerations in the treatment condition. It is unnecessary (or impossible) to consider group-level or aggregated variables with a single-level data structure.

Moderator Effect and Error Variance

The moderated effect is still estimated using differences between the treatment and control model coefficients associated with the moderator. The $\Delta_2^{(t)}$ coefficient captures the total moderated effect under our formulation as it includes the moderated effect at the individual level (e.g., an individual's optimism M_{ij}) and group levels (e.g., average optimism $\bar{M}_j$). Our expression of the moderated effect is then

$$\text{ME}_{W+B} = \Delta_2^{(t)} - \Delta^{(c)} \tag{7}$$

The $\Delta_2^{(t)}$ term in our example captures the relationship between a student's optimism and weight loss in the treatment condition at the individual level. It also captures the relationship between the average optimism in an intensive lifestyle intervention group and weight loss in the treatment

condition. The moderated effect is estimated by finding the difference between $\Delta_2^{(t)}$ and the relationship between participant optimism and weight loss in the control condition ($\Delta^{(c)}$). Note that the parameter of importance from the control condition, $\Delta^{(c)}$, has not changed in substance or interpretation justifying our use of the same models from the within-only moderator variance section.

The statistical significance test for the moderated effect and power formula (see Test Statistic and Power Formula in Technical Supplemental Appendix) does not differ under the new analytic model. They do require the new moderation effect (see Equation 7) and a formulation of moderator effect variance that reflects $\Delta_2^{(t)}$ such that

$$\sigma^2_{\mathrm{ME}_{W+B}} = \sigma^2_{\Delta_2^{(t)}} + \sigma^2_{\Delta^{(c)}} \tag{8}$$

The formulation of $\sigma^2_{\Delta^{(c)}}$ remains unchanged. The $\sigma^2_{\Delta_2^{(t)}}$ term is new and requires new considerations to reflect between-group variance components (see Between Group Moderator Coefficient Variance in Technical Supplemental Appendix). We again unpack the variance of the moderated effect as a function of parameters that are accessible in the design stage. This is necessary for its implementation in a power analyses (see Technical Supplemental Appendix for details). The error variance of the moderated effect ($\sigma^2_{\mathrm{ME}_{W+B}}$) now reflects a moderator that varies within and between groups such that

$$\begin{aligned}\sigma^2_{\mathrm{ME}_{W+B}} = &\frac{\tau^2_{Y^{(t)}}\left(1 - R^2_{Y^{L2}_{(t)}}\right) + \sigma^2_{Y^{(t)}}\left(1 - R^2_{Y^{L1}_{(t)}}\right)/n_1^{(t)}}{(n_2^{(t)} - C_{(t)} - 1)\left(\tau^2_{M^{(t)}}\left(1 - R^2_{M^{L2}_{(t)}}\right) + \sigma^2_{M^{(t)}}\left(1 - R^2_{M^{L1}_{(t)}}\right)\Big/ n_1^{(t)}\right)} \\ &+ \frac{\sigma^2_{Y^{(c)}}\left(1 - R^2_{Y^{L1}_{(c)}}\right)}{(n^{(c)} - C_{(c)} - 1)\sigma^2_{M^{(c)}}(1 - R^2_{M_{(c)}})}\end{aligned} \tag{9}$$

Several of the terms in the expanded version of $\sigma^2_{\Delta_2^{(t)}}$ parallel those previously discussed including variance in the outcome at Level 1 or the student level ($\sigma^2_{Y^{(t)}}$), variance explained by predictors (R^2), and sample sizes ($n_1^{(t)}$, $n_2^{(t)}$, and $n^{(c)}$). These parallel terms operate and can be interpreted similarly to their counterparts discussed as under the previous model assuming within group moderator variation only.

There are, however, two new variance components from the group level of the outcome and moderator models in the treatment arm ($\tau^2_{Y^{(t)}}$ and $\tau^2_{M^{(t)}}$). We standardize the variance such that $\tau^2_{\cdot^{(t)}} + \sigma^2_{\cdot^{(t)}} = 1$. This standardization aids in the interpretation of the formula as $\tau^2_{\cdot^{(t)}}$ is equivalent to the unconditional ICC. The $\tau^2_{M^{(t)}}$ term represents the group-level moderator variance in the treatment condition. It is obtained from the multilevel model now necessary to reflect variation in the moderator across groups (see Multilevel Moderator Model in Technical Supplemental Appendix). In our example, $\tau^2_{M^{(t)}}$ indicates the variance in optimism attributable to the group level or differences in optimism across intervention groups. The $\tau^2_{Y^{(t)}}$ term represents the group-level outcome variance in the treatment condition. The inclusion of this term is now necessary in the moderation effect variance formula because we have a moderator ($\bar{M}_j$) and moderator coefficient term ($\Delta_2^{(t)}$) in the group level of the outcome model (see Equation 6). In terms of our example, $\tau^2_{Y^{(t)}}$ indicates the variance in weight loss attributable to the group level or differences in weight loss across intervention groups.

