The Journal of Experimental Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/vjxe20

Power to Detect Moderated Effects in Studies with Three-Level Partially Nested Data

Kyle Cox, Ben Kelcey & Hannah Luce

To cite this article: Kyle Cox, Ben Kelcey & Hannah Luce (2022): Power to Detect Moderated Effects in Studies with Three-Level Partially Nested Data, The Journal of Experimental Education, DOI: 10.1080/00220973.2022.2130130

To link to this article: https://doi.org/10.1080/00220973.2022.2130130

Published online: 20 Oct 2022.

THE JOURNAL OF EXPERIMENTAL EDUCATION
https://doi.org/10.1080/00220973.2022.2130130

Power to Detect Moderated Effects in Studies with Three-Level Partially Nested Data

Kyle Cox[a], Ben Kelcey[b], and Hannah Luce[a]

[a]Department of Educational Leadership, University of North Carolina at Charlotte, Charlotte, North Carolina; [b]Department of Educational Studies, University of Cincinnati, Cincinnati, Ohio

ABSTRACT

Comprehensive evaluation of treatment effects is aided by considerations for moderated effects. In educational research, the combination of natural hierarchical structures and prevalence of group-administered or shared facilitator treatments often produces three-level partially nested data structures. Literature details planning strategies for a variety of experimental designs when moderation effects are of interest but has yet to establish power formulas for detecting moderation effects in three-level partially nested designs. To address this gap, we derive and assess the accuracy of power formulas for detecting the different types of moderation effects possible in these designs. Using Monte Carlo simulation studies, we probe power rates and adequate sample sizes for detecting the different moderation effects while varying common influential factors including variance in the outcome explained by covariates, magnitude of the moderation effect, and sample sizes. The power formulas developed improve the planning of experimental studies with partial nesting and encourage the inclusion of moderator variables to capture for whom and under what conditions a treatment is effective. Educational researchers also have some initial guidance regarding adequate sample sizes and the factors that influence detecting moderation effects in three-level partially nested designs.

KEYWORDS

Cluster randomized trial; HLM; moderation effect; Monte Carlo Simulation; partially nested design; simulation studies; statistical power

Experimental designs in educational research provide the most robust causal evidence for program, intervention, or policy effectiveness (i.e., treatment effectiveness). Inclusion of moderator variables is a prevalent technique to capture treatment effect heterogeneity in these educational experiments (Dong et al., 2018; Dong et al., 2021a; Dong et al., 2021b; MacKinnon, 2011; Spybrook et al., 2016). Investigating differences in program, intervention, or policy effectiveness across group- (e.g., school size, school location, teacher experience, teacher training) or individual-characteristics (e.g., student race, student sex, pretest score) improves generalization of results through a better understanding of for whom or under what conditions a treatment is effective. Put differently, consideration of supplementary effects can elucidate individual and contextual factors that influence treatment effectiveness.

A long-standing emphasis on understanding treatment effect heterogeneity by professional organizations and funding agencies (e.g., Institute of Education Sciences (IES),) 2016; Society for Research on Educational Effectiveness, 2012) is reflected in a growing literature base detailing study design and analyses that includes moderation effects. For example, literature has developed

CONTACT Kyle Cox kyle.cox@uncc.edu; kcox59@uncc.edu Department of Educational Leadership, University of North Carolina at Charlotte, 266 Cato College of Education Building, Charlotte, NC 28223, USA

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00220973.2022.2130130.

techniques for the planning various experimental designs that include moderated effects and are well suited for educational settings (Cox & Kelcey, 2022; Dong et al., 2018; Dong et al., 2021a; Dong et al., 2021b; Jaciw et al., 2016; MacKinnon, 2011; Spybrook et al., 2016; Tong et al., 2022; Yang et al., 2020). In applied educational research, considerations for treatment effect heterogeneity were included in Chambers et al. (2008) who used a multisite randomized trial to examine the effect of a computer assisted tutoring program on reading achievement while considering the moderating effects of school average pretest and Weidinger et al. (2020) who considered a utility-value intervention on mathematics outcomes while considering the moderating effects of student migration and parent education. Beyond these individual studies, the Journal of Research on Educational Effectiveness and Exceptional Children published special issues focusing on treatment effect heterogeneity in rigorous intervention studies (Fuchs & Fuchs, 2019; Reardon & Stuart, 2017).

We focus on moderation effects in experimental designs with partially nested data. Partial nesting occurs when the intervention and control arm have different nesting or grouping structures. For example, in a study comparing a remote asynchronous online instruction intervention to typical classroom instruction, the intervention arm (i.e., remote instruction) has a two-level data structure (i.e., students nested within teachers) while the control arm (i.e., typical in-person classroom instruction) has a three-level data structure (i.e., students nested within teachers within schools). Partial nesting commonly arises when a treatment induces nesting through a group-administered (e.g., whole-school reform, classroom-based intervention, small-group instruction) or shared facilitator treatment (principal-, teacher-, tutor-, counselor-led activity). The natural hierarchical structure of educational settings lends itself to group-administered and shared facilitator treatments. The combination of natural hierarchical structures in educational settings and treatments that induce nesting produce a variety of partially nested data structures in educational experiments (e.g., Bauer et al., 2008; Lohr et al., 2014). These partially nested studies can be identified using the number of levels in each treatment arm. For example, 2/1 partially nested studies have a two-level data structure in one treatment arm and single-level data in the other arm. For 3/1 and 3/2 partially nested studies there is a three-level data structure in one treatment arm and a single-level or two-level data structure in the remaining treatment arm.

Unfortunately, analytic approaches suitable for individual randomized trials or cluster randomized trials produce inaccurate results when applied to partially nested data. Specifically, these misspecifications produce bias in standard errors of the treatment effect, and bias in estimates of variance components (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). In response, literature has mapped out design and analytic strategies for partially nested studies to avoid these issues (e.g., Lachowicz et al., 2015; Lohr et al., 2014; Moerbeek & Wong, 2008; Roberts & Roberts, 2005). More recent literature has established power formulas and related design strategies for detecting mediation (Kelcey et al., 2020) and moderated effects (Cox & Kelcey, 2022; Cox et al., under review) in partially nested designs. These recent advancements allow for more efficient and effective study planning and more comprehensive understanding of treatment effects when partially nested data are present.

While power formulas for detecting main effects in partially nested designs with two- or three-levels have been established (e.g., Heo et al., 2017), power formulas for moderated effects remain limited to partially nested designs with only two-levels (see Cox et al., under review; Cox & Kelcey, 2022). To address this gap, we derive and assess the accuracy of power formulas for detecting moderation effects in 3/1 and 3/2 partially nested designs as well as 3/2 partial nesting as part of a cluster randomized trial (see Figure 1). To supplement these results, we conduct an initial probe into power and adequate sample sizes for detecting moderation effects in these designs while varying common influential factors (e.g., variance in the outcome explained by covariates, magnitude of moderation effect, group and individual per group sample size). The

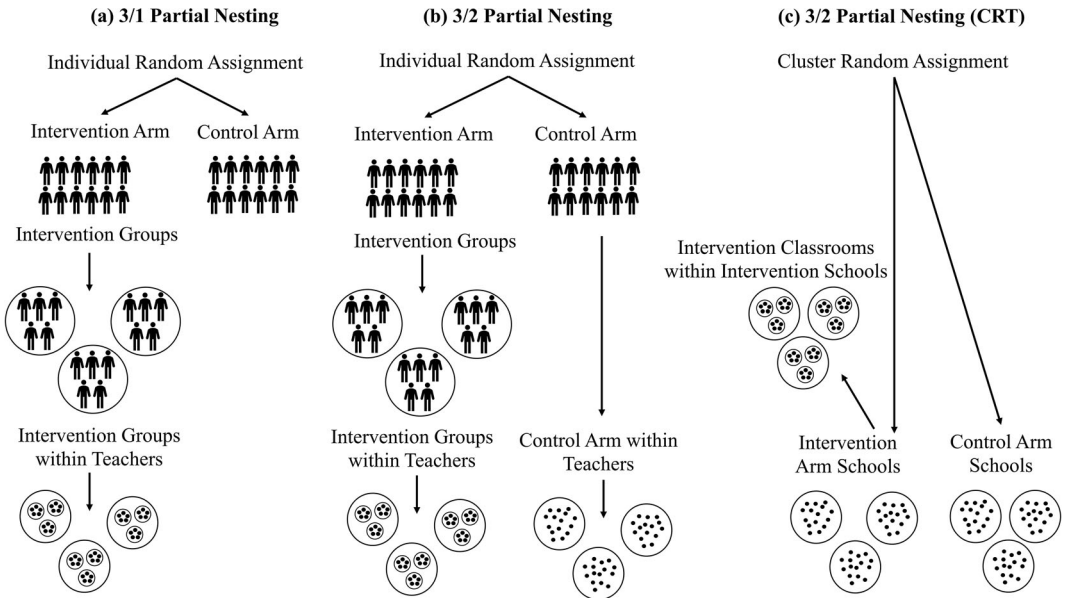

Figure 1. Example three-level partially nested designs with (a) 3/1 partial nesting, (b) 3/2 partial nesting, and (c) 3/2 partial nesting in a cluster randomized trial.

remainder of the paper is organized around the moderation effects possible in three-level partially nested designs. We derive power formulas for detecting moderation effects with 3/1 partial nesting, lower-level moderation with 3/2 partial nesting, upper-level moderation with 3/2 partial nesting, and cluster randomized designs with 3/2 partial nesting. Each section includes an introduction to the specific type of three-level partially nested design, applicable analytic model, and moderator effect variance formulas suitable for a priori power analysis. We utilize four Monte Carlo simulation studies (one for each moderation effect considered) to demonstrate accuracy of moderator effect variance and power formulas and provide an initial assessment of typical sample sizes required to detect these effects and the influence of several design parameters. Results of each simulation study are then shared and implications to design are discussed. Illustrative examples from educational research are utilized throughout each section to aid in the description of the design, formula, and results.

