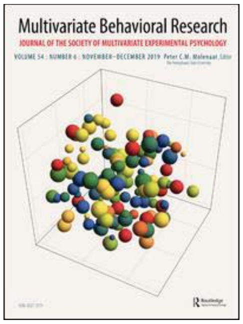

Multivariate Behavioral Research

ISSN: 0027-3171 (Print) 1532-7906 (Online) Journal homepage: https://www.tandfonline.com/loi/hmbr20

Power and Sample Size Determination for Multilevel Mediation in Three-Level Cluster-Randomized Trials

Ben Kelcey, Yanli Xie, Jessaca Spybrook & Nianbo Dong

MULTIVARIATE BEHAVIORAL RESEARCH
https://doi.org/10.1080/00273171.2020.1738910

Routledge
Taylor & Francis Group

Power and Sample Size Determination for Multilevel Mediation in Three-Level Cluster-Randomized Trials

Ben Kelcey[a], Yanli Xie[a], Jessaca Spybrook[b], and Nianbo Dong[c]

[a]College of Education, Criminal Justice, Human Services and Information Technology, University of Cincinnati; [b]College of Education, Criminal Justice, Human Services and Information Technology, Western Michigan University; [c]College of Education, University of North Carolina Chapel Hill

ABSTRACT
Mediation analyses supply a principal lens to probe the pathways through which a treatment acts upon an outcome because they can dismantle and test the core components of treatments and test how these components function as a coordinated system or theory of action. Experimental evaluation of mediation effects in addition to total effects has become increasingly common but literature has developed only limited guidance on how to plan mediation studies with multi-tiered hierarchical or clustered structures. In this study, we provide methods for computing the power to detect mediation effects in three-level cluster-randomized designs that examine individual- (level one), intermediate- (level two) or cluster-level (level three) mediators. We assess the methods using a simulation and provide examples of a three-level clinic-randomized study (individuals nested within therapists nested within clinics) probing an individual-, intermediate- or cluster-level mediator using the *R* package *PowerUpR* and its *Shiny* application.

KEYWORDS
Mediation; power; indirect effects; multilevel models; sample size determination; experimental design

Mediation analyses supply a principal lens to probe the pathways through which a treatment acts upon an outcome because they can dismantle and test the core components of treatments and test how these components function as a coordinated system or theory of action (Gottfredson et al., 2015; Kazdin, 2007; Windgassen et al., 2016). Across a broad range of disciplines, mediation analyses now represent a core standard of evidence in establishing treatment effects and how these effects take root while advancing foundational theories within a discipline (e.g., Eden, 2017; Imai et al., 2013; MacKinnon, 2008; Wilt, 2012; Windgassen et al., 2016). For example, the field of organizational research has established guidelines designed to promote the careful planning and incorporation of mediators in studies and the Society for Prevention Research has introduced specific standards that advocate for and guide mediational analyses (e.g., Aguinis et al., 2017; Gottfredson et al., 2015; Eden et al., 2015).

In many disciplines, mediation guidelines often include considerations for the multilevel structures that typically arise in social and psychological settings in order to examine individual and group behaviors amidst the flow of a treatment effect to an outcome (e.g., Krull & MacKinnon, 1999). Multilevel mediation methods have been applied to a wide variety of disciplines and topics including, for example, those focused on psychological well-being (Van Mierlo, Rutte, Vermunt, Kompier, & Doorewaard, 2007), mental health (Kozlowski & Klein, 2000), neighborhood effects (e.g., VanderWeele, 2010), stress and coping processes (Brincks et al., 2010), early childhood education (Curenton et al., 2015), youth delinquency interventions (e.g., Brown et al., 2014), provider-patient communication (e.g., Cegala & Post, 2009), and teacher effects (e.g., Kelcey & Carlisle, 2013; Kelcey et al., 2019).

Similarly, these types of multilevel considerations have expanded to accommodate more complicated structures that involve three levels of nesting. Prior research has, for example, drawn on three-level structures to investigate neighborhood effects using children nested within families nested within neighborhoods (e.g., National Longitudinal Survey of Children & Youth & Statistics Canada, 2019), anxiety and depression within the context of students in classes in schools (Kozina, 2018a; 2018b; VanderWeele et al., 2013), health outcomes with

CONTACT Ben Kelcey ben.kelcey@gmail.com College of Education, Criminal Justice, Human Services and Information Technology, University of Cincinnati, Cincinnati, OH, USA.

individuals nested in sites nested within regions or states (Livert et al., 2001), and organizational research questions addressing leadership effects within the context of employees in teams in organizations (e.g., Černe et al., 2013; Xu et al., 2017).

Although a diverse range of disciplines routinely call for and draw upon multilevel mediation analyses, a key gap in this literature is the development of expressions and tools that detail the statistical power of multilevel designs to detect mediation effects and identify sample sizes that provide a desired level of power. Prior research has widely documented the foundational value of establishing guidelines and accessible planning tools in improving study design, the quality of evidence garnered from a study, and the capacity of a discipline (e.g., Gottfredson et al., 2015; Spybrook, Shi, et al., 2016; Spybrook, Kelcey, et al., 2016; Windgassen et al., 2016). From a study planning perspective, however, literature provides little guidance or accessible tools regarding how to understand and estimate the statistical power with which hierarchical designs can detect mediation effects and the scale needed to ensure a desired level of power. Recent literature has probed and detailed these considerations for two-level designs (e.g., Kelcey, Dong, et al., 2017), however, such planning considerations for comparable three-level designs have not been well-studied and delineated in a similar manner despite the increasing prevalence of three-level designs (e.g., Hedges & Hedberg, 2013; Jacob et al., 2010; Kelcey et al., 2017; Pituch & Stapleton, 2008; Pituch et al., 2006; Spybrook et al., 2016).

In this study, we developed expressions and software to estimate the statistical power with which three-level cluster-randomized designs can detect mediation effects using cluster-, intermediate-, or individual-level mediators. The remainder of our study is broken up into five additional sections. The next section outlines a working illustration to give a more concrete context to our formulations, structures and expressions. We then use three sections that are separated by the level of the mediator (i.e., cluster-, intermediate-, and individual-level mediator) to outline the resulting power expressions, illustrate the results using the *PowerUpR* Shiny application (Bulus et al., 2019), and outline the accuracy of the expressions through a simulation. We end with a discussion.

Working example

To frame our subsequent analyses, we first develop an illustrative example through which we examine three types of mediation (defined by level of the mediator) in three-level experiments. Our example context focuses on studies investigating the effective use of evidence-based psychological treatments on patient mental health outcomes. Recent reviews have suggested that the number of people facing mental disorders continues to rise (e.g., Harvey & Gumport, 2015). Although research has established a number of effective evidence-based psychological treatments for a broad range of mental disorders (e.g., Layard & Clark, 2014), this same research has also indicated that the majority of people suffering from mental disorders are not receiving appropriate or effective care (e.g., SAMSHA, 2007).

Recent analyses have in part sought to identify critical barriers that constrain the effective use of evidence-based psychological treatments and have identified several key hurdles that stagnate the effective use of evidence-based treatments across multiple structural or organizational levels (Harvey & Gumport, 2015). Barriers arise along multiple levels including the patient-level (e.g., adherence, motivation, engagement), therapist-level (e.g., beliefs, preferences, cognitive biases, fidelity), the clinic-level (e.g., leader skepticism, lack of training time), and government-level (e.g., policies; Harvey & Gumport, 2015).

In part, these findings call for increased research into implementation practices that better provision the take up and use of evidence-based therapies to improve patient outcomes (e.g., Harvey & Gumport, 2015; Schwartz, 2010). In our working examples, we considered three different types of programs that provide supportive training directed toward improvements in clinic-, therapist-, or patient-level processes. The first type of program we considered included those aimed at supporting key clinic-level processes to reduce barriers to implementation. In our context, one example is the level of administrative support provisioned by a clinic leader (e.g., Roche & Freeman, 2004; Willenbring et al., 2004). Prior literature has found that material support by administrative leaders (e.g., in service training, dedicated preparation time for staff, expert consultation, group discussions) can afford or constrain an organization's take up of a practice and the eventual impact of the training on patient outcomes (e.g., Willenbring et al., 2004).

The second type of program we considered was those providing material support for therapists (level two). In clinical settings, therapists are often the principal delivery agents of psychological treatments and research suggests that, for example, their beliefs about a treatment serve as key intermediate mechanisms in the pathway to improved patient outcomes (Roche & Freeman, 2004). That is, clinical theories frequently submit that effective use of a new treatment is preceded by transformations in beliefs regarding the clinical utility and validity of that approach (e.g., Sikorski et al., 2012).

The third type of program we considered included those that scaffold key patient processes thought to strengthen outcomes. One common example is the degree of patient motivation or treatment favoring behavior cultivated in therapy sessions (e.g., Keeley et al., 2016). Prior research in several areas including depression suggests that weak outcomes owe in part to patient resistance to or apathy toward evidence-based treatments (e.g., Keeley et al., 2016; Van Voorhees et al., 2005). Clinical training programs often focus on strategies such as motivational interviewing to nurture and promote patient motivation (e.g., Keeley et al., 2016).

Within this context, our example analyses examined a cluster- or clinic-randomized study designed to evaluate how and the extent to which exposure to a training program on effective structural or behavioral supports of psychological therapies improves patient quality of life (e.g., Mowbray et al., 2009; Slade et al., 2015). We examined experimental designs that nest patients (level one) within therapists (level two) within clinics (level three) and randomly assign clinics to participate in a training program or a control condition. In turn, we draw on multilevel mediation analyses to examine the extent to which participation in the training (clinic-level treatment) affects patient quality of life (patient-level outcome) by acting on different types of intermediate variables (clinic-, therapist- or patient-level mediator).

The first type of mediator we considered was a clinic-level (or level three) that describes variation across clinics and is constant across therapists and patients within clinics (e.g., level of administrative support provisioned by a clinic leader). When tracking the clinic-level pathways through which a treatment impacts an outcome, the resulting three-level mediation design is often described as 3-3-1 mediation. More specifically, the acronym is used to denote that the treatment is assigned at level three (clinic-level), the mediator captures differences among level three units (clinic-level), while the outcome is assessed at level one (patient-level). In 3-3-1 mediation, our analyses track the influence of the clinical training on a patient-level outcome as it operates through a clinic-level mediator.

The second type of mediator we examined was a therapist-level variable (level two) such as therapists' beliefs about the clinical utility of a treatment. When probing a therapist-level mediator in a three-level design, literature has described the analysis as a type of 3-2-1 mediation because such studies investigate how a clinic-level treatment (level three) impacts a therapist-level (level two) mediator in ways that advance a patient-level outcome (level one). Although therapist beliefs ostensibly target variation among therapists, the collection of beliefs at a clinic also holds the potential to describe the larger social context and capital of a clinic in ways that may shape patient outcomes (Aarons et al., 2012). Past research has demonstrated that organizational social context including the collective beliefs of therapists toward a psychological treatment can buttress or undermine the facility with which therapists take up and implement the treatment (Aarons et al., 2012). In this way, many multilevel analyses purposefully probe the collective beliefs of therapists as supplementary vehicles for change.