Including covariates (that explain outcome variance) is an effective strategy for increasing power to detect the moderated effect. Reducing $\tau^2_{Y^{(t)}}$ and $\sigma^2_{Y^{(t)}}$ reduces moderation effect variance ($\sigma^2_{\mathrm{ME}_{W+B}}$), thus increasing the likelihood of detecting statistically significant moderation effects. Under the

analytic model for moderators that vary within and between groups (see Equation 6), outcome variance can be explained at different levels. This is illustrated by the Levels 1 and 2 in the R^2 terms. Covariates may be more or less effective at explaining variance at different levels. For example, covariates capturing prescription medication use or demographics (e.g., education, race, gender) may explain variation in the weight loss outcome ($\sigma^2_{Y(t)}$) and aggregated covariates (i.e., average medication use or demographics in an intervention group) may also explain variation in aggregated weight loss ($\tau^2_{Y(t)}$). Group-level covariates (W) can only explain variance in the outcome at the group level (the $R^2_{Y^{L2}_{(t)}}$ associated with $\tau^2_{Y(t)}$) because a group-level variable cannot vary within groups. We suggest prioritizing individual-level covariates because they can explain variance at the individual and group level.

To summarize, partially nested designs with moderators that only vary within groups ($\sigma^2_{\text{ME}_{\text{within}}}$) share several similarities with designs in which moderators vary within and between groups ($\sigma^2_{\text{ME}_{W+B}}$). For example, variance of the moderator effect decreases as sample sizes increase and/or a greater proportion of outcome variance is explained by covariates. The model allowing moderator variance between groups does include additional variance components. The source of these additional components is the variance of the aggregated moderator coefficient ($\sigma^2_{\Delta_2^{(t)}}$). This term includes variance components from the second level of the outcome and moderator models ($\tau^2_{Y(t)}$ and $\tau^2_{M(t)}$). Estimates of $\tau^2_{Y(t)}$ and $\tau^2_{M(t)}$ are needed for evaluation planning in addition to those mentioned when moderators vary within groups only.

Design Implications

After gathering evidence of formula accuracy (see Power Formula Accuracy Simulation in Technical Supplemental Appendix), we investigated the feasibility of detecting moderator effects in two/one partially nested designs when (a) the moderator varied within groups only (i.e., induced nesting design) and (b) the moderator varied within and between groups (i.e., extant nesting design). We considered moderator effects of $\text{ME}_{\text{within}} = .2$ and .1, $\text{ME}_{W+B} = .25$ (composed of $\text{ME}_W = .15$ and $\text{ME}_B = .1$), and $\text{ME}_{W+B} = .15$ (composed of $\text{ME}_W = .1$ and $\text{ME}_B = .05$) to represent evaluations that expected moderation effects of various magnitudes. We also varied two other factors that commonly influence power: sample size and model variance structure in the treatment. Feasible sample sizes are dependent on the context of the evaluation so group ($n_2^{(t)}$) and individual per group sample size ($n_1^{(t)}$) in the treatment arm ranged from 10 to 100 with the product of these two sample sizes dictating the sample size of the control arm ($n^{(c)}$; i.e., balanced designs). For example, in an educational context, an evaluation of an early childhood literacy intervention in a large school district included a treatment sample of 49 classrooms and 1,229 students (Zvoch et al., 2007) while a more localized evaluation of an art education program included a treatment sample of only 366 students and approximately 11 teachers (Smith et al., 2010).

Variance structure conditions include individual-level variance of the outcome and moderator set at $\sigma^2_Y = \sigma^2_M = 0.8$ with group-level variance of the outcome and moderator $\tau^2_Y = \tau^2_M = .2$ and a second condition with $\sigma^2_Y = \sigma^2_M = .6$ and $\tau^2_Y = \tau^2_M = .4$ (τ^2_M is set to 0 when the moderator only varies within groups). These conditions represent evaluation contexts with more ($\tau^2_Y = \tau^2_M = .4$) or less ($\tau^2_Y = \tau^2_M = .2$) variance attributable to the group level. Group-level variance fluctuates based on several factors including the type of variable (e.g., academic, psychosocial, health), group setting (e.g., classroom, school, hospital, neighborhood), and geographic location (e.g., state or country; Hedges & Hedberg, 2007; Kelcey et al., 2016; Shackleton et al., 2016).

