

## Statistical power in partially nested designs probing multilevel mediation


Ben Kelcey, Fangxing Bai & Yanli Xie

To cite this article: Ben Kelcey, Fangxing Bai & Yanli Xie (2020): Statistical power in partially nested designs probing multilevel mediation, *Psychotherapy Research*, DOI: [10.1080/10503307.2020.1717012](https://doi.org/10.1080/10503307.2020.1717012)

To link to this article: <https://doi.org/10.1080/10503307.2020.1717012>

 View supplementary material 

---

 Published online: 09 Feb 2020.

---

 Submit your article to this journal 

---

 Article views: 46

---

 View related articles 

---

 View Crossmark data 

---

## METHOD PAPER

# Statistical power in partially nested designs probing multilevel mediation

BEN KELCEY, FANGXING BAI, & YANLI XIE

*College of Education, Criminal Justice, Human Services and Information Technology, University of Cincinnati, Cincinnati, OH, USA*

*(Received 16 September 2019; revised 18 December 2019; accepted 19 December 2019)*

### Abstract

**Objective:** Analysis of the intermediate behaviors and mechanisms through which innovative therapies come to shape outcomes is a critical objective in many areas of psychotherapy research because it supports the iterative exploration, development and refinement of theories and therapies. Despite widespread interest in the intermediate behaviors and mechanisms that convey treatment effects, there is limited guidance on how to effectively and efficiently design studies to detect such mediated effects in the types of partially nested designs that commonly arise in psychotherapy research. In this study, we develop statistical power formulas to identify requisite sample sizes and guide the planning of studies probing mediation under two- and three-level partially nested designs.

**Method:** We investigate multilevel mediation in partially nested structures and models for two- and three-level designs.

**Results:** Well-powered studies probing mediation using partially nested designs will typically require moderate to large sample sizes or moderate to large effects.

**Discussion:** We implement these formulas in the R package PowerUpR and a simple Shiny web application (<https://poweruprshiny.shinyapps.io/PartiallyNestedMediationPower/>) and demonstrate their use to plan studies using partially nested designs.

**Keywords:** statistical methodology; process research; partially nested; mediation

**Clinical or methodological significance of this article:** Despite widespread interest in the intermediate behaviors and mechanisms that convey treatment effects, there is limited guidance on how to effectively and efficient design studies to detect such mediated effects in the types of partially nested designs that commonly arise in psychotherapy research. we develop statistical power formulas to identify requisite sample sizes and guide the planning of studies probing mediation under two- and three-level partially nested designs.

Analysis of the intermediate behaviors and mechanisms through which innovative therapies come to shape individual outcomes is a central aim in psychotherapy research because it supports the iterative exploration, development and refinement of theories and therapies (e.g., Gottfredson et al., 2015; Lachowicz, Sterba, & Preacher, 2015). Mediation analyses now represent a core standard of evidence across a diverse range of fields and provide a general platform for the iterative testing and development of innovative therapies, the execution and assessment of

measurement-based care approaches, and the evaluation of program fidelity and engagement (e.g., Eden, 2017; Kelcey, Hill, & Chin, 2019; Nohe, Michaelis, Menges, Zhang, & Sonntag, 2013; Scott & Lewis, 2015). For instance, mediation has been leveraged to delineate the intermediate role of ineffective parental discipline in youth treatments for attention-deficit/hyperactivity disorder (e.g., Hinshaw, 2007), to evaluate how the development of coping skills scaffolds the effects of depression treatments (e.g., Christensen, Haugen, Sirpal, & Haavet, 2015; Liu, Chang, Wu, & Tsai, 2015), to

---

Correspondence concerning this article should be addressed to Ben Kelcey, Office 638-O, 2610 McMicken Circle, University of Cincinnati, Cincinnati, OH 45221, USA. Email: ben.kelcey@gmail.com

probe the mediating role of regulatory skills in mode deactivation therapy (e.g., Apsche, Bass, & Backlund, 2012), to examine client engagement as a critical intermediate process of measurement-based care programs and targeted outcomes (e.g., Lindsey et al., 2014; Scott, 2018), and to detail and optimize therapeutic change across different types of therapies (Kazdin, 2007; Kazdin & Nock, 2003).

More generally, there has been a sustained emphasis in the literature on the study of treatment mechanisms—the how and why of treatment effectiveness—in the psychotherapy literature (e.g., Kazdin, 2006, 2007; Gelfand, Mensinger, & Tenhave, 2009; Sterba, 2017; Windgassen, Goldsmith, Moss-Morris, & Chalder, 2016). This literature has detailed and built up specific statistical and conceptual principles to support the design, analysis, and interpretation of mediation studies (e.g., Candlish, Teare, Cohen, & Bywater, 2019; Kazdin, 2006, 2007; Gelfand et al., 2009; Sterba, 2017; Windgassen et al., 2016). Many other fields have also adopted a comparable focus on mediation and the mechanisms for change (e.g., Gottfredson et al., 2015). The Society for Prevention Research, for instance, has introduced specific standards that advocate for and guide mediational analyses (e.g., Aguinis, Edwards, & Bradley, 2017; Eden, Stone-Romero, & Rothstein, 2015; Gottfredson et al., 2015).

Despite widespread interest in and emphasis on the intermediate behaviors and mechanisms that convey treatment effects, there is limited guidance on how to effectively and efficiently design studies to detect such processes in the types of designs that commonly arise in psychotherapy research (Bauer, Sterba, & Hallfors, 2008; Sterba, 2017). More specifically, psychotherapy research frequently leverages partially nested designs because the nature of many treatments induces clustering through the delivery mode of the treatment (e.g., through therapy groups and/or mental health professionals) that may not exist in many control conditions. A common example arises when individuals are randomly assigned to remain on a control waitlist or participate in a treatment that is delivered by mental health professionals. In this instance, individuals in the treatment condition are clustered or nested within mental health professionals whereas individuals in the control condition are unclustered or independent.

Although recent reviews have delineated and underscored the distinct design and analysis considerations associated with partially nested designs, there has been a lack of guidance on how to plan partially nested studies that target mediation effects (Sterba, 2017). In this study, we address this gap by developing statistical power formulas and software for partially nested designs that identify requisite

sample sizes and guide the planning of studies probing mediation (e.g., Fritz & MacKinnon, 2007; Kelcey, Dong, Spybrook, & Shen, 2017). A critical question in planning a study or assessing the likely sensitivity of a particular design and sample size is the statistical power or probability with it can detect an effect if it exists. Power analyses assess the sensitivity of a design under different sample sizes. More generally, in conjunction with other constraints (e.g., financial, geographical, practical), power analyses can be used as a starting guideline for what might be a reasonable scale for a particular study. The results detailed in this study are intended to help researchers identify how large of a sample size is needed to have a reasonably high chance of detecting mediation effects in two- and three-level partially nested studies. The results can also be used to better understand the likely power or capacity of studies with a fixed sample size. For instance, based on unwritten rules of thumb, we might presume that a study with 50 therapists serving 10 patients each would be large enough to detect mediation effects with a high probability (e.g., 80% chance). Without a power analysis framework, however, we cannot know even the approximate validity of such a guideline. In these ways, the present study seeks to develop power analyses to gain a better understanding of the scale needed to conduct multilevel mediation studies and to assist researchers with the identification and selection of appropriate sample sizes when planning a study.

We investigated two- and three-level partially nested designs that commonly arise in psychotherapy research (e.g., Roberts & Roberts, 2005): (a) the treatment arm has a two-level structure but the control condition has only a one-level structure (denoted as two-/one-level), (b) the treatment arm has a three-level structure but the control condition has only a one-level structure (denoted as three-/one-level), and (c) the treatment arm has a three-level structure but the control condition has only a two-level structure (denoted as three-/two-level). Below we consider each of the three focal designs in turn by outlining the nature of the design, the error variances governing mediation effects and sampling considerations, and three statistical tests and their associated power formulas. We then assess the precision of the formulas using simulations and end sections with an example using the *PowerUpR* Shiny application for simplified calculations.