The most widespread summary of collective beliefs in the clinical context is the clinic-level average of therapists' beliefs (e.g., Brincks et al., 2017; Pituch & Stapleton, 2012; VanderWeele et al., 2013). Inclusion of the collective beliefs allows researchers to track the total association between beliefs and patient outcomes—that is, it tracks both the association between collective (clinic-level) beliefs and patient outcomes (i.e., the contextual relationship) and the association between individual therapist deviations from their clinic averages and patient outcomes (i.e., the individual-level relationship).

Last, we examined patient-level intermediate variables (level one) that delineate differences among patients (e.g., motivation). Literature has commonly described this type of analysis as a 3-1-1 mediation model because the analyses study how the effect of a clinic-level treatment (level three) impacts a patient-level (level one) mediator in ways that advance a patient-level outcome (level one). Like therapist-level mediators, our 3-1-1 multilevel mediation analysis considered aggregates of the patient-level mediators to summarize the environment at the therapist- and clinic-levels that emerges by a collection of patients. For example, when therapy involves multiple patients (e.g., group-based therapy), the average motivation of patients within a therapist can be used to describe the social context. A high level of motivation across all patients served by a therapist in a group session may have a galvanizing effect on each of the patients in terms of strengthening their commitment to the treatment and eventually positively impact outcomes. Similar aggregate processes are included at the clinic-level. For example, persistently high levels of motivation across all patients in a clinic may set expectations for patient investment and commitment that introduce positive contextual effects on patient outcomes.

Below, we examine each type of mediator in turn (i.e., clinic-, therapist-, and then patient-level mediator). Each section introduces the analytic models, outlines key results for the error variances and the power expressions for several mediation tests, assesses these expressions using simulations, and then ends with an illustrative application.

Clinic-level mediators

We begin by examining the power with which clinic- or cluster-randomized studies can detect indirect effects operating through clinic-level mediators (i.e., 3-3-1 mediation). More specifically, we examine experiments that plan to delegate clinics at random to participate in either a control or treatment condition (T) and evaluate the treatment's effect on a patient-level outcome through a clinic-level mediator (M). Within this context, we draw on the typical set of linear multilevel models to estimate model parameters (e.g., Pituch et al., 2009; Pituch & Stapleton, 2012; VanderWeele et al., 2013; VanderWeele & Vansteelandt, 2009; Zhang et al., 2009). We structure our model to employ group-mean centered variables (i.e., centered within context) as this approach is widespread across disciplines due to its ability to decompose and deconfound effects at different levels (e.g., Brincks et al., 2017; Pituch & Stapleton, 2012; Zhang et al., 2009). However, other methods including no centering or grand-mean centering will produce equivalent parameter estimates when drawing on random intercept models with the hierarchical group means entered as fixed effects at each appropriate level (Brincks et al., 2017; Enders & Tofighi, 2007; Kreft et al., 1995).

For the 3-3-1 analyses, we first delineate the mediator as a function of the treatment and covariates such that

$$\begin{aligned} M_k &= \zeta_{00} + aT_k + \zeta_{01}\bar{X}_k + \zeta_{02}\bar{W}_k + \zeta_{03}Z_k + \varepsilon_k^M \\ &\quad \varepsilon_k^M \sim N(0, \sigma^2_{M|\boldsymbol{\theta}}) \end{aligned} \tag{1}$$

Here, M_k is used as the mediator value for clinic k, $\bar{X}_k$ as its clinic-level mean aggregate of a patient-level covariate (X) with ζ_{01} as its path coefficient, $\bar{W}_k$ as its clinic-level mean aggregate of therapist covariate (W) with ζ_{02} as its path coefficient, T_k as the treatment assignment coded as ±½ with path coefficient a, Z_k as a clinic-level covariate (with ζ_{03} as its path coefficient), ε_k^M as the clinic-level error term with variance $\sigma^2_{M|\boldsymbol{\theta}}$ that is conditional on the fixed effects (collectively denoted as $\boldsymbol{\theta}$).

Applied to our clinical example, the mediation model examines the extent to which a clinic's participation in a training program influences a clinic's level of administrative support offered to its therapists and patients. Covariates in the model may include clinic-level variables such as clinic policies, clinic-level aggregates of practitioner-level variables such as average prior experience of practitioners, and clinic -level aggregates of patient-level variables such as average therapy history.

The use of an experimental design balances the observed and unobserved characteristics of the clinics, practitioners, and patients between the treatment and control conditions such that it alleviates the potential for confounding for the a path—that is, the possibility that the observed treatment-mediator association is due to an alternative explanation rather than the effect of the training program. However, many design plans call for covariance adjustment on covariates that are predictive of the mediator because it can improve the statistical precision with which we can track the treatment-mediator relationship (a path) and detect the mediation effect.

The associated model for patient outcomes is

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_1(X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon_{ijk}^Y \\ &\quad \varepsilon_{ijk}^Y \sim N(0, \sigma^2_{Y|\boldsymbol{\theta}}) \\ \beta_{0jk} &= \gamma_{00k} + \gamma_{01}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{02}(W_{jk} - \bar{W}_k) + \gamma_{03}Q_{jk} + u_{0jk}^Y \\ &\quad u_{0jk}^Y \sim N(0, \tau^2_{Y|\boldsymbol{\theta}}) \\ \gamma_{00k} &= \zeta_0 + BM_k + c'T_k + \xi_1\bar{X}_k + \xi_2\bar{W}_k + \xi_3 Z_k + \upsilon_{00k}^Y \\ &\quad \upsilon_{00k}^Y \sim N(0, \nu^2_{Y|\boldsymbol{\theta}}) \end{aligned} \tag{2}$$

Here, we use Y_{ijk} as the outcome for patient i served by therapist j in clinic k, X_{ijk} as a patient-level covariate (with β_1 as its path coefficient), V_{ijk} as a patient-level covariate that retains variance only among patients within therapists (no variation across therapists or clinics) with β_2 as its path coefficient, M_k as the mediator for clinic k with path coefficient B, c' as the treatment-outcome conditional path coefficient, γ as coefficients for therapist-level covariates, ξ as coefficients for clinic-level covariates, and υ_{00k}^Y, u_{0jk}^Y and ε_{ijk}^Y as the clinic, therapist, and patient error terms with respective variances $\nu^2_{Y|\boldsymbol{\theta}}$, $\tau^2_{Y|\boldsymbol{\theta}}$, and $\sigma^2_{Y|\boldsymbol{\theta}}$ that are conditional on the fixed effects (collectively denoted as $\boldsymbol{\theta}$). We can further augment this model to incorporate a treatment by mediator interaction (e.g., adding a $T\bar{M}_k$ term) if we anticipate participation in the training to strengthen or weaken the relationship between the mediator and outcome. For ease, we detail the expression and analyses without this additional effect but including it is a straightforward addition (Kelcey et al., 2017).

Returning to our clinical illustration, the outcome model describes how changes in administrative support are conditionally associated with patient outcomes. Similar to the mediation model, the outcome model conditions on clinic, therapist, and patient covariates. The use of covariates in the outcome model is critical in order to obtain an unbiased estimate of the mediator-outcome path coefficient because randomization of clinics to treatment conditions does not address the potential for confounding in the mediator-outcome path (i.e., sequential ignorability).

When assumptions regarding causal identification and the model specification (Vander Weele, 2010) are correct, the 3-3-1 multilevel mediation effect (ME) is quantified as the product of the treatment-mediator (a) and mediator-outcome (B) paths $ME_{331} = aB$.[1] This indirect or mediation effect delineates the impact of the treatment on the outcome as it works through changes in a clinic-level mediator. Placed within our working example, the mediation effect describes how changes in administrative support brought about by participation in a training program produce improvements in a patient outcome.

Standardizing the outcome and mediator to have a total unconditional variance of one (i.e., $\sigma^2_M = 1$, $\upsilon^2_Y + \tau^2_Y + \sigma^2_Y = \rho_{Y^{L3}} + \rho_{Y^{L2}} + (1 - \rho_{Y^{L3}} - \rho_{Y^{L2}}) = 1$), the clinic- and therapist-level components of the outcome variance can be viewed as the respective proportion of variance owing to each level or the variance partitioning coefficients ($\rho_{Y^{L3}}, \rho_{Y^{L2}}$). As a result, the scale of the a (treatment-mediator) path and c' (direct effect of the treatment on the outcome) path can be understood on a standardized mean difference scale while the B (mediator-outcome) path can be placed on a standardized regression coefficient for a clinic-level variable.

Error variance

Our analyses focus on three tests that can be used in the planning phases before data are collected: (a) the Sobel test, (b) the joint test, and (c) the Monte Carlo interval test. The power associated with each of these tests is governed by the error variances (i.e., the square of the standard error) of the treatment-mediator and mediator-outcome path coefficients. For this reason, we first outline the resulting error variances and then turn to the power of each test.

Our derivations show that the error variance of the treatment-mediator path (σ^2_a) is

$$\sigma^2_a = \frac{\sigma^2_{M|\boldsymbol{\theta}}}{(n_3 - C_M - 1)p(1-p)} = \frac{\sigma^2_M(1 - R^2_{M^{L3}})}{(n_3 - C_M - 1)p(1-p)} \tag{3}$$

with C_M as the number of level three covariates (see Supplemental Material for outline). We use σ^2_M as the unconditional clinic-level variance of the mediator, p as the proportion of clinics assigned to the treatment condition, $R^2_{M^{L3}}$ as the clinic-level mediator variance explained by predictors (i.e., T, $\bar{X}$, $\bar{W}$, and Z) in the mediator model (expression 1), and n_3 as the clinic-level sample size. To link the expression with the path coefficients that are primary to the mediation effect, we can further specify the clinic-level mediator variance explained ($R^2_{M^{L3}}$) as

$$R^2_{M^{L3}} = R^2_{M^{L3}_{\vec{Z}}} + \frac{p(1-p)a^2}{\sigma^2_M} \tag{4}$$

Here, we additionally introduce $R^2_{M^{L3}_{\vec{Z}}}$ as the variance explained by covariates (i.e., the variance explained by the covariates $\bar{X}$, $\bar{W}$, and Z when not conditioning on or including the treatment indicator T). Values for the predictive capacity of the covariates ($R^2_{M^{L3}_{\vec{Z}}}$) can often be drawn from empirical literature documenting the correlations between common covariates and outcomes (e.g., Baldwin et al., 2011; Cosby et al., 2003).

For the mediator-outcome path coefficient, we can track the error variance (σ^2_B) using

$$\begin{aligned}\sigma^2_B &= \frac{\nu^2_{Y|\boldsymbol{\theta}} + \tau^2_{Y|\boldsymbol{\theta}}/n_2 + \sigma^2_{Y|\boldsymbol{\theta}}/(n_2 n_1)}{(n_3 - C_Y - 1)\sigma^2_{M|\boldsymbol{\theta}}} \\ &= \frac{\nu^2_Y(1 - R^2_{Y^{L3}}) + \tau^2_Y(1 - R^2_{Y^{L2}})/n_2 + (1 - R^2_{Y^{L1}})\sigma^2_Y/(n_2 n_1)}{(n_3 - C_Y - 1)\sigma^2_M(1 - R^2_{M^{L3}})}\end{aligned} \tag{5}$$

Here we use υ^2_Y, τ^2_Y, and σ^2_Y to represent the unconditional variances for the outcome at the clinic-, therapist- and patient-levels, $R^2_{M^{L3}}$ to represent the clinic-level mediator variance explained by all other predictors (i.e., $T, \bar{X}, \bar{W}, Z$), and $R^2_{Y^{L3}}$, $R^2_{Y^{L2}}$, and $R^2_{Y^{L1}}$ to represent the outcome explained by predictors at the clinic- (M, T, $\bar{X}_k$, $\bar{W}_k$ and Z), therapist-

[1] Core assumptions include (a) stable unit treatment value assumption, (b) sequential ignorability, (c) consistency, (d) no downstream confounders, and (e) no treatment-by-mediator interaction.

