Original Article

Robustness of Statistical Power in Group-Randomized Studies of Mediation Under an Optimal Sampling Framework

Kyle Cox and Benjamin Kelcey

Quantitative and Mixed Methods Research Methodologies, University of Cincinnati, OH, USA

Abstract: When planning group-randomized studies probing mediation, effective and efficient sample allocation is governed by several parameters including treatment-mediator and mediator-outcome path coefficients and the mediator and outcome intraclass correlation coefficients. In the design stage, these parameters are typically approximated using information from prior research and these approximations are likely to deviate from the true values eventually realized in the study. This study investigates the robustness of statistical power under an optimal sampling framework to misspecified parameter values in group-randomized designs with group- or individual-level mediators. The results suggest that estimates of statistical power are robust to misspecified parameter values across a variety of conditions and tests. Relative power remained above 90% in most conditions when the incorrect parameter value ranged between 50% and 150% of the true parameter.

Keywords: optimal sample allocation, power, multilevel mediation, group-randomized study

When investigating a treatment that operates within a hierarchical structure, experimental designs that assign groups to treatment conditions often ameliorate ethical concerns, more appropriately suit extant organizational structures, and better reflect the contexts in which the results would be generalized while maintaining a rigorous basis for causal inference (Raudenbush, 1997; Spybrook & Raudenbush, 2009). One of the primary questions in the planning stages of these studies is the sample size required to detect the expected treatment effect with a reasonably high probability. Sampling more groups almost always increases the power or probability to detect an effect in group-randomized studies (see Figure 1A) but this strategy typically incurs substantial additional costs (Kelcey & Phelps, 2013, 2014; Raudenbush, 1997). Optimal sample allocation frameworks balance these concerns by introducing cost parameters in the design phase that incorporate cost considerations directly into study planning. In turn, such frameworks identify a sample of individuals per group and groups that maximize power given study-specific conditions and a designated budget and cost structure (see Figure 1B; e.g., Hedges & Borenstein, 2014; Kelcey, Phelps, Spybrook, Jones, & Zhang, 2017).

Despite the conceptual utility of the optimal sampling framework, a key constraint in its practical implementation is a priori knowledge of the parameter values that govern power for a particular estimand. In the planning phase, the values of these parameters are typically not precisely available, so researchers may utilize previous empirical results to infer a range of plausible values. It is likely that these initial parameter estimates will deviate to some extent from the true parameter values because estimates from the literature are almost always based on outcomes and populations that differ in overt and subtle ways from the outcome and sample to be used in the new study (e.g., Korendijk, Moerbeek, & Maas, 2010).

For these reasons, prior literature has investigated the robustness of designs in terms of their relative efficiency to various types of parameter value misspecifications in the planning phase (e.g., Korendijk et al., 2010). This research has largely been confined to the loss of efficiency regarding estimation of the main effect. Recent research has, however, promoted studies that are intentionally designed to probe a more comprehensive set of effects that unpack the theory of action underlying the treatment effects using mediation (e.g., Gottfredson et al., 2015;

https://doi.org/10.1027/1614-2241/a000169

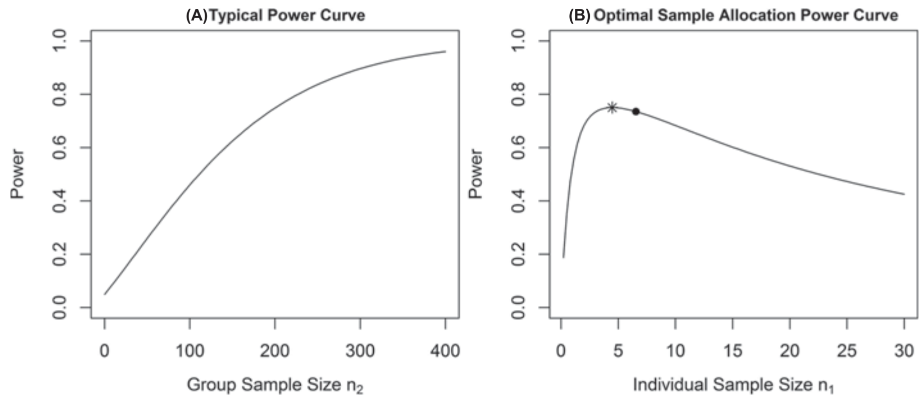

Figure 1. Power of a group-randomized study to detect group-level mediation using the Sobel test as a function of (A) group sample size (n_2) and (B) individual sample size (n_1) under an optimal sample allocation framework. *True optimal individual sample size while a dot marks the optimal individual sample size based on a misspecified value.

Institute of Education Sciences, US Department of Education, & National Science Foundation, 2013; Authors, B **[Author: Please provide full reference details]**). Yet, little is known regarding the robustness of power in these designs when parameter values are misspecified.

In this study, we advance the literature base regarding the robustness of multilevel designs to parameter value misspecifications in the context of mediation and optimal sampling. More specifically, we incorporate study costs and budget into the planning phase using an optimal sampling framework in order to contrast the power of two competing designs – one design that yields the maximum level of power and an alternative design that yields a sub-optimal level of power because of parameter value misspecifications. We limit our considerations regarding sampling design to the relative sample sizes at the individual- and group-level.

To illustrate these differences, consider Figure 1B, it presents power as a function of the individual-level sample size (n_1) under an optimal sampling framework. The peak of the power curve represents the most efficient use of resources (i.e., sample of individuals per group and sample of groups) which maximizes attainable power given study-specific conditions. An "*" identifies this optimal design and serves as an appropriate benchmark for comparing alternative designs because any deviation from the optimal design is an inefficient use of limited resources and achieves less power. This study tracks the consequences of using misspecified parameter values in the study planning phase by comparing power rates for an optimally designed study (* in Figure 1B) to one planned with misspecified parameter values (• in Figure 1B). Consequences are represented by the distance between the true (*) and misspecified (•) optimal individual sample allocation in Figure 1.

Specifically, we estimate the power of a study design when true parameter values were used to determine the optimal sample allocation and compare it to power rates when the study employs a sample allocation based on misspecified parameter values. We probe studies examining the effect of a group-level treatment on an individual-level outcome through a group-level mediator or an individual-level mediator with different sampling cost structures and three tests of the mediation effect. We refer to these mediated effects by the level of the treatment, mediator, and outcome (i.e., 2-2-1 and 2-1-1 mediated effects). Sampling design considerations are limited to the relative sample sizes at the group- and individual-level and assume an equal number of groups in each of the two treatment conditions. The paper is divided into two major sections based on these design types (i.e., 2-2-1 and 2-1-1). Both represent investigations of multilevel mediation as the mediated effect operates across levels of a hierarchy with the numerical notation identifying the level of the treatment-mediator-outcome variables, respectively. For each design, we outline the key components of the optimal sampling framework, describe the scope of our investigation, and detail our results. We conclude with a brief discussion.