### Two-Level Partially Nested Designs

We first considered an individually-randomized study that induces one degree of clustering (two

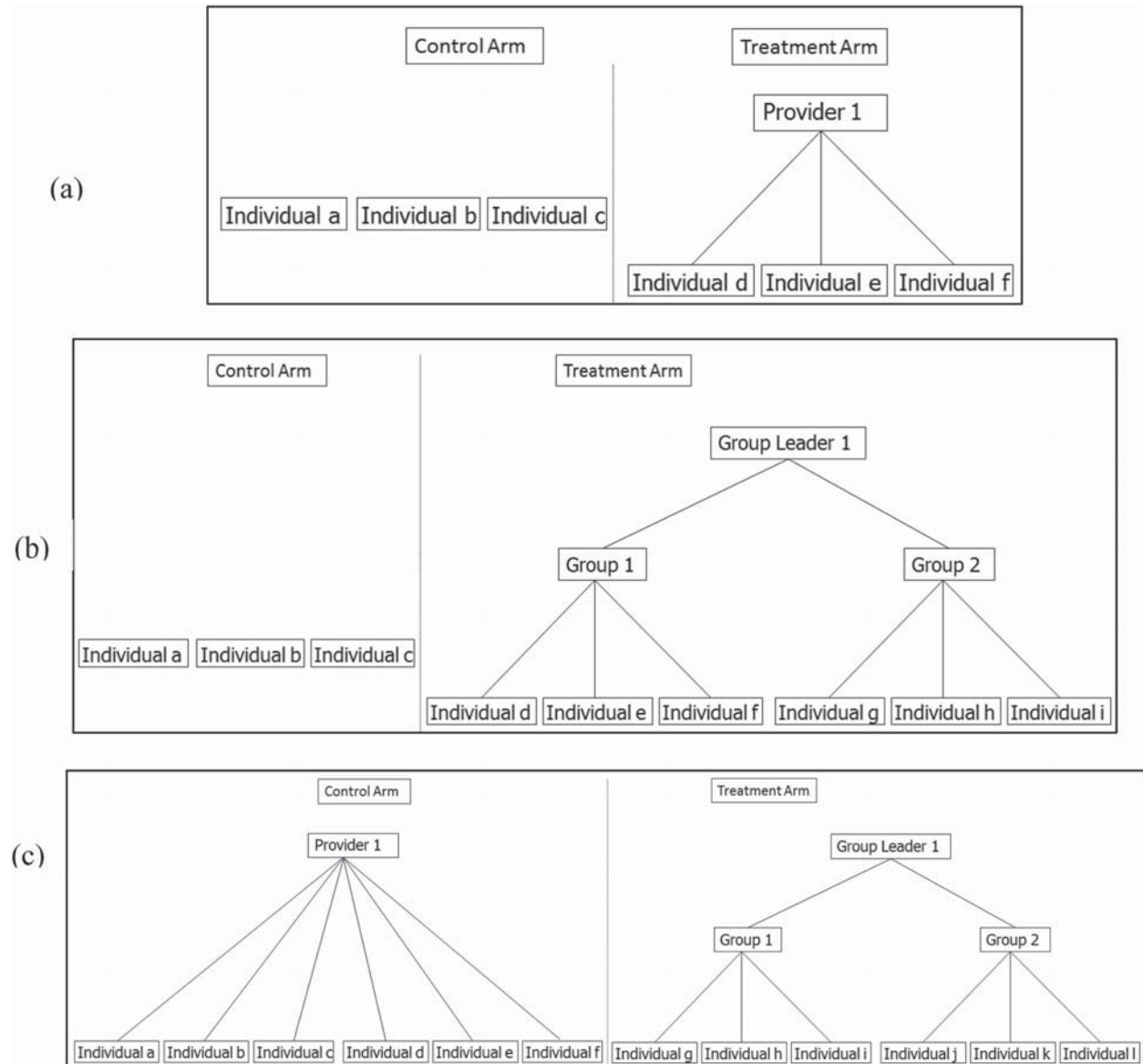


Figure 1. Organizational structures in partially nested designs for: (a) two-/one-level design such that the control arm is unclustered and the treatment arm has individuals clustered within providers; (b) three-/one-level design such that the control arm is unclustered and the treatment arm has individuals clustered within therapy groups which are then clustered within group leaders; and (c) three-/two-level design such that the control arm has individuals clustered within providers and the treatment arm has individuals clustered within therapy groups which are then clustered within group leaders.

levels) for those assigned to a treatment condition but no clustering for those assigned to a control or comparison condition (denoted as two-/one-level; Figure 1). As previously noted, a prototypical example of this type of design arises when individuals are assigned at random to either participate in an innovative type of therapy led by a mental health professional or act as a control case by remaining on a waitlist (unclustered). In these settings, individuals assigned to the innovative therapy become clustered within a mental health professional such that the eventual outcomes are not necessarily independent among

individuals served by the same provider. By contrast, the outcomes of the individuals on the waitlist are expected to remain independent because they do not participate in a therapy that plausibly introduces common provider effects.

To provide a working example, consider the design of a study examining the extent to which the impact of the Problem Solving Skills Training program (PSST; treatment) on disruptive behavior (outcome) operates through improvements in coping skills (mediator) in children with oppositional defiant disorder (e.g., Kazdin, 2010; Suldo, Parker, Shaunessy-Dedrick, &

O’ Brennan, 2019). Oppositional defiant disorder is in part described by patterns of irritability and disruptive behavior (e.g., stubbornness, tantrums) and prior research has established consistent predictive associations between this disorder in children and later psychiatric disorders including, for example, depression and anxiety (Lochman et al., 2015).

The PSST program is a cognitive-behavioral approach designed to reduce disruptive behavior in children by establishing constructive methods for children to manage thoughts, feelings and their interactions with others (e.g., Kazdin, 2010). In part, the theory of action that undergirds the PSST program is predicated on the notion that disruptive behaviors arise because of difficulties in emotion regulation. That is, many children with oppositional defiant disorder lack constructive ways to deal and cope with negative and positive emotions, thoughts and feelings (Dunsmore, Booker, & Ollendick, 2013). Our example analysis considers a two-/one-level partially nested study that randomly assigns individuals (level one) to a waitlist control condition or to participate in a PSST program delivered by a mental health provider (level two).

## Models

A common approach to estimating these effects draws on a system of linear mixed effect models to concurrently account for the nesting structure and delineate the relationships among the treatment, mediator and outcome (e.g., Kelcey, Dong, Spybrook, & Cox, 2017; Sterba, 2017; VanderWeele, 2010). In this study, we adopt the flexible and common multiple-arm multilevel framework for partially nested data that accommodates the potential for heteroscedasticity and differential relationships across treatment conditions through treatment condition-specific models (e.g., Lachowicz et al., 2015). To explain variation in the mediator, we draw on two complementary models—a multilevel model for the treatment arm and single-level model for the control arm. For the treatment arm we have

$$\begin{aligned} M_{ij} &= \pi_{0j}^{(t)} + \pi_1^{(t)}(X_{ij} - \bar{X}_j) + \pi_2^{(t)}V_{ij} + \varepsilon_{ij}^{M(t)} \\ \varepsilon_{ij}^{M(t)} &\sim N(0, \sigma_{M_i^{(t)}}^2) \\ \pi_{0j}^{(t)} &= a^{(t)} + \zeta_1^{(t)}\bar{X}_j + \zeta_2^{(t)}W_j + u_{0j}^{M(t)} \\ u_{0j}^{M(t)} &\sim N(0, \tau_{M_i^{(t)}}^2) \end{aligned} \quad (1)$$

Here, we use  $M_{ij}$  as used as the mediator (e.g., coping skills) value for individual  $i$  served by provider  $j$ ,  $a^{(t)}$  as the overall intercept that represents the conditional

average of the mediator value in the treatment condition,  $X_{ij}$  is an individual-level pretreatment covariate (e.g., baseline measure of disruptive behavior) with  $\bar{X}_j$  as its average across all individuals served by the same provider (with coefficient  $\zeta_1^{(t)}$ ),  $(X_{ij} - \bar{X}_j)$  as the provider-mean centered version of the individual-level pretreatment covariate (with coefficient  $\pi_1^{(t)}$ ),  $V_{ij}$  as an individual-level pretreatment covariate that varies only across individuals within providers (no variation across providers) with coefficient  $\pi_2^{(t)}$ ,  $W_j$  as a provider-level pretreatment covariate (e.g., experience level of provider) with coefficient  $\zeta_2^{(t)}$ ,  $\varepsilon_{ij}^{M(t)}$  as the individual-level error term and  $u_{0j}^{M(t)}$  is the provider-level random effects for the treatment condition.