$((\bar{X}_{jk} - \bar{X}_k), (W_{jk} - \bar{W}_k),$ and $Q)$, and patient-level $((X_{ijk} - \bar{X}_{jk})$ and $V)$ and C_Y as the number of clinic-level covariates in the outcome $(C_Y = 5)$. We further detail $R^2_{Y^{L3}}$ in order to tie it to the core parameters defining the effects as (e.g., Kelcey et al., 2017)

$$R^2_{Y^{L3}} = R^2_{Y^{L3}\vec{Z}} + \frac{p(1-p)(ab+c')^2}{\nu_Y^2} + \frac{\sigma_M^2}{\nu_Y^2} B^2 (1 - \frac{p(1-p)a^2}{\sigma_M^2} - R^{L3}_{M_{\vec{z}}} 2) \qquad (6)$$

with $R^2_{Y^{L3}\vec{Z}}$ as the outcome variance explained by covariates (i.e., $\bar{X}_k$, $\bar{W}_k$ and Z) and $R^2_{M^{L3}_{\vec{Z}}}$ mediator variance explained by covariates (i.e., $\bar{X}, \bar{W},$ and Z).

Statistical power

Having outlined the error variance of the two principal path coefficients that quantify mediation, we extend three test statistics and develop expressions to track their power. We take up two asymptotic-based tests and one resampling-based test that can be readily used in the design phase before data collection: (a) the Sobel test, (b) the joint test, and (c) the Monte Carlo interval test.

Sobel test

Our first test is the classic Sobel test that evaluates the ratio of the mediation effect to its (asymptotic) standard error (Sobel, 1982)

$$z^{Sobel}_{aB} = aB/\sigma_{aB} \qquad (7)$$

with σ_{aB} as the asymptotic standard error of the mediation effect such that

$$\sigma^2_{ab} = \sigma^2_a b^2 + a^2 \sigma^2_b + \sigma^2_a \sigma^2_b \qquad (8)$$

Hypothesis tests using the Sobel test are developed by relating the test statistic to a normal distribution on the asymptotic theory that the maximum likelihood estimates are normal with the centrality parameter equal to the mediation effect and dispersion parameter equal to the square root of expression (8) above (Sobel, 1982). The statistical power of the test can be estimated as

$$P(|z^{Sobel}_{aB}| > z_{critical}) = 1 - \Phi(z_{critical} - z^{Sobel}_{aB}) + \Phi(-z_{critical} - z^{Sobel}_{aB}) \qquad (9)$$

with z^{Sobel}_{aB} as the test statistic in expression (7), Φ as the cumulative density of the normal distribution, and $z_{critical}$ as a chosen critical value such as 1.96 from the normal distribution that captures a pre-specified type one error rate.

Joint test

Although the normal distribution used in the Sobel test serves as a convenient reference distribution for larger multilevel samples (e.g., 100 clusters), it does not typically capture the distribution of the null in studies using small to moderate sample sizes (Kisbu-Sakarya et al., 2014). A popular alternative is the joint test (e.g., Cohen & Cohen, 1983; MacKinnon et al., 2002). The joint test develops inferences using sub-tests that target the path coefficients that compose a mediation effect—a sub-test for the intervention-mediator path coefficient and, separately, a sub-test for the mediator-outcome path coefficient. When employing the joint test, the null hypothesis of no indirect effect is rejected only when both sub-tests are rejected. Past research has shown that the joint test yields levels of power and type one error rates that rival more complicated approaches such as resampling-based bootstrap tests (Hayes & Scharkow, 2013).

For the intervention-mediator path coefficient (a), we form the test statistic

$$t_a = a/\sigma_a \qquad (10)$$

where σ_a is the standard error in expression (3). For the mediator-outcome path coefficient we have

$$t_B = B/\sigma_B \qquad (11)$$

For each of these sub-tests we use a referent t-distribution with degrees of freedom equal to n_3-C-1 where C is the number of level three or clinic-level predictors in the model (Raudenbush & Bryk, 2002; Kenny & Judd, 2014).

The power of two-sided tests to detect the indirect effect (i.e., both paths simultaneously nonzero) is the product of the power to detect the intervention-mediator path and the power to detect the corresponding mediator-outcome path

$$P(|t_a| > t_{critical} \& |t_B| > t_{critical}) = (1 - t(t_{critical} - t_a) + t(-t_{critical} - t_a)) * (1 - t(t_{critical} - t_B) + t(-t_{critical} - t_B)) \qquad (12)$$

where t is the appropriate cumulative t density function and $t_{critical}$ is the critical value for n_3-C-1 degrees of freedom.

Monte Carlo interval test

Last, we considered the resampling-based Monte Carlo interval test (Preacher & Selig, 2012). For this test, samples are drawn for the treatment-mediator and the mediator-outcome path coefficients from normal distributions centered on their estimated values ($\hat{a}, \hat{B}$) with variances set to their expected error variances based on the aforementioned expressions. Collectively, the product of a^* and B^* from these samples approximate the sampling distribution of the indirect effect (Preacher & Selig, 2012)

$$\begin{pmatrix} a^* \\ B^* \end{pmatrix} \sim \boldsymbol{t}_{n_3-C-1}\left[\begin{pmatrix} \hat{a} \\ \hat{B} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\hat{a}} & \hat{\sigma}_{\hat{a},\hat{B}} \\ \hat{\sigma}_{\hat{a},\hat{B}} & \hat{\sigma}^2_{\hat{B}} \end{pmatrix}\right] \qquad (13)$$

Subsequently, we develop inferences regarding the indirect effect by assessing whether the simulated asymmetric confidence intervals leave out zero. As a result, the statistical power of this test is estimated as the proportion of confidence intervals that leave out zero. Recent literature has also presented a fairly sizeable body of results that have supported the efficacy of this approach (e.g., Kelcey et al., 2017; Preacher & Selig, 2012).

Simulation

To assess the accuracy of our results, we drew on Monte Carlo simulations that compare the simulated type one error and power rates for detecting the indirect effect with our formula-based power predictions. We drew on 1000 simulated datasets based on the models above and considered 35 different conditions that varied the clinic-, therapist-, and patient-level sample sizes and the size of each of the path coefficients. The results are outlined in the Supplemental Material and are subsequently summarized using the 3-1-1 mediation (see Table 1 below).

Collectively, the proposed formulas predicted the power of the three tests well across the conditions. The potential exception is that the Sobel test can become inaccurate for extremely small samples (e.g., 10 clinics, 3 therapists/clinic and 3 patients/therapist). For type one error rates, all three tests demonstrated rates lower than the nominal level but these rates were well predicted under the Monte Carlo formulas, moderately predicted for the joint test and poorly predicted by the Sobel test. Overall, the results strongly supported the accuracy of the formulas when predicting power, even when the number of clinics, therapists per clinic and patients per therapist were relatively small.

Table 1. 3-1-1 simulation results.

Sample sizes			Path coefficients				Power or Type 1 Error				
n_3	n_2	n_1	a	B	c'	Sobel	$\widehat{\text{Sobel}}$	Joint	$\widehat{\text{Joint}}$	MC	$\widehat{\text{MC}}$
Power											
10	3	3	0.8	0.5	0.1	0.03	0.13	0.06	0.02	0.05	0.07
20	3	3	0.8	0.5	0.1	0.26	0.34	0.43	0.34	0.42	0.40
30	3	3	0.8	0.5	0.1	0.58	0.53	0.72	0.62	0.71	0.67
40	6	6	0.5	0.3	0.1	0.35	0.28	0.38	0.41	0.41	0.41
80	6	6	0.5	0.3	0.1	0.64	0.67	0.72	0.72	0.73	0.71
40	12	6	0.5	0.3	0.1	0.36	0.32	0.39	0.42	0.41	0.42
80	12	6	0.5	0.3	0.1	0.65	0.67	0.72	0.71	0.73	0.71
40	6	12	0.5	0.3	0.1	0.36	0.30	0.39	0.43	0.41	0.43
80	6	12	0.5	0.3	0.1	0.64	0.66	0.71	0.70	0.72	0.69
40	12	12	0.5	0.3	0.1	0.35	0.31	0.38	0.41	0.41	0.41
80	12	12	0.5	0.3	0.1	0.66	0.70	0.73	0.74	0.73	0.74
Type 1 Error											
40	6	6	0	0.3	0.1	0.05	0.00	0.02	0.00	0.00	0.00
40	6	6	0.5	0	0.1	0.05	0.00	0.05	0.01	0.02	0.00
40	6	6	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
80	6	6	0	0.3	0.1	0.06	0.00	0.04	0.02	0.03	0.02
80	6	6	0.5	0	0.1	0.05	0.03	0.05	0.04	0.02	0.04
80	6	6	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
40	12	6	0	0.3	0.1	0.06	0.00	0.03	0.02	0.00	0.02
40	12	6	0.5	0	0.1	0.05	0.01	0.05	0.03	0.04	0.03
40	12	6	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
80	12	6	0	0.3	0.1	0.05	0.00	0.04	0.02	0.02	0.02
80	12	6	0.5	0	0.1	0.05	0.02	0.05	0.02	0.02	0.02
80	12	6	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
40	6	12	0	0.3	0.1	0.05	0.00	0.02	0.02	0.00	0.01
40	6	12	0.5	0	0.1	0.05	0.00	0.05	0.01	0.02	0.01
40	6	12	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
80	6	12	0	0.3	0.1	0.05	0.01	0.04	0.06	0.02	0.06
80	6	12	0.5	0	0.1	0.05	0.00	0.05	0.01	0.03	0.01
80	6	12	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
40	12	12	0	0.3	0.1	0.05	0.00	0.02	0.02	0.01	0.02
40	12	12	0.5	0	0.1	0.05	0.00	0.05	0.00	0.02	0.01
40	12	12	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00
80	12	12	0	0.3	0.1	0.05	0.01	0.04	0.01	0.01	0.01
80	12	12	0.5	0	0.1	0.05	0.01	0.05	0.01	0.03	0.01
80	12	12	0	0	0.1	0.05	0.00	0.00	0.00	0.00	0.00

Note: a circumflex or hat indicates predictions made by formulas developed in this study.