Table 1. Power to Detect a Moderated Effect When the Moderator Varies Only Within Groups.

Sample Size			Power	
Treatment		Control	Moderated Effect	
Groups ($n_2^{(t)}$)	Individuals per Group ($n_1^{(t)}$)	Individuals ($n^{(c)}$)	0.1	0.2
100	100	10,000	1.00	1.00
100	50	5,000	1.00	1.00
100	20	2,000	0.86	1.00
100	10	1,000	0.58	0.99
50	100	5,000	1.00	1.00
50	50	2,500	0.93	1.00
50	20	1,000	0.59	0.99
50	10	500	0.30	0.86
25	100	2,500	0.93	1.00
25	50	1,250	0.65	1.00
25	20	500	0.31	0.83
25	10	250	0.18	0.54
10	100	1,000	0.46	0.93
10	50	500	0.22	0.69
10	20	200	0.08	0.35
10	10	100	0.05	0.20

Note. Individual-level variance of the outcome and moderator are $\sigma^2_Y = \sigma^2_M = .8$ with group-level variance of the outcome $\tau^2_Y = .2$.

The remaining parameters were held constant. Variance explained by predictors was set at 50% (R^2=.5) to represent effective covariates and typical relationships between the moderator, treatment, and outcome. Coefficients for each path were .3 ($\pi_{\cdot} = Z_{\cdot} = .3$) and all intercepts were set to zero. These values are not directly used in the moderator variance and power formulations so their values were set to ease interpretation of results. The moderator coefficient in the control arm was set to zero ($\Delta^{(c)} = 0$) and the model variance components in the control arm were set to one ($\sigma^2_{Y^{(c)}} = \sigma^2_{M^{(c)}} = 1.0$), again to ease interpretation of results.

A fully crossed design with these factors produced 64 conditions. Results can help address common evaluation planning questions such as: How many intervention groups and individuals are necessary to consistently detect moderated treatment effects? Is the evaluation sample size large enough to consider moderators that vary within and between groups? and How will the size of the moderated effect influence the adequate sample of groups and individuals per group?

Results

Moderator Varies Only Within Groups

Table 1 presents power rates to detect a moderated effect in a two/one partially nested design when the individual-level continuous moderator only varies within groups. Larger moderation effects and larger sample sizes substantially (and predictably) increased power. These power rates remained constant under the different variance structures considered. Results indicate that adequate power to detect ME_{within} is achievable with feasible sample sizes for some experimental evaluations (e.g., Schochet, 2011). For example, with $ME_{within} = .1$ and larger individual per group sample sizes ($n_1 \geq 50$), the number of groups required in the treatment condition was often <30. This suggests that evaluations considering group entities capable of including more than 50 individuals per group (e.g., schools, hospitals, companies) have a greater capacity to consistently identify ME_{within}.