For the control arm we have a simpler single-level or unclustered model

$$\begin{aligned} M_i &= a^{(c)} + \pi_1^{(c)}X_i + \pi_2^{(c)}V_{ij} + \varepsilon_i^{M(c)} \\ &\sim N(0, \sigma_{M_i^{(c)}}^2) \end{aligned} \quad (2)$$

Here we use  $M_i$  as mediator for independent individual  $i$ ,  $a^{(c)}$  as the overall intercept capturing the conditional average of the mediator in the control condition,  $X_i$  as the uncentered version of the pretreatment individual-level covariate (with coefficient  $\pi_1^{(c)}$ ), as  $V_{ij}$  as a potential second individual-level pretreatment covariate (with coefficient  $\pi_2^{(c)}$ ), and  $\varepsilon_i^{M(c)}$  as the individual-level error term for control individuals.

To track how improvements in the mediator manifest as improvements in an outcome, we draw on an outcome model for the treatment condition is

$$\begin{aligned} Y_{ij} &= \beta_{0j}^{(t)} + b_1^{(t)}(M_{ij} - \bar{M}_j) + \beta_1^{(t)}(X_{ij} - \bar{X}_j) \\ &\quad + \beta_2^{(t)}V_{ij} + \varepsilon_{ij}^{Y(t)} \quad \varepsilon_{ij}^{Y(t)} \sim N(0, \sigma_{Y_i^{(t)}}^2) \\ \beta_{0j}^{(t)} &= \zeta_{00}^{(t)} + B^{(t)}\bar{M}_j + \zeta_1^{(t)}\bar{X}_j + \zeta_2^{(t)}W_j + u_{0j}^{Y(t)} \\ u_{0j}^{Y(t)} &\sim N(0, \tau_{Y_i^{(t)}}^2) \end{aligned} \quad (3)$$

Here we expand our notation to include  $Y_{ij}$  as the outcome (e.g., disruptive behavior) for individual  $i$  served by provider  $j$ ,  $\beta_{0j}^{(t)}$  as the conditional average outcome value across all individuals served by provider  $j$ ,  $\bar{M}_j$  as the average mediator (e.g., coping skills) across individuals served by provider  $j$  with path coefficient  $B$ ,  $(M_{ij} - \bar{M}_j)$  as the provider-mean centered individual mediator value (with coefficient  $b_1$ ) and  $u_{0j}^{Y(t)}$  and  $\varepsilon_{ij}^{Y(t)}$  as the provider and individual error terms.

In the specification of the outcome and mediator models, we draw on cluster- (i.e., provider-) mean centered variables such that individual-level predictor variables that vary both within and among providers (i.e.,  $X$  and  $M$ ) are centered based on their provider-level means (e.g., Raudenbush & Bryk, 2002). This type of centering is widely used because it conceptually deconfounds the within and between level relationships (e.g., Pituch & Stapleton, 2012). Alternative approaches including grand mean centering or no centering, however, will produce equivalent results in our models (e.g., Enders & Tofighi, 2007; Kelcey, Dong, Spybrook, & Shen, 2017). In our outcome model (3), for instance, the provider-level coefficient ( $B$ ) attached to the average mediator score ( $\bar{M}_j$ ) captures the total (i.e., within plus between) relationship between the mediator and outcome. In complement, the individual-level coefficient ( $b_1$ ) attached to the cluster-mean centered mediator ( $M_{ij} - \bar{M}_j$ ) captures the unique within cluster relationship between the mediator and outcome while the difference between the provider-level and individual-level coefficients ( $b_2 = B - b_1$ ) tracks the unique contextual or between cluster relationship (denoted as  $b_2$ ) between the mediator and outcome.

Returning to our example,  $b_1$  conceptually describes how changes in an individual's coping skills (mediator) correlate with changes in the disruptive behavior (outcome). In contrast,  $b_2$  describes how changes in the average level of coping skills across individuals served by a provider (i.e.,  $\bar{M}_j$ ) are additionally correlated with changes in disruptive behavior. In our example, there may be little reason to suspect such a contextual effect (i.e.,  $b_2 = 0$ ) because individuals served by a provider do not directly interact with each other. However, in alternative treatments that leverage, for example, group therapy such that individuals served by a provider directly interact, there may be a basis for such contextual effects ( $b_2 > 0$ ; see three-/two-level design section below for an example).

Under this model specification, the mediation effect is captured by

$$ME = (a^{(t)} - a^{(c)})B^{(t)} = aB \quad (4)$$

The first term ( $a^{(t)} - a^{(c)}$ ) in expression (4) captures how exposure to a treatment produces changes in the mediator relative to a control or comparison condition. The second term ( $B^{(t)}$ ) then quantifies how changes in the mediator translate into improvements in an outcome. Collectively, the  $aB$  product captures how changes in the mediator produced by exposure to a treatment manifest as improvements in an outcome.

## Statistical Power

To track the statistical power with which partially nested designs can detect the production of mediation effects in partially nested designs with maximum likelihood estimation, we extended three tests that can be used in the design phase before data has been collected. We outline these tests below and detail their developments and derivations in Section 3 of the Supplemental Material.

The first test we considered was the traditional the Sobel test that approximates the sampling distribution of the mediation effect using its asymptotic normality (Sobel, 1982). The comparison of the Sobel test to a normal distribution can be reasonable under sample sizes that are large. With small to moderate sample sizes, however, the normal distribution can serve as a poor referent distribution because the sampling distribution of the mediation effect can be heavily skewed (Kisbu-Sakarya, MacKinnon, & Miočević, 2014). The practical implication of this approximation is that the Sobel test tends to be underpowered and requires larger sample sizes relative to other tests.

A conventional but high-powered alternative that relaxes the normality assumption is the joint test (MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The joint test constructs inferences using sub-tests that target the individual path coefficients that summarize a mediation effect. More specifically, the joint test first employs a sub-test to examine the intervention-mediator path coefficient and then, separately, employs a sub-test to examine the mediator-outcome path coefficient. In turn, the joint test infers the presence of a mediation effect only when there is evidence of a nonzero coefficient for both paths. Despite its simplicity, prior research has demonstrated that the joint test functions well and yields inferences like resampling-based tests such as those based on bootstrap methods (e.g., Hayes & Scharkow, 2013; Kelcey, Dong, Spybrook, & Shen, 2017).

For the third test, we extended a contemporary resampling-based test—the Monte Carlo interval test (Preacher & Selig, 2012). Rather than approximate the sampling distribution of the mediation effect with theory, this test tracks the distribution by drawing plausible values for the treatment-mediator and the mediator-outcome path coefficients using the anticipated path coefficients ( $\hat{a}, \hat{B}$ ) and the expected error variances (Preacher & Selig, 2012). Prior research on single-level and multilevel mediation has consistently suggested that the Monte Carlo interval test performs well under a variety of conditions and is comparable to bootstrap-based methods that are only available after data collection (e.g., Kelcey, Dong, Spybrook, &

Cox, 2017; Kelcey, Spybrook, & Dong, 2019; Preacher & Selig, 2012).

The results of our derivations suggested that statistical power in two/one partially nested designs is governed by four categories of parameters that are outlined in Table I. The first category of parameters is the sample size. With two/one partially nested designs, we must consider three different sample sizes: (a) the number of providers in the treatment condition ( $\mathcal{J}_{(t)}$ ), (b) the number of individuals per each provider in the treatment condition ( $n_{(t)}$ ), and (c) the number of individuals in the control condition ( $n_{(c)}$ ).

Analysis of the roles of the different sample sizes across levels and conditions underscores that many different sampling plans can produce a desired level of power. For instance, the results suggested that when large reservoirs of waitlist control individuals are available, researchers can to some extent oversample from those reservoirs to reduce the requisite number of treatment providers and individuals. In other settings, the results suggest that when the sampling plan is constrained by cost considerations, it is advantageous to seek an optimal blend

of treatment and control units (e.g., Cox & Kelcey, 2019; Kelcey & Shen, 2016). More generally, under many design considerations, the number of providers in the treatment condition ( $\mathcal{J}_{(t)}$ ) will have the most influence on power followed by the number of individuals in the control condition ( $n_{(c)}$ ), and last the number of individuals per each provider in the treatment condition ( $n_{(t)}$ ). In settings where the control sample size ( $n_{(c)}$ ) is particularly small (e.g., less than about 25), increasing the control sample size ( $n_{(c)}$ ) can yield gains in power on par with increasing the number of providers in the treatment condition ( $\mathcal{J}_{(t)}$ ).

The second category of parameters is the path coefficients that detail the production of the mediation effect ( $a$ ,  $B$ , and  $b_1$ ). As outlined above, the  $a$  parameter describes the degree to which exposure to a treatment shifts the average mediator value relative to the average mediator value in the control condition. In our formulation, we standardize the mediator values on the total standard deviation (individual- plus cluster-level) pooled (across treatment and control arms) such that the  $a$  coefficient captures the standardized mean difference (e.g., Cohen,

Table I. Summary of parameters governing power in a two/one partial nested design.