Illustration

Let us now consider a hypothetical study that targets a clinic-level mediator (i.e., 3-3-1 mediation). Our example context focuses on planning a clinic-randomized study to assess the impact of a psychological treatment training program (treatment) on patient quality of life (outcome) as it is facilitated through administrative support (mediator). To apply the power expressions and identify a requisite sample size, we draw on empirical estimates of the governing parameters such as those pulled from pilot studies or prior research in the field (e.g., Baldwin et al., 2011; Cosby et al., 2003). In our example, we presume that based on pilot data the variance decomposition of the patient quality of life outcome is about 50% owing to clinics, 20% owing to therapists, and 30% owing to patients. Our design also calls for covariates at each level (e.g., patient history, therapist training, clinic location and population served) such that we expect that the covariates will explain about 75% of

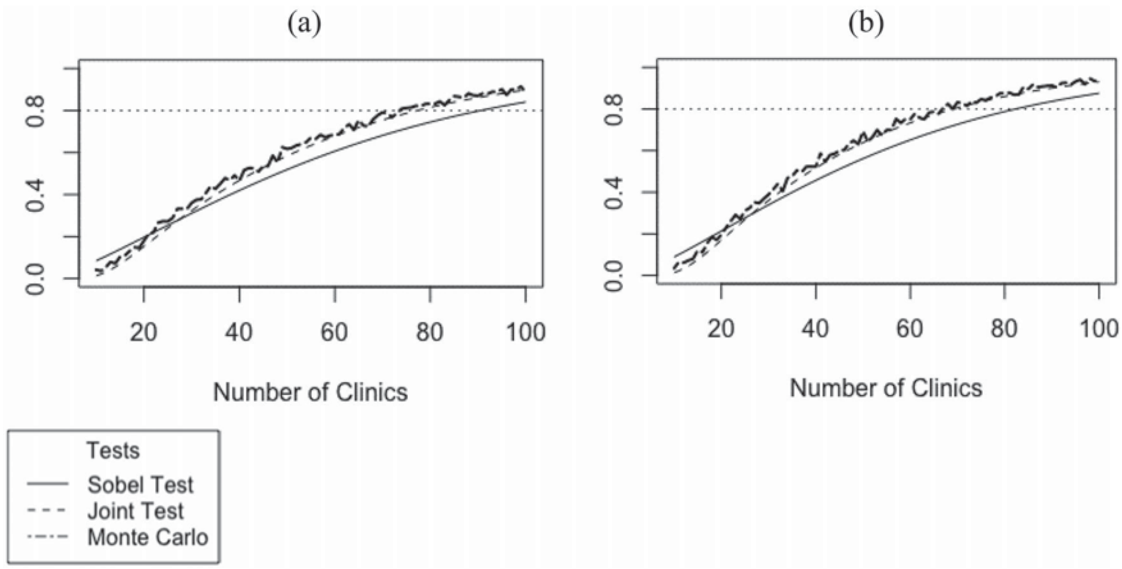

Figure 1. Power as a function of the number of clinics by test for a 3-3-1 mediation effect of 0.24 when (a) the number of patients per therapist (n1) is 5 and the number of therapists per clinic (n2) is 2 and (b) the number of patients per therapist (n1) is 10 and the number of therapists per clinic (n2) is 10.

the outcome variance at the patient- and therapist-level but only about 50% at the clinic-levels. Similarly, for the mediator, covariates are expected to explain 75% of the variance. Presume that prior research on this treatment suggests that the total effect of the training on patient quality of life is about 0.35 (standardized differences across conditions) with the participation in the training shifting administrative support (mediator) by $a = 0.60$ (standardized differences across conditions) and the conditional association between administrative support and patient quality of life (standardized regression coefficient scale) as $B = 0.40$ such that the indirect effect is $0.6 \times 0.40 = 0.24$. Further, assume that the direct effect of the treatment on patient quality of life is 0.11 such that a simple decomposition of the total effect (c) is $0.35 = 0.6 \times 0.4 + 0.11$ ($c = aB + c'$). That is, let $a = 0.6$, $B = 0.40$ $c' = 0.11$ $R^2_{Y_z^{L3}} = 0.5$, $R^2_{Y_z^{L2}} = R^2_{Y_z^{L1}} = R^2_{M_z^{L3}} = 0.75$. If we sample 5 patients per therapist (n_1), 2 therapists per clinic (n_2), how many clinics do we need to achieve roughly 80% power?

To estimate the number of clinics needed, we used the *R* package *PowerUpR* and the associated *Shiny* application. We plot statistical power of the Sobel, joint, and Monte Carlo interval tests for the indirect effect against the clinic-level sample size (n_3) in Figure 1a. With the Monte Carlo and joint tests approximately 78 clinics yield an 80% level of power whereas the Sobel test would require 90 clinics to produce a similar level of power.

It can be instructive to concurrently consider complementary effects and alternative sampling schemes. For instance, under this example the sample size needed for an 80% chance of detecting the total effect is approximately 75 clinics. Similarly, if we are able to inflate the sample of patients per therapist to 10 (n_1) and the number of therapists per clinic to 10 (n_2), we would need about 66 clinics to achieve roughly 80% power (Figure 1b).

Therapist-level mediators

We next examine the power with which clinic-randomized studies can detect indirect effects operating through an intermediate- or therapist-level (level two) mediators (i.e., 3-2-1 mediation; see Figure 1b). We modify the previous models (expressions (1) and (2)) to allow for a therapist-level (level two) mediator such that (e.g., Pituch et al., 2009)

$$
\begin{aligned}
M_{jk} &= \pi_{0k} + \pi_1(\bar{X}_{jk} - \bar{X}_k) + \pi_2(W_{jk} - \bar{W}_k) + \pi_3 Q_{jk} + \varepsilon^M_{jk} \qquad \varepsilon^M_{jk} \sim N(0, \sigma^2_{M|\boldsymbol{\theta}}) \\
\pi_{0k} &= \zeta_{00} + aT_k + \zeta_{01}\bar{X}_k + \zeta_{02}\bar{W}_k + \zeta_{03}Z_k + u^M_{0k} \qquad u^M_{0k} \sim N(0, \tau^2_{M|\boldsymbol{\theta}})
\end{aligned}
\tag{14}
$$

We use M_{jk} as the mediator value for therapist j in clinic k, $\bar{X}_{jk}$ as a therapist-level mean of a patient-level covariate (with π_1 as its path coefficient) and $\bar{X}_k$ as its clinic-level average (with ζ_{01} as its path coefficient), W_{jk} as a therapist-level covariate (with π_2 as its path coefficient) and $\bar{W}_k$ as its clinic-level average (with ζ_{02} as its path coefficient), Q_{jk} as a therapist-level variable that varies only among therapist within clinics (no variation among clinics) with π_3 as its path coefficient,T_k as the treatment assignment coded as ±½ with associated path coefficient a, Z_k as a clinic-level covariate (with ζ_{03} as its path coefficient), and ε^M_{jk} and u^M_{0k} as the therapist-, and clinic-level errors with variances $\sigma^2_{M|\boldsymbol{\theta}}$ and $\tau^2_{M|\boldsymbol{\theta}}$ that are conditional on the fixed effects ($\boldsymbol{\theta}$). We can further adjust the outcome model such that

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + \beta_1(X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon^Y_{ijk} \qquad \varepsilon^Y_{ijk} \sim N(0, \sigma^2_{Y|\boldsymbol{\theta}}) \\ \beta_{0jk} &= \gamma_{00k} + b_2(M_{jk} - \bar{M}_k) + \gamma_{01}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{02}(W_{jk} - \bar{W}_k) + \gamma_{03}Q_{jk} + u^Y_{0jk} \qquad u^Y_{0jk} \sim N(0, \tau^2_{Y|\boldsymbol{\theta}}) \\ \gamma_{00k} &= \zeta_0 + B\bar{M}_k + c'T_k + \xi_1\bar{X}_k + \xi_2\bar{W}_k + \xi_3 Z_k + \upsilon^Y_{00k} \qquad \upsilon^Y_{00k} \sim N(0, \nu^2_{Y|\boldsymbol{\theta}}) \end{aligned} \tag{15}$$

Y_{ijk} is employed as the outcome for patient i in therapist j in clinic k, X_{ijk} as a patient-level covariate (with β_1 as its path coefficient), V_{ijk} as a patient-level covariate that only varies across patients within therapists (no variation across therapists or clinics) with β_2 as its path coefficient, $M_{jk} - \bar{M}_k$ as the clinic-centered therapist-level mediator with coefficient b_2, $\bar{M}_k$ as the average of the mediator in clinic k with path coefficient B, c' as the treatment-outcome conditional path coefficient, γ as coefficients for therapist-level covariates, ξ as coefficients for clinic-level covariates, and υ^Y_{00k}, u^Y_{0jk} and ε^Y_{ijk} as the clinic, therapist, and patient error terms with variances $\nu^2_{Y|\boldsymbol{\theta}}$, $\tau^2_{Y|\boldsymbol{\theta}}$, and $\sigma^2_{Y|\boldsymbol{\theta}}$ conditional on the fixed effects ($\boldsymbol{\theta}$). By using group-mean centering for the mediator, the clinic-level mediator-outcome coefficient (B) captures the between or total association between the mediator and outcome.

When assumptions are met, the 3-2-1 mediation effect (ME) is summarized as the product of the treatment-mediator (a) path coefficient and clinic-level mediator-outcome path coefficient (B) paths: $ME_{321} = aB$. This mediation effect describes the impact of the treatment on a patient outcome as it works through the collective changes in therapist beliefs.

Error variance & power

For the a path, derivations suggest the error variance (σ^2_a) can be estimated as (see Supplemental Material for outline)

$$\begin{aligned} \sigma^2_a &= \left(\frac{\tau^2_{M|\boldsymbol{\theta}} + \sigma^2_{M|\boldsymbol{\theta}}/n_2}{(n_3 - C_M - 1)p(1-p)} \right) \\ &= \left(\frac{\tau^2_M(1 - R^2_{M^{L3}}) + (1 - R^2_{M^{L2}})\sigma^2_M/n_2}{(n_3 - C_M - 1)p(1-p)} \right) \end{aligned} \tag{16}$$

In expression (16), we use τ^2_M and σ^2_M as the unconditional clinic- and therapist-level mediator variances, p as the proportion of clinics assigned to the treatment condition, $R^2_{M^{L3}}$ and $R^2_{M^{L2}}$ as the clinic- and therapist-level mediator variance explained by predictors in the mediator model (expression 14), and n_3 and n_2 capture the clinic- and therapist-level sample sizes with C_M as the number of clinic-level covariates in the mediator model ($C_M = 4$ here). We can further detail the variance explained in the mediator. At the therapist-level the mediator variance explained ($R^2_{M^{L2}}$) simply reduces to the variance explained by covariates ($R^2_{M^{L2}} = R^2_{M^{L2}_{\bar{Z}}}$). At the clinic-level the variance explained ($R^2_{M^{L3}}$) can be approximated as

$$R^2_{M^{L3}} = R^2_{M^{L3}_{\bar{Z}}} + \frac{a^2 p(1-p)}{\tau^2_M} \tag{17}$$

For error variance of the B path, similar analyses provide the following result

$$\begin{aligned} \sigma^2_B &= \frac{\nu^2_{Y|\boldsymbol{\theta}} + \tau^2_{Y|\boldsymbol{\theta}}/n_2 + \sigma^2_{Y|\boldsymbol{\theta}}/(n_2 n_1)}{(n_3 - C_Y - 1)(\tau^2_{M|\boldsymbol{\theta}} + \sigma^2_{M|\boldsymbol{\theta}}/n_2)} \\ &= \frac{\nu^2_Y(1 - R^2_{Y^{L3}}) + \tau^2_Y(1 - R^2_{Y^{L2}})/n_2 + (1 - R^2_{Y^{L1}})\sigma^2_Y/(n_2 n_1)}{(n_3 - C_Y - 1)(\tau^2_M(1 - R^2_{M^{L3}}) + (1 - R^2_{M^{L2}})\sigma^2_M/n_2)} \end{aligned} \tag{18}$$

For the B path, the expression tracking the error variance additionally draws on, τ^2_M, and σ^2_M as the unconditional mediator variances at the clinic- and therapist- levels and $R^2_{M^{L3}}$ as the variance in the clinic-level average mediator ($\bar{M}$) explained by the other clinic-level predictors in the outcome model (i.e., $T, \bar{X}, \bar{W}, Z$) and $R^2_{M^{L2}} = R^2_{M^{L2}_{\vec{Z}}}$ as therapist-level variance explained in the (centered) mediator by just the covariates (i.e., $(\bar{X}_{jk} - \bar{X}_k), (W_{jk} - \bar{W}_k), Q$).