Three/one partial nesting

Partially nested designs with a 3/1 data structure have a three-level data structure in one treatment arm and a single-level data structure in the other treatment arm. Consider an illustrative example as we describe 3/1 partial nesting, associated analytic models, moderator effect variance and power formulas, and simulation study results. Our example is adapted from Lowrie et al. (2021) such that the hypothetical study examines the effectiveness of a summer school spatial reasoning intervention program aimed at improving student mathematical performance. Individual random assignment is utilized to place students in the intervention arm or waitlist control arm (see Figure 1a). Students in the waitlist control arm continue with summer as usual, thus avoiding any grouping (single-level data structure). Conversely, the intervention arm has students in spatial reasoning intervention groups and teachers instructing multiple groups (three-level data structure with students nested within groups nested within teachers). Student's math anxiety has a well-established detrimental effect on mathematics performance (e.g., Ashcraft & Krause, 2007; Ashcraft & Moore, 2009). To investigate heterogeneous intervention effects, student math anxiety

is included as a possible moderator between the spatial reasoning intervention program and student performance in mathematics.

Analytic model

We utilize the common multiple-arm multilevel framework for partially nested data (MA-PN). This approach eases extensions to moderation effects and accommodates the heteroscedastic variances of the different treatment arms common in partially nested designs. The analytic model in the intervention arm of a three/one partially nested design with a continuous outcome ($y_{ijk}^{(t)}$), continuous level-one moderator, $m_{ijk}^{(t)} \sim N(0, \sigma_m^{2(t)})$, and level-one covariate $x_{ijk}^{(t)} \sim N(0, \sigma_x^{2(t)})$ is

Level 1:

$$y_{ijk}^{(t)} = \beta_{0jk}^{(t)} + \beta_{1jk}^{(t)} m_{ijk}^{(t)} + \beta_{2jk}^{(t)}(x_{ijk}^{(t)} - \bar{x}_{.jk}^{(t)}) + \varepsilon_{ijk}^{(t)}, \quad \varepsilon_{ijk}^{(t)} \sim N(0, \sigma^2_{y|^{(t)}}) \tag{1}$$

Level 2:

$$\beta_{0jk}^{(t)} = \beta_{00k}^{(t)} + \beta_{02k}^{(t)}(\bar{x}_{.jk}^{(t)} - \bar{x}_{..k}^{(t)}) + u_{0jk}^{(t)}, \quad u_{0jk}^{(t)} \sim N(0, \tau^2_{y|^{(t)}})$$

Level 3:

$$\beta_{00k}^{(t)} = \delta_{000}^{(t)} + \beta_{002}^{(t)} \bar{x}_{..k}^{(t)} + v_{00k}^{(t)}, \quad v_{00k}^{(t)} \sim N(0, \phi^2_{y|^{(t)}})$$

Level-one or the student-level in our example includes a mathematics performance score ($y_{ijk}^{(t)}$) for student i, in group j, instructed by teacher k. The student's math anxiety score, which serves as a possible moderator of the intervention effect, is represented by $m_{ijk}^{(t)}$ with $\beta_{1jk}^{(t)}$ capturing the moderator-outcome relationship in the intervention arm. Level-two represents the spatial reasoning intervention groups with $u_{0jk}^{(t)}$ capturing group-specific deviations and level-three represents the teacher or instructor level where $\delta_{000}^{(t)}$ is the mean mathematics performance score in the intervention arm and $v_{00k}^{(t)}$ captures teacher-specific deviations. Our intervention model includes the assumption that the moderator does not systematically vary across intervention group or teacher. This is a tenable assumption in the presence of random assignment because students randomly assigned to the intervention and control arm form groups that will not systematically differ on average math anxiety (in expectation). Additionally, our assumption of uncorrelated outcomes across treatment arms requires individuals and groups are not located in the same school building and thus error terms for the intervention and control arms do not covary.

Covariates (e.g., $x_{\cdot}^{(\cdot)}$) that explain variance in the outcome are a common and effective design strategy to increase the likelihood of detecting main, mediation, and moderation effects or, relatedly, decreasing the sample size necessary to consistently detect these effects (e.g., Cox & Kelcey, 2022; Cox et al., under review; Raudenbush et al., 2007; Spybrook et al., 2016). A covariate or its aggregate (e.g., $\bar{x}_{.jk}^{(\cdot)}$ and $\bar{x}_{..k}^{(\cdot)}$) may be included at any level in the analytic model of the outcome for the intervention or control arm. Variance components are then reduced based on variance explained in the outcome at that level by the covariate or its aggregate.

The outcome model for the single-level control arm with a continuous moderator, $m_i^{(c)} \sim N(0, \sigma^2_{m^{(c)}})$ and continuous covariate $x_i^{(c)} \sim N(0, \sigma^2_{x^{(c)}})$ is

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

Mathematics performance score for student i in the single-level control arm is captured with $y_i^{(c)}$ and $\delta^{(c)}$ captures the mean mathematics performance score for the control arm students. The control arm model uses $m_i^{(c)}$ to represent a student's math anxiety score and $\beta_1^{(c)}$ to capture the moderator-outcome relationship.

Moderator effects and error variance

Contrasting the coefficients in the treatment arms that capture the relationship between the moderator and outcome provides an estimate of the moderation effect (*ME*) such that

$$ME = \left[(\delta_{000}^{(t)} + \beta_{1jk}^{(t)}) - (\delta^{(c)} + \beta_1^{(c)})\right] - \left[(\delta_{000}^{(t)}) - (\delta^{(c)})\right] \tag{3}$$

which simplifies to

$$ME = \beta_{1jk}^{(t)} - \beta_1^{(c)}. \tag{4}$$

In terms of our example, the moderation effect estimate is the difference between the math anxiety-math performance relationship in the intervention and control arms. If this relationship is different in the presence of the intervention ($\beta_{1jk}^{(t)}$), then the intervention effect on math performance is dependent on student math anxiety.

In the MA-PN framework, assuming independence and a covariance term that is zero by design, the sum of the moderator coefficient variances forms the variance of *ME* such that

$$\sigma^2_{ME} = \sigma^2_{(\beta_{1jk}^{(t)} - \beta_1^{(c)})} = \sigma^2_{\beta_{1jk}^{(t)}} + \sigma^2_{\beta_1^{(c)}}. \tag{5}$$

The σ^2_{ME} term is used in calculations of the non-centrality parameter and subsequent statistical tests making it crucial to developing statistical power formulations. Conceptually, σ^2_{ME} is simple. It is the combined error variance of the moderator coefficients in the intervention ($\sigma^2_{\beta_{1jk}^{(t)}}$) and control ($\sigma^2_{\beta_1^{(c)}}$) arms. However, our formulation of σ^2_{ME} requires components easily predicted during the design phase of a study. We begin with a reformulated error variance of the moderator coefficient in the intervention arm (i.e.,$\sigma^2_{\beta_{1jk}^{(t)}}$) such that (Dong et al., 2021b; Snijders, 2001; Snijders, 2005)

$$\sigma^2_{\beta_{1jk}^{(t)}} = \frac{\sigma^{2(t)}_{y|}(1 - R^2_{y_{L1}^{(t)}})/n_1^{(t)}}{(n_2^{(t)} n_3^{(t)} - C^{(t)} - 1)\sigma^2_{m^{(t)}}}. \tag{6}$$

The $\sigma^2_{\beta_{1jk}^{(t)}}$ term includes $\sigma^{2(t)}_{y|}$, variance of the outcome (e.g., mathematics performance scores) in the intervention arm, which is reduced by the proportion of outcome variance at level-one explained by the covariate at level-one ($R^2_{y_{L1}^{(t)}}$, see Raudenbush and Bryk (2002) for details on the calculation of multilevel R^2) and then divided by $n_1^{(t)}$, the sample of individuals per group (e.g., students per intervention group). The sample size terms for level -two and -three ($n_2^{(t)}$ and $n_3^{(t)}$) and variance of the moderator ($\sigma^2_{m^{(t)}}$) form the denominator of the $\sigma^2_{\beta_{1jk}^{(t)}}$ term along with $C^{(t)}$, the number of predictor variables in the outcome model of the treatment arm. This suggests that group sample size and teacher sample size along with greater variance in student math anxiety scores reduce $\sigma^2_{\beta_{1jk}^{(t)}}$. The second component of $\sigma^2_{ME}$is the error variance for the moderator

coefficient in the control arm (i.e.,$\sigma^2_{\beta_1^{(c)}}$) which we formulate as

$$\sigma^2_{\beta_1^{(c)}} = \frac{\sigma^2_{y|^{(c)}}(1 - R^2_{y^{(c)}_{L1}})}{(n^{(c)} - C^{(c)} - 1)\sigma^2_{m^{(c)}}}. \tag{7}$$

The terms in the formulation of $\sigma^2_{\beta^{(t)}_{1jk}}$ and $\sigma^2_{\beta_1^{(c)}}$ are similar with total sample size in the control arm represented with $n^{(c)}$ (e.g., waitlisted students). Substituting the expanded formulations of $\sigma^2_{\beta^{(t)}_{1jk}}$ and $\sigma^2_{\beta_1^{(c)}}$ into the σ^2_{ME}formula from Equation 5 we have

$$\sigma^2_{ME} = \frac{\sigma^{2(t)}_{y|}(1 - R^2_{y^{(t)}_{L1}})/n_1^{(t)}}{(n_2^{(t)}n_3^{(t)} - C^{(t)} - 1)\sigma^2_{m^{(t)}}} + \frac{\sigma^2_{y|^{(c)}}(1 - R^2_{y^{(c)}_{L1}})}{(n^{(c)} - C^{(c)} - 1)\sigma^2_{m^{(c)}}}. \tag{8}$$

Noncentrality parameter, statistical test, and power

With an estimate of the moderated effect (*ME*) and its variance (σ^2_{ME}), we can determine statistical significance using a *t* test. Assuming the alternative hypothesis is true, the *t* statistic will follow a noncentral *t* distribution with a noncentrality parameter of

$$\frac{ME}{\sigma_{ME}} \tag{9}$$

with $n_3^{(t)} - 2$ degrees of freedom. The statistical power for the two-sided test is then

$$P(|t_{ME}| > t_{critical}) = (1 - t(t_{critical} - t_{ME}) + t(-t_{critical} - t_{ME})) \tag{10}$$

where *t* is the cumulative *t* density described above. While we assume a continuous moderator, our formulas are easily adapted to accommodate binary moderators (see Binary Moderators in the Supplemental Material).