Parameter	Parameter value	Label	Example
$a$	0.40	Standardized mean difference in mediator values between treatment and control groups ( $a = a^{(t)} - a^{(c)}$ )	Impact of PSST on coping skills
$B$	0.40	The standardized regression coefficient for the outcome on the mediator that captures the cluster- and individual-level relationship ( $B = b_1 + b_2$ )	Conditional association between coping skills and disruptive behavior
$b_1$	0.40	The standardized regression coefficient for the outcome on the mediator at the individual-level	Conditional association between coping skills and disruptive behavior at the individual-level
ICCy	0.20	Proportion of variance in outcome that is between clusters ( $\rho_Y$ ) for the treatment condition	Proportion of variation in disruptive behavior attributable to differences among providers
ICCM	0.20	Proportion of variance in mediator that is between clusters ( $\rho_M$ ) for the treatment condition	Proportion of variation in coping attributable to differences among providers
R2y1_Treatment	0.50	Proportion of outcome variance at individual-level explained by covariates for the treatment condition	Proportion of variance among individuals explained by pretreatment variables
R2y2_Treatment	0.50	Proportion of outcome variance at cluster-level explained by covariates for the treatment condition	Proportion of variance among providers explained by pretreatment variables
R2m1_Treatment	0.60	Proportion of mediator variance explained by individual-level covariates for the treatment condition	Proportion of coping variance among individuals explained by pretreatment variables
R2m2_Treatment	0.60	Proportion of mediator variance explained by cluster-level covariates for the treatment condition	Proportion of mediator variance providers explained by pretreatment variables
R2m_Control		Proportion of mediator variance explained by individual-level covariates for the control condition	Proportion of coping variance explained by pretreatment variables
n_treatment	6	Number of individuals per cluster	Number of individuals served by each provider
n_control		Total number of individuals in control condition	Total number of individuals in control condition
J_treatment	50	Number of clusters	Number of providers

1988). Holding other parameters constant, increases in the  $a$  coefficient tend to return higher power.

The second and third path coefficients are  $B$  and  $b_1$ . The  $B$  and  $b_1$  parameters describe the total and individual-level conditional association between the mediator and outcome. In our approach, we standardize the outcome values on the total standard deviation (individual- plus cluster-level) so that these coefficients are placed on the typical standardized regression coefficient scale (Cohen, 1988; Phelps, Kelcey, Liu, & Jones, 2016). Holding other parameters constant, increases in these coefficients produce higher power.

The third category of parameters captures the intraclass correlation coefficients (or variance decomposition) of the mediator ( $\rho_M$ ) and the outcome ( $\rho_Y$ ) in the treatment condition. More specifically, we can define intraclass correlation coefficients as

$$\rho_Y = \frac{\tau_{Y_1^{(o)}}^2}{(\tau_{Y_1^{(o)}}^2 + \sigma_{Y_1^{(o)}}^2)} \quad (5a)$$

$$\rho_M = \frac{\tau_{M_1^{(o)}}^2}{(\tau_{M_1^{(o)}}^2 + \sigma_{M_1^{(o)}}^2)} \quad (5b)$$

Holding constant other parameters, increases in the intraclass correlation coefficient for the outcome ( $\rho_Y$ ) tend to diminish power. That is, increased similarity among individuals within a providers will typically require a greater number of providers to reach a desired level of power (e.g., Raudenbush, 1997). In contrast, the role of the intraclass correlation coefficient for the mediator ( $\rho_M$ ) can serve to increase or decrease power depending on other parameter values (e.g., Beasley, 2014; Kelcey & Shen, 2016).

The fourth category of parameters governing power in this context is the variance explained in the mediator and outcome by covariates. For the outcome, we use  $R_{Y_{(o)}}^{L1}$  and  $R_{Y_{(o)}}^{L2}$  as the variance explained at the individual- and provider-levels by covariates under the treatment condition (e.g.,  $M, X, W, V$  in expression (3)). Increases in the outcome variance explained by covariates at either level improve power. This relationship suggest a design strategy that parallels study design for main and moderator effects—covariate adjusted designs that include variables predictive of the outcome can quickly reduce the sample size needed to achieve a certain power level (e.g., Kelcey & Phelps, 2013a, 2013b; Moss, Kelcey, & Showers, 2014; Raudenbush, Martinez, & Spybrook, 2007; Spybrook, Kelcey, & Dong, 2016).

For the mediator, we use  $R_{M_{(o)}}^{L1}$  and  $R_{M_{(o)}}^{L2}$  for the variance explained at the individual- and provider-

levels by covariates (e.g.,  $X, V, W$  in expression (1)) in the treatment condition. Further, we use  $R_{M_{(o)}}^{L1}$  as the individual-level variance explained by covariates under the control condition (e.g.,  $X, V$ , in expression (2)). Comparable to the role of the mediator intraclass correlation coefficient, the influence of the variance explained in the mediator by covariates on power depends heavily on the values of other parameters.

### Simulation

We assessed the accuracy of our results using a Monte Carlo simulation that compared the power and type one error rates predicted by our formulas with the observed power and type one error rates across 1000 draws. We outline the results from 33 different conditions with details provided in Table S1 in the Supplemental Material. Overall the simulation results substantiated our developments in that the formula-based predictions demonstrated a strong correspondence with the observed power and type one error rates. Moreover, the results replicated prior literature in suggesting that the Monte Carlo interval test tended to demonstrate the most power, followed closely by the joint test, and last the Sobel test.

### Illustration with a Two-/One-Level Study

Returning to our example, consider a study examining the extent to which the impact of the PSST on disruptive behavior operates through improvements in coping skills in children with oppositional defiant disorder. We are interested in planning a study using a two-/one-level partially nested design that randomly assigns individuals (level one) to a waitlist control condition or to participate in a PSST program delivered by a mental health provider (level two).

To implement the statistical power formulas developed above and determine a requisite sample size, we draw on the *R* shiny application for partial mediation designs (<https://poweruprshiny.shinyapps.io/PartiallyNestedMediationPower/>) in *PowerUpR* package (Dong, Reinke, Herman, Bradshaw, & Murray, 2016). We begin by specifying the anticipated parameter values based on, for example, pilot studies or published studies (e.g., Baldwin et al., 2011; David, David, & Dobrea, 2014). In our illustration, we anticipate the proportion of variance in children’s disruptive behavior (outcome) and coping skills (mediator) attributable to clustering across providers to be approximately 20%, (i.e.,  $\rho_{Y_{(o)}}^2 = \rho_{M_{(o)}}^2 = 0.20$ ).



Let us further assume that we plan to collect baseline measures of children's disruptive behavior that explains roughly 50% of the outcome variation at each level in the treated group ( $R_{Y(i)}^{2L2} = R_{Y(i)}^{2L1} = 0.50$ ) and 60% of the mediator variation at each level in the treatment and control groups ( $R_{M(i)}^{2L2} = R_{M(i)}^{2L1} = R_{M(i)}^{2L1} = 0.60$ ). We intend to standardize all variables except for the treatment indicator to have unit variance and expect participation in the program to improve coping skills (mediator) by roughly  $a = 0.40$  (standardized differences scale) and expect the  $B$  path coefficient connecting coping skills and disruptive behavior to be roughly  $B = 0.40$  (standardized regression coefficient scale). Based on the nature of the treatment, we expect that decomposition of the  $B$  path coefficient is  $b_1 = 0.40$  and  $b_2 = 0$ . That is, because the treatment is an individualized therapy, we suspect that the relationship between changes in coping skills and disruptive behavior is solely driven by individual changes in coping rather than group-based contextual changes.

If the mental health providers serve roughly 6 children each ( $n_{(t)}$ ), how large of a sample of mental health providers ( $\mathcal{J}_{(t)}$ ) and control waitlist children ( $n_{(c)}$ ) will we need to achieve a roughly 75% level of power to detect the mediation effect? The resulting power curve for each test of mediation is graphed in the Supplemental Material Figure S1 as a function of the number of mental health providers holding constant a waitlist sample of  $n_{(c)} = 50$ . With the most powerful test (Monte Carlo interval test), we would need to sample roughly  $\mathcal{J}_{(t)} = 50$  mental health providers to have almost a three-in-four chance of detecting the anticipated mediation effect. If instead we had access to a larger waitlist sample of  $n_{(c)} = 500$ , the number of mental health providers we would need to sample would be only slightly reduced to roughly  $n_{(t)} = 48$ . In contrast, reducing the waitlist sample to  $n_{(c)} = 25$  would necessitate a sample of about  $n_{(t)} = 70$  mental health providers.