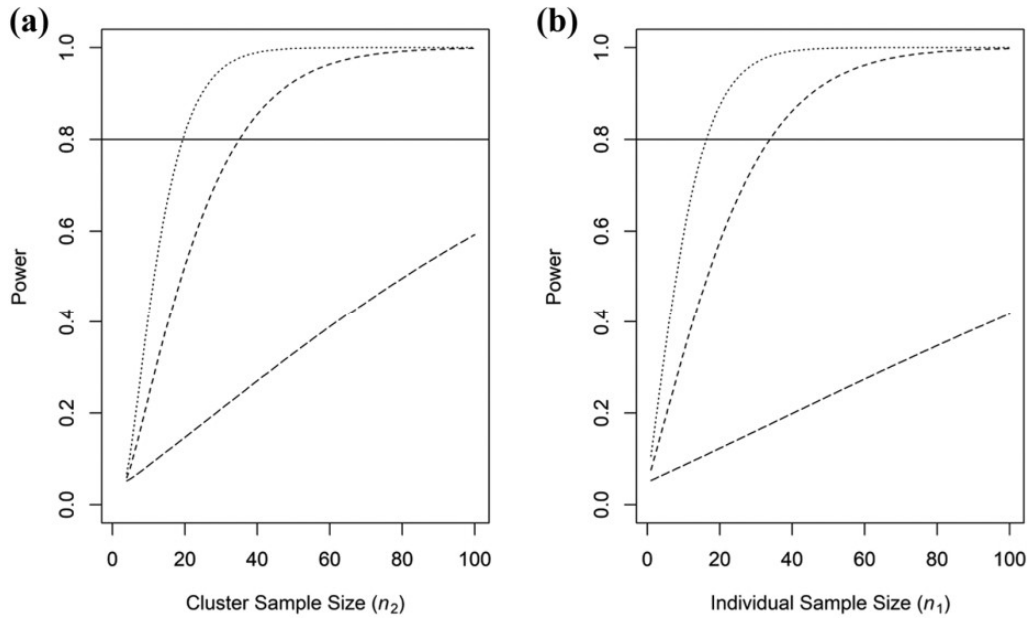

Figure 2. Power to detect a moderated effect when the moderator varies within groups only by group and individual per group sample size.
Note. Power to detect a .1 moderation effect by (a) group sample size and (b) individual per group sample size with a sample of 10 (long dash), 50 (dash), and 100 (dot) (a) individuals per group or (b) groups. A solid horizontal line marks 80% power.

Individual per group sample size has a substantial influence on power to detect the moderated effect (see Figure 2). A reasonable result, given the individual per group sample size, directly reduces outcome variance, which reduces the variance of the moderation effect (see Equation 5). This relationship is noteworthy because power to detect main effects in a typical group-randomized trial is driven by group sample size (e.g., Raudenbush, 1997). The result implies increasing individual per group sample size is an effective design strategy for detecting these moderation effects. This is often less expensive than sampling additional groups. In our working example evaluation, we could sample more students per intervention group to increase the likelihood of detecting a moderation effect.

Moderator Varies Within and Between Groups

Table 2 presents power rates to detect a moderated effect in a two/one partially nested design when the individual-level continuous moderator varies within and between groups. Larger total moderation effects (ME_{W+B}) and sample sizes again led to increased power rates. However, adequate sample sizes were much larger than those for similar studies with a moderator that varied within groups only. For example, with a $\mathrm{ME}_{W+B} = .15$ and $n_1^{(t)} = 50$, a sample of over 400 groups is required to achieve 80% power to detect a moderated effect (see Figure 3).

Increasing the sample of individuals per group does little to alleviate the need for a large sample of groups (see Figure 3). The power to detect ME_{W+B} under the probed conditions stems primarily from the group sample size. We noted that this was likely because variance of the outcome at the group level $\tau^2_{Y(t)}$, which increases moderator effect variance and decreases power, is only reduced by group sample size (see Equation 9). The $\tau^2_{Y(t)}$ value has a substantial influence on the power to detect ME_{W+B} through its influence on moderator effect variance. The example above requiring 400

Table 2. Power to Detect a Moderated Effect When the Moderator Varies Within and Between Groups.

Sample Size			Power	
Treatment		Control	Moderated Effect	
Groups ($n_2^{(t)}$)	Individuals per Group ($n_1^{(t)}$)	Individuals ($n^{(c)}$)	.15	.25
100	100	10,000	.31	.68
100	50	5,000	.30	.68
100	20	2,000	.30	.66
100	10	1,000	.29	.64
50	100	5,000	.17	.39
50	50	2,500	.17	.38
50	20	1,000	.17	.37
50	10	500	.16	.36
25	100	2,500	.10	.20
25	50	1,250	.10	.20
25	20	500	.10	.20
25	10	250	.10	.19
10	100	1,000	.06	.09
10	50	500	.06	.09
10	20	200	.06	.09
10	10	100	.06	.09

Note. Individual-level variance of the outcome and moderator are $\sigma_Y^2 = \sigma_M^2 = .8$ with group-level variance of the outcome and moderator $\tau_Y^2 = \tau_M^2 = .2$.