Three/two partial nesting

A more complex partially nested design involves a three-level data structure in one treatment arm and a two-level data structure in the remaining treatment arm (see Figure 1b). Multilevel structure in both the intervention and control arms allows a moderator at the lower-level and/or the upper-level to affect the intervention-outcome relationship. We utilize a new illustrative example to describe lower-level and upper-level moderation in 3/2 partially nested designs, their associated analytic models, moderator effect variance and power formulas, and simulation study results.

The example is adapted from Weidinger et al. (2020) and investigates the effect of a utility-value intervention on student mathematics performance while considering student-level and teacher- or classroom-level moderation effects. Our hypothetical 3/2 partially nested study design has a three-level data structure in the intervention arm and a two-level data structure in the control arm and we again assume error terms for the intervention and control arms do not covary. Individual random assignment places high-school students into the intervention or control arm. Students are then assigned to a mathematics teacher. Students in the intervention arm classrooms are also placed in small groups for a utility-value intervention in which students evaluate and discuss interview quotations from other high school students describing situations in which mathematics was useful (Gaspard et al., 2015; Weidinger et al., 2020). Utility-value interventions that prompt students to better grasp the usefulness of mathematics have improved student motivation and achievement (e.g., Brisson et al., 2017; Gaspard et al., 2015) but it is unclear if these benefits

(i.e., intervention effects) are consistent across all students, teachers, and/or classrooms (e.g., Rosenzweig et al., 2019). Our example study considers a composite measure of student's socio-economic status (family income, parental educational attainment, parental occupation) as a lower-level moderator of the utility-value intervention-mathematics performance relationship. To review, the intervention arm has students nested within utility-value intervention groups, nested within classrooms while the control arm includes students nested within classrooms. Note the treatment arms have a corresponding lower-level (students) and a corresponding upper-level (teachers) but the intervention arm has an additional middle-level induced by the intervention (utility-value intervention groups). In summary, the example study is examining the effect of a utility-value intervention on high-school student mathematics performance while considering treatment effect heterogeneity across student socio-economic status.

Utilizing the MA-PN framework, the analytic model for the three-level intervention arm in a 3/2 partially nested design is unchanged from the model for the 3/1 design. However, a 3/2 partially nested design requires a two-level outcome model for the control arm. With an individual-level moderator, $m_{ijk}^{(c)} \sim N(0, \sigma^2_{m^{(c)}})$ and covariate $(x_{ijk}^{(c)} \sim N(0, \sigma^2_{x^{(c)}})$ it can be represented using

$$\begin{gathered} y_{ijk}^{(c)} = \beta_{00k}^{(c)} + \beta_{1jk}^{(c)} m_{ijk}^{(c)} + \beta_{2jk}^{(c)}(x_{ijk}^{(c)} - \bar{x}_{.jk}^{(c)}) + \varepsilon_{ijk}^{(c)} \quad \varepsilon_{ijk}^{(c)} \sim N(0, \sigma^2_{y|^{(c)}}) \\ \beta_{00k}^{(c)} = \delta_{000}^{(c)} + \beta_{002}^{(c)} \bar{x}_{..k}^{(c)} + \nu_{00k}^{(c)}, \quad \nu_{00k}^{(c)} \sim N(0, \phi^2_{y|^{(c)}}). \end{gathered} \tag{11}$$

This model reflects the similarity between corresponding levels in the intervention and control arms. Level-one or the lower-level of the analytic models for the intervention and control arms are nearly identical with the superscript *c* indicated the control arm. The upper-level of both analytic models is also nearly identical (i.e., level-three in the intervention arm and level-two in the control arm). The middle-level is absent from the control arm as it is induced by the group-administered intervention. Contrasting the coefficients capturing the relationship between the moderator and outcome from the different treatment arms again provides an estimate of the moderation effect and the sum of the moderator coefficient variances again forms the variance of *ME*.

The second component of σ^2_{ME} for 3/2 partially nested designs is the error variance for the moderator coefficient in the control arm (i.e., $\sigma^2_{\beta_{1jk}^{(c)}}$). This term reflects the two-level data structure of the control group such that

$$\sigma^2_{\beta_{1jk}^{(c)}} = \frac{\sigma^2_{y|^{(c)}}(1 - R^2_{y_{L1}^{(c)}})/n_1^{(c)}}{(n_3^{(c)} - C^{(c)} - 1)\sigma^2_{m^{(c)}}}. \tag{12}$$

The terms in the formulation $\sigma^2_{\beta_{1jk}^{(c)}}$ now include individual-level variance of the outcome and individual sample size in the numerator and upper-level sample size and the variance of the moderator in the denominator. In our example, individual-level variance of mathematics performance score reduced by variance explained by covariates, and divided by students per classroom would form the numerator while sample of classrooms and individual-level variance of student socio-economic status composite scores are the key terms in the denominator. Substituting the expanded formulations of $\sigma^2_{\beta_{1jk}^{(t)}}$ and $\sigma^2_{\beta_{1jk}^{(c)}}$ into the general σ^2_{ME}formula from Equation 5 we have

$$\sigma^2_{ME} = \frac{\sigma^2_{y|^{(t)}}(1 - R^2_{y_{L1}^{(t)}})/n_1^{(t)}}{(n_2^{(t)} n_3^{(t)} - C^{(t)} - 1)\sigma^2_{m^{(t)}}} + \frac{\sigma^2_{y|^{(c)}}(1 - R^2_{y_{L1}^{(c)}})/n_1^{(c)}}{(n_3^{(c)} - C^{(c)} - 1)\sigma^2_{m^{(c)}}}. \tag{13}$$

The non-centrality parameter, t-test for determining a significant moderated effect and power formula remain unchanged from the 3/1 partially nested design formulations.

Upper-level moderator

Thus far we have considered moderation effects stemming from moderators located at level-one (e.g., student anxiety and student socio-economic status). For 3/2 partially nested designs a variable located at the upper-level (level-three in the intervention arm and level-two in the control arm) can moderate the relationship between the intervention and outcome. Upper-level moderators may include variables measured at the upper-level or aggregates of lower-level variables. We can revise our current working example such that the moderator is a composite score of teacher effectiveness (e.g., Martinez et al., 2016) with intervention effect heterogeneity across teacher effectiveness levels under examination. The moderator is now located at the upper-level (i.e., teacher-level) while other study design features remain the same.

To examine upper-level moderation effects we make a slight change to the analytic models. In the intervention arm of a 3/2 partially nested design with a continuous outcome ($y_{ijk}^{(t)}$), individual-level covariate, $x_{ijk}^{(t)} \sim N(0, \sigma_x^{2(t)})$, and upper-level moderator, $m_k^{(t)} \sim N(0, \phi^2_{m|^{(t)}})$ the analytic model is

Level 1:

$$y_{ijk}^{(t)} = \beta_{0jk}^{(t)} + \beta_{2jk}^{(t)}(x_{ijk}^{(t)} - \bar{x}_{.jk}^{(t)}) + \varepsilon_{ijk}^{(t)}, \quad \varepsilon_{ijk}^{(t)} \sim N(0, \sigma^2_{y|^{(t)}}) \tag{14}$$

Level 2:

$$\beta_{0jk}^{(t)} = \beta_{00k}^{(t)} + \beta_{02k}^{(t)}(\bar{x}_{.jk}^{(t)} - \bar{x}_{..k}^{(t)}) + u_{0jk}^{(t)}, \quad \mu_{0jk}^{(t)} \sim N(0, \tau^2_{y|^{(t)}})$$

Level 3:

$$\beta_{00k}^{(t)} = \delta_{000}^{(t)} + \beta_{001}^{(t)} m_k^{(t)} + \beta_{002}^{(t)} \bar{x}_{..k}^{(t)} + u_{00k}^{(t)}, \quad \mu_{00k}^{(t)} \sim N(0, \phi^2_{y|^{(t)}})$$

The moderator ($m_k^{(t)}$) is now located at the upper-level (level-three) of the intervention arm outcome model. The two-level control arm outcome model is now

$$\begin{aligned} y_{ijk}^{(c)} &= \beta_{00k}^{(c)} + \beta_{2jk}^{(c)}(x_{ijk}^{(c)} - \bar{x}_{.jk}^{(c)}) + \varepsilon_{ijk}^{(c)} \quad \varepsilon_{ijk}^{(c)} \sim N(0, \sigma^2_{y|^{(c)}}) \\ \beta_{00k}^{(c)} &= \delta_{000}^{(c)} + \beta_{001}^{(c)} m_k^{(c)} + \beta_{002}^{(c)} \bar{x}_{..k}^{(c)} + u_{00k}^{(c)}, \quad \mu_{00k}^{(c)} \sim N(0, \phi^2_{y|^{(c)}}) \end{aligned} \tag{15}$$

with the moderator, $m_k^{(c)} \sim N(0, \phi^2_{m|^{(c)}})$, located at the upper-level (level-two). While the moderator is located at different levels in the respective outcome models (intervention and control) both reflect the classroom-or teacher-level.