We can additionally decompose the variance explained at each level in terms of the principal path coefficients. For the clinic-level outcome variance explained ($R^2_{Y^{L3}}$) we have

$$R^2_{Y^{L3}} = R^2_{Y^{L3}\vec{Z}} + \frac{p(1-p)(aB+c')^2}{\nu_Y^2} + \frac{\tau_M^2 + \sigma_M^2/n_2}{\nu_Y^2} B^2 \left(1 - \frac{p(1-p)a^2}{\tau_M^2} - R^2_{M^{L3}_{\vec{Z}}}\right) \tag{19}$$

For the therapist-level outcome variance explained ($R^2_{Y^{L2}}$) we can approximate it as

$$R^2_{Y^{L2}} = R^2_{Y^{L2}_{\vec{z}}} + \left(\frac{\sigma_M^2}{\tau_Y^2}\right) b_2^2 (1 - R^2_{\tilde{M}^{L2}_{\vec{z}}}) \tag{20}$$

where $\tilde{M}$ captures the clinic-mean centered therapist values of the mediator ($\tilde{M} = M_{jk} - \bar{M}_k$) and $R^2_{Y^{L2}_{\vec{z}}}$ represents the total therapist-level outcome variance explained by covariates. Last, the outcome variance explained at the individual-level ($R^2_{Y^{L1}}$) reduces to just the variance explained by covariates (i.e., $R^2_{Y^{L1}} = R^2_{Y^{L1}_{\vec{Z}}}$).

Standardizing the variance so that $\tau_M^2 + \sigma_M^2 = \rho_M + (1 - \rho_M) = 1$ and $\nu_Y^2 + \tau_Y^2 + \sigma_Y^2 = \rho_{Y_3} + \rho_{Y_2} + (1 - \rho_{Y_3} - \rho_{Y_2}) = 1$) places the paths on the standardized mean difference scale (for *a* and *c'*) and a standardized regression coefficient scale (*B*).

With these results, the statistical power of our three tests can be tracked in a manner analogous to the 3-3-1 analyses. We can simply substitute the abovementioned error variances for the paths under the 3-2-1 design for those in the power formulas for the 3-3-1 design.

Simulation

Similar to the 3-3-1 case, we assessed the accuracy of our results using a comparable Monte Carlo simulation. The results are outlined in the Supplemental Material. The power and type one error levels of the joint and the Monte Carlo tests were well predicted by our formulas. However, predictions based on the normal distribution for the Sobel test performed poorly. Collectively, the results strongly supported the use of the formulas for the joint and Monte Carlo tests to identify requisite sample sizes and plan studies but much less so for the Sobel test formulas.

Illustration

Now take on the design of a study that plans to use a clinic-randomized design to unpack the impact of a training program (treatment) on patient quality of life (outcome) as it operates through a therapist-level mediator such as therapist beliefs in the value and validity of the treatment. Drawing on our pilot data, we expect that the variance decomposition of the patient quality of life outcome is about 20% owing to clinics, 50% owing to therapists, and the remaining 30% owing to patients. In addition, we must now also consider the variance decomposition of the mediator—in this example assume we anticipate that therapist beliefs on the treatment are substantially clustered within clinics such that the variance owing to clinic differences in beliefs is about 40% while the remaining 60% owes to differences among therapists within clinics. We still expect covariates explain about 75% of the outcome variance at the patient- and therapist-level but only about 50% at the clinic-levels. However, for the beliefs mediator, we now anticipate that only 40% of the variance at each level will be explained by covariates. If we retain the same effects sizes, we have: $a = 0.6$, $B = 0.40$ $c' = 0.11$, $R^2_{Y^{L3}_z} = 0.5, R^2_{Y^{L2}_z} = R^2_{Y^{L1}_z} = 0.75, R^2_{M^{L3}_z} = R^2_{M^{L2}_z} = 0.40$. If we sample 5 patients per therapist (n_1), 2 therapists per clinic (n_2), how many clinics do we need to achieve a roughly 80% power?

The power curves produced by our formulas and *PowerUpR* are displayed in Figure 2. The analyses indicated that the Monte Carlo interval test tended to provide the most power and required only 35 clinics for an 80% chance of detecting the indirect effect. The joint test was a close second and required two additional clinics whereas the Sobel test required 49 clinics. For the total or main effect, an 80% level power was produced with about 46 clinics.

A critical prediction in this and other examples is the predictive capacity of the covariates. Recent empirical research has begun to compile plausible empirical values of such parameters for an increasing range of outcomes and mediators (e.g., Hedges & Hedberg, 2007; Phelps, et al., 2016). If the variances explained in the mediator and outcome by covariates were reduced, this would likely require many more clinics. For example, if the variances explained in the mediator and outcome by covariates were severely reduced to, for example, 25% at each level (i.e., $R^2_{Y^{L3}_z} = R^2_{Y^{L2}_z} = R^2_{Y^{L1}_z} = R^2_{M^{L3}_z} = R^2_{M^{L2}_z} = 0.25$), 80% power would only be reached once we sampled about 50 clinics under the joint and Monte Carlo intervals tests (Figure 2).

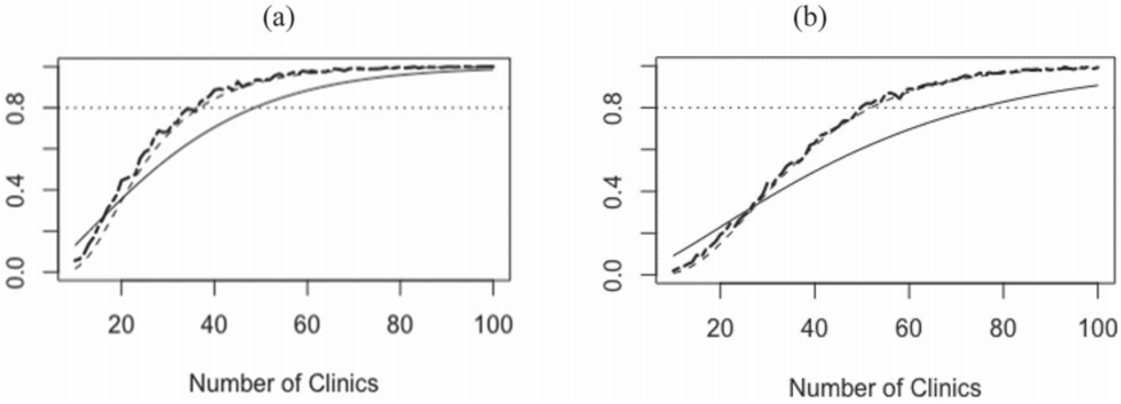

Figure 2. Power as a function of the number of clinics by test for a 3-2-1 mediation effect of 0.24 when the number of patients per therapist (n1) is 5 and the number of therapists per clinic (n2) is 2 and (a) $R^2_{Y_z^{L3}} = 0.5, R^2_{Y_z^{L2}} = R^2_{Y_z^{L1}} = 0.75, R^2_{M_z^{L3}} = R^2_{M_z^{L2}} = 0.40$ and (b) $R^2_{Y_z^{L3}} = R^2_{Y_z^{L2}} = R^2_{Y_z^{L1}} = R^2_{M_z^{L3}} = R^2_{M_z^{L2}} = 0.25$.

Patient-level mediators

Last, we consider a study design with a patient-level mediator (i.e., 3-1-1 mediation). We begin by first updating our mediator and outcome models to incorporate the patient-level mediation. The mediator model for the 3-1-1 design becomes

$$\begin{aligned} M_{ijk} &= \pi_{0jk} + \pi_1(X_{ijk} - \bar{X}_{jk}) + \pi_2 V_{ijk} + \varepsilon^M_{ijk} \\ &\quad \varepsilon^M_{ijk} \sim N(0, \sigma^2_{M|\boldsymbol{\theta}}) \\ \pi_{0jk} &= \zeta_{00k} + \zeta_1(\bar{X}_{jk} - \bar{X}_k) + \zeta_2(W_{jk} - \bar{W}_k) + \zeta_3 Q_{jk} + u^M_{0jk} \\ &\quad u^M_{0jk} \sim N(0, \tau^2_{M|\boldsymbol{\theta}}) \\ \zeta_{00k} &= \varsigma_0 + aT_k + \varsigma_1 \bar{X}_k + \varsigma_2 \bar{W}_k + \varsigma_3 Z_k + \upsilon^M_{00k} \\ &\quad \upsilon^M_{00k} \sim N(0, \nu^2_{M|\boldsymbol{\theta}}) \end{aligned} \tag{21}$$

In this extension, we use M_{ijk} as the mediator value for patient i served by therapist j in clinic k, X_{ijk} as a patient-level covariate (with π_1 as its path coefficient), $\bar{X}_{jk}$ as a therapist-level mean of a patient-level covariate and $\bar{X}_k$ as its clinic-level average (with ς_1 as its path coefficient), V_{ijk} as a patient-level covariate that only varies across patients within therapists (no variation among therapists or clinics) with π_2 as its path coefficient, W_{jk} as a therapist-level covariate (with ζ_2 as its path coefficient) and $\bar{W}_k$ as its clinic-level average (with ς_2 as its path coefficient), Q_{jk} as a therapist-level variable that varies only across therapists within clinics (no variation across clinics) with ζ_3 as its path coefficient, T_k as the treatment indicator with coefficient a, Z_k as a clinic-level covariate (with ς_3 as its path coefficient), and ε^M_{ijk} as the patient-level error term, u^M_{jk} as the therapist-level random effects, and υ^M_k as the clinic-level random effects with respective variances $\sigma^2_{M|\boldsymbol{\theta}}$, $\tau^2_{M|\boldsymbol{\theta}}$, and $\nu^2_{M|\boldsymbol{\theta}}$ that are conditional on the fixed effects ($\boldsymbol{\theta}$). Similarly, we can modify the outcome model to become

$$\begin{aligned} Y_{ijk} &= \beta_{0jk} + b_1(M_{ijk} - \bar{M}_{jk}) + \beta_1(X_{ijk} - \bar{X}_{jk}) + \beta_2 V_{ijk} + \varepsilon^Y_{ijk} \\ &\quad \varepsilon^Y_{ijk} \sim N(0, \sigma^2_{Y|\boldsymbol{\theta}}) \\ \beta_{0jk} &= \gamma_{00k} + b_2(\bar{M}_{jk} - \bar{M}_k) + \gamma_{01}(\bar{X}_{jk} - \bar{X}_k) + \gamma_{02}(W_{jk} - \bar{W}_k) \\ &\quad + \gamma_{03} Q_{jk} + u^Y_{0jk} \\ &\quad u^Y_{0jk} \sim N(0, \tau^2_{Y|\boldsymbol{\theta}}) \\ \gamma_{00k} &= \zeta_0 + B\bar{M}_k + c'T_k + \xi_1 \bar{X}_k + \xi_2 \bar{W}_k + \xi_3 Z_k + \upsilon^Y_{00k} \\ &\quad \upsilon^Y_{00k} \sim N(0, \nu^2_{Y|\boldsymbol{\theta}}) \end{aligned} \tag{22}$$