Group-Level Mediator

Analytic Model

In 2-2-1 designs groups are randomized to one of two treatment conditions (T) to examine treatment impacts on an individual-level outcome (Y) operating through a group-level mediator (M). We adopt a typical multilevel mediation path formulation with parameters estimated using maximum likelihood (e.g., Authors, D**[Author: Please provide full reference details]**; Zhang, Zyphur, & Preacher, 2009). The mediator path model at the group-level can be expressed as

$$\text{Mediator model (Level 2)}: M_j = \pi_0 + aT_j + \varepsilon_j \;\; \varepsilon_j \sim N\left(0, \sigma^2_{M|}\right). \quad (1)$$

The mediation model (i.e., Equation 1) captures the treatment-mediator relationship and describes variation in the group-level mediator as a function of a group-level treatment. We use M_j as the mediator for group j, T_j as the treatment assignment coded as 0.5 and -0.5 for the treatment and control condition, respectively. Coefficient a represents the treatment-mediator relationship with e_j as the normally distributed mediator error with a mean of zero and variance $\sigma^2_{M|}$ conditional on T_j.

The outcome model captures the relationship between the group-level mediator and individual-level outcome such that

$$\text{Outcome model (Level 1)}: \; Y_{ij} = \beta_{0j} + e_{ij} \;\; e_{ij} \sim N\left(0, \sigma^2_Y\right), \quad (2a)$$

$$\text{(Level 2)}: \; \beta_{0j} = \gamma_{00} + bM_j + c'T_j + u_{0j} \;\; u_{0j} \sim N\left(0, \tau^2_{Y|}\right). \quad (2b)$$

Here, Y_{ij} is the outcome for individual i in group j. At the individual-level, e_{ij} represents the normally distributed error term with a mean of zero and variance σ^2_Y. At the group-level, b is the conditional relationship between the mediator and the outcome, while c' captures the direct effect of the treatment on the outcome while controlling for the mediator, and u_{0j} is the normally distributed group-specific random effect with a mean of zero and variance $\tau^2_{Y|}$ conditional on T_j and M_j. A combined mediator and outcome model from Equations 1 and 2can be expressed as

$$Y_{ij} = (\gamma_{00} + b\pi_0) + \left(ba + c'\right)T_j + b\varepsilon_j + u_{0j} + e_{ij}. \quad (3)$$

Finally, we utilize a standardized outcome and mediator such that each has a mean of zero and unit variance. Under this formulation $\sigma^2_M = 1$ and $\tau^2_Y + \sigma^2_Y = 1$. This implies $\tau^2_Y = \rho$ and $\tau^2_Y + \sigma^2_Y = \rho + (1 - \rho)$ when ρ is the unconditional intraclass correlation coefficient for the outcome defined as

$$\rho = \frac{\tau^2_Y}{\sigma^2_Y + \tau^2_Y}. \quad (4)$$

Conditional and unconditional ρ values capture variance in the outcome attributable to the group-level or the correlation among individuals on the outcome within the same group.

This formulation places the a and c' path coefficients on a Cohen's d or standardized group differences scale and formats the b path coefficient as a standardized regression coefficient.

Mediation Test Statistics and Power

We estimate the 2-2-1 multilevel mediation effect (ME) using the typical product of coefficients method: ME = ab with the a coefficient representing the treatment-mediator path (Equation 1 and the b coefficient representing the mediator-outcome path (Equation 2b. To determine the significance of this effect we consider the Sobel, joint, and Monte-Carlo interval tests because they can be employed before the collection of data (i.e., during study planning). Performance of these tests in terms of power converges with large sample sizes but under sample sizes typical for studies in the social sciences there are substantial differences (e.g., Authors D**[Author: Please provide full reference details]**). Additionally, the optimal sample allocation for a study can differ based solely on the selected mediation test (Authors A**[Author: Please provide full reference details]**). These factors necessitate the inclusion of multiple mediation tests representing various approaches.

The historically popular Sobel Test compares the ratio of the estimated mediation effect to its estimated standard error. In recent years, the test has been heavily criticized because of its imprecision in small-to-moderate sample sizes. We outline the test for illustrative and comparative purposes and recommend the use of alternative tests subsequently outlined. The Sobel test statistic is

$$z^{\text{Sobel}}_{ab} = ab/\sqrt{\sigma^2_{ab}}. \quad (5)$$

Here, σ^2_{ab} represents the error variance of the mediated effect and ab is the ME defined above. Based on prior literature we can estimate this error variance as a function of individual paths and their individual error variances and form the Sobel Test statistic for the mediation effect as (Sobel, 1982)

$$z^{\text{Sobel}}_{ab} = ab/\sqrt{b^2\sigma^2_a + a^2\sigma^2_b}. \quad (6)$$

The test statistic has an asymptotically normal distribution allowing inferences with large sample sizes to be drawn based on a comparison of the test statistic (e.g., z^{Sobel}_{ab}) to a critical value (e.g., z_{critical} = 1.96) in a standard normal distribution (Φ) at the associated type one error rate (e.g., 0.05 for 1.96).

When the alternative hypothesis is true, the test statistics follow a non-central distribution with the ratios as the noncentrality parameter. Power to detect the mediation effect is then determined with

$$P\left(\left|z_{ab}^{\text{Sobel}}\right| > z_{\text{critical}}\right) = 1 - \Phi\left(z_{\text{critical}} - z_{ab}^{\text{Sobel}}\right) + \Phi\left(-z_{\text{critical}} - z_{ab}^{\text{Sobel}}\right). \quad (7)$$

An alternative to the Sobel test is the joint test which avoids the direct estimation of the mediation effect and its distribution. Rather, it uses a composite null approach that considers the treatment-mediator and mediator-outcome paths separately to determine the significance of a mediated effect. The two concurrent sub-tests compare the ratio of the path estimate and its standard error to a normal or *t*-distribution. Only after rejecting the null hypotheses of each individual path is the mediation effect considered statistically significant. Avoiding the distributional assumptions of the Sobel test allows the joint test to perform well in terms of power and type one error rate in a variety of settings (e.g., Hayes & Scharkow, 2013; Kelcey, Dong, Spybrook, & Shen, 2017).

The statistical test of the a path representing the treatment-mediator association is

$$z_a = a/\sigma_a, \quad (8)$$

and the test of the b path representing the mediator-outcome association is

$$z_b = b/\sigma_b. \quad (9)$$

With σ_a^2 and σ_b^2 indicating the error variances of the a path and b path respectively. Power to detect the mediation effect using the joint test is simply the product of the power to detect each path which we formulate as

$$\begin{aligned} P(|z_a| > z_{\text{critical}} \text{ and } |z_b| > z_{\text{critical}}) &= (1 - \Phi(z_{\text{critical}} - z_a) + \Phi(-z_{\text{critical}} - z_a)) \\ &\times (1 - \Phi(z_{\text{critical}} - z_b) + \Phi(-z_{\text{critical}} - z_b)), \end{aligned} \quad (10)$$

with $\Phi()$ as the normal cumulative density function.