### Three-Level Designs

Under the purview of three-level designs, we examined two types of partially nested structures that commonly arise in psychotherapy research. The first three-level type of study was again an individually-randomized design but one that now induces two degrees of clustering (three levels) for those treated based on the nature of the treatment (Figure 1b). This type of design arises when, for example, individuals (level one) are assigned at random to either participate in a pioneering type of group therapy (level two) led by a

group leader (level three) or act as a control case by remaining on a waitlist (unclustered). In these instances, treated individuals (level one) are nested within a therapy group (level two) and further nested within a group leader (level three). Like the two-level design outlined above, however, the individuals assigned to the waitlist control condition remain independent from such structures and nesting.

The third type of design we examined adopts the same type of three-level nesting structure for individuals assigned to the treatment condition but altered the nesting structure for those in the control condition. Specifically, we probed individually-randomized designs that induce two degrees of clustering (three levels) for those treated (same as previous design above) but now introduce an additional single degree of clustering (two levels) for those in the control condition (Figure 1b). This type of design arises when, for instance, individuals (level one) are assigned at random to either participate in a group therapy (level two) led by a group leader (level three) or to participate in a control condition that uses a type individual therapy that is led by a mental health provider (level two). In studies using this type of design, treatment individuals (level one) are nested within a therapy group (level two) and nested within a group leader (level three) whereas control individuals (level one) are nested only within mental health provider (level two) because therapy groups are not employed.

### Three-/One-Level Design

#### Models

When individual-randomization induces two degrees of clustering (three levels) for those in the treatment condition but no nesting for a comparison condition (e.g., waitlist; Figure 1b), we expand the prior two-level results to incorporate an additional level of nesting. Let us first consider a three-/one-level partially nested study designed to examine how the impact of the Stop Now and Plan (SNAP) training program (treatment) on irritability (outcome) is conveyed through the enhancement of emotion regulation skills (e.g., Derella, Johnston, Loeber, & Burke, 2019). The SNAP program is a group-format cognitive behavioral program that incorporates psychosocial treatment strategies through weekly small-group treatment sessions that include a variety of structured practice experiences (e.g., Burke & Loeber, 2015, 2016).

Past research has suggested the effectiveness of this program across both behavioral and affective symptom domains (e.g., Burke & Loeber, 2015, 2016). A key component of the SNAP program is

the direct discussion about and development of emotion regulation skills. Past research has hypothesized that the measurable development of emotion regulation skills acts as a core mediator of SNAP effects (Achenbach & Rescorla, 2001; Burke & Loeber, 2015, 2016). Recent research has pointed out, however, that there is an absence of work regarding if and how the SNAP program can improve youth irritability. More generally, researchers have noted that studies probing the value and role of emotion regulation in conveying the effects of SNAP on irritability may have important implications as to the key components of the program and pathways for improvement (Derella et al., 2019).

Because the SNAP program draws on small group sessions, we describe the study using a three-/one-level partially nested structure. The three-/one-level partially nested design randomly assigns individuals (level one of control condition) to a waitlist control condition (unclustered) or to participate in the SNAP program that nests individuals (level 1 of treatment condition) within small therapy groups (level 2 of treatment condition) that are led by a group leader (level three of treatment condition; see Figure 1b). Under this design, we can describe the formation of mediator values in the treatment arm as

$$\begin{aligned}
 M_{ijk} &= \pi_{0jk}^{(t)} + \pi_1^{(t)}(X_{ijk} - \bar{X}_{jk}) + \pi_2^{(t)}V_{ijk} + \varepsilon_{ijk}^{M(t)} \\
 \varepsilon_{ijk}^{M(t)} &\sim N(0, \sigma_{M_1^{(t)}}^2) \\
 \pi_{0jk}^{(t)} &= \zeta_{00k}^{(t)} + \zeta_1^{(t)}(\bar{X}_{jk} - \bar{X}_k) + \zeta_2^{(t)}(W_{jk} - \bar{W}_k) \\
 &\quad + \zeta_3^{(t)}Q_{jk} + u_{0jk}^{M(t)} \quad u_{0jk}^{M(t)} \sim N(0, \tau_{M_1}^2) \\
 \zeta_{00k}^{(t)} &= a^{(t)} + s_1^{(t)}\bar{X}_k + s_2^{(t)}\bar{W}_k + s_3^{(t)}Z_k + v_{00k}^{M(t)} \\
 v_{00k}^{M(t)} &\sim N(0, v_{M_1^{(t)}}^2)
 \end{aligned} \tag{6}$$

Here,  $M_{ijk}$  is used as the mediator (e.g., emotion regulation) value for individual  $i$  in group  $j$  led by group leader  $k$ ,  $X_{ijk}$  is an individual-level pretreatment covariate (e.g., baseline irritability) with  $\bar{X}_{jk}$  as its mean across all individuals within the same therapy group (e.g., average baseline irritability of individuals within a therapy group) and  $\bar{X}_k$  as its mean across all individuals served by the same group leader (e.g., average baseline irritability of individuals and all groups served by a leader),  $V_{ijk}$  is an individual-level pretreatment covariate that varies only across individuals (no variation across groups or group leaders),  $W_{jk}$  is a group-level pretreatment covariate with  $\bar{W}_k$  as its mean across all groups served by the same group leader,  $Q_{jk}$  is a group-level pretreatment covariate that varies only across

groups (no variation across group leaders),  $Z_k$  is a leader-level pretreatment covariate, and  $\varepsilon_{ijk}^{M(t)}$ ,  $u_{0jk}^{M(t)}$ , and  $v_{00k}^{M(t)}$  are the individual-, group-, and leader-level error terms for treatment condition. For the control arm we can retain the single-level model described in expression (2) above.

The associated model for individual outcomes under the treatment is

$$\begin{aligned}
 Y_{ijk} &= \beta_{0jk}^{(t)} + b_1^{(t)}(M_{ijk} - \bar{M}_{jk}) + \beta_1^{(t)}(X_{ijk} - \bar{X}_{jk}) \\
 &\quad + \beta_2^{(t)}V_{ijk} + \varepsilon_{ijk}^{Y(t)} \quad \varepsilon_{ijk}^{Y(t)} \sim N(0, \sigma_{Y_1^{(t)}}^2) \\
 \beta_{0jk}^{(t)} &= \gamma_{00k}^{(t)} + b_2^{(t)}(\bar{M}_{jk} - \bar{M}_k) + \gamma_{01}^{(t)}(\bar{X}_{jk} - \bar{X}_k) \\
 &\quad + \gamma_{02}^{(t)}(W_{jk} - \bar{W}_k) + \gamma_{03}^{(t)}Q_{jk} + u_{0jk}^{Y(t)} \\
 u_{0jk}^{Y(t)} &\sim N(0, \tau_{Y_1^{(t)}}^2) \\
 \gamma_{00k}^{(t)} &= \zeta_0^{(t)} + B^{(t)}\bar{M}_k + \zeta_1^{(t)}\bar{X}_k + \zeta_2^{(t)}\bar{W}_k + \zeta_3^{(t)}Z_k + v_{00k}^{Y(t)} \\
 v_{00k}^{Y(t)} &\sim N(0, v_{Y_1^{(t)}}^2)
 \end{aligned} \tag{7}$$

We use  $Y_{ijk}$  as the outcome for individual  $i$  in group  $j$  led by group leader  $k$ ,  $M_{ijk}$  as the mediator with  $\bar{M}_{jk}$  as its average mediator value for all individuals in the same therapy group and  $\bar{M}_k$  as its average for all individuals served by the same group leader. Further we use  $(M_{ijk} - \bar{M}_{jk})$  as the group-centered individual-level mediator with coefficient  $b_1$ ,  $(\bar{M}_{jk} - \bar{M}_k)$  as the leader-centered mediator with coefficient  $b_2$ ,  $\bar{M}_k$  as the average mediator value for leader  $k$  with coefficient  $B$ ,  $\zeta$ ,  $\gamma$ , and  $\beta$  as coefficients for other leader-, group- and individual-level covariates, and  $v_{00k}^{Y(t)}$ ,  $u_{0jk}^{Y(t)}$  and  $\varepsilon_{ijk}^{Y(t)}$  as the leader, group, and individual errors.