We now subtract the coefficients paired with the upper-level moderator in the treatment and control arms to estimate of the moderation effect (*ME*) such that

$$ME = \beta_{001}^{(t)} - \beta_{001}^{(c)}. \tag{16}$$

The sum of the moderator coefficient variances still forms the variance of *ME* such that

$$\sigma^2_{ME} = \sigma^2_{(\beta_{001}^{(t)} - \beta_{001}^{(c)})} = \sigma^2_{\beta_{001}^{(t)}} + \sigma^2_{\beta_{001}^{(c)}}. \tag{17}$$

However, the error variance of moderator coefficient in the intervention arm (i.e.,$\sigma^2_{\beta^{(t)}_{001}}$) must now include outcome variance across all levels such that

$$\sigma^2_{\beta^{(t)}_{001}} = \frac{\phi^2_{y|^{(t)}}(1-R^2_{y^{(t)}_{L3}}) + \tau^2_{y|^{(t)}}(1-R^2_{y^{(t)}_{L2}})/n^{(t)}_2 + \sigma^2_{y|^{(t)}}(1-R^2_{y^{(t)}_{L1}})/(n^{(t)}_2 n^{(t)}_1)}{(n^{(t)}_3 - C^{(t)} - 1)\sigma^2_{m^{(t)}}} \tag{18}$$

In terms of our example, the error variance of the moderation effect includes the variance of mathematics performance scores at the student-level ($\sigma^2_{y|^{(t)}}$), at the utility-value intervention group level ($\tau^2_{y|^{(t)}}$), and the classroom/teacher-level ($\phi^2_{y|^{(t)}}$). All of these outcome error variance terms can be reduced by the variance explained by covariates ($R^2_{y^{(t)}}$) and sample size at the corresponding level. For example, outcome error variance at the utility-value intervention group level ($\tau^2_{y|^{(t)}}$) is reduced by intervention group sample size ($n^{(t)}_2$).

The error variance for the moderator coefficient in the control arm (i.e.,$\sigma^2_{\beta^{(c)}_{001}}$) also includes outcome variance across levels but reflects a two-level data structure. We formulate $\sigma^2_{\beta^{(c)}_{001}}$ such that

$$\sigma^2_{\beta^{(c)}_{001}} = \frac{\phi^2_{y|^{(c)}}(1-R^2_{y^{(c)}_{L3}}) + \sigma^2_{y|^{(c)}}(1-R^2_{y^{(c)}_{L1}})/n^{(c)}_1}{(n^{(c)}_3 - C_{(c)} - 1)\sigma^2_{m^{(c)}}} \tag{19}$$

with terms and notation retaining similar meaning from the intervention arm outcome model. We utilize level-one and level-three notation in the control arm outcome model to emphasize the correspondence between levels in both outcome models.

Substituting these formulations of $\sigma^2_{\beta^{(t)}_{001}}$ and $\sigma^2_{\beta^{(c)}_{001}}$ into the σ^2_{ME}formula we have

$$\begin{aligned}\sigma^2_{ME} = &\frac{\phi^2_{y|^{(t)}}(1-R^2_{y^{(t)}_{L3}}) + \tau^2_{y|^{(t)}}(1-R^2_{y^{(t)}_{L2}})/n^{(t)}_2 + \sigma^2_{y|^{(t)}}(1-R^2_{y^{(t)}_{L1}})/(n^{(t)}_2 n^{(t)}_1)}{(n^{(t)}_3 - C^{(t)} - 1)\sigma^2_{m^{(t)}}} + \\ &\frac{\phi^2_{y|^{(c)}}(1-R^2_{y^{(c)}_{L3}}) + \sigma^2_{y|^{(c)}}(1-R^2_{y^{(c)}_{L1}})/n^{(c)}_1}{(n^{(c)}_3 - C_{(c)} - 1)\sigma^2_{m^{(c)}}}.\end{aligned} \tag{20}$$

The non-centrality parameter, t-test for determining a significant moderated effect and power formula remain unchanged from the previous designs and analytic models.

Cluster randomized trials with partial nesting

Thus far we have only considered partially nested designs with individual randomization to treatment arm. Cluster randomized designs or cluster randomized trials (CRTs) randomize intact groups or clusters to treatment arm and are commonly employed in educational experiments. For 3/2 partially nested designs, this equates to randomization at the upper-level (see Figure 1c). This is not a trivial change when considering moderation effects because cluster randomization allows the moderator to vary within and between groups. Recall, for individual randomized trials the randomization of individuals allows us to assume a moderator at the lower- or upper-level does not systematically vary across groups. For CRTs, it is plausible, if not likely, that moderator values vary systematically between groups. For CRTs with 3/2 partial nesting, we must account for moderator variance within and between groups in the analytic models and subsequent moderator effect variance formulations. Put differently, randomization of groups still ensures asymptotically unbiased estimates of the main treatment effect but individual-level moderators may vary systematically between these extant groups.

For our example we adapt Lawrence et al. (2017) to illustrate CRTs with 3/2 partially nested data. The study examines Word Generation, a whole school intervention for middle school student vocabulary (Lawrence et al., 2017). This two-level CRT investigates a group-administered treatment such that an additional level of grouping is added to the intervention arm. Middle schools are randomly assigned to implement the Word Generation intervention or be placed on a two-year waitlist before implementation to serve as a control. The intervention itself involves specific instructional tasks implemented by English Language Arts teachers with the outcome of interest student academic vocabulary. Differential treatment effects across baseline student vocabulary ability are of interest so it is included as a possible moderator of the Word Generation-student academic vocabulary treatment effect.

In summary, intact schools (level-3) are randomly assigned to adopt the Word Generation intervention with ELA teachers (level-2) implementing the intervention in treatment classrooms and middle school student (level-1) academic vocabulary serving as the outcome of interest. The intervention arm includes students nested within teachers, nested within schools but the control arm only includes students nested within schools. This hypothetical investigation includes student-level baseline vocabulary scores ($m_{ijk}^{(t)}$) as a possible moderator of the Word Generation-student academic vocabulary treatment effect. Teachers may have systematic differences in aggregated levels of the moderator ($\bar{m}_{.jk}^{(t)}$) that influence the treatment effect, and the use of extant schools may also produce aggregated levels of the moderator ($\bar{m}_{..k}^{(t)}$) that influence the treatment effect.

To consider this systematic variation and capture the moderation effects possible at each level, we include the aggregate or average moderator in the analytic model of the outcome in the intervention arm at levels two and three ($\bar{m}_{.jk}^{(t)}$ and $\bar{m}_{..k}^{(t)}$) such that

Level 1:

$$y_{ijk}^{(t)} = \beta_{0jk}^{(t)} + \beta_{1jk}^{(t)} m_{ijk}^{(t)} + \beta_{2jk}^{(t)}(x_{ijk}^{(t)} - \bar{x}_{.jk}^{(t)}) + \varepsilon_{ijk}^{(t)}, \quad \varepsilon_{ijk}^{(t)} \sim N(0, \sigma^2_{y|^{(t)}}) \tag{21}$$

Level 2:

$$\beta_{0jk}^{(t)} = \beta_{00k}^{(t)} + \beta_{01k}^{(t)} \bar{m}_{.jk}^{(t)} + \beta_{02k}^{(t)}(\bar{x}_{.jk}^{(t)} - \bar{x}_{..k}^{(t)}) + u_{0jk}^{(t)}, \quad u_{0jk}^{(t)} \sim N(0, \tau^2_{y|^{(t)}})$$

Level 3:

$$\beta_{00k}^{(t)} = \delta_{000}^{(t)} + \beta_{001}^{(t)} \bar{m}_{..k}^{(t)} + \beta_{002}^{(t)} \bar{x}_{..k}^{(t)} + v_{00k}^{(t)}, \quad v_{00k}^{(t)} \sim N(0, \phi^2_{y|^{(t)}}).$$

In terms of our example, we have academic vocabulary scores as the outcome ($y_{ijk}^{(t)}$) for student i, instructed by teacher j, in school k. Baseline vocabulary ability ($m_{ijk}^{(t)}$) for student i, instructed by teacher j, in school k is included to track possible moderation effects ($\beta_{1jk}^{(t)}$). The average student baseline vocabulary ability for a teacher and school ($\bar{m}_{.jk}^{(t)}$ and $\bar{m}_{..k}^{(t)}$) are also included to track possible moderation effects at level-two and level-three. For example, a student's baseline vocabulary ability may influence the effectiveness of the Word Generation intervention but the average

student baseline vocabulary ability for a teacher ($\bar{m}^{(t)}_{.jk}$) may also influence the effectiveness of the Word Generation. Perhaps, the intervention is easier to implement with a student group possessing higher baseline vocabulary or, conversely, a group of students with lower baseline vocabulary may be more receptive to the intervention. These moderation effects would be captured with $\beta^{(t)}_{01k}$ while school-level moderation effects are captured with $\beta^{(t)}_{001}$.

The outcome model in the control arm also includes the aggregated moderator at the upper-level to consider moderator variance within and between groups such that

$$\begin{aligned} y^{(c)}_{ijk} &= \beta^{(c)}_{00k} + \beta^{(c)}_{1jk} m^{(c)}_{ijk} + \beta^{(c)}_{2jk}(x^{(c)}_{ijk} - \bar{x}^{(c)}_{.jk}) + \varepsilon^{(c)}_{ijk} \quad \varepsilon^{(c)}_{ijk} \sim N(0, \sigma^2_{y|^{(c)}}) \\ \beta^{(c)}_{00k} &= \delta^{(c)}_{000} + \beta^{(c)}_{001}\bar{m}^{(c)}_{..k} + \beta^{(c)}_{002}\bar{x}^{(c)}_{..k} + v^{(c)}_{00k}, \quad v^{(c)}_{00k} \sim N(0, \phi^2_{y|^{(c)}}). \end{aligned} \tag{22}$$

Terms and interpretations for the outcome model of the control arm retain similar meaning from the intervention arm.