The Monte-Carlo (MC) interval test resamples a and b path values with sampling variability equal to the error variance of the respective paths (Preacher & Selig, 2012). Maximum likelihood estimation identifies path coefficient estimates with an assumed multivariate normal distribution with means, variances, and covariances based on the maximum likelihood estimates (Preacher & Selig, 2012). For the 2-2-1 analytic models described above, we draw a and b path values using

$$\begin{pmatrix} a^{\#} \\ b^{\#} \end{pmatrix} \sim \text{MVN}\left(\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}_{\hat{a}}^2 & \hat{\sigma}_{\hat{a},\hat{b}} \\ \hat{\sigma}_{\hat{a},\hat{b}} & \hat{\sigma}_{\hat{b}}^2 \end{pmatrix}\right). \quad (11)$$

An approximation of the sampling distribution of the mediated effect is formed using the product of each set of $a^{\#}$ and $b^{\#}$. Inferences regarding the mediated effect are drawn based on the inclusion of zero in the asymmetric confidence intervals constructed from the sampling distribution. Power of the MC interval test is simply the proportion of asymmetric confidence intervals that exclude zero.

Optimal Sample Allocation

Theoretically, optimal sample allocation provides researchers a means to identify the sampling strategy that maximizes power under specific constraints on design and budget (e.g., Kelcey, Phelps, et al., 2017). The process begins with the identification of the optimal sample of individuals per group (n_1^{opt}) and then a simple function of this value, budget, and sampling costs identifies the optimal number of groups n_2^{opt}.

Throughout this study, we apply the conventional linear cost formulation (Raudenbush, 1997) such that

$$T = c_2 n_2 + c_1 n_2 n_1. \quad (12)$$

where T is the total funds available to collect data for a study, c_1 is the cost to enroll each individual after sampling the group, and c_2 is sampling cost for each additional group. Each is typically measured in monetary units (e.g., dollars). For our optimal sample allocation formulas, we assume an equal number of groups in two treatment conditions (i.e., treatment and control conditions) and equal cost across conditions.

Like group-randomized studies of main effects, optimal sample allocation under the Sobel test is derived by minimizing the error variance (Raudenbush, 1997). With a standardized mediator and outcome, the optimal individual sample size for the Sobel test in terms of path coefficients is (Authors, A[**Author: Please provide full reference details**])

$$n_1^{\text{opt}} = \sqrt{\left(\frac{c_2}{c_1}\right) \frac{a^2(1-\rho)}{4b^2\left(1-\frac{a^2}{4}\right)^2 + a^2\left(\rho - \frac{\left(ab+c'\right)^2}{4} - b^2\left(1-\frac{a^2}{4}\right)\right)}}, \quad (13)$$

with the optimal sample of groups determined by substituting n_1^{opt} in $n_2^{\text{opt}} = T/(c_2 + c_1 n_1)$.

The structure of the joint test complicates the calculation of an optimal individual sample size because there is no mediation effect error variance to minimize. Rather, one must directly maximize power to determine n_1^{opt} under the joint test. There is no simple closed-form expression, but one is possible through numerical methods (see Electronic Supplemental Material, ESM 1; Authors, A [**Author: Please provide full reference details**]).

Like the joint test, the MC interval test does not have a closed form solution for n_1^{opt} but can be approximated by the Sobel and joint test formulations in larger samples

(Authors, A). However, there are typical conditions (e.g., large sampling cost ratios; c_2/c_1) in which the n_1^{opt} values under the Sobel and joint test substantially diverge from the values for the MC interval test (Authors, A). Such conditions would introduce a confounding effect into this investigation because we could not determine if differences in power from inefficient sample allocation were due to parameter value misspecification or a misalignment between the Sobel and joint test n_1^{opt} and the true n_1^{opt} for the MC interval test. We avoid these confounding effects by numerically estimating n_1^{opt} for the MC interval test using a linear search algorithm under the specified design and budget.

Power of 2-2-1 Mediation Studies

Path coefficients (a, b, c'), the unconditional intraclass correlation coefficient (ρ), and cost structure (c_2/c_1) all influence optimal sample allocation. Given that the parameter values employed in the study design phase will not precisely match their true values, we compare power rates when optimal sample allocation is determined using true and misspecified parameter values. Power is determined analytically using the tests and formulations above and path coefficient error variance formulations that utilize budget and sampling cost values. Assuming balanced random assignment, we formulate these error variances such that (Authors, A; Authors D **[Author: Please provide full reference details]**)

$$\sigma_a^2 = \frac{4(1-(a^2+4))}{T/(c_2+c_1 n_1)} \quad (14)$$

and

$$\sigma_b^2 = \frac{\left(\rho - \frac{(ab+c')^2}{4} - b^2\left(1-\frac{a^2}{4}\right)\right) + (1-\rho)/n_1}{\left(1-\frac{a^2}{4}\right)\left(T/(c_2+c_1 n_1)\right)}. \quad (15)$$

These formulations employ parameters common to planning group-randomized mediation studies (Authors, B; Authors D) and, under an optimal sampling framework, allow power analyses using only n_1^{opt}. Cost and budget information has replaced n_2 values.

Borrowing from Korendijk et al. (2010), we refer to the true but unknown parameter values as population values, denoting them with * (e.g., a^*, b^*, c'^*, ρ^*), and labeling the true optimal individual per group sample size as $n_1^{\text{opt}*}$. The estimated or predicted misspecified values used in the study design phase are referred to as the initial values (e.g., a, b, c', ρ) with the optimal individual per group sample size based on these values retaining the n_1^{opt} notation.

We defined the robustness of statistical power under an optimal sampling framework against misspecified parameter values in terms of the relative loss of statistical power:

$$\text{Relative Power} = \frac{\text{Initial Design Study Power}}{\text{Population Design Study Power}}. \quad (16)$$

To determine the power of a design under the population values, we identified the optimal individual sample size using the true population values ($n_1^{\text{opt}*}$) and then calculated the implied power to detect the mediation effect. Similarly, to determine the power of a design under the initial values, we identified the optimal individual sample size using the initial misspecified parameter values (n_1^{opt}) and then calculated the implied power to detect the mediation effect. Relative power examines the loss of power associated with an n_1^{opt} based on a misspecified parameter value and has a maximum value of one when the initial and population parameters are equal (i.e., $n_1^{\text{opt}*} = n_1^{\text{opt}}$). Imprecision in our linear search algorithm for the MC interval test did produce some relative power values exceeding one but these erroneous values only reflect noise in the estimation of initial and population MC interval test power. They do not bias the overall robustness of power results or interpretations.

Without clear benchmarks regarding what constitutes a scale for relative power, we borrowed from the optimal design literature regarding relative efficiency and interpret a relative power value of .9 and above (i.e., 10% loss of power or less) to be good and values between .8 and .9 as acceptable (e.g., Korendijk et al., 2010). Our general descriptions of robustness are based on relative power meeting or exceeding these benchmarks when the misspecified parameter values are between 50% and 150% of the true population value. Specific levels of acceptable power loss and parameter misspecification are a study-specific consideration.