Similar to the two-level models, when using cluster-mean centering (e.g.,  $(M_{ijk} - \bar{M}_{jk})$  or  $(\bar{M}_{jk} - \bar{M}_k)$ ) for the mediator, the leader-level mediator-outcome coefficient ( $B$ ) captures the total association between the mediator and outcome (i.e.,  $B = b_1 + b_2 + b_3$ ). In this context,  $B$  thus captures the individual-level mediator-outcome association ( $b_1$ ; level one), group-level mediator-outcome association ( $b_2$ ; level two), and the leader-level mediator-outcome association ( $b_3$ ; level three). The mediation effect (ME) can then be estimated using expression (4) above.

Applied to our example, the difference between the treatment ( $a^{(t)}$ ) and control ( $a^{(c)}$ ) intercepts ( $a$ ) again captures how participation in the treatment improves emotion regulation skills. Likewise, the  $B$  coefficient describes how any changes in emotion regulation (i.e., at the individual-, group, or leader-level) are

associated with changes in irritability. More specifically,  $b_1$  captures how changes in an individual's emotion regulation skills (mediator) correlate with changes in irritability (outcome);  $b_2$  captures how changes in the average level of emotion regulation skills across individuals in a small therapy group correlate with changes on irritability; and  $b_3$  captures how changes in the average level of emotion regulation skills across individuals (and groups) served by a group leader correlate with changes on irritability. In our example, it is quite reasonable to suspect that there may be a contextual effect at the group-level (i.e.,  $b_2 > 0$ ) because group members directly interact in sessions and high levels of emotion regulation by group-mates may support the improvement of all in a group. In contrast, it is also likely that there will be less of a basis for such contextual effects at the leader level ( $b_3 = 0$ ) because individuals served by different small groups but the same leader have no interaction.

### Statistical Power

To track the statistical power with which three-level partially nested designs can detect mediation, we further extended the same three tests as developed for the two-level designs. We outline their developments and derivations in Section 4 of the Supplemental Material. The results of our derivations suggested that statistical power in three/one partially nested designs is again governed by four categories of parameters. The first category of parameters is the sample size and includes: (a) the number of group leaders in the treatment condition ( $K_{(t)}$ ), (b) the number of therapy groups served by each group leader in the treatment condition ( $\mathcal{J}_{(t)}$ ), (c) the number of individuals per each therapy group in the treatment condition ( $n_{(t)}$ ), and (d) the number of individuals in the control condition ( $n_{(c)}$ ). The roles of the different sample sizes across levels and conditions parallel those of the two-level designs. Many different types of sampling plans can be formulated to achieve a desired level of power. However, under many settings the number of group leaders in the treatment condition ( $K_{(t)}$ ) will have the most influence on power followed by the number of therapy groups served by each leader ( $\mathcal{J}_{(t)}$ ), the control condition sample ( $n_{(c)}$ ), and the number of individuals in each therapy group ( $n_{(t)}$ ).

The second category of parameters is again the path coefficients that detail the production of the mediation effect ( $a$ ,  $B$ ,  $b_2$  and  $b_1$ ). The role of these parameters is analogous to their two-level counterparts and, holding other parameters constant, increases return higher power.

The third category of parameters captures the intraclass correlation coefficients of the mediator and the outcome in the treatment condition. With three hierarchical levels, the results now describe the variance decomposition using intraclass correlation coefficients for individuals nested within therapy groups and therapy groups nested within leaders. For the outcome we have

$$\rho_{Y^{L3}} = \frac{v_{Y^{(t)}}^2}{(v_{Y^{(t)}}^2 + \tau_{Y^{(t)}}^2 + \sigma_{Y^{(t)}}^2)} \quad (8a)$$

$$\rho_{Y^{L2}} = \frac{\tau_{Y^{(t)}}^2}{(v_{Y^{(t)}}^2 + \tau_{Y^{(t)}}^2 + \sigma_{Y^{(t)}}^2)} \quad (8b)$$

For the mediator, we have

$$\rho_{M^{L3}} = \frac{v_{M^{(t)}}^2}{(v_{M^{(t)}}^2 + \tau_{M^{(t)}}^2 + \sigma_{M^{(t)}}^2)} \quad (9a)$$

$$\rho_{M^{L2}} = \frac{\tau_{M^{(t)}}^2}{(v_{M^{(t)}}^2 + \tau_{M^{(t)}}^2 + \sigma_{M^{(t)}}^2)} \quad (9b)$$

Holding constant other parameters, increases in either of the intraclass correlation coefficients for the outcome tend to diminish power. In contrast, the role of both intraclass correlation coefficients for the mediator can serve to increase or decrease power depending on other parameter values (e.g., Beasley, 2014; Kelcey & Shen, 2016).

The fourth category of parameters governing power in this context is the variance explained in the mediator and outcome by covariates. For the outcome, we use  $R_{Y^{(t)}}^{L1}$ ,  $R_{Y^{(t)}}^{L2}$  and  $R_{Y^{(t)}}^{L3}$  as the variance explained at the individual-, therapy group- and leader-levels by covariates under the treatment condition (e.g.,  $M$ ,  $X$ ,  $W$ ,  $V$ ,  $Q$ ,  $Z$  in expression (7)). Increases in the outcome variance explained by covariates at any level improve power. For the mediator, we use  $R_{M^{(t)}}^{L1}$ ,  $R_{M^{(t)}}^{L2}$  and  $R_{M^{(t)}}^{L3}$  for the variance explained at the individual-, therapy group-, and leader-levels by covariates expression (18) in the treatment condition.  $R_{M^{(t)}}^{L1}$  is the individual-level variance explained by covariates under the control condition (e.g.,  $X$ ,  $V$ , in expression (2)). Just like the two-level designs, the influence of the variance explained in the mediator by covariates on power depends heavily on the values of other parameters.

### Three-/Two-Level Design

#### Models

When individual-randomization induces two degrees of clustering (three levels) for those treated based on the nature of the treatment but only one degree of nesting in the comparison condition (e.g., group therapy led by group leaders versus individual therapy led by mental health provider; Figure 1c), we can expand on the previous three/one design to incorporate the additional level of nesting for the control group. For illustration, we continue with our previous example that examined examine the extent to which the impact of the Stop Now and Plan (SNAP) training program (treatment) on irritability (outcome) is conveyed through the enhancement of emotion regulation skills. However, in this design we change the structure of the control condition—it no longer consists of an unclustered wait-list but rather draws on an alternative approach that uses an individualized treatment. This example considers the design of a three-/two-level partially nested study that randomly assigns individuals (level one of control condition) to an individualized treatment control condition that is led by providers (level two of control condition) or to participate in the SNAP program that nests individuals (level 1 of treatment condition) within small therapy groups (level 2 of treatment condition) that are led by a group leader (level three of treatment condition; see Figure 1c).

In this setting we can specify the mediator model in the treatment arm as the three-level model detailed above in expression (18). For the control arm we have a two-level model

$$\begin{aligned}
 M_{ik} &= \pi_{0k}^{(c)} + \pi_1^{(c)}(X_{ik} - \bar{X}_k) + \pi_2^{(c)}V_{ik} + \varepsilon_{ik}^{M(c)} \\
 \varepsilon_{ik}^{M(c)} &\sim N(0, \sigma_{M_1}^{2(c)}) \\
 \pi_{0k}^{(c)} &= a^{(c)} + s_1^{(c)}\bar{X}_k + s_3^{(c)}Z_k + v_{00k}^{M(c)} \\
 v_{00k}^{M(c)} &\sim N(0, v_{M_1}^{2(c)})
 \end{aligned}
 \tag{10}$$

with  $\pi_{0k}^{(c)}$  as average mediator value for provider  $k$ ,  $X_{ik}$  as the individual-level covariate with  $\bar{X}_k$  as its average across all individuals served by provider  $k$  (with coefficient  $s_1^{(c)}$ ),  $V_{ik}$  as an individual-level covariate that varies only across individuals (no provider-level variance),  $a^{(c)}$  as the control conditional average mediator value across all providers,  $Z_k$  as a provider-level covariate with coefficient  $s_3^{(c)}$ , and  $v_{00k}^{M(c)}$  and  $\varepsilon_{ik}^{M(c)}$  as the errors for the respective levels.

The associated model for individual outcomes under the treatment does not change from the

three/one design above (see expression (19)). The TIE can be obtained using expression (4).