Our moderation effect variance formulations for CRTs with 3/2 partially nested data must accommodate moderator variation within and between groups. This requires a multilevel model for the moderator to reflect variation across the three-levels such that

Level 1:

$$m^{(t)}_{ijk} = \beta^{(m,t)}_{0jk} + \varepsilon^{(m,t)}_{ijk}, \quad \varepsilon^{(m,t)}_{ijk} \sim N(0, \sigma^2_{m^{(t)}}) \tag{23}$$

Level 2:

$$\beta^{(m,t)}_{0jk} = \beta^{(m,t)}_{00k} + u^{(m,t)}_{0jk}, \quad u^{(m,t)}_{0jk} \sim N(0, \tau^2_{m^{(t)}})$$

Level 3:

$$\beta^{(m,t)}_{00k} = \delta^{(m,t)}_{000} + v^{(m,t)}_{00k}, \quad v^{(m,t)}_{00k} \sim N(0, \phi^2_{m^{(t)}}).$$

The key terms to map out the variance of moderation effects in CRTs with 3/2 partially nested data are: $\sigma^2_{m^{(t)}}$, the variance of the moderator or baseline student vocabulary scores, $\tau^2_{m^{(t)}}$, the variance of the moderator at level-two or the variance of baseline student vocabulary scores at the teacher-level, and $\phi^2_{m^{(t)}}$, the variance of the moderator at level-three or variance of baseline student vocabulary scores at the school-level.

A similar multilevel model for the moderator is utilized in the control arm such that

$$\begin{aligned} m^{(c)}_{ijk} &= \beta^{(m,c)}_{00k} + \varepsilon^{(m,c)}_{ijk} \quad \varepsilon^{(m,c)}_{ijk} \sim N(0, \sigma^2_{m^{(c)}}) \\ \beta^{(m,c)}_{00k} &= \delta^{(m,c)}_{000} + v^{(m,c)}_{00k}, \quad v^{(m,c)}_{00k} \sim N(0, \phi^2_{m^{(c)}}). \end{aligned} \tag{24}$$

Like the moderator model in the intervention arm, the key terms for estimating moderator effect error variance are $\sigma^2_{m^{(c)}}$ and $\phi^2_{m^{(c)}}$. These terms capture the variance of the moderator in the control arm at the lower- and upper-level. In our example, $\sigma^2_{m^{(c)}}$ and $\phi^2_{m^{(c)}}$ capture variance of baseline vocabulary scores at the student- and school-level in the control arm.

Moderator effects and error variance

We consider the total moderation effect of $m^{(.)}_{ijk}$ because it is unlikely that a study will be designed with power to detect a level-specific moderation effect as the focal point. In our example, we are investigating treatment effect heterogeneity across different student baseline vocabulary ability.

The actual moderation effect may stem from differences in student baseline vocabulary ability, aggregated student baseline vocabulary ability at the teacher-level, aggregated student baseline vocabulary ability at the school-level or some combination (i.e., $\beta^{(t)}_{1jk} + \beta^{(t)}_{01k} + \beta^{(t)}_{001}$). Without group mean centering, the coefficients paired with the moderator and aggregated moderator at each level capture the total moderation effect such that

$$ME_T = (\beta^{(t)}_{1jk} + \beta^{(t)}_{01k} + \beta^{(t)}_{001}) - (\beta^{(c)}_{1jk} + \beta^{(c)}_{001}). \tag{25}$$

This is akin to finding the total moderation effect of student baseline vocabulary ability, aggregated student baseline vocabulary ability at the teacher-level and aggregated student baseline vocabulary ability at the school-level on the Word Generation-student academic vocabulary relationship.

With a total moderation effect estimated using coefficients paired with each moderator (or aggregated moderator value), the error variance of the total moderation effect is the sum of all coefficient variances such that

$$\sigma^2_{ME_T} = (\sigma^2_{\beta^{(t)}_{1jk}} + \sigma^2_{\beta^{(t)}_{01k}} + \sigma^2_{\beta^{(t)}_{001}}) + (\sigma^2_{\beta^{(c)}_{1jk}} + \sigma^2_{\beta^{(c)}_{001}}). \tag{26}$$

Variance of coefficient estimates for moderators that vary within groups remain the same (see Equations 6 and 15 for $\sigma^2_{\beta^{(t)}_{1jk}}$ and $\sigma^2_{\beta^{(c)}_{1jk}}$ respectively). The variance of coefficient estimates related to the aggregated moderators ($\sigma^2_{\beta^{(t)}_{001}}$, $\sigma^2_{\beta^{(t)}_{01k}}$ and $\sigma^2_{\beta^{(c)}_{001}}$) are new terms. Under stated assumptions, we formulate these new coefficient variance terms as

$$\sigma^2_{\beta^{(t)}_{01k}} = \frac{\tau^2_{y|^{(t)}}(1 - R^2_{y^{(t)}_{L2}}) + (\sigma^2_{y|^{(t)}}(1 - R^2_{y^{(t)}_{L1}})/n^{(t)}_1)}{(n^{(t)}_3 n^{(t)}_2 - 2)(\tau^2_{m|^{(t)}} + \sigma^2_{m^{(t)}}/n^{(t)}_1)} \tag{27}$$

$$\sigma^2_{\beta^{(t)}_{001}} = \frac{\phi^2_{y|^{(t)}}(1 - R^2_{y^{(t)}_{L3}}) + (\tau^2_{y|^{(t)}}(1 - R^2_{y^{(t)}_{L2}})/n^{(t)}_2) + (\sigma^2_{y|^{(t)}}(1 - R^2_{y^{(t)}_{L1}})/n^{(t)}_2 n^{(t)}_1)}{(n^{(t)}_3 - 2)(\phi^2_{m|^{(t)}} + \tau^2_{m|^{(t)}}/n^{(t)}_2 + \sigma^2_{m|^{(t)}}/n^{(t)}_2 n^{(t)}_1)} \tag{28}$$

and

$$\sigma^2_{\beta^{(c)}_{001}} = \frac{\phi^2_{y|^{(c)}}(1 - R^2_{y^{(c)}_{L3}}) + (\sigma^2_{y|^{(c)}}(1 - R^2_{y^{(c)}_{L1}})/n^{(c)}_1)}{(n^{(c)}_3 - 2)(\phi^2_{m|^{(c)}} + \sigma^2_{m^{(c)}}/n^{(c)}_1)}. \tag{29}$$

These formulations can be substituted into the total moderation effect variance formula (e.g., $\sigma^2_{ME_T}$) with the non-centrality parameter, t-test for determining a significant moderated effect, and power formula remaining unchanged.

These variance terms are a bit more complicated than those found in partially nested designs using individual random assignment due to the additional moderator variance terms (e.g., $\phi^2_{m|^{(t)}}$, $\phi^2_{m|^{(c)}}$, and $\tau^2_{y|^{(t)}}$). However, they follow a very similar structure to the previous coefficient variance formulas. Generally, outcome variance terms occupy the numerator of the formulas and are reduced by variance explained by covariates and the sample size at the corresponding level. The denominator now includes variance of the moderator at all applicable levels and these variance terms are reduced by sample size at the corresponding level. For example, Equation 32 is the formula for variance of the coefficient associated with the upper-level aggregated moderator ($\beta^{(c)}_{001}$). It includes variance of the outcome in the control arm at the upper-level and lower-level, variance explained by covariate terms, and lower-level per group sample size in the numerator. This combination of components matches pervious two-level coefficient variance formulas (see Equation

22). The denominator includes group and per group sample size terms along with variance of the moderator at the upper- and lower-levels. To summarize in terms of our example, when a moderator such as baseline student vocabulary ability may vary systematically within and between groups, which can be expected in a CRT randomizing schools, variance of the baseline student vocabulary moderator must be modeled and included in moderation effect error variance formulas.

Simulation studies

We conducted four simulation studies to establishment the accuracy of our four moderation effect variance and power formulas with three-level partially nested data (see Supplemental Materials for R code). Conditions were purposefully selected to provide some initial indication of sample size requirements to consistently detect these moderated effects and the influence of key design parameters on power rates (see Table 1). We generated data sets in R (R Core Team, 2021) with sample sizes of 10 and 20 at each level. Control sample size was set to ensure a balanced design with $n^{(c)} = n_1^{(t)} \times n_2^{(t)} \times n_3^{(t)}$ for 3/1 partial nesting and $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$ for 3/2 partial nesting. Moderation effects of $ME = 0.1$ and $ME = 0.05$ were included with a total $ME = 0.15$ for CRTS with 3/2 partial nesting ($\beta_{1jk}^{(t)} = 0.1, \beta_{01k}^{(t)} = 0.0, \beta_{001}^{(t)} = 0.05$). In the intervention arm, three values of individual-level variance of the outcome were considered,$\sigma^2_{y^{(t)}} = 0.9,\ 0.8,$ and 0.6 with variance of the outcome at level-two and level-three held equal with corresponding values of $\tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.05,\ 0.1,$ and 0.2. In the control arm, $\sigma^2_{y^{(c)}} = 1$ and $\phi^2_{y^{(c)}} = 0$ for the single-level control condition in a 3/1 partially nested design. In all other control arm conditions $\sigma^2_{y^{(c)}} = 0.95,\ 0.9,$ and 0.8 with corresponding values at the upper-level of $\phi^2_{y^{(c)}} = 0.05,\ 0.1,$ and 0.2. Finally, we considered variance explained in the outcome at each level by covariates. Across all levels in both the intervention and control arm, we examined $R^2 = 0.0,\ 0.4,$ and 0.7. Simulation conditions were guided by previous simulation literature examining partially nested designs (Cox & Kelcey, 2022; Roberts, 2021; Roberts et al., 2016; Heo et al., 2017; Snijders, 2005) and moderation in group-randomized trials (Spybrook et al., 2016; Dong et al., 2016; Dong et al., 2021b; Mathieu et al., 2012) with all analyses conducted in R (R Core Team, 2021).