Our investigation included three different path coefficient ratios (a/b = 0.3/0.3, 0.5/0.3, 0.6/0.2), three intraclass correlation coefficient values (ρ = .3, .4, .5), and three group to individual sampling cost structures (c_2/c_1 = 5/1, 10/1, 100/1). We employed a fully crossed design resulting in a total of 27 conditions with various power rates and n_1^{opt} values. Under each condition, we found relative power when incorrect initial a values ranged from 0.05 to 0.9, incorrect initial b values ranged from 0.05 to 0.5, and incorrect initial intraclass correlation coefficient values ranged from 0.15 to 0.9 all using an interval of 0.025. While not comprehensive, these explicit comparisons reflect a range of substantive applications. It is also possible to examine the robustness of statistical power to optimal sample allocation based on misspecified parameters values analytically. We found the analytic expressions to be complex and unable to provide clear, intuitive, and descriptive context for the properties of statistical power robustness in these settings when compared to our explicit comparisons. We provide an example analytic analysis in ESM 1.

While the direct effect or c' is also included in the optimal sample allocation formulas it has little influence on n_1^{opt} and relative power. It is held constant at $c' = 0.1$ throughout the analyses and not discussed further (see ESM 1). Additionally, the total funds (T) do not play a role in determining the optimal individual sample size so T was set as a function of group cost ($T = 100c_2$; see Authors, A).

Results

Results are presented by misspecified parameter (*a, b,* and ρ) with a focus on the general patterns involving robustness of power to capture the essences of the analysis. The complete results for the fully crossed design can be found in ESM 1, Tables 2, 3, and 4.

Incorrect Initial Intraclass Correlation Coefficient

We examined power when n_1^{opt} was determined using ρ values ranging from .15 to .9 with true population values of ρ* = .3, .4, and .5. Overall, power was very robust to incorrect ρ values under stated conditions. With ρ from 50% to 150% of ρ*, relative power remained greater than .9 in nearly every condition (see Table 1 for an example of complete results when ρ* = .5). This indicates designs with a misspecified ρ value suffered less than a 10% loss in power due to inefficient sample allocation.

A few noteworthy patterns emerged from these results. First, larger a/b path coefficient ratios decreased the robustness of power to incorrect ρ values. Second, over- and underestimated ρ values had similar influence on relative power. In other words, the loss of power associated with using incorrect ρ values to determine n_1^{opt} was fairly symmetric. This symmetric power loss was consistent across each test. Lastly, when the group-to-individual cost ratio (c_2/c_1) was small, power was more susceptible to inefficient sample allocation from misspecified ρ values (see Figure 2B). As c_2/c_1 increased, designs became more robust with relative power remaining well above .9 across a wide range of predicted ρ values in designs with the highest cost ratio.

Incorrect Initial *a* Path Coefficient

Our examination found relative power when n_1^{opt} was determined using *a* values ranging from .05 to .9 with true population values of *a** = .3, .5, and .6. Overall, power was very robust to incorrect initial *a* path coefficient values under stated conditions. When *a* ranged from 50% to over 150% of *a**, relative power remained greater than .9 in almost every condition. While results are similar, power was generally more robust when misspecified *a* values were used to determine n_1^{opt} and less robust using misspecified ρ values.

Effects of misspecified *a* values (see Figure 3B) varied by mediation test with designs using the MC interval test providing the greatest degree of robustness followed closely by the joint test. Relative power in designs utilizing these tests remained above .95 across nearly the full range of *a* values. One exception were the reductions in relative power under the joint test when *a* overestimated *a**. The inverse was true when using the Sobel test, inefficient sample allocation caused by *a* values that underestimated *a** were more detrimental to power.

Similar to our examination of incorrect ρ values, power was more robust to incorrect *a* values when the design had larger (c_2/c_1) values (see Figure 3B). Additionally, power in designs with greater ρ values was more robust to inefficient sample allocation based on incorrect *a* values.

Incorrect Initial *b* Path Coefficient

We examined power when n_1^{opt} was determined with *b* values from .05 to .5 and true population values of *b** = .2 and .3. Like previous results, power under the optimal sampling framework was robust to incorrect *b* values. When *b* was anywhere from 50% to 150% of *b** relative power remained above .9.

Overall, c_2/c_1 and ρ values did little to influence the robustness of power when incorrect *b* values were used to determine n_1^{opt} but we again found relative power varied by test (see Figure 4B). Relative power using the Sobel test and MC interval test was similar with notable differences occurring mostly in designs with smaller ρ values and larger c_2/c_1 values. Under those conditions power using the Sobel test decreased at a greater rate when the *b* value used to determine n_1^{opt} underestimated *b**. Designs using the joint test also demonstrated robustness to incorrect *b* values but when *b* significantly overestimated *b** inefficient sample allocation caused power rates to fall sharply.

Individual-level Mediator

Analytic Model

We next consider a group-randomized design with an individual-level mediator. Here, we again have intact groups assigned to a treatment condition (T) and an individual-level outcome (Y) but the relationship under examination flows through an individual-level mediator (M). For this design parameters are also estimated using maximum likelihood but a multilevel model is needed for the mediator such that (Authors, B; Pituch & Stapleton, 2012;

Table 1. Relative power of group-randomized studies of 2-2-1 mediation with misspecified intraclass correlation coefficients when ρ^*=.5

Conditions			ρ							
c_2:c_1	*a*:*b*	Test	.2	.3	.4	.5*	.6	.7	.8	.9
		S	.993	.997	.999	1.000	.999	.993	.978	.935
	0.3:0.3	JT	.986	.992	.998	1.000	.996	.981	.943	.840
		MC	.976	.955	1.006	.980	1.000	1.012	.918	.828
		S	.986	.993	.998	1.000	.997	.988	.963	.898
5	0.5:0.3	JT	.986	.992	.998	1.000	.996	.982	.945	.848
		MC	.993	.995	1.011	1.006	.994	1.005	.952	.848
		S	.966	.985	.996	1.000	.995	.979	.942	.854
	0.6:0.2	JT	.937	.973	.993	1.000	.993	.969	.919	.809
		MC	.923	.944	.985	1.004	.992	.984	.943	.786
		S	.995	.997	.999	1.000	.999	.994	.983	.948
	0.3:0.3	JT	.996	.997	.999	1.000	.998	.989	.963	.885
		MC	1.025	1.014	1.042	.997	1.031	1.015	.987	.879
		S	.989	.995	.999	1.000	.998	.990	.971	.918
10	0.5:0.3	JT	.997	.998	.999	1.000	.998	.990	.965	.893
		MC	1.003	1.009	1.002	1.012	.991	.985	.983	.876
		S	.972	.988	.997	1.000	.996	.983	.952	.876
	0.6:0.2	JT	.953	.979	.995	1.000	.994	.974	.930	.830
		MC	.955	.993	1.006	1.016	1.014	.997	.974	.839
		S	.998	.999	1.000	1.000	1.000	.998	.993	.980
	0.3:0.3	JT	.997	.999	1.000	1.000	1.000	1.000	.997	.984
		MC	1.002	1.031	1.011	1.051	1.017	1.023	1.023	1.006
		S	.996	.998	.999	1.000	.999	.996	.989	.969
100	0.5:0.3	JT	.996	.999	1.000	1.000	1.000	1.000	.997	.985
		MC	1.012	1.002	.999	1.003	1.018	1.008	1.003	.991
		S	.989	.995	.999	1.000	.998	.993	.980	.945
	0.6:0.2	JT	.996	.995	.998	1.000	.998	.990	.972	.926
		MC	1.008	1.006	1.008	1.015	.998	1.017	.988	.949

Notes. S = Sobel; JT = Joint, MC = Monte-Carlo. *True population parameter value. Full results with an interval of 0.025 between incorrect initial values are available in ESM 1, Table 2.