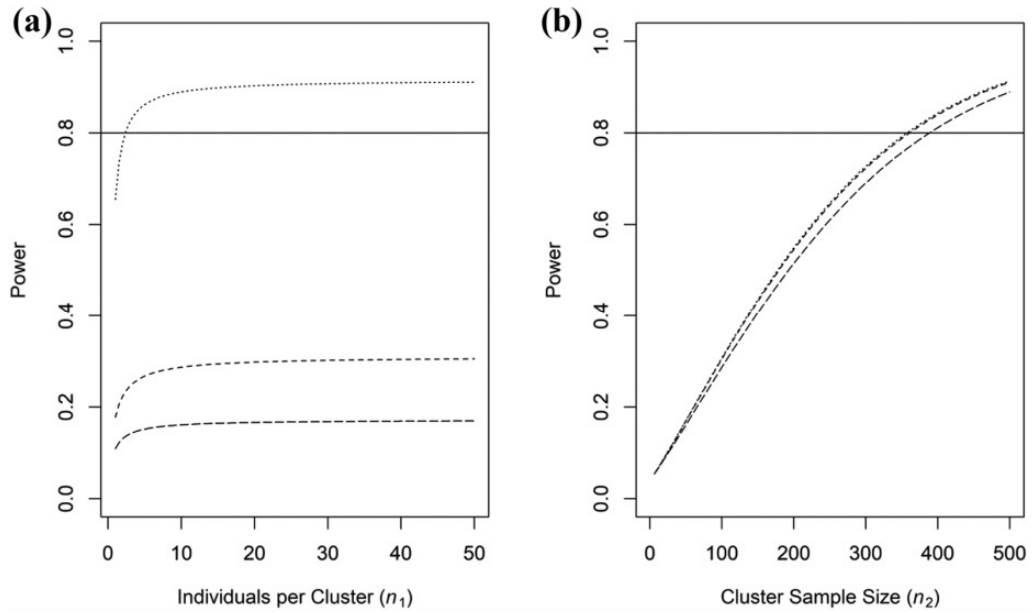

Figure 3. Power to detect a moderated effect when the moderator varies within and between groups by group and individual per group sample size.

Note. Power to detect a .15 moderation effect (ME_{W+B}) when the moderator varies within and between groups by (a) individual per group sample size and (b) group sample size with a sample of (a) 50 (long dash), 100 (dash), and 500 (dot) groups and (b) 10 (long dash), 50 (dash), and 100 (dot) individuals per group. A solid horizontal line marks 80% power.

groups to achieve adequate power would only require 125 groups if $\tau^2_{Y(t)} = .05$ and $\sigma^2_{Y(t)} = .95$. Individual per group sample size also becomes more influential on power rates under this variance structure. The feasibility of an evaluation including a moderator that varies within and between groups is therefore heavily dependent on the variance of the outcome attributable to individuals or the group (i.e., unconditional ICC of the outcome). Our evaluation would require an exorbitant number of intensive lifestyle intervention groups to detect a moderated effect if optimism varied among participants and across intervention groups while variance in the weight loss outcome attributable to the group was around $\tau^2_{Y(t)} \geq .1$. If weight loss was less clustered (e.g., $\tau^2_{Y(t)} \leq .05$), then the feasibility of the evaluation increases (i.e., adequate sample sizes become more reasonable).

In summary, simulation study results indicated that our formulas produced appropriate Type I error rates and accurately predicted power (see simulation results in Technical Supplemental Appendix). A probe of these formulas found that power to detect both $\text{ME}_{\text{within}}$ and ME_{W+B} increased as their magnitude increased and as sample sizes increased. The influence of outcome variance on power to detect moderated effects indicates prognostic covariates that explain variance in the outcome represent an effective and important design strategy for consistently detecting moderated effects. It is worth noting that when covariates explain similar amounts of variance in the outcome and moderator, power to detect the moderated effect remains relatively constant. Put differently, inclusion of covariates that explain approximately equal amounts of variance in the outcome and moderator will not improve the detection of moderated effects.

Adequate power (e.g., 80%) to detect moderator effects when the moderator only varied within groups was achievable using typical sample sizes for studies planned to detect main effects. Increasing individual per group sample size substantially influenced power to detect these effects. When the moderator varied within and between groups, achieving adequate power required larger sample sizes or larger moderated effects. A primary driver of these power rate differences was increased moderator effect variance caused by the inclusion of group-level outcome variance ($\tau^2_{Y(t)}$). As $\tau^2_{Y(t)}$ increases, increasing individual per group sample also becomes less effective at increasing power with group sample size becoming increasingly influential.

Illustrations

We now illustrate our results and the use of our formulas in the planning of evaluations with a two/one partially nested design. Our first evaluation example examined the effect of a spatial intervention program on math performance while considering the moderating effects of student math anxiety. Treatment assignment induced nesting (i.e., novel intervention groups were created) so we can assume intervention groups will have approximately equal levels of student math anxiety (i.e., the moderator varies within groups only).

Using the R-Shiny application *PowerUpRShiny* for moderated effects in partially nested designs (Bai et al., 2020; Bulus et al., 2019) , we can predict the power rate to detect a moderated effect for this evaluation with a specific sample size or determine the adequate sample to detect a moderated effect with acceptable power (e.g., >80%). Several design parameter estimates are needed to conduct this type of evaluation planning. We assume a variance structure based on educational evaluations considering academic outcomes (e.g., Hedges & Hedberg, 2007) such that variance of math performance and math anxiety at the individual level is set at $\sigma^2_Y = \sigma^2_M = .8$ with variance at the group level set at $\tau^2_Y = .2$ ($\tau^2_M = 0$ in this case because the moderator does not vary across groups). Variance in the outcome and moderator for the control arm is standardized with a value of 1.0. We include several covariates and assume these variables along with the moderator explain approximately 50% of the variance in the outcome and moderator in both the treatment and control arms (R^2=.5). Given strong evidence of spatial intervention program effectiveness, we assume it will have an approximately .2 standardized effect on math performance with a moderated effect from student math anxiety of .1 ($\text{ME}_{\text{within}} = .1$).