Results

Power to detect a lower-level moderation effect in a three/one partially nested design

Formula based predicted power rates to detect a moderated effect from a lower-level moderator in a 3/1 partially nested design closely approximated empirical rejection rates in our simulation study. The accuracy of the power formulas held across various decompositions of outcome variance, variance explained by covariates, and sample sizes (see Table 2 for selected results and Supplementary Materials for all conditions). Slightly overestimated power rates in several conditions stem from the small sample sizes considered at the second and third levels of the intervention arm (e.g., $n_2^{(t)}$ and $n_3^{(t)}$). These discrepancies dissipate as $n_2^{(t)}$ and $n_3^{(t)}$ increase and a comparison of formula-based moderation effect error variance and the observed variance of the moderation effect across simulation runs supports power formula accuracy.

While limited, results do provide some initial indication of the feasibility of detecting a moderation effect in a 3/1 partially nested design. Typical sample sizes found in planned educational experiments with multilevel structures (e.g., Schochet, 2011) will often be sufficient to consistently detect these moderation effects. For example, power to detect the moderation effect of student math anxiety on the summer school spatial reasoning intervention would be >80% in all but the

Table 1. Monte Carlo simulation conditions by moderated effect and partially nested data structure.

Partial Nesting (Effect)	*ME*	R^2	$n_3^{(t)}$	$n_2^{(t)}$	$n_1^{(t)}$	Outcome Variance Decomposition		
3/1	0.1	0.0	10	10	10	$\sigma^2_{y^{(t)}} = 0.9,\ \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.05$	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$	$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$
(Lower-level)	0.05	0.4	10	10	20			
			10	20	20	$\sigma^2_{y^{(c)}} = 1.0,\ \phi^2_{y^{(c)}} = \text{na}$	$\sigma^2_{y^{(c)}} = 1.0,\ \phi^2_{y^{(c)}} = \text{na}$	$\sigma^2_{y^{(c)}} = 1.0,\ \phi^2_{y^{(c)}} = \text{na}$
			20	20	20			
3/2	0.1	0.0	10	10	10	$\sigma^2_{y^{(t)}} = 0.9,\ \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.05$	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$	$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$
(Lower-level)	0.05	0.4	10	10	20			
			10	20	20	$\sigma^2_{y^{(c)}} = 0.95,\ \phi^2_{y^{(c)}} = 0.05$	$\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$	$\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$
			20	20	20			
3/2	0.1	0.0	10	10	10	$\sigma^2_{y^{(t)}} = 0.9,\ \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.05$	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$	$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$
(Upper-level)	0.05	0.4	10	10	20			
		0.7	10	20	20	$\sigma^2_{y^{(c)}} = 0.95,\ \phi^2_{y^{(c)}} = 0.05$	$\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$	$\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$
			20	20	20			
3/2 CRT	0.15	0.0	10	10	10	$\sigma^2_{y^{(t)}} = 0.9,\ \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.05$	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$	$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$
(Total)		0.4	10	10	20			
		0.7	10	20	20	$\sigma^2_{y^{(c)}} = 0.95,\ \phi^2_{y^{(c)}} = 0.05$	$\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$	$\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$
			20	20	20			

Note. $n^{(c)} = n_1^{(t)} \times n_2^{(t)} \times n_3^{(t)}$ for 3/1 partial nesting and $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$ for 3/2 partial nesting. For 3/2 CRT (Total) $\sigma^2_{y|^{(t)}} = \sigma^2_{m^{(t)}}$, $\tau^2_{y|^{(t)}} = \tau^2_{m^{(t)}}$, and $\phi^2_{y|^{(t)}} = \phi^2_{m^{(t)}}$.

Table 2. Selected comparisons of formula based statistical power rate and Monte Carlo simulation rejection rate for a lower-level moderation effect in a 3/1 partially nested design.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$				$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$				$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1,$				$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$			
$n_3^{(t)}$	10	10	10	20	10	10	10	20	10	10	10	20	10	10	10	20
$n_2^{(t)}$	10	10	20	20	10	10	20	20	10	10	20	20	10	10	20	20
$n_1^{(t)}$	10	20	20	20	10	20	20	20	10	20	20	20	10	20	20	20
$R^2_{\cdot}$	0	0	0	0	0	0	0	0	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Rejection rate	0.51	0.84	0.99	1.00	0.57	0.89	1.00	1.00	0.76	0.98	1.00	1.00	0.81	0.99	1.00	1.00
Power rate	0.54	0.83	0.98	1.00	0.59	0.87	0.99	1.00	0.75	0.96	1.00	1.00	0.81	0.98	1.00	1.00
Difference	−0.03	0.01	0.01	0.00	−0.02	0.02	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.01	0.00	0.00

Note. Results were based on 5,000 replications with $ME = 0.1$ and in the control condition $n^{(c)} = n_1^{(t)} \times n_2^{(t)} \times n_3^{(t)}$, and $\sigma^2_{y^{(c)}} = 1$.

Table 3. Selected comparisons of formula based statistical power rate and Monte Carlo simulation rejection rate for a lower-level moderation effect in a 3/2 partially nested design.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
	$\sigma^2_{y(t)} = 0.8, \tau^2_{y(t)} = \phi^2_{y(t)} = 0.1$				$\sigma^2_{y(t)} = 0.6, \tau^2_{y(t)} = \phi^2_{y(t)} = 0.2$				$\sigma^2_{y(t)} = 0.8, \tau^2_{y(t)} = \phi^2_{y(t)} = 0.1$				$\sigma^2_{y(t)} = 0.6, \tau^2_{y(t)} = \phi^2_{y(t)} = 0.2$			
	$\sigma^2_{y(c)} = 0.9,\ \phi^2_{y(c)} = 0.1$				$\sigma^2_{y(c)} = 0.8,\ \phi^2_{y(c)} = 0.2$				$\sigma^2_{y(c)} = 0.9,\ \phi^2_{y(c)} = 0.1$				$\sigma^2_{y(c)} = 0.8,\ \phi^2_{y(c)} = 0.2$			
$n_3^{(t)}$	10	10	10	20	10	10	10	20	10	10	10	20	10	10	10	20
$n_2^{(t)}$	10	10	20	20	10	10	20	20	10	10	20	20	10	10	20	20
$n_1^{(t)}$	10	20	20	20	10	20	20	20	10	20	20	20	10	20	20	20
R^2	0	0	0	0	0	0	0	0	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
Rejection rate	0.53	0.86	0.99	1.00	0.62	0.92	1.00	1.00	0.79	0.98	1.00	1.00	0.86	0.99	1.00	1.00
Power rate	0.52	0.80	0.98	1.00	0.58	0.87	0.99	1.00	0.73	0.95	1.00	1.00	0.80	0.98	1.00	1.00
Difference	0.02	0.06	0.02	0.00	0.03	0.06	0.01	0.00	0.05	0.03	0.00	0.00	0.05	0.02	0.00	0.00

Note. Results were based on 5,000 replications with $ME = 0.1$ and in the control condition $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$.

Table 4. Selected comparisons of formula based statistical power rate and Monte Carlo simulation rejection rate for an upper-level moderation effect in a 3/2 partially nested design.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12
	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$ $\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$			$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$ $\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$			$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$ $\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$			$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$ $\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$		
$n_3^{(t)}$	10	10	20	10	10	20	10	10	20	10	10	20
$n_2^{(t)}$	10	20	20	10	20	20	10	20	20	10	20	20
$n_1^{(t)}$	20	20	20	20	20	20	20	20	20	20	20	20
$R^2_{\cdot}$	0	0	0	0	0	0	0.4	0.4	0.4	0.4	0.4	0.4
Rejection rate	0.07	0.07	0.14	0.05	0.05	0.08	0.09	0.08	0.19	0.06	0.06	0.12
Power rate	0.08	0.08	0.14	0.06	0.07	0.09	0.10	0.10	0.19	0.07	0.08	0.12
Difference	−0.01	−0.01	0.00	−0.02	−0.01	0.00	−0.01	−0.01	0.00	−0.01	−0.01	0.00

Note. Results were based on 5,000 replications with $ME = 0.1$ and in the control condition $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$.

smallest sample size considered when $ME = 0.1$. Even smaller moderation effects could be consistently detected if student, group, and teacher sample size at each level exceeded 20. Other design parameters that influenced power rates included outcome variance decomposition and variance explained by covariates. While minor (<10%), increased share in outcome variance at levels two and three in the intervention arm, increased power to detect the moderation effect. Put differently, if more variance in mathematics performance scores is attributable to group- and teacher-levels, power to detect the moderation effect of student math anxiety would increase. The use of covariates that explain variance in the outcome also increased power to detect moderation paralleling results found in previous literature (e.g., Cox & Kelcey, 2022; Cox et al., under review). In our example, a pretest of student mathematics performance would be a valuable covariate to include in the outcome model of both intervention and control arms.

Power to detect a lower-level moderation effect in a three/two partially nested design

Our second simulation study found formula based predicted power rates for lower-level moderation effects in a 3/2 partially nested design closely approximated empirical rejection rates (see Table 3 for selected results and Supplementary Materials for all conditions). We again found minor discrepancies between predicted power and simulation rejection rates at the smallest sample sizes. As with moderated effects in 3/1 partially nested designs, no systematic bias was found in the error variance formula and power rates matched reject rates as sample sizes increased.