Raudenbush & Bryk, 2002; VanderWeele, 2010; Zhang et al., 2009)

$$M_{ij} = \pi_{0j} + \varepsilon^M_{ij}\ \ \varepsilon^M_{ij} \sim N\left(0, \sigma^2_M\right),$$
$$\pi_{0j} = \zeta_{00} + aT_j + u^M_{0j}\ \ u^M_{0j} \sim N\left(0, \tau^2_{M|}\right), \tag{17}$$

with a combined mediator model formulation expressed as

$$M_{ij} = \zeta_{00} + aT_j + u^M_{0j} + \varepsilon^M_{ij}, \tag{18}$$

where M_{ij} represents the mediator value for individual i in group j, T_j as the treatment assignment coded as 0.5 and −0.5 for the treatment and control condition, respectively with associated path coefficient a, ε^M_{ij} as the normally distributed error term with a mean of zero and variance σ^2_M, and u^M_{0j} as the group-specific random effects that follow a normal distribution with mean zero and variance $\tau^2_{M|}$ conditional on T_j. Conceptually, a maps out how exposure to the treatment produces changes in the individual-level mediator. For the 2-1-1 design, the outcome model is

$$Y_{ij} = \beta_{0j} + b_1\left(M_{ij} - \bar{M}_j\right) + \varepsilon^Y_{ij}\ \ \varepsilon^Y_{ij} \sim N\left(0, \sigma^2_{Y|}\right),$$
$$\beta_{0j} = \gamma_{00} + B\bar{M}_j + c'T_j + u^Y_{0j}\ \ u^Y_{0j} \sim N\left(0, \tau^2_{Y|}\right), \tag{19}$$

with the combined outcome model formulation expressed as

$$Y_{ij} = \left(\gamma_{00} + B\bar{M}_j + c'T_j + u^Y_{0j}\right) + b_1\left(M_{ij} - \bar{M}_j\right) + \varepsilon^Y_{ij},$$
$$Y_{ij} = \gamma_{00} + (B - b_1)\bar{M}_j + b_1 M_{ij} + c'T_j + u^Y_{0j} + \varepsilon^Y_{ij}. \tag{20}$$

We retain Y_{ij} as the outcome for individual i in group j, and use $M_{ij} - \bar{M}_j$ as the group-centered individual-level mediator with coefficient b_1, $\bar{M}_j$ as the mean of the mediator in group j with path coefficient B, c' as the treatment-outcome conditional path coefficient, and u^Y_{0j} and ε^Y_{ij} as

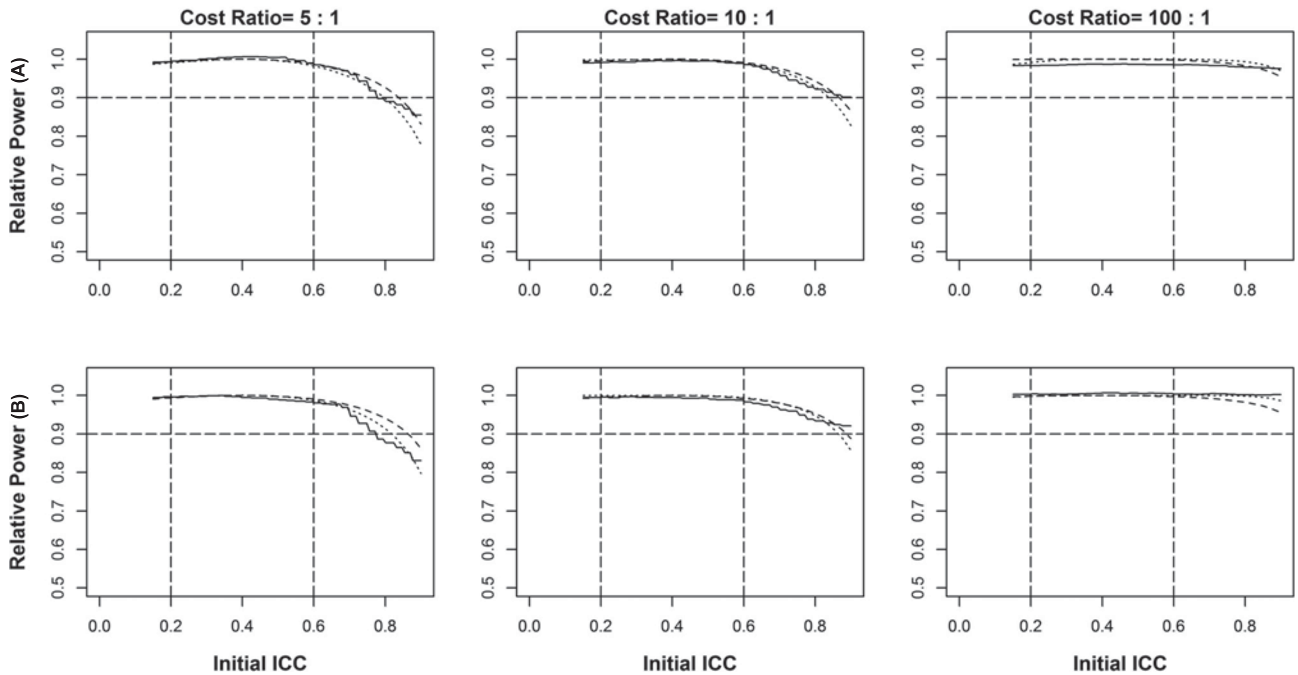

Figure 2. Relative power of group-randomized studies of 2-1-1 (A) and 2-2-1 (B) mediation using the Sobel (dash), joint (dot), and MC interval test (solid) as a function of incorrect intraclass correlation coefficient values with different group to individual cost ratios and a total budget of 100 times the cost ratio when a = .5, b (or B) = .3, c' = 1, and the true intraclass correlation value is a .4. Vertical lines mark incorrect initial parameter values that are 50% and 150% of the true value and a horizontal line marks the 90% relative power benchmark.