Our evaluation takes place in a large school district so we have access to a large number of qualifying students and resources for approximately 50 intervention groups (i.e., $n_2^{(t)}$=50). Our formulas indicate that approximately 30 students per group with 1,500 wait-listed students are necessary to achieve 80% power to detect a moderated effect under the specified conditions (see Figure 3). Thirty students per group may not be practical so let us examine the same evaluation with the number of students per group capped at 10 (i.e., $n_1^{(t)}$=10). An evaluation with this sample structure would require well over 100 groups. This result is indicative of the substantial influence individual per group sample size has on the ability to consistently detect $\mathrm{ME}_{\mathrm{within}}$.

Our second evaluation example examined the effect of an intensive lifestyle intervention program on weight loss while considering the moderating effects of optimism. Treatment participants were assigned to an existing intervention group (i.e., use of extant intervention groups) so intervention groups may vary on aggregate levels of optimism (i.e., the moderator varies within and between groups). Using the described parameter estimates and conditions, but with $\mathrm{ME}_{W+B} = .15$, an evaluation including 50 groups with 30 participants per group and 1,500 participants continuing with their current care would achieve approximately 17% power to detect the moderated effect. This is far from adequate; in fact, nearly 400 intervention groups are required to achieve power rates approaching 80%. However, health and psychosocial outcomes may have a much smaller $\tau^2_{Y(t)}$ (e.g., Shackleton et al., 2016). With $\tau^2_{Y(t)} = .07$ and $\sigma^2_{Y(t)} = .93$, the same evaluation requires around 170 intervention groups.

Discussion

Experimentally designed evaluations provide strong evidence regarding average effectiveness of a treatment but treatment effects may depend on individual and contextual factors. Including moderators in evaluation design planning can help evaluators identify these differential treatment effects. Planning experimental evaluations that include moderation effects, however, has been limited in some cases. Specifically, partially nested designs pose a challenge because statistical power formulas for moderation effects have not yet been available.

In response, this study develops these formulas and investigates their properties and implications for practice. Our aim is to encourage the use of adequate sample sizes, to identify typical sample sizes necessary to detect moderated effects, and to determine the factors that influence these sample sizes. We first considered moderators that varied within groups only which are likely in partially nested designs with treatment-induced nesting. Random assignment serves to eliminate systematic differences across groups on pretreatment moderator variables under this design. In a second set of formulas, we relaxed this assumption and allowed moderators to vary within and between groups, which is likely when using extant groups in the treatment condition.

The power formulas presented here improve evaluation efficiency. Evaluators can now determine the sample sizes necessary to detect a moderated effect in a proposed evaluation, avoiding wasted resources from oversampling. Evaluations that consider moderation effects also produce better evidence regarding treatment effects. When a significant effect is present, evaluators can examine for whom and under what conditions it is applicable. Conversely, evaluators can investigate nonsignificant results and determine whether the treatment was effective for some groups while counter-effective for others resulting in a combined effect near zero (i.e., counteracting treatment effects).

The initial probe of power formulas for moderated effects in partially nested designs found that the sample sizes required to achieve adequate power are similar to those required to detect main and mediated effects when the moderator only varies within groups (e.g., Kelcey et al., 2017). Evaluations examining this type of moderation effect are feasible as the sample size is likely to be adequate based on planning for other effects. We also found a particularly strong relationship between individual per group sample size and power. This implies that evaluations with a limited number of groups can still consistently detect moderated effects if a large individual per group sample size is

available. For example, an evaluation limited to 25 treatment groups only has an 18% chance to detect a .1 magnitude moderation effect with 10 individuals per group. However, if the group contains many easily accessible individuals (e.g., school) such that individual per group sample size can be 100, the power to detect the moderation effect is greater than 90%.

In comparison, when the moderator varies within and between groups, the additional variance from the group level can overpower any additional moderated effect from the aggregated moderator. Evaluations with this type of moderator will often require large group sample sizes with increases to individual per group sample size doing little to increase power once $n_1^{(t)}>10$. These problematic conditions do dissipate as outcome variance at the group level decreases.