Power rates and the influence of design parameters were very similar for lower-level moderation effects across designs with 3/1 and 3/2 partially nested data. In terms of our illustration, it would be feasible to detect the moderating effect of student socio-economic status on the Utility-value intervention treatment effect with typical sample sizes. Greater variation in mathematics performance at the intervention-group (level-two) and classroom-level (level-three) and inclusion of a student's pretest mathematics performance score in the outcome models would increase power to detect the moderation effects.

Power to detect an upper-level moderation effect in a three/two partially nested design

Our power formulas for detecting upper-level moderation effects in designs with 3/2 partially nested data again closely matched simulation study rejection rates (see Table 4 for selected results and Supplementary Materials for all conditions). With moderators located at the upper-level of the treatment and control models, power to detect moderation effects was much more dependent on upper-level sample size (i.e., $n_3^{(t)}$ and $n_3^{(c)}$). Logistical and financial constraints inherently limit these sample sizes thus reducing the power to detect upper-level moderation effects under typical conditions. We adjusted some conditions in our simulation study to reflect differences in detecting upper-level moderation effects. First, we eliminated the small ME condition ($ME = 0.05$) because power rates would be consistently and extraordinarily low. We then added an $R^2 = 0.7$ condition and varied $n_3^{(t)}$ sample sizes from 30 to 100 to consider power formula accuracy across a wider range of power rates. These additional conditions also provide supplementary evidence regarding the feasibility of detecting upper-level moderation effects.

The shift from lower-level moderation to upper-level moderation effects produced substantially lower power rates, as noted, and led to changes in the influence of design parameters. Results suggest we would have great difficulty detecting moderation effects from a teacher effectiveness composite score or any other teacher-level characteristic on the Utility-value intervention-student mathematics relationship. Power to detect upper-level moderation effects under the original simulation conditions never exceeded 20%. Even with $R^2 = 0.7$, power to detect the upper-level moderation effect did not exceed 60%. This suggests that even inclusion of a very effective covariate such as mathematics performance pretest score would not result in consistently detectible upper-

Table 5. Selected comparisons of formula based statistical power rate and Monte Carlo simulation rejection rate for the total moderation effect in a cluster randomized trial with 3/2 partially nested data.

Scenario	1	2	3	4	5	6	7	8	9	10	11	12
	$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$ $\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$			$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$ $\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$			$\sigma^2_{y^{(t)}} = 0.8, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.1$ $\sigma^2_{y^{(c)}} = 0.9,\ \phi^2_{y^{(c)}} = 0.1$			$\sigma^2_{y^{(t)}} = 0.6, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$ $\sigma^2_{y^{(c)}} = 0.8,\ \phi^2_{y^{(c)}} = 0.2$		
$n_3^{(t)}$	10	10	20	10	10	20	10	10	20	10	10	20
$n_2^{(t)}$	10	20	20	10	20	20	10	20	20	10	20	20
$n_1^{(t)}$	20	20	20	20	20	20	20	20	20	20	20	20
R^2	0	0	0	0	0	0	0.4	0.4	0.4	0.4	0.4	0.4
Rejection rate	0.03	0.04	0.06	0.03	0.04	0.06	0.04	0.04	0.07	0.04	0.05	0.08
Power rate	0.06	0.06	0.07	0.06	0.06	0.07	0.06	0.06	0.08	0.06	0.06	0.09
Difference	−0.03	−0.02	−0.01	−0.03	−0.02	−0.01	−0.02	−0.02	−0.01	−0.03	−0.02	−0.01

Note. Results were based on 5,000 replications with $ME = 0.15$ and in the control condition $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$.

level moderation effects. In the expanded $n_3^{(t)}$ sample size conditions we found adequate power (i.e., 80%) was not achievable without covariates when $n_3^{(t)} \leq 100$. The only scenarios in which upper-level moderation effects were consistently detectable (i.e., >80%) included effective covariates (i.e., $R^2_{\cdot} = 0.7$) and larger upper-level sample sizes (i.e., $n_3^{(t)} \leq 60$; see complete results in Supplementary materials). It is possible that well resourced educational experiments could reach these sample sizes but for many it may be financially or practically difficult to recruit more than >100 teachers, schools or classrooms (e.g., $n_3^{(t)} \geq 50$ and $n_3^{(c)} \geq 50$).

As for the influence of other design parameters on power, detecting upper-level moderation effects was more susceptible to changes in outcome variance decomposition. Specifically, increased share in outcome variance at levels two and three in the intervention arm, substantially decreased power to detect the moderation effect. For example, when $n_3^{(t)} = n_2^{(t)} = n_1^{(t)} = 20$, $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$, $R^2_{\cdot} = 0.7$, and $ME = 0.1$, power to detect an upper-level moderation effect is $\approx$ 60% with $\sigma^2_{y^{(t)}} = 0.9, \tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.05, \sigma^2_{y^{(c)}} = 0.95$, and $\phi^2_{y^{(c)}} = 0.05$but $< 20\%$ under the same conditions with $\sigma^2_{y^{(t)}} = 0.6$, $\tau^2_{y^{(t)}} = \phi^2_{y^{(t)}} = 0.2$, $\sigma^2_{y^{(c)}} = 0.8$, and $\phi^2_{y^{(c)}} = 0.2$. Note that this is the inverse of results found for lower-level moderation. If we are interested in moderated effects of a teacher or classroom characteristic (i.e., moderator at the upper-level), it is advantageous for power rates if variance in mathematics performance outcome is concentrated at the student-level.

Power to detect the total moderation effect in a cluster randomized trials with three/two partial nesting

In our final simulation study, we examined the accuracy of our power formulas for moderated effects in cluster randomized trials with 3/2 partially nested data. Formula power rates approximated empirical rejection rates for moderated effects in this design (see Table 5 for selected results and Supplementary Materials for all conditions). We did note formula predicted power consistently overestimated the empirical rejection rate when $n_2^{(t)}$ and $n_1^{(t)}$ were small (e.g., ≤ 10). However, these discrepancies were small (i.e., ~2%) and disappeared as $n_2^{(t)}$ and $n_1^{(t)}$ exceeded 20.

We found the total moderated effect (i.e., moderated effects that include effects from the aggregated moderator at levels-2 and −3) was overwhelmed by additional error variance components. Put differently, it would be difficult to detect moderation effects from varying student baseline vocabulary ability on the Word Generation intervention. Randomization of schools would require the inclusion of student academic vocabulary variance at the student-, teacher-, and school-level to be included in components of the moderation effect variance. This additional variance would likely exceed any additional moderation effects found at the teacher- and school-level. For example, power to detect a total moderated effect of $ME = 0.15$ ($\beta^{(t)}_{1jk} = 0.1$, $\beta^{(t)}_{001} = 0.05$) was ~8% even with $R^2_{\cdot} = 0.4$. Sample sizes at the upper-levels must be increased to achieve power rates even approaching adequate (i.e., 80%). For example, power to detect the total moderation effect in these designs was 45% with $n_3^{(t)} = 100$, $n_2^{(t)} = n_1^{(t)} = 10$, $n_1^{(c)} = n_1^{(t)} \times n_2^{(t)}$ and $n_3^{(c)} = n_3^{(t)}$, $R^2_{\cdot} = 0.7$, and ME_T=0.15. There are, of course, conditions in which one could consistently detect total moderation effects in these designs (e.g., large $n_3^{(t)}, n_2^{(t)}, n_1^{(t)}$ values, large ME values) but acquiring these sample sizes in planned educational experiments is difficult. For example, it would be difficult to recruit more than 200 schools (intervention and control arm $n_3^{(.)}$ sample sizes ≥ 100) for our example study. We conclude that under common conditions detecting total moderation effects in cluster randomized studies with 3/2 partially nested data is not practically feasible.

Discussion

Comprehensive evaluation of treatment effects is aided by considering moderated effects. In educational research, the combination of natural hierarchical structures (e.g., students nested within classrooms) and prevalence of group-administered or shared facilitator treatments often produces

three-level partially nested data. Literature details planning strategies for a variety of experimental designs when moderation effects are of interest but had yet to establish power formulas for detecting moderation effects in three-level partially nested designs. The lack of planning strategies and tools (e.g., power formulas) made efficient planning of these studies difficult. To address this gap, we developed moderation effect variance formulas and the subsequent power formulas for detecting moderation effects in three/one and three/two partially nested designs. We conducted simulation studies to both assess the accuracy of the newly developed formulas and provide an initial probe into power and adequate sample sizes for moderation effects in these designs. Simulation conditions included different decompositions of outcome variance across levels, differing amounts of variance in the outcome explained by covariates, different magnitudes for the moderation effect, and different sample sizes across all levels.

The two primary contributions of this work are then (a) the moderation effect variance and subsequent power formulas for detecting moderation effects in three-level partially nested designs and (b) increased understanding of power rates and adequate sample sizes for detecting moderation effects in partially nested designs. These contributions improve planning of educational experiments with partial nesting. Use of resources is likely to be more efficient because estimates of the adequate sample sizes needed to consistently detect moderation effects will be more accurate. Availability of power formulas and other study planning guidance (e.g., use of covariates) also promotes inclusion of moderator variables that elucidate important features of an intervention (i.e., for whom and under what conditions it is effective). To promote adoption and implementation of the formulas they have been implemented in the online R-Shiny application PowerUpRShiny (https://poweruprshiny.shinyapps.io/PartiallyNestedPower/).

More specific implications of this research involve the feasibility of detecting moderation effects in studies with different partially nested data structures. It is unlikely that a planned educational experiment will be designed for the sole purpose of investigating moderated effects. Rather, intervention effectiveness is likely to be the primary focus. It is then beneficial to consider the power to detect moderation effects under sample sizes typically achieved in planned educational experiments. Detecting moderation effects stemming from individual-level moderators in both 3/1 and 3/2 partially nested designs is feasible under a variety of scenarios with samples sizes typically seen in planned educational experiments. We encourage the inclusion of individual-level moderators in these studies because sample sizes are likely adequate to detect any moderation effects and these effects provide a more comprehensive understanding of treatment effects. Conversely, the detection of moderation effects from upper-level moderators in 3/2 partially nested designs is not feasible with samples sizes typically seen in planned educational experiments. Detection of these upper-level moderation effects rely on upper-level sample sizes that have logistical (limited availability) and financial (limited budgets) constraints.