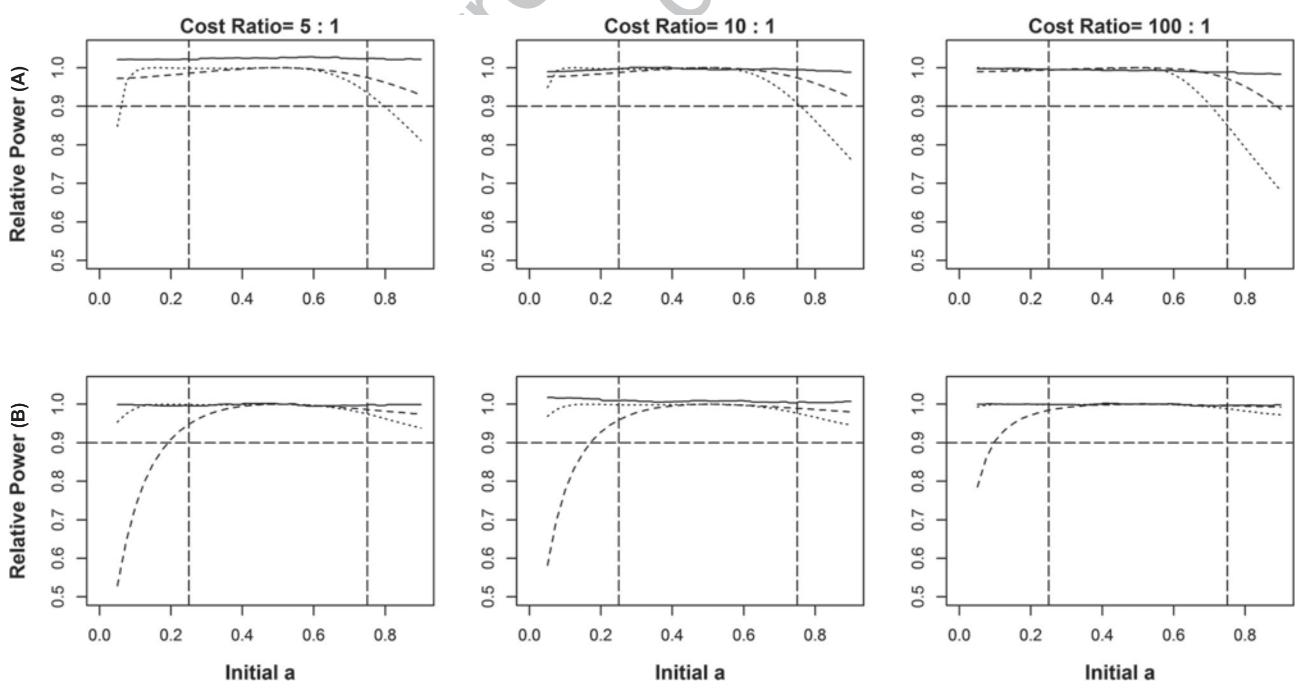

Figure 3. Relative power of group-randomized studies of 2-1-1 (A) and 2-2-1 (B) mediation using the Sobel (dash), joint (dot) and MC interval test (solid) as a function of incorrect a path coefficient values with different group to individual cost ratios and a total budget of 100 times the cost ratio when b (or B) = .3, c' = 1, all intraclass correlation values are .5, and the true a path coefficient value is a^* = .5. Vertical lines mark incorrect initial parameter values that are 50% and 150% of the true value (a^*) and a horizontal line marks the 90% relative power benchmark.

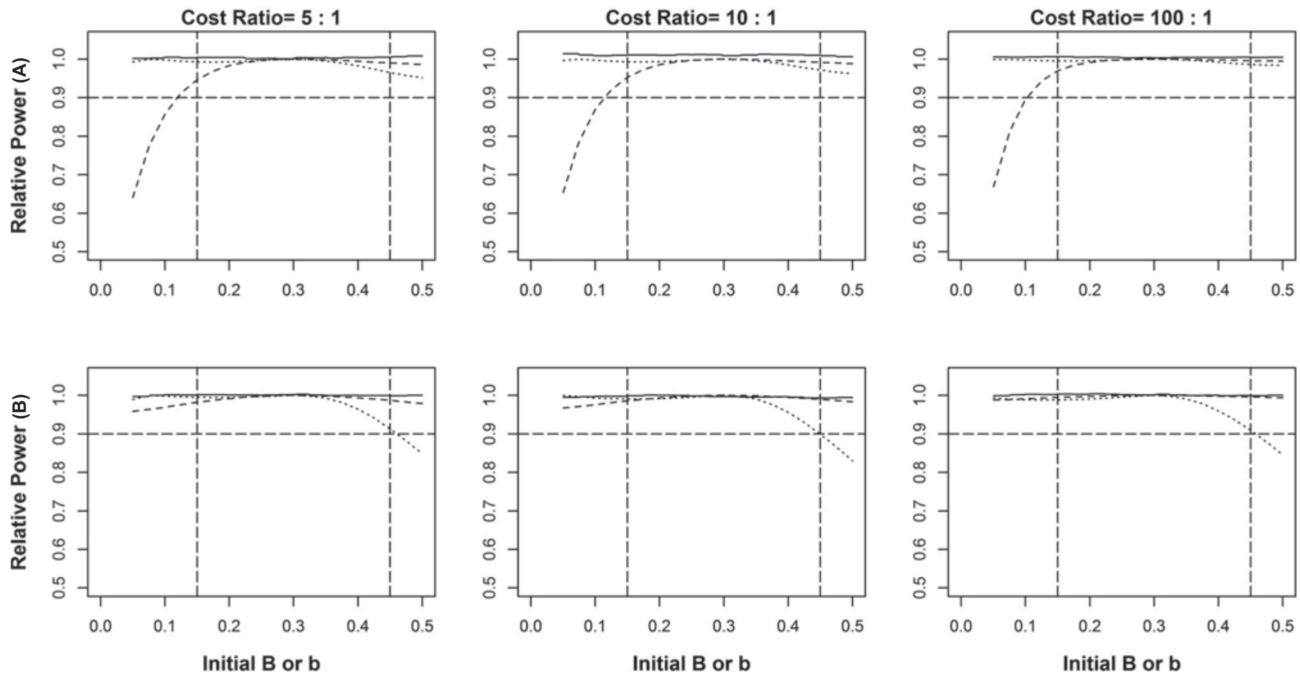

Figure 4. Relative power of group-randomized studies of 2-1-1 (A) and 2-2-1 (B) mediation using the Sobel (dash), joint (dot) and MC interval test (solid) as a function of incorrect *b* (or *B*) path coefficient values with different group to individual cost ratios and a total budget of 100 times the cost ratio when *a* = .5, *c′* = 1, all intraclass correlation values are .5, and the true b (or B) path coefficient value is 0.3. Vertical lines mark incorrect initial parameter values that are 50% and 150% of the true value (*b** or *B**) and a horizontal line marks the 90% relative power benchmark.

the level-two and level-one error terms. Both error terms are normally distributed with variances of $\sigma^2_{Y|}$ conditional on $(M_{ij} - \bar{M}_j)$ for ε^Y_{ij} and $\tau^2_{Y|}$ conditional on $\bar{M}_j$ and T_j for u^Y_{0j}. Conceptually, the mediator-outcome relationship occurring at the individual-level is represented by b_1 and at the group-level by $(B - b_1)$ with B representing the total individual- and group-level relationship. A combined mediator and outcome model for 2-1-1 mediation can be expressed as

$$Y_{ij} = (\gamma_{00} + b_1\zeta_{00}) + (B - b_1)\bar{M}_j + \left(b_1 a + c'\right)T_j + b_1 u^M_{0j} + b_1\varepsilon^M_{ij} + u^Y_{0j} + \varepsilon^Y_{ij}, \quad (21)$$

which allows a clear comparison of the 2-2-1 and 2-1-1 analytic models.