#### Statistical Power

We can leverage the same tests as before for the three/two designs (see Supplemental Material for outline). The results largely replicate those of the three/one design in that power is governed by the same four parameter categories. With the introduction of clustering in the control arm, however, we must add or replace several parameters. For the control condition sample size, we replace the total number of individuals in the control condition ( $n_{(c)}$ ) with the number of providers ( $J_{(c)}$ ) and the number of individuals per provider ( $n_{(c)}$ ). For the variance decompositions, we must now introduce a parameter quantifying the intraclass correlation coefficients of the mediator ( $\rho_{M_{(c)}^{L2}}$ ) such that

$$\rho_{M_{(c)}^{L2}} = \frac{\tau_{M_1}^{2(c)}}{(\tau_{M_1}^{2(c)} + \sigma_{M_1}^{2(c)})}
 \tag{11}$$

Last, we use  $R_{M_{(c)}^{L2}}^2$  as the variation in the mediator explained at the provider-level by covariates

#### Simulation

Similar to the two-level case, we appraised the accuracy of our formulas in terms of their ability to correctly predict the observed power and type one error rate across 1000 simulation draws. The simulation results for the three-level designs are provided in the Supplemental Material in Tables S2 and S3 for 33 conditions. The simulation results again substantiated our developments in that the formula-based predictions demonstrated a strong correspondence with the observed power and type one error rates. The correspondence between the formulas and simulation were strongest for the joint and Monte Carlo interval tests—the average (absolute) discrepancies were 0.02 (0.04) for the Sobel test, 0.01 (0.01) for the joint test and well less than 0.01 (<0.01) for the Monte Carlo interval test. Our findings here also suggested that the Monte Carlo interval and joint tests tended to be the most powerful.

#### Illustration with a Three-/Two-Level Study

We illustrate our results under three-level designs using the aforementioned three-/two-level example examining how emotion regulation (mediator) conveys the effects of the SNAP program (treatment)

on irritability (outcome). The design uses a three-/two-level partially nested structure that randomly assigns individuals (level one of control condition) to a control condition that involves individualized treatment led by a providers (level two of control condition) or to participate in the SNAP program that nests individuals (level 1 of treatment condition) within small therapy groups (level 2 of treatment condition) that are led by a group leader (level three of treatment condition; see Figure 1c)

To determine a reasonable sample size, we again used the *PowerUpR* Shiny application (<https://poweruprshiny.shinyapps.io/PartiallyNestedMediationPower/>). Let us presume that we will standardize all variables except for the treatment indicator to have unit variance and that we anticipate the proportion of variance in children's irritability (outcome) and emotion regulation skills (mediator) in the treatment condition attributable to clustering across therapy groups and group leaders is approximately 10% each (i.e.,  $v_{Y^{(t)}}^2 = \tau_{Y^{(t)}}^2 = v_{M^{(t)}}^2 = \tau_{M^{(t)}}^2 = 0.10$ ) while the remaining variance (80%) for both variables is attributable to children (i.e.,  $\sigma_{Y^{(t)}}^2 = \sigma_{M^{(t)}}^2 = 0.80$ ). Moreover, assume that the proportion of variance in emotion regulation skills (mediator) attributable to clustering across providers in the control condition is approximately 15% (i.e.,  $\rho_{M^{(c)}}^{L2} = 0.15$ ). Let us further assume that we plan to collect baseline measures of children's disruptive behavior that explains roughly 70% of the outcome variation at each level in the treated group ( $R_{Y^{(t)Z}}^{L3} = R_{Y^{(t)Z}}^{L2} = R_{Y^{(t)Z}}^{L1} = 0.70$ ) and 60% of the mediator variation at each level in the treatment and control groups ( $R_{M^{(t)}}^{L3} = R_{M^{(t)}}^{L2} = R_{M^{(t)}}^{L1} = R_{M^{(c)}}^{L2} = R_{M^{(c)}}^{L1} = 0.60$ ). We expect participation in the program to improve emotion regulation skills (mediator) by roughly  $a = 0.30$  (standardized differences scale) and expect the  $B$  path coefficient connecting emotion regulation skills and disruptive behavior to be roughly  $B = 0.50$  (standardized regression coefficient scale) with  $b_1 = 0.30$ ,  $b_2 = 0.10$ , and  $b_3 = 0.10$ .

If leaders in the treatment condition serve 2 small groups each ( $\mathcal{F}_{(t)}$ ) with roughly 6 children each ( $n_{(t)}$ ), while control condition providers serve roughly 6 children each ( $n_{(c)}$ ), how many treatment condition group leaders ( $K_{(t)}$ ) and control providers ( $\mathcal{F}_{(c)}$ ) will we need to achieve a roughly 75% level of power? Under the Monte Carlo interval test, we would need to sample roughly  $K_{(t)} = 26$  small group leaders in the treatment condition and  $\mathcal{F}_{(c)} = 26$  control condition providers to have a three-in-four chance of detecting the anticipated mediation effect.

## Discussion

Past research has emphasized the value of well-planned mediation studies and has detailed the sampling requirements for a variety of designs. In this way, expressions detailing power and requisite sample size emerge as important preconditions to the effective and efficient design of such studies. However, expressions and frameworks guiding the design of partially nested studies have lagged behind developments for other designs (Sterba, 2017). In this study, we broaden the tools available for planning partially nested studies by developing expressions and software that support their careful and judicious design when probing mediation.

A central challenge detailed in the literature on clustered and partially nested designs is the degree to which requisite sample sizes are reasonable under the typical study constraints. A detailed analysis of this line of inquiry across content areas is outside the purview of our study; however, the simulations and our illustrative case studies offer a preliminary (but limited) summary of the potential scale needed to produce sufficient levels of power under partially nested designs. Although any qualitative labeling of the magnitude of a sample size is fundamentally limited, our case studies and simulations suggest that high levels of power will often necessitate a fairly large number of clusters (e.g., providers). In our the two-/one-level illustration, even with a waitlist sample of 50 individuals we required upwards of 50 mental health providers serving 6 individuals each to maintain a 75% chance of detecting the effect.

There are multiple practical constraints and considerations in the design of any study, but one potentially useful lens through which we can judge the relative magnitude of the sample size required for detecting mediation effects is to compare it to the sample size required for main effects. Returning to our two-/one-level illustration with 50 waitlist individuals, if we assume that the main effect ( $C$ ) is on the order of  $C = 0.3$  (e.g., with  $C = ab + C' = 0.4 \times 0.4 + 0.14$  and  $c'$  as the direct effect of the treatment on the outcome), then the number of providers required for a three-in-four chance of detecting the main effect is about 70. That is, we would need about 70 providers each serving 6 individuals to detect a main effect of  $C = 0.3$  but only 50 providers to detect a mediation effect of  $aB = 0.16$ . These results suggests that in the sample size required for mediation effects will often be somewhat comparable to that required for main effects and in some instances less. More generally, the sample sizes needed for detecting mediation effects in partially

nested designs can be less, more, or about the same as that needed for the main effect.

Given the potentially large sampling requirements, an important next step is to develop design strategies that help to reduce the sampling burden. In our examples, we considered one such strategy—conditioning on baseline covariates. In most mediation analyses, controlling for pretreatment variables that may confound the mediator–outcome relationship will be required in order to control for preexisting differences that may confound that relationship. However, the inclusion of variables that are prognostic of the outcome and mediator may also serve to improve the power with which we can detect mediation effects. Returning to our two-/one-level example, if we eliminated covariates from our models, the number of providers required for a three-in-four chance of detecting the mediation effects would nearly double. This example suggests that covariance adjustment can be a useful strategy for reducing the requisite sample size—and in many cases this is a valuable strategy. However, the power to detect mediation effects is complicated by the composite nature of mediation effects (i.e., the product of two coefficients). Covariates that are predictive of the outcome will often yield gains in power but the advantages are typically tempered by the degree to which those covariates are predictive of the mediator because the error variance of the mediation effect is governed by the error variance of both coefficients (e.g., Beasley, 2014). More generally, an important implication of this study is that the effective and efficient use of partially nested designs to detect mediation effects will require the development of design strategies that help minimize requisite sample sizes.

### Supplemental data

Supplemental data for this article can be accessed here [10.1080/10503307.2020.1717012](https://doi.org/10.1080/10503307.2020.1717012).

### Funding

This work was supported by National Science Foundation [Grant Numbers 1760884 and 1552535].