These findings are similar to those for detecting moderation effects in fully nested three-level designs (i.e., three-level cluster randomized trials, Dong et al., 2018). Specifically, in both fully and partially nested three-level designs the location of the moderator or moderation effect was the most influential factor in determining the feasibility of detecting the effect. For both designs, power to detect lower-level moderation effects was typically adequate with sample sizes that achieve adequate power to detect main effects. Conversely, power to detect upper-level moderation effects was typically inadequate with sample sizes that achieve adequate power to detect main effects.

We also found typical sample sizes resulted in inadequate power to detect total moderation effects in cluster randomized trials with 3/2 partially nested data. In this context, it is not necessarily limited sample sizes that prevent the detection of moderation effects but the study design structure and resulting moderator effect variance. With randomization occurring at the upper-level, individual-level moderators may vary in systematic ways across groups and this variance must be incorporated into moderation effect error variance formulas. Our results indicate these

additional variance components typically overwhelm the total moderation effect producing inadequate power across the conditions considered. Exceptions are certainly possible but detecting total moderation effects is likely to be difficult in typical cluster randomized education experiments with partial nesting.

Conclusion

This study, like all simulation studies, had a limited scope of conditions that was practically considerable. A more comprehensive examination of the different moderation effects possible in these partially nested designs is recommended with a particular focus on conditions for sufficiently detecting upper-level moderation and moderation in cluster randomized trials with partial nesting. These comprehensive examinations should expand on the sample size conditions and variance of the outcome decompositions.

Along with subsequent simulation studies, we encourage exploration and documentation of the design parameter values required for the moderation effect power formulas. Accuracy of these parameters directly influences the accuracy of ensuing power formulas. Design parameters have been compiled for several types of experimental designs and outcomes but need further development for partially nested designs that include moderators. Not only will empirically based design parameters increase study design accuracy, they can be used to better delineate appropriate conditions in future simulation studies.

To close, let us summarize two key takeaways from the results of this study. First, three-level partially nested studies that randomize at the individual-level will often be sufficiently powered to detect a moderated effect when utilizing typical sample sizes for detecting main effects. Increasing individual-level sample sizes and incorporating covariates that explain variance in the outcome represent two effective strategies for increasing the power to detect moderation effects in these settings. Second, upper-level moderation in 3/2 partially nested designs and total moderation in cluster randomized designs often require upper-level sample sizes that are not feasible in educational experiments. In these settings, outcome variance explained by covariates does increase power but not dramatically enough to ensure study feasibility.

Funding

This article is based on work funded by the National Science Foundation [#1552535 and #1760884].

ORCID

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

References

Ashcraft, M., & Krause, J. (2007). Working memory, math performance, and math anxiety. *Psychonomic Bulletin & Review*, *14*(2), 243–248. https://doi.org/10.3758/bf03194059

Ashcraft, M., & Moore, A. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment*, *27*(3), 197–205. https://doi.org/10.1177/0734282908330580

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, *16*(2), 149–165. https://doi.org/10.1037/a0023464

Bauer, D. J., Sterba, S. B., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, *43*(2), 210–236. https://doi.org/10.1080/00273170802034810

Brisson, B. M., Dicke, A.-L., Gaspard, H., H€afner, I., Flunger, B., Nagengast, B., & Trautwein, U. (2017). Short intervention, sustained effects: Promoting students' math competence beliefs, effort, and achievement. *American Educational Research Journal*, *54*(6), 1048–1078. https://doi.org/10.3102/0002831217716084

Candlish, J., Teare, M. D., Munyaradzi, D., Flight, L., Mandefield, L., & Walters, S. J. (2018). Appropriate statistical methods for analysing partially nested randomized controlled trials with continuous outcomes: A simulation study. *BMC Medical Research Methodology*, *18*(1), 105. https://doi.org/10.1186/s12874-018-0559-x

Chambers, B., Abrami, P., Tucker, B., Slavin, R. E., Madden, N. A., Cheung, A., & Gifford, R. (2008). Computer-assisted tutoring in success for all: Reading outcomes for first graders. *Journal of Research on Educational Effectiveness*, *1*(2), 120–137. https://doi.org/10.1080/19345740801941357

Cox, K., & Kelcey, B. (2022). Statistical power for detecting moderation in partially nested designs. *American Journal of Evaluation*. https://doi.org/10.1177/1098214020977692

Cox, K., Kelcey, B., & Luce, H. (under review). Power to detect moderated effects with random slopes in partially nested designs. *Journal of Educational and Behavioral Statistics*.

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. https://doi.org/10.1080/00220973.2017.1315714

Dong, N., Kelcey, B., & Spybrook, J. (2021a). Design considerations in multisite randomized trials probing moderated treatment effects. *Journal of Educational and Behavioral Statistics*, *46*(5), 527–559. https://doi.org/10.3102/1076998620961492

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. https://doi.org/10.5964/meth.4003

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/.

Fuchs, D., & Fuchs, L. S. (2019). On the importance of moderator analysis in intervention research: An introduction to the special issue. *Exceptional Children*, *85*(2), 126–128. https://doi.org/10.1177/0014402918811924

Gaspard, H., Dicke, A., Flunger, B., Brisson, B., Hafner, I., Nagengast, B., & Trautwein, U. (2015). Fostering adolescents' value beliefs for mathematics with a relevance intervention in the classroom. *Developmental Psychology*, *51*(9), 1226–1240.

Hedges, L. V., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods*, *47*(4), 1295–1308. https://doi.org/10.3758/s13428-014-0538-z

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. https://doi.org/10.1177/0962280214547381

Institute of Education Sciences (IES). (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. P., Lin, L., & Ma, B. (2016). An empirical study of design parameters for assessment differential impacts for students in group randomized trials. *Evaluation Review*, *40*(5), 410–443. https://doi.org/10.1177/0193841X16659600

Kelcey, B., Bai, F., & Xie, Y. (2020). Statistical power in partially nested designs probing multilevel mediation. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *30*(8), 1061–1074. https://doi.org/10.1080/10503307.2020.1717012

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 (Dissertation) (pp. 59–94). 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*(3), 274–289. https://doi.org/10.1177/1368430214550343

Lawrence, J. F., Francis, D., Paré-Blagoev, J., & Snow, C. E. (2017). The poor get richer: Heterogeneity in the efficacy of a school-level intervention for academic language. *Journal of Research on Educational Effectiveness*, *10*(4), 767–793. https://doi.org/10.1080/19345747.2016.1237596

Lee, K. J., & Thompson, S. G. (2005). The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials (London, England)*, *2*(2), 163–173. https://doi.org/10.1191/1740774505cn082oa

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

Lowrie, T., 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. https://doi.org/10.1080/00220973.2019.1684869

MacKinnon, D. P. (2011). Integrating mediators and moderators in research design. *Research on Social Work Practice*, *21*(6), 675–681. https://doi.org/10.1177/1049731511414148

Martinez, J. F., Schweig, J., & Goldschmidt, P. (2016). Approaches for combining multiple measures of teacher performance: Reliability, validity, and implications for evaluation policy. *Educational Evaluation and Policy Analysis*, *38*(4), 738–756. https://doi.org/10.3102/0162373716666166

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. *The Journal of Applied Psychology*, *97*(5), 951–966.

Moerbeek, M., & Wong, W. (2008). Sample size formulae for trials com- paring group and individual treatments in a multilevel model. *Statistics in Medicine*, *27*(15), 2850–2864. https://doi.org/10.1002/sim.3115

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

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

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. https://doi.org/10.3102/0162373707299460

Reardon, S. F., & Stuart, E. A. (2017). Editor's introduction: Theme issue on variation in treatment effects. *Journal of Research on Educational Effectiveness*, *10*(4), 671–674. https://doi.org/10.1080/19345747.2017.1386037

Roberts, C. (2021). The implications of noncompliance for randomized trials with partial nesting due to group treatment. *Statistics in Medicine*, *40*(2), 349–368. https://doi.org/10.1002/sim.8778

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. https://doi.org/10.1002/sim.6828

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials (London, England)*, *2*(2), 152–162. https://doi.org/10.1191/1740774505cn076oa

Rosenzweig, E. Q., Hulleman, C. S., Barron, K. E., Kosovich, J. J., Priniski, S. J., & Wigfield, A. (2019). Promises and pitfalls of adapting utility-value interventions for online math courses. *The Journal of Experimental Education*, *87*(2), 332–352. https://doi.org/10.1080/00220973.2018.1496059

Sanders, E. A. (2011). *Multilevel analysis methods for partially nested cluster randomized trials*. (No. 3452760). ProQuest LLC.

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*(4), 441–471. https://doi.org/10.3102/1076998610375840

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*, *39*(3), 268–286.

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 & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1570–1573). Wiley.

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. https://doi.org/10.3102/1076998616655442

Tong, G., Esserman, D., & Li, F. (2022). Accounting for unequal cluster sizes in designing cluster randomized trials to detect treatment effect heterogeneity. *Statistics in Medicine*, *41*(8), 1376–1396. https://doi.org/10.1002/sim.9283

Weidinger, A. F., Gaspard, H., Harackiewicz, J. M., Paschke, P., Bergold, S., & Steinmayr, R. (2020). Utility-value intervention in school: Students' migration and parental educational backgrounds as moderators. *The Journal of Experimental Education*, *90*(2), 364–382.

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. https://doi.org/10.1002/sim.8721