We use Y_{ijk} as the outcome for patient i in therapist j in clinic k, $M_{ijk} - \bar{M}_{jk}$ as the therapist-centered patient-level mediator with coefficient b_1,$\bar{M}_{jk} - \bar{M}_k$ as the clinic-centered therapy-level mediator with coefficient b_2, $\bar{M}_k$ as the average mediator value in clinic k with coefficient B, c' as the direct relationship between the treatment and -outcome, γ as coefficients for therapist-level covariates, ξ as coefficients for clinic-level covariates, and υ^Y_{00k}, u^Y_{0jk} and ε^Y_{ijk} as the clinic, therapist, and patient error terms with variances $\nu^2_{M|\boldsymbol{\theta}}$, $\tau^2_{M|\boldsymbol{\theta}}$, and $\sigma^2_{M|\boldsymbol{\theta}}$ conditional on the fixed effects ($\boldsymbol{\theta}$). When using group-mean centering for the mediator, the clinic-level mediator-outcome coefficient (B) captures the total association between the mediator and outcome—in this context this indicates that B thus captures the patient-level mediator-outcome association, therapist-level mediator-outcome association, and the clinic-level mediator-outcome association.

When assumptions are met, the 3-1-1 mediation effect (ME) is once again summarized as the product of the treatment-mediator (a) path coefficient and clinic-level mediator-outcome path coefficient (B) paths: $ME_{311} = aB$. Like the previous 3-2-1 effect, the product of a and B describes the impact of the treatment on a patient outcome as it works through the collective changes in patient levels of motivation.

Error variance & power

Similar to prior analyses we first delineate the error variances of the individual path coefficients that comprise the indirect effect. For the a path error variance (σ_a^2), the results indicate that

$$\begin{aligned}\sigma_a^2 &= \left(\frac{v^2_{M|\boldsymbol{\theta}} + \tau^2_{M|\boldsymbol{\theta}}/n_2 + \sigma^2_{M|\boldsymbol{\theta}}/(n_2 n_1)}{(n_3 - C_M - 1)p(1-p)}\right)\\ &= \left(\frac{v^2_M(1 - R^2_{M^{L3}}) + \tau^2_M(1 - R^2_{M^{L2}})/n_2 + (1 - R^2_{M^{L1}})\sigma^2_M/(n_2 n_1)}{(n_3 - C_M - 1)p(1-p)}\right)\end{aligned} \tag{23}$$

Extending previous explications, we use v_M^2, τ_M^2 and σ_M^2 as the unconditional clinic-, therapist-, and patient-level variances of the mediator, $R^2_{M^{L3}}$, $R^2_{M^{L2}}$, and $R^2_{M^{L1}}$ represent the clinic-, therapist-, and patient-level mediator variance explained by predictors in the mediator model (expression 22).

We can again restructure the error variance to draw on the primary path coefficients. The clinic-level mediator variance explained expands to

$$R^2_{M^{L3}} = R^2_{M^{L3}_{\vec{Z}}} + \frac{p(1-p)a^2}{v^2_M} \tag{24}$$

where $R^2_{M^{L3}_{\vec{Z}}}$ is the variance explained by covariates at the clinic-level. Similarly, the total therapist- ($R^2_{M^{L2}}$) and patient-level ($R^2_{M^{L1}}$) mediator variance explained reduces to just the variance explained by covariates; that is, $R^2_{M^{L2}} = R^2_{M^{L2}_{\vec{Z}}}$ and $R^2_{M^{L1}} = R^2_{M^{L1}_{\vec{Z}}}$ with $R^2_{M^{L2}_{\vec{Z}}}$ and $R^2_{M^{L1}_{\vec{Z}}}$ as the variance explained by covariates at each level.

Similarly, the error variance of the B path now becomes

$$\begin{aligned}\sigma_B^2 &= \frac{v^2_{Y|\boldsymbol{\theta}} + \tau^2_{Y|\boldsymbol{\theta}}/n_2 + \sigma^2_{Y|\boldsymbol{\theta}}/(n_2 n_1)}{(n_3 - C_Y - 1)(v^2_{M|\boldsymbol{\theta}} + \tau^2_{M|\boldsymbol{\theta}}/n_2 + \sigma^2_{M|\boldsymbol{\theta}}/(n_2 n_1))}\\ &= \frac{v^2_Y(1 - R^2_{Y^{L3}}) + \tau^2_Y(1 - R^2_{Y^{L2}})/n_2 + (1 - R^2_{Y^{L1}})\sigma^2_Y/(n_2 n_1)}{(n_3 - C_Y - 1)(v^2_M(1 - R^2_{M^{L3}}) + \tau^2_M(1 - R^2_{M^{L2}})/n_2 + (1 - R^2_{M^{L1}})\sigma^2_M/(n_2 n_1))}\end{aligned} \tag{25}$$

The error variance now includes v_M^2, τ_M^2, and σ_M^2 as the unconditional mediator variances at the clinic-, therapist-, and patient-levels and $R^2_{M^{L3}}$ as the variance in the clinic-level average mediator ($\bar{M}$) explained by the other clinic-level predictors in the outcome model (i.e., $T, \bar{X}, \bar{W}, Z$), $R^2_{M^{L2}}$ as the therapist-level variance explained in the (centered) mediator by the predictors (i.e., $(\bar{X}_{jk} - \bar{X}_k), (W_{jk} - \bar{W}_k), Q$), and $R^2_{M^{L1}}$ as the patient-level variance explained in the (centered) mediator by the predictors (i.e., $(\bar{X}_{ijk} - \bar{X}_{jk}), V$).

Consonant with the previous designs, we can also delineate the variance explained terms in order to link them to the core paths defining mediation. The total outcome variance explained at the clinic-level ($R^2_{Y^{L3}}$) decomposes into

$$R^2_{Y^{L3}} = R^2_{Y^{L3}_{\vec{Z}}} + \frac{p(1-p)(aB+c')^2}{v^2_Y} + \frac{v^2_M + \tau^2_M/n_2 + \sigma^2_M/(n_2 n_1)}{v^2_Y} B^2\left(1 - \frac{p(1-p)a^2}{v^2_M} - R^{L3}_{M_{\vec{z}}}{}^2\right) \tag{26}$$

with $R^2_{Y^{L3}_{\vec{Z}}}$ as the clinic-level variance explained by covariates.

For the outcome variance explained at the therapist-level ($R^2_{Y^{L2}}$), we can re-express it as

$$R^2_{Y^{L2}} = R^2_{Y^{L2}_{\vec{Z}}} + \left(\frac{(\tau^2_M + \sigma^2_M/n_1)(1 - 1/n_2)}{\tau^2_Y}\right) b_2^2(1 - R^2_{M^{L2}_{\vec{Z}}}) \tag{27}$$

with $R^2_{Y^{L2}_{\vec{Z}}}$ as the total therapist-level outcome variance explained by covariates (e.g., $\bar{X}_{jk} - \bar{X}_k$, $W_{jk} - \bar{W}_k$, and Q_{jk}).

Applying a similar analysis at the patient-level returns to the following expansion for the total patient-level variance explained ($R^2_{Y^{L1}}$)

$$R^2_{Y^{L1}} = R^{L1}_{Y\vec{z}}2 + \left(\frac{\sigma^2_M}{\sigma^2_Y}\right) b_1^2\left(1 - R^2_{M^{L1}_{\vec{Z}}}\right) \tag{28}$$

with $R^2_{M^{L1}_{\vec{Z}}}$ as the variance explained by patient-level covariates (e.g., $X_{ijk} - \bar{X}_{jk}$ and V_{ijk}).

Once again standardization of the mediator and outcome by the total variance places them on a standardized mean difference scale (a and c') and a standardized regression coefficient scale (B). Like prior

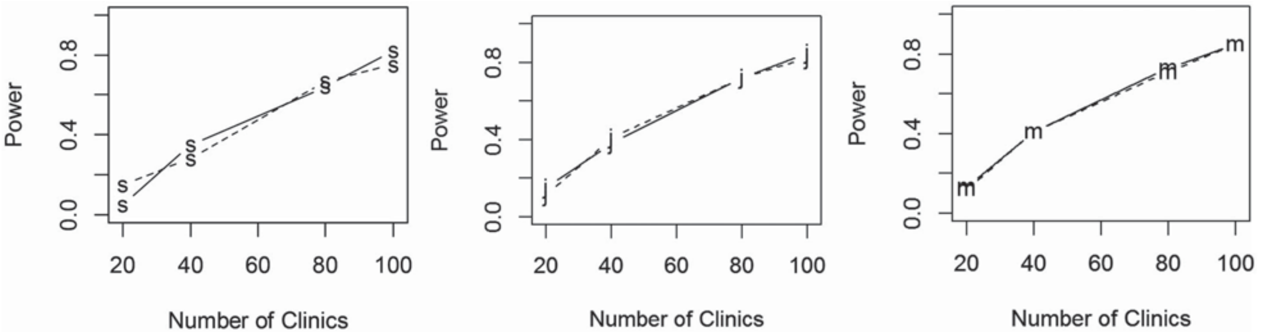

Figure 3. Power for 3-1-1 mediation as a function of the number of clinics as observed in the simulation (solid line) and predicted using formulas (dash) by test (s for Sobel, j for joint, and m for Monte Carlo interval tests) when the number of therapists per clinic is 6 (n2) and the number of patients per therapist is 6 (n1).

analyses, the statistical power of the three tests can be traced in a manner parallel to the 3-3-1 and 3-2-1 analyses. We simply substitute the abovementioned error variances for the paths under the 3-1-1 design for those in the power formulas for the previous designs.

Simulation

We assessed the accuracy of our results using a Monte Carlo simulation. Our simulation results regarding our formula-based predictions and the observed power and type 1 error rates for 3-1-1 mediation are provided in Table 1. The results are consistent with those regarding 3-3-1 and 3-2-1 mediation. Comparisons between the simulated and predicted rates are well aligned for the joint and Monte Carlo tests while such comparisons for the Sobel test suggested that with small multilevel samples the resulting expressions incurred some error. The comparisons between the predicted and observed power rates are further illustrated in Figure 3 under a sample of 6 clinics (n_2) and 6 patients (n_1). Evident from Figure 3, the power of the Monte Carlo and joint tests are well captured by the proposed formulas but predictions of the Sobel test are error prone.