Conclusion

To conclude, we highlight some limitations, opportunities for future research, and summarize implications for practice. We limited the scope of our study to designs with a two-level data structure in one treatment arm and a single-level data structure in the other treatment arm (i.e., two/one partially nested designs). Many evaluations take place across a single entity (e.g., school district, state, company) with the treatment inducing nesting or using extant groups. However, many settings will have additional levels of nesting that should be considered for statistical or substantive reasons (e.g., students within schools within districts). These conditions require considerations beyond two levels. We encourage future research examining power in three/one and three/two partially nested designs.

Limited design conditions were used in our probe of the newly developed power formulas (e.g., sample sizes from 10 to 100). A more comprehensive set of sample size, moderator variance structure, and outcome variance structure combinations is needed. It would be informative to establish expected power rates for total and specific moderation effects (i.e., within-group moderation effects and between-group moderation effects) across a wider range of design conditions.

We noted that making assumptions about inputs to the power analysis was necessary because of the sparsity of literature reporting such values. Pilot studies are an excellent source for these values but are not always practical. We encourage future research to report the empirical values required for the power formulas presented here. Additionally, investigations into the robustness of predicted power to misspecified parameter values could indicate the degree of precision required for accurate predictions of power. Despite these limitations, this study enhances the set of tools that evaluators can use to plan evaluations. Specifically, considering moderated effects is increasingly relevant to policy and practice. Better planning evaluations to generate evidence about for whom and under what conditions an intervention, program, or policy is effective is coveted across evaluation settings.

To close, we highlight several takeaway recommendations for evaluators interested in moderation effects with partially nested designs. First, consult existing evidence (e.g., literature or pilot study results) to identify moderators that are likely to have a large effect on the treatment–outcome relationship. This evidence should also be consulted to identify other parameter values required for the power analysis. Second, include covariates that reduce outcome variation. This is especially important when the moderator varies within and between groups as reductions in group-level outcome variance typically have a substantial influence on power. Third, when a moderator only varies within groups, increasing the sample of individuals per group is an effective strategy to increase power. Finally, if the moderator varies within and between groups, carefully consider the variance structure of the outcome as it has a substantial influence on evaluation feasibility and the relationship between power and sample size.

Authors' Note

The opinions expressed herein are those of the authors and not the funding agency.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This article is based on work funded by the National Science Foundation [#1012665 and #1760884].

ORCID iD


Kyle Cox https://orcid.org/0000-0002-7173-4701

Supplemental Material

Supplemental Appendix is available in the online version of this article at http://journals.sagepub.com/home/aje

References

Aguinis, H., Beaty, J., Boik, R., & Pierce, C. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94–107. https://doi.org/10.1037/0021-9010.90.1.94

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., Cox, K., & Kelcey, B. (2020). PowerUpR Shiny App for Designing Partially Nested Moderation Studies (Version 0.1) [Software]. https://poweruprshiny.shinyapps.io/PartiallyNestedModerationPower/

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

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

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision: Empirical guidelines for studies that randomize schools to measure the impacts of educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30–59.

Bulus, M., Dong, N., Kelcey, B., & Spybrook, J. (2019). PowerUpR: Power Analysis Tools for Multilevel Randomized Experiments. R package version 1.0.4. https://CRAN.R-project.org/package=PowerUpR

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

Dishion, T. J., 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., Kelcey, B., Spybrook, J., & Maynard, R. A. (2016). *PowerUp!-Moderator: A tool for calculating statistical power and minimum detectable effect size differences of the moderator effects in cluster randomized trials* [Software]. http://www.causalevaluation.org/

Hall, P., Zehr, C., Paulitzki, J., Rhodes, R., & Hall, P. (2014). Implementation intentions for physical activity behavior in older adult women: An examination of executive function as a moderator of treatment effects. *Annals of Behavioral Medicine: A Publication of the Society of Behavioral Medicine*, *48*(1), 130–136.

Hedges, L., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods*, *47*, 1295–1308.

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

Institute of Education Sciences. (2016). *Request for application: Statistical and research methodology in education*. U.S. Department of Education. http://ies.ed.gov/funding/pdf/2017_84305D.pdf

Jaciw, A., Lin, L., Ma, B., Kelcey, B., & Spybrook, J. (2016). An empirical study of design parameters for assessing differential impacts for students in group randomized trials. *Evaluation Review*, *40*(5), 410–443.

Kelcey, B., Dong, N., Spybrook, J., & Cox, K. (2017). Statistical power for causally-defined indirect effects in group-randomized trials with individual-level mediators. *Journal of Educational and Behavioral Statistics*, *42*, 499–530.

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.