In group-randomized studies of 2-1-1 mediation, there are two intraclass correlation coefficients. The first, ρ_Y, corresponds to the ρ described for group-randomized studies of 2-2-1 mediation and captures the correlation among individuals on the outcome within the same group. The second, ρ_M, is introduced as a result of the multilevel nature of the mediator model in the 2-1-1 design. It captures the correlation among individuals on the mediator within the same group and we formulate the unconditional intraclass correlation coefficient of the mediator as

$$\rho_M = \frac{\tau^2_M}{\sigma^2_M + \tau^2_M}. \quad (22)$$

We use a standardized outcome and mediator similar to that described for ρ under the 2-2-1 model resulting in similar scaling for the a, c', b_1, and B paths (i.e., a and c' as standardized group differences and b_1 and B as standardized regression coefficients).

Our utilization of group-mean centering operationalizes B to capture the total influence of the mediator on the outcome. We focus on this cumulative or overall mediation estimated using the typical product of coefficients method: ME = aB (Pituch & Stapleton, 2012, VanderWeele, 2010; VanderWeele & Vansteelandt, 2009). The 2-1-1 model also captures the relationship between the treatment and outcome as it operates through a mediator at the group- $(B - b_1)$ and individual-level (b_1). The $a(B - b_1)$ estimate of the mediated effect in the 2-1-1 model is conceptually equivalent to the ab mediated effect of the 2-2-1 model (Pituch & Stapleton, 2012).

Power and Optimal Sample Allocation

The general form of the mediation tests and subsequent power formulations presented for the 2-2-1 design are similar for 2-1-1 design. The error variances associated with each path and the error variance of the mediated effects do change substantially but formulations utilizing path coefficients, budget, and sampling cost are available in the literature along with related optimal sample allocation formulations (see ESM 1; Authors, B; Authors C).

Unlike 2-2-1 designs, there is no simple closed form solution for n_1^{opt} under the Sobel test. We can, however, determine n_1^{opt} for the Sobel test numerically (see ESM 1; Authors, C **[Author: Please provide full reference details]**). Similar processes are used to determine n_1^{opt} under the joint test for 2-1-1 and 2-2-1 designs. Using the cost function formulations for n_1 and the error variance path formulations, we directly maximize the power function (see ESM 1; Authors, C). The n_1^{opt} values from the Sobel and joint test formulations approximate those of the MC interval test for 2-1-1 designs but we again use a linear search algorithm to identify n_1^{opt} values for the MC interval test to avoid possible confounding effects.

Power of 2-1-1 Mediation Studies

We now repeat our investigation but for group-randomized studies of 2-1-1 mediation. For this design, we investigate the robustness of power against misspecified a, B, ρ_M, and ρ_Y values in the n_1^{opt} formulation. Procedures and conditions are similar to those described for the 2-2-1 design with the B parameter assuming the same values as those assigned to b, and the unconditional intraclass correlation coefficient of the mediator (ρ_M) and outcome (ρ_Y) assuming the same values as those assigned to ρ. Results from an initial simulation with differing values for the misspecified values were not qualitatively different from those presented below so we constrained ρ_M and ρ_Y to be equal. That is, $\rho_M{}^* = \rho_Y{}^*$ when we set the population values and $\rho_M = \rho_Y$ as we varied the initial intraclass correlation coefficient values. The initial simulation did suggest that the ρ_Y value is likely the more influential parameter in terms of relative power. Complete results for the 2-1-1 investigation can be found in ESM 1, Tables 2, 3, and 4.

Results

Incorrect Initial Intraclass Correlation Coefficients

We examined power when n_1^{opt} was determined using ρ_M and ρ_Y values ranging from .15 to .9 with true population values of $\rho_M{}^*$ and $\rho_Y{}^*$ = .3, .4, and .5. Overall, power was very robust to incorrect ρ_M and ρ_Y values under stated conditions. In nearly every situation, relative power remained above .9 when ρ_M and ρ_Y ranged from 50% to nearly 200% of $\rho_M{}^*$ and $\rho_Y{}^*$.

Over- or underestimating $\rho_M{}^*$ and $\rho_Y{}^*$ led to similar amounts of lost power due to inefficient sample allocation with power loss slightly greater in designs using the joint test. Power was more robust to n_1^{opt} based on misspecified parameter values when the design had higher sampling cost ratios (i.e., c_2/c_1; see Figure 2A) or larger path coefficient ratios (i.e., a/b). The influence of the path coefficient ratio was less pronounced when using the joint test. Additionally, higher c_2/c_1 muted the benefits of larger path coefficient ratios.

Incorrect Initial *a* Path Coefficient

Our examination found relative power when n_1^{opt} was determined using a values ranging from 0.05 to 0.9 with true population values of a^* = .3, .5, and .6. Overall, power was robust to misspecified a values that led to inefficiencies in sample allocation. Across the different mediation tests and conditions, a values ranging from 50% to 150% of a^* almost always maintained relative power above .8. While robust when compared to our 80% relative power benchmark, misspecifying a values when determining the n_1^{opt} for a group-randomized 2-1-1 mediation study led to the largest reductions in power across the conditions and designs considered here.

Results indicated that power decreased at a smaller rate when a underestimated a^* in the determination of n_1^{opt}. Conversely, power rates were less robust when a overestimated a^* when determining n_1^{opt}, especially in designs that utilized the joint test (see Figure 3A). This relationship became stronger in designs with larger ρ_M and ρ_Y values and a/b ratios. Misspecified a values in the n_1^{opt} formulation had similar detrimental effects on power across designs with different a/b ratios but effects varied across designs with different ρ_M and ρ_Y values. Here, relative power was greater in designs with larger ρ_M and ρ_Y values.

For the a parameter, we found a unique relationship between relative power and c_2/c_1. Power was more robust to misspecified a values in the n_1^{opt} formulation, when 2-1-1 designs had smaller c_2/c_1 ratios (see Figure 3A). In all other conditions and with different parameters, the converse was true.

Incorrect Initial *B* Path Coefficient

We examined power when n_1^{opt} was determined with B values ranging from 0.05 to 0.5 and true population values of B^* = 0.2 and 0.3. Following previous results, power under the optimal sample allocation framework was very robust to incorrect B values under stated conditions. Relative power remained at or above .9 across tests and under nearly every condition when B was 50–150% of B^*.

When using the joint test or MC interval test, power was even more robust to misspecified B values in the n_1^{opt} formulation and over a greater range of B values. However, in designs that used the Sobel test, power decreased rapidly due to inefficient sample allocation when B was less than B^*. Larger c_2/c_1, ρ_M and ρ_Y values did offset some of the detrimental effects of incorrect B values on Sobel test power. Designs using the joint or MC interval test had such high relative power across the range of B values that their parameter-power relationships were practically concealed (see Figure 4A).