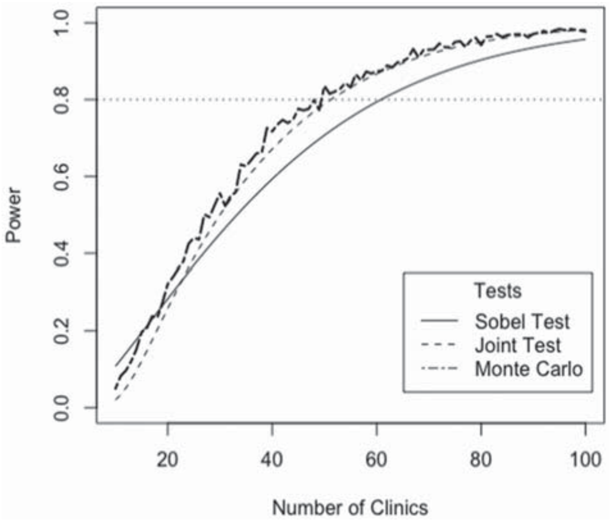

Figure 4. Power as a function of the number of clinics by test for a 3-1-1 mediation effect of 0.24 when the number of patients per therapist (n1) is 5 and the number of therapists per clinic (n2) is 2.

Illustration

Continuing with clinic-randomized design, we can now consider power analyses for a study that probes a patient-level mediator (i.e., 3-1-1 mediation). That is, we are now interested in identifying a sample size that achieves an 80% level of power to detect the extent to which the impact of the clinic-based training program on patient quality of life (outcome) operates through patient motivation (mediator). We anticipate similar parameter values for the outcome and effects as before. However, we presume the variance decomposition of the patient quality of life (outcome) to be approximately 20% owing to clinics, 30% owing to therapists, and the remaining 50% owing to patients. For the patient motivation (mediator), we assume that the variance decomposition is approximately 25% owing to clinics, 25% owing to therapists, and the remaining 50% owing to patients. Furthermore, we expect covariates to explain about 50% of the variation in patient motivation at each level. That is, we anticipate the following parameter values: $a = 0.6$, $B = 0.40$ $c' = 0.11$ $R^2_{Y_z^{L3}} = 0.5, R^2_{Y_z^{L2}} = R^2_{Y_z^{L1}} = 0.75, R^2_{M_z^{L3}} = R^2_{M_z^{L2}} = R^2_{M_z^{L1}} = 0.5.$ If we sample 5 patients per therapist (n_1), 2 therapists per clinic (n_2), how many clinics do we need to achieve a roughly 80% power?

Our power analyses indicates that under the Monte Carlo interval test roughly 48 clinics would yield a power level close to 0.80 (Figure 4). For the joint test we would need about three more clinics for a total of 51 while for the Sobel test we would need about 61 clinics. Under the same conditions, by contrast, we would need around 42 clinics to have a power level of 0.80 to detect the total effect.

Discussion

Prospective planning of multilevel studies is increasingly common across a broad array of disciplines (e.g., Evans, 2003; Gottfredson et al., 2015; IES, 2019; Kelcey et al., 2017; Spybrook et al., 2016; Windgassen et al., 2016). In this study, we expanded the basis for the design of such studies by developing methods and software to compute the statistical power associated with mediation in three-level cluster-randomized trials for mediators at each hierarchical level. Our results identify the parameters governing power, outline the potential capacity of multilevel studies to detect mediation effects, and guide researchers in selecting sample sizes that ensure a chosen level of power.

A historically important question of multilevel designs is the extent to which the sample sizes they demand are practically feasible within the context of different fields. Although a careful analysis of this question within and across different fields is beyond the scope of this study, our simulations and example applications provide an initial (but qualified and complex) outline of the scale that may be required to achieve common levels of power under a three-level cluster-randomized design. Even though regarding a sample as large- or small-scale depends heavily on context, our examples collectively suggest that adequate power will frequently demand clinic-level samples that are somewhat large but comparable to that needed to detect total effects. In our applications, for instance, designs with predictive covariates required samples on the order of 30-80 clinics with 2 therapists/clinic and 5 patients/therapist to produce an 80% chance of detecting mediation effects (see Table 1 and Figures 2–4). In each of these examples, the number of clinics needed to produce a similar level of power for the total effect was only moderately higher or lower than that for mediation effects—and more generally the sample sizes required for both effect were on the same order of magnitude.

Further analysis of the resulting error variance and power expressions, however, suggests that the development of prospective design strategies for detecting mediation effects will be complicated and qualified. Many of the design principles known for improving the power to detect total effects will remain useful in some contexts but may also fail in other contexts. For instance, parameters associated with the outcome will typically influence mediation power in ways that parallel power for the total effect—e.g., covariance adjustment on variables that are prognostic of the outcome will improve efficiency and the power to detect both total and mediation effects.

However, applying a similar covariance strategy to the mediator model may fail. Conditioning on covariates that explain mediator variance can diminish or improve the power to detect mediation effects because it decreases the error variance for the *a* path but it also possibly increases the error variance of the *B* path through collinearity effects (e.g., Beasley, 2014). More generally, parameters associated with or involving the mediator can have complicated and at times paradoxical relationships with power.

As a result, a pressing conclusion from our preliminary survey of parameter values and sample sizes is that power is a complicated function of the full array of parameter values rather than principally driven by just a few parameters as is typically the case with total effects. As another illustrative case in point, consider the relationship between power and effect size. Although power has a simple monotonic relationship with the magnitude of the total effect, the relationship between power and the magnitude of the mediation effect is much more complicated and influenced by the decomposition of the mediation effect and the values of concomitant parameters. For example, increasing the magnitude of the mediation effect by increasing the *a* path coefficient and holding the *B* path coefficient constant can actually reduce power because power is dependent on both the magnitude of the mediation effect and its decomposition (e.g., Beasley, 2014).

The net implication of these results is that well-designed studies probing mediation will require good empirical estimates of parameter values (e.g., Hedges & Hedberg, 2007; Kelcey & Shen, 2016). To some extent, many fields have been actively developing and compiling empirical estimates for design purposes (e.g., Hedges & Hedberg, 2007; Westine et al., 2013; Dong et al., 2016; Kelcey & Shen, 2016). These types of studies have, however, largely limited their scope to parameters associated with total effects and two-level designs—although there are some recent exceptions that have expanded these efforts (see e.g., Hedges et al., 2012; Kelcey et al., 2019; Kelcey & Phelps, 2013a; Kelcey & Phelps, 2013b; Roth et al., 2019; Westine, 2016).

Despite the potential accessibility and utility of the formulas developed in this study, they have several limitations. Our analyses did not consider more complicated designs and structures that draw on, for example, partially nested assignment mechanisms, cross-classified nesting structures, random slopes, moderated mediation or those that intend to probe the collective contribution of multiple mediators. Similarly, our results do not address the broad range of outcomes types (e.g., ordinal or count mediators and outcomes) often found in psychological and behavioral research (e.g., Sterba, 2017). These are important considerations and areas for subsequent research.

In conclusion, the intersection of our results and the increasing interest in appraising more comprehensive sets of effects suggests there is a growing need to expand the scope of the design literature to include empirical assessments of parameters that govern the detection of a broad range of effects. For instance, there is an increasing press for studies to routinely examine not just whether a treatment works (i.e., total effect) but also how a treatment works (i.e., mediation effects) and for whom and under what conditions it works (i.e., moderation effects; e.g., IES, 2019). Our work delineates the parameters that govern the effective and efficient design of three-level cluster-randomized trials probing mediation—however, effective use of these results involves the development of empirical values for a diverse set of mediators. In this way, empirical investigation and compilation of parameter values for diverse sets of effects and designs surfaces as a critical hurdle to the widespread implementation of rigorous and efficient study designs.

Article Information

Conflict of interest disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This study was supported by grants from the National Science Foundation (Award Nos. 1437679, 1552535, 1437745, and 1437692).

Role of the funders/sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions is not intended and should not be inferred.

References

Aarons, G. A., Glisson, C., Green, P. D., Hoagwood, K., Kelleher, K. J., & Landsverk, J. A. (2012). The organizational social context of mental health services and clinician attitudes toward evidence-based practice: a United States national study. *Implementation Science*, *7*(1), 56. doi:10.1186/1748-5908-7-56

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. doi:10.1177/1094428115627498

Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., Andersson, G., Atkins, D. C., Carlbring, P., Carroll, K. M., Christensen, A., Eddington, K. M., Ehlers, A., Feaster, D. J., Keijsers, G. P. J., Koch, E., Kuyken, W., Lange, A., Lincoln, T., … Watson, J. (2011). Intraclass correlation associated with therapists: estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy*, *40*(1), 15–33. doi:10.1080/16506073.2010.520731

Beasley, T. (2014). Test of mediation: paradoxical decline in statistical power as a function of mediator collinearity. *The Journal of Experimental Education*, *82*(3), 283–306. doi:10.1080/00220973.2013.813360

Brincks, A., Enders, C., Llabre, M., Bulotsky-Shearer, R., Prado, G., & Feaster, D. (2017). Centering predictor variables in three-level contextual models. *Multivariate Behavioral Research*, *52*(2), 149–163. doi:10.1080/00273171.2016.1256753

Brincks, A. M., Feaster, D. J., & Mitrani, V. B. (2010). A multilevel mediation model of stress and coping for women with HIV and their families. *Family Process*, *49*(4), 517–529. doi:10.1111/j.1545-5300.2010.01337.x

Brown, E. C., Hawkins, J. D., Rhew, I. C., Shapiro, V. B., Abbott, R. D., Oesterle, S., Arthur, M. W., Briney, J. S., & Catalano, R. F. (2014). Prevention system mediation of communities that care effects on youth outcomes. *Prevention Science*, *15*(5), 623–632. doi:10.1007/s11121-013-0413-7

Bulus, M., Dong, N., Kelcey, B., Spybrook, J. (2019). *PowerUpR* Shiny App for Experimental and Quasi-Experimental Study Design (Version 1.0.4) [Software]. https://powerupr.shinyapps.io/index/

Cegala, D. J., & Post, D. M. (2009). The impact of patients' participation on physicians' patient-centered communication. *Patient Education and Counseling*, *77*(2), 202–208. doi:10.1016/j.pec.2009.03.025

Černe, M., Jaklič, M., & Škerlavaj, M. (2013). Authentic leadership, creativity, and innovation: A multilevel perspective. *Leadership*, *9*(1), 63–85. doi:10.1177/1742715012455130

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Erlbaum.