### References

- Achenbach, T. M., & Rescorla, L. A. (2001). *Manual for the ASEBA school-age forms & profiles*. Burlington: University of Vermont, Research Center for Children, Youth, & Families.
- Aguinis, H., Edwards, J. R., & Bradley, K. J. (2017). Improving our understanding of moderation and mediation in strategic management research. *Organizational Research Methods*, 20(4), 665–685.
- Apsche, J. A., Bass, C. K., & Backlund, B. (2012). Mediation analysis of Mode Deactivation Therapy, (MDT). *The Behavior Analyst Today*, 13(2), 2–10.
- Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., ... Watson, J. (2011). Intraclass correlation associated with therapists: Estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy*, 40(1), 15–33.
- Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, 43(2), 210–236.
- Beasley, T. M. (2014). Tests of mediation: Paradoxical decline in statistical power as a function of mediator collinearity. *The Journal of Experimental Education*, 82(3), 283–306.
- Burke, J. D., & Loeber, R. (2015). The effectiveness of the stop now and Plan (SNAP) Program for boys at risk for violence and delinquency. *Prevention Science*, 16(2), 242–253.
- Burke, J. D., & Loeber, R. (2016). Mechanisms of behavioral and affective treatment outcomes in a cognitive behavioral intervention for boys. *Journal of Abnormal Child Psychology*, 44(1), 179–189.
- Candlish, J., Teare, M. D., Cohen, J., & Bywater, T. (2019). Statistical design and analysis in trials of proportionate interventions: A systematic review. *Trials*, 20(1), 151.
- Christensen, K. S., Haugen, W., Sirpal, M. K., & Haavet, O. R. (2015). Diagnosis of depressed young people—criterion validity of WHO-5 and HSCL-6 in Denmark and Norway. *Family Practice*, 32(3), 359–363. doi:10.1093/fampra/cmz011
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cox, K., & Kelcey, B. (2019). Optimal sample allocation in group-randomized mediation studies with a group-level mediator. *The Journal of Experimental Education*, 87(4), 616–640.
- David, O. A., David, D., & Dobrean, A. (2014). Efficacy of the rational positive parenting program for child externalizing behavior: Can an emotion-regulation enhanced cognitive-behavioral parent program be more effective than a standard one? *Journal of Evidence-Based Psychotherapies*, 14(2), 159.
- Derella, O. J., Johnston, O. G., Loeber, R., & Burke, J. D. (2019). CBT-enhanced emotion regulation as a mechanism of improvement for childhood irritability. *Journal of Clinical Child & Adolescent Psychology*, 48(Suppl. 1), S146–S154.
- Dong, N., Reinke, W. M., Herman, K. C., Bradshaw, C. P., & Murray, D. W. (2016). Meaningful effect sizes, intraclass correlations, and proportions of variance explained by covariates for planning two- and three-level cluster randomized trials of social and behavioral outcomes. *Evaluation Review*, 40(4), 334–377.
- Dunsmore, J. C., Booker, J. A., & Ollendick, T. H. (2013). Parental emotion coaching and child emotion regulation as protective factors for children with oppositional defiant disorder. *Social Development*, 22(3), 444–466.
- Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, 4, 91–122.
- Eden, D., Stone-Romero, E. F., & Rothstein, H. R. (2015). Synthesizing results of multiple randomized experiments to establish causality in mediation testing. *Human Resource Management Review*, 25(4), 342–351.
- Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, 12(2), 121–138.
- Fritz, M. S., & MacKinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, 18(3), 233–239.
- Gelfand, L. A., Mensinger, J. L., & Tenhave, T. (2009). Mediation analysis: A retrospective snapshot of practice and more recent directions. *The Journal of General Psychology*, 136(2), 153–178.

- Gottfredson, D. C., Cook, T. D., Gardner, F. E., Gorman-Smith, D., Howe, G. W., Sandler, I. N., & Zafft, K. M. (2015). Standards of evidence for efficacy, effectiveness, and scale-up research in prevention science: Next generation. *Prevention Science, 16*(7), 893–926.
- Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: Does method really matter? *Psychological Science, 24* (10), 1918–1927.
- Hinshaw, S. P. (2007). Moderators and mediators of treatment outcome for youth with ADHD: Understanding for whom and how interventions work. *Journal of Pediatric Psychology, 32* (6), 664–675.
- Kazdin, A. E. (2006). Arbitrary metrics: Implications for identifying evidence-based treatments. *American Psychologist, 61*(1), 42–49.
- Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology, 3*, 1–27.
- Kazdin, A. E. (2010). Problem-solving skills training and parent management training for opposition a defiant disorder and conduct disorder. In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence-based psychotherapies for children and adolescents* (pp. 211–226). New York: The Guilford Press.
- Kazdin, A. E., & Nock, M. K. (2003). Delineating mechanisms of change in child and adolescent therapy: Methodological issues and research recommendations. *Journal of Child Psychology and Psychiatry, 44*(8), 1116–1129.
- 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*(5), 499–530.
- Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Experimental power for indirect effects in group-randomized studies with group-level mediators. *Multivariate Behavioral Research, 52*(6), 699–719.
- Kelcey, B., Hill, H., & Chin, M. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: A multilevel mediation analysis. *School Effectiveness & School Improvement, 30*(4), 398–431.
- Kelcey, B., & Phelps, G. (2013a). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis, 35*(3), 370–390.
- Kelcey, B., & Phelps, G. (2013b). Strategies for improving power in school-randomized studies of professional development. *Evaluation Review, 37*(6), 520–554.
- Kelcey, B., & Shen, Z. (2016). Multilevel design of school effectiveness studies in sub-Saharan Africa. *School Effectiveness and School Improvement, 27*(4), 492–510.
- Kelcey, B., Spybrook, J., & Dong, N. (2019). Sample size planning in cluster-randomized studies of multilevel mediation. *Prevention Science, 20*, 407–418.
- Kisbu-Sakarya, Y., MacKinnon, D. P., & Miočević, M. (2014). The distribution of the product explains normal theory mediation confidence interval estimation. *Multivariate Behavioral Research, 49*(3), 261–268.
- 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.
- Lindsey, M., Brandt, N., Becker, K., Lee, B., Barth, R., Daleiden, E., & Chorpita, B. (2014). Identifying the common elements of treatment engagement interventions in children’s mental health services. *Clinical Child and Family Psychology Review, 17*(3), 283–298.
- Liu, J. C., Chang, L. Y., Wu, S. Y., & Tsai, P. S. (2015). Resilience mediates the relationship between depression and psychological health status in patients with heart failure: A cross-sectional study. *International Journal of Nursing Studies, 52*(12), 1846–1853. doi:10.1016/j.ijnurstu.2015.07.005
- Lochman, J. E., Evans, S. C., Burke, J. D., Roberts, M. C., Fite, P. J., Reed, G. M., ... Elena Garralda, M. (2015). An empirically based alternative to DSM-5’s disruptive mood dysregulation disorder for ICD-11. *World Psychiatry, 14*(1), 30–33.
- MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*(1), 83–104.
- Moss, B., Kelcey, B., & Showers, N. (2014). Classrooms as moderators of development education effectiveness and student achievement. *Community College Review, 42*, 201–220.
- Nohe, C., Michaelis, B., Menges, J. I., Zhang, Z., & Sonntag, K. (2013). Charisma and organizational change: A multilevel study of perceived charisma, commitment to change, and team performance. *The Leadership Quarterly, 24*(2), 378–389.
- Phelps, G., Kelcey, B., Liu, S., & Jones, N. (2016). Informing estimates of program effects for studies of mathematics professional development using teacher content knowledge outcomes. *Evaluation Review, 40*, 383–409.
- Pituch, K. A., & Stapleton, L. M. (2012). Distinguishing between cross- and cluster-level mediation processes in the cluster randomized trial. *Sociological Methods & Research, 41*, 630–670.
- Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures, 6*(2), 77–98.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*(2), 173.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). New York: Sage.
- Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis, 29*, 5–29.
- Roberts, C., & Roberts, S. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials, 2* (2), 152–162.
- Scott, K. (2018). *Identifying the mechanisms of change and in-session therapist fidelity in measurement based care for depression* (Unpublished Doctoral Dissertation). Retrieved from <http://hdl.handle.net/2022/22500>
- Scott, K., & Lewis, C. C. (2015). Using measurement-based care to enhance any treatment. *Cognitive and Behavioral Practice, 22* (1), 49–59.
- Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290–312.
- Spybrook, J., Kelcey, B., & Dong, N. (2016). Power analyses for detecting treatment by moderator effects in cluster randomized trials. *Journal of Educational and Behavioral Statistics, 41*, 605–627.
- Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research, 27*(4), 425–436.
- Suldo, S. M., Parker, J. S., Shaunessy-Dedrick, E., & O’Brennan, L. M. (2019). Mental health interventions. In Jennifer A. Fredricks, Amy L. Reschly, & Sandra L. Christenson (Eds.), *Handbook of student engagement interventions* (pp. 199–215). San Diego: Academic Press.
- VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research, 38*, 515–544.
- Windgassen, S., Goldsmith, K., Moss-Morris, R., & Chalder, T. (2016). Establishing how psychological therapies work: The importance of mediation analysis.