Influence of Mediation Test and Analytic Model

Results revealed several differences in the robustness of power across analytic model, mediation test, and type of misspecified parameter value which deserve some attention. First, the overall robustness of power under an optimal sampling framework is less surprising when considering n_1^{opt} is typically small and relatively stable across these models and the typical conditions included in this study (Author A, Author C). Put differently, n_1^{opt} is often small (e.g., < 10) across a variety of typical conditions and true parameter values. Therefore, using misspecified parameter values, even with large discrepancies, still results in ~~a~~ small n_1^{opt} values, likely < 10. Because the actual n_1^{opt} value remains fairly stable, it is reasonable to find only minor change in power rates. Additionally, previous research has shown that study design conditions with less stable n_1^{opt} values and therefore an increased likelihood of large differences between n_1^{opt} and n_1^{opt*}, are the same conditions in which these differences have little influence on power (Author A, Author C). In other words, when minor changes in parameter values influence n_1^{opt} values, deviations from n_1^{opt} do not influence power and under conditions where deviations from n_1^{opt} are detrimental to power, n_1^{opt} values tend to be so similar (e.g., < 10) power is not substantially influenced.

Another consistent result across models and tests was the influence of unconditional intraclass correlation coefficient values (see Figure 2). The consistent influence of ρ (or ρ_M and ρ_Y) on the robustness of power under the optimal sampling framework is directly related to the consistent influence of ρ on n_1^{opt} values (Author A, Author C). If the ρ value is wrong (i.e., misspecified) then it is likely we have poorly estimated n_1^{opt} resulting in inefficiencies and a substantial loss of statistical power. This is especially true in designs with larger ρ values but less so in designs with large c_2/c_1 ratios.

Relative power did vary across models and tests when the path coefficients were misspecified with the exception of the MC interval test for which power demonstrated consistent robustness. The MC interval test result can be traced back to stable n_1^{opt} values when using the test even with different a and b path coefficient values (Author A, Author C). Conversely, power under the optimal sampling framework for the joint and Sobel test varied in robustness to misspecified a and b path coefficient values across models reflecting the conditions in which these parameters influenced n_1^{opt} values for that specific test. For example, in a group-randomized study of 2-2-1 mediation using the joint test, changes to the b path coefficient resulted in substantial changes to n_1^{opt} values suggesting that misspecified b path coefficient values will influence optimal sample allocation and therefore power (Authors A).

Illustrative Example

Pulling from an investigation of the Every Classroom, Everyday (ECED) program by Early et al. (2016), we illustrate the process described above. ECED aims to improve student academic achievement, an individual-level outcome, through improvements in teaching practice and curriculum alignment. The ECED program represents a group-level treatment as it is implemented across whole schools. For our illustration, we also include a group-level mediator that captures the degree of program implementation within a school, a crucial factor in the success of these programs (Desimone, Porter, Garet, Yoon, & Birman, 2002). We now have a two-level group-randomized study examining a 2-2-1 mediation effect. To test the significance of this effect, we employ the joint test and for study planning purposes predict the relationship between the ECED program and our measure of program implementation is $a = 0.5$, the relationship between program implementation and student outcomes is $b = 0.2$, the conditional direct effect of the ECED program on student outcomes is $c' = 0.1$, the correlation of student outcomes within a school is $\rho = 0.3$, with a budget of T = US$175,000 and a school to student sampling cost ratio of $c_2/c_1 = 2{,}000$. Under these conditions, the optimal individual sample size is 38 students per school with a sample of 86 schools (i.e., $n_1^{opt} = 38$ and $n_2^{opt} = 86$). However, in the empirical results of our hypothetical study the true a path value was $a^* = 0.65$ indicating we underestimated the a path during study planning and therefore did not use the most efficient sample. The true optimal design would have included a sample of 48 students per school in 85 schools (i.e., $n_1^{opt*} = 48$ and $n_2^{opt*} = 85$). As conducted, power to detect the 2-2-1 mediated effect was $\approx$ 81.99% using the incorrect $n_1^{opt} = 38$. If we had perfectly predicted parameters during study planning and employed the true optimal design, study power would have been $\approx$ 82.07%. Mirroring our results above, the misspecified a value influenced n_1^{opt} but the inefficiencies were inconsequential to the relative power of the study.

Discussion

Group-randomized studies of mediation effects probe the mechanisms presumed to operate within treatment conditions that are implemented in extant hierarchical structures.

These studies can be efficiently planned by utilizing optimal sample allocation to identify the sample of individuals per group and sample of groups that maximizes power while considering costs. Determining the optimal sample allocation depends on the estimation of several parameters during study planning and these estimates are likely to deviate from the true population values observed once data are collected. To understand the consequences of these deviations, we examined the robustness of power under an optimal sampling framework to parameter value misspecification in group-randomized studies of mediation. We found power rates in 2-2-1 and 2-1-1 designs to be robust to misspecified parameters ranging from 50% to 150% of their true value although results varied by mediation test, design cost structure, path coefficient values, and the unconditional intraclass correlation coefficient values.

For example, across all conditions group-randomized studies of 2-1-1 mediation were the most susceptible to power loss when n_1^{opt} was identified with a misspecified *a* path coefficient value. In group-randomized studies of 2-2-1 mediation, utilizing a misspecified ρ value when determining n_1^{opt} was the most detrimental to power. In these conditions, researchers need stronger theoretical and empirical guidance to predict parameter values in order to ensure accurate power analyses. Conversely, in group-randomized studies of 2-1-1 mediation misspecified ρ_M and ρ_Y values had the least detrimental influence on power. Minimal theoretical and empirical guidance is sufficient to estimate these parameter values because inaccuracies will have only minor consequences to subsequent power analyses.

An extension of this implication is a call to prioritize investigations and empirically based collections (e.g., Hedges & Hedberg, 2007) of those parameters that are crucial to accurate power analyses. If the theoretical and empirical support to accurately predict a parameter is lacking in a substantive area, the scope of study design possibilities is limited. Given the utility and feasibility of group-randomized studies of multilevel mediation this presents a serious limitation to research.

This study, like all simulation studies, is limited by the number and combination of factors that can be manipulated and examined. We are confident in the robustness of power under the optimal sampling framework to misspecified parameter values when detecting mediated effects in 2-2-1 and 2-1-1 group-randomized studies, but this is a relatively narrow set of models and conditions. The broader takeaway from this investigation is the blueprint for considering the robustness of power under the optimal sampling framework in studies of multilevel mediation. The process involves identifying the two statistical power rates used to determine relative power (see Equation 16). First, researchers identify the optimal sample allocation and power to detect the effect in question using the most plausible parameter estimates available. This power rate can serve as the true or population study power (i.e., denominator in Equation 16. Second, the researcher considers alternative parameter values to determine another optimal sample allocation and uses this sample with the original