Cosby, R., Howard, M., Kaczorowski, J., Willan, A., & Sellors, J. (2003). Randomizing patients by family practice: sample size estimation, intracluster correlation and data analysis. *Family Practice*, *20*(1), 77–82. doi:10.1093/fampra/20.1.77

Curenton, S. M., Dong, N., & Shen, X. (2015). Does aggregate school-wide achievement mediate fifth grade outcomes for former early childhood education participants? *Developmental Psychology*, *51*(7), 921–934. doi:10.1037/a0039295

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. doi:10.1177/0193841X16671283

Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 91–122. doi:10.1146/annurev-orgpsych-041015-062400

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. doi:10.1016/j.hrmr.2015.02.001

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. doi:10.1037/1082-989X.12.2.121

Evans, D. (2003). Hierarchy of evidence: a framework for ranking evidence evaluating healthcare interventions. *Journal of Clinical Nursing*, *12*(1), 77–84. doi:10.1046/j.1365-2702.2003.00662.x

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. doi:10.1007/s11121-015-0555-x

Harvey, A. G., & Gumport, N. B. (2015). Evidence-based psychological treatments for mental disorders: Modifiable barriers to access and possible solutions. *Behaviour Research and Therapy*, *68*, 1–12. doi:10.1016/j.brat.2015.02.004

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. doi:10.1177/0956797613480187

Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning two-and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*(6), 445–489. doi:10.1177/0193841X14529126

Hedges, L., & Hedberg, E. (2007). Intraclass correlation values for planning school-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87. doi:10.3102/0162373707299706

Hedges, L., Hedberg, E., & Kuyper, A. (2012). The variance of intraclass correlations in three- and four-level models. *Educational and Psychological Measurement*, *72*(6), 893–909. doi:10.1177/0013164412445193

Imai, K., Tingley, D., & Yamamoto, T. (2013). Experimental designs for identifying causal mechanisms. *Journal of the Royal Statistical Society: Series A (Statistics in Society))*, *176*(1), 5–51. doi:10.1111/j.1467-985X.2012.01032.x

Institute of Education Sciences (IES). (2019). *Institute of Education Sciences Request for Applications: Education Research Grants CFDA Number: 84.305A*. https://ies.ed.gov/funding/pdf/2019_84305A.pdf

Jacob, R., Zhu, P., & Bloom, H. (2010). New empirical evidence for the design of group randomized trials in education. *Journal of Research on Educational Effectiveness*, *3*(2), 157–198. doi:10.1080/19345741003592428

Kazdin, A. E. (2007). Mediators and mechanisms of change in psychotherapy research. *Annual Review of Clinical Psychology*, *3*(1), 1–27. doi:10.1146/annurev.clinpsy.3.022806.091432

Keeley, R. D., Brody, D. S., Engel, M., Burke, B. L., Nordstrom, K., Moralez, E., Dickinson, L. M., & Emsermann, C. (2016). Motivational interviewing improves depression outcome in primary care: A cluster randomized trial. *Journal of Consulting and Clinical Psychology*, *84*(11), 993–1007. https://psycnet.apa.org/doi/10.1037/ccp0000124 doi:10.1037/ccp0000124

Kelcey, B., & Carlisle, J. F. (2013). Learning about teachers' literacy instruction from classroom observations. *Reading Research Quarterly*, *48*(3), 301–317. doi:10.1002/rrq.51

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. doi:10.3102/1076998617695506

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. C., & Chin, M. J. (2019). Teacher mathematical knowledge, instructional quality, and student outcomes: a multilevel quantile mediation analysis. *School Effectiveness and School Improvement*, *30*(4), 398–431. doi:10.1080/09243453.2019.1570944

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. doi:10.3102/0162373713482766

Kelcey, B., & Phelps, G. (2013b). Strategies for improving power in school-randomized studies of professional development. *Evaluation Review*, *37*(6), 520–554. doi:10.1177/0193841X14528906

Kelcey, B., & Shen, Z. (2016). Multilevel design of school effectiveness studies in sub-Saharan Africa. *School Effectiveness and School Improvement*, *27*(4), 492–510. doi:10.1080/09243453.2016.1168855

Kenny, D. A., & Judd, C. M. (2014). Power anomalies in testing mediation. *Psychological Science*, *25*(2), 334–339. doi:10.1177/0956797613502676

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. doi:10.1080/00273171.2014.903162

Kozina, A. (2018a). School-based prevention of anxiety using the "MyFRIENDS" emotional resilience program: Six-month follow-up. *International Journal of Psychology*, *55*(1), 1–8. doi:10.1002/ijop.12553

Kozina, A. (2018b). Can the "My FRIENDS" anxiety prevention programme also be used to prevent aggression?

A six-month follow-up in a school. *School Mental Health*, *10*(4), 500–509. doi:10.1007/s12310-018-9272-5

Kozlowski, S. W., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes.

Kreft, I. G., De Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research*, *30*(1), 1–21. doi:10.1207/s15327906mbr3001_1

Krull, J. L., & MacKinnon, D. P. (1999). Multilevel mediation modeling in group-based intervention studies. *Evaluation Review*, *23*(4), 418–444. doi:10.1177/0193841X9902300404

Layard, R., & Clark, D. M. (2014). *Thrive: The power of evidence-based psychological therapies*. Penguin.

Livert, D., Rindskopf, D., Saxe, L., & Stirratt, M. (2001). Using multilevel modeling in the evaluation of community-based treatment programs. *Multivariate Behavioral Research*, *36*(2), 155–184. doi:10.1207/S15327906MBR3602_02

MacKinnon, D. P. (2008). *Multivariate applications series. Introduction to statistical mediation analysis*. Taylor & Francis Group/Lawrence Erlbaum Associates.

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. doi:10.1037/1082-989X.7.1.83

Mierlo, H., Rutte, C. G., Vermunt, J. K., Kompier, M. A. J., & Doorewaard, J. A. C. M. (2007). A multi-level mediation model of the relationships between team autonomy, individual task design and psychological well-being. *Journal of Occupational and Organizational Psychology*, *80*(4), 647–664. doi:10.1348/096317907X196886

Mowbray, C. T., Woodward, A. T., Holter, M. C., MacFarlane, P., & Bybee, D. (2009). Characteristics of users of consumer-run drop in centers versus clubhouses. *The Journal of Behavioral Health Services & Research*, *36*(3), 361–371. doi:10.1007/s11414-008-9112-8

National Longitudinal Survey of Children and Youth, Statistics Canada. (2019). Government of Canada, Statistics. *National Longitudinal Survey of Children and Youth (NLSCY)*. http://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=4450

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., Murphy, D. L., & Tate, R. L. (2009). Three-level models for indirect effects in school- and class-randomized experiments in education. *The Journal of Experimental Education*, *78*(1), 60–95. doi:10.1080/00220970903224685

Pituch, K. A., & Stapleton, L. M. (2008). The performance of methods to test upper-level mediation in the presence of nonnormal data. *Multivariate Behavioral Research*, *43*(2), 237–267. doi:10.1080/00273170802034844

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. doi:10.1177/0049124112460380

Pituch, K. A., Stapleton, L. M., & Kang, J. Y. (2006). A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. *Multivariate Behavioral Research*, *41*(3), 367–400. doi:10.1207/s15327906mbr4103_5

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98. doi:10.1080/19312458.2012.679848

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

Roche, A. M., & Freeman, T. (2004). Brief interventions: good in theory but weak in practice. *Drug and Alcohol Review*, *23*(1), 11–18. doi:10.1080/09595230410001645510

Roth, K. J., Wilson, C. D., Taylor, J. A., Stuhlsatz, M. A., & Hvidsten, C. (2019). Comparing the Effects of Analysis-of-Practice and Content-Based Professional Development on Teacher and Student Outcomes in Science. *American Educational Research Journal*, *56*(4), 1217–1253. doi:10.3102/0002831218814759

SAMSHA. (2007). *National Survey on Drug Use and Health.* http://www.oas.samhsa.gov/NSDUHlatest.htm

Schwartz, R. (2010). Motivational interviewing (patient-centered counseling) to address childhood obesity. *Pediatric Annals*, *39*(3), 154–158. doi:10.3928/00904481-20100223-06

Sikorski, C., Luppa, M., König, H. H., van den Bussche, H., & Riedel-Heller, S. G. (2012). Does GP training in depression care affect patient outcome?-A systematic review and meta-analysis. *BMC Health Services Research*, *12*(1), 10. doi:10.1186/1472-6963-12-10

Slade, M., Bird, V., Clarke, E., Le Boutillier, C., McCrone, P., Macpherson, R., Pesola, F., Wallace, G., Williams, J., & Leamy, M. (2015). Supporting recovery in patients with psychosis through care by community-based adult mental health teams (REFOCUS): a multisite, cluster, randomised, controlled trial. *The Lancet Psychiatry*, *2*(6), 503–514. doi:10.1016/S2215-0366(15)00086-3

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology*, *13*, 290–312. doi:10.2307/270723

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, 6*, 605–627.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: an examination of the precision of cluster randomized trials funded by the US Institute of Education Sciences. *International Journal of Research & Method in Education*, *39*(3), 255–267. doi:10.1080/1743727X.2016.1150454

Sterba, S. K. (2017). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research*, *27*(4), 425–436. doi:10.1080/10503307.2015.1114688

Van Voorhees, B. W., Fogel, J., Houston, T. K., Cooper, L. A., Wang, N.-Y., & Ford, D. E. (2005). Beliefs and attitudes associated with the intention to not accept the diagnosis of depression among young adults. *Ann Fam Med.*, *3*(1), 38–46. doi:10.1370/afm.273

VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research*, *38*, 515–544. doi:10.1177/0049124110366236

VanderWeele, T. J., Hong, G., Jones, S. M., & Brown, J. L. (2013). Mediation and spillover effects in group-randomized trials: A case study of the 4Rs educational intervention. *Journal of the American Statistical Association*, *108*(502), 469–482. doi:10.1080/01621459.2013.779832

VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface*, *2*(4), 457–468. doi:10.4310/SII.2009.v2.n4.a7

Westine, C. D. (2016). Finding efficiency in the design of large multisite evaluations: Estimating variances for science achievement studies. *American Journal of Evaluation*, *37*(3), 311–325. doi:10.1177/1098214015624014

Westine, C., Spybrook, J., & Taylor, J. (2013). An empirical investigation of variance design parameters for planning cluster-randomized trials of science achievement. *Evaluation Review*, *37*(6), 490–519. doi:10.1177/0193841X14531584

Willenbring, M. L., Kivlahan, D., Kenny, M., Grillo, M., Hagedorn, H., & Postier, A. (2004). Beliefs about evidence-based practices in addiction treatment: A survey of Veterans Administration program leaders. *Journal of Substance Abuse Treatment*, *26*(2), 79–85.(03)00161-2 doi:10.1016/S0740-5472(03)00161-2

Wilt, J. (2012). Mediators and mechanisms of psychotherapy: Evaluating criteria for causality. *Graduate Student Journal of Psychology*, *14*, 53–60. https://www.researchgate.net/profile/Joshua_Wilt/publication/263041559_Mediators_and_Mechanisms_of_Psychotherapy_Evaluating_Criteria_for_Causality/links/00b495399d3778fc88000000.pdf

Windgassen, S., Goldsmith, K., Moss-Morris, R., & Chalder, T. (2016). Establishing how psychological therapies work: the importance of mediation analysis. *Journal of Mental Health*, *25*(2), 93–99. doi:10.3109/09638237.2015.1124400

Xu, B. D., Zhao, S. K., Li, C. R., & Lin, C. J. (2017). Authentic leadership and employee creativity: testing the multilevel mediation model. *Leadership & Organization Development Journal*, *38*(3), 482–498. doi:10.1108/LODJ-09-2015-0194

Zhang, Z., Zyphur, M., & Preacher, K. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods*, *12*(4), 695–719. doi:10.1177/1094428108327450