Kenny, D., Calsyn, R., Morse, G., Klinkenberg, W., Winter, J., & Trusty, M. (2004). Evaluation of treatment programs for persons with severe mental illness: Moderator and mediator effects. *Evaluation Review*, *28*(4), 294–324.

Korendijk, E., Maas, C., Hox, J., & Moerbeek, M. (2012). The robustness of the parameter and standard error estimates in trials with partially nested data: A simulation study. In E. Korendijk (Ed.), *Robustness and optimal design issues for cluster randomized trials* (pp. 59–94) [Dissertation]. Utrecht University.

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

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

Lohr, S., Schochet, P. Z., & Sanders, E. (2014). *Partially nested randomized controlled trials in education research: A guide to design and analysis* (NCER 2014-2000). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. http://ies.ed.gov/

Lowrie, H., Harris, D., Logan, T., & Hegarty, M. (2019). The impact of a spatial intervention program on students' spatial reasoning and mathematics performance. *The Journal of Experimental Education*, 1–19.

Lüdtke, T., Westermann, S., Pult, L., Schneider, B., Pfuhl, G., & Moritz, S. (2018). Evaluation of a brief unguided psychological online intervention for depression: A controlled trial including exploratory moderator analyses. *Internet Interventions*, *13*, 73–81.

Mackinnon, D., Soydan, H., & Sundell, K. (2011). Integrating mediators and moderators in research design. *Research on Social Work Practice*, *21*(6), 675–681.

Mayer-Davis, E., D'Antonio, A., Smith, S., Kirkner, G., Martin, S., Parra-Medina, D., & Schultz, R. (2004). Pounds off with empowerment (POWER): A clinical trial of weight management strategies for Black and White adults with diabetes who live in medically underserved rural communities. *The American Journal of Public Health*, *94*(10), 1736–17342.

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

Morrell, C., Walters, S., Dixon, S., Collins, K., Brereton, L., Peters, J., & Brooker, C. (1998). Cost effectiveness of community leg ulcer clinics: Randomised controlled trial. *British Medical Association*, *316*(7143), 1487–1491.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

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

Rivera, R., Santos, D., Brändle, G., & Cárdaba, M. (2016). Design effectiveness analysis of a media literacy intervention to reduce violent video games consumption among adolescents: The relevance of lifestyles segmentation. *Evaluation Review*, *40*, 142–161.

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

Roberts, J., 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*, 1553–1566.

Sanders, E. (2011). *Multilevel Analysis Methods for Partially Nested Cluster Randomized Trials* (Doctoral dissertation). ProQuest LLC.

Scheier, M., & Carver, C. (1992). Effects of optimism on psychological and physical well-being: Theoretical overview and empirical update. *Cognitive Therapy and Research*, *16*(2), 201–228.

Schochet, P. Z. (2011). Do typical RCTs of education interventions have sufficient statistical power for linking impacts on teacher practice and student achievement outcomes? *Journal of Educational and Behavioral Statistics*, *36*, 441–471.

Schweig, J., & Pane, J. (2016). Intention-to-treat analysis in partially nested randomized controlled trials with real-world complexity. *International Journal of Research & Method in Education: Special Issue (PART 1): Is the Educational "What Works" Agenda Working? Critical Methodological Developments*, *39*(3), 268–286.

Shackleton, N., Hale, D., Bonell, C., & Viner, R. (2016). Intraclass correlation values for adolescent health outcomes in secondary schools in 21 European countries. *SSM—Population Health*, *2*(C), 217–225.

Smith, N., Brandon, P., Lawton, B., & Krohn-Ching, V. (2010). Evaluation exemplar: Exemplary aspects of a small group-randomized local educational program evaluation. *American Journal of Evaluation*, *31*(2), 254–265.

Society for Research on Educational Effectiveness. (2012). *Spring 2012 Conference: Understanding variation in treatment effects*. https://www.sree.org/assets/conferences/2012s/program.pdf

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.

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

Stone-Romero, E. F., & Liakhovitski, D. (2002). Strategies for detecting moderator variables: A review of conceptual and empirical issues. *Research in Personnel and Human Resources Management*, *21*, 333–372.

Van Nguyen, H., Huang, H., Wong, M., Yang, Y., Huang, T., & Teng, C. (2018). Moderator roles of optimism and weight control on the impact of playing exergames on happiness: The perspective of social cognitive theory using a randomized controlled trial. *Games for Health Journal*, *7*(4), 246–252.

Zvoch, K., Letourneau, L., & Parker, R. (2007). A multilevel multisite outcomes-by-implementation evaluation of an early childhood literacy model. *American Journal of Evaluation*, *28*(2), 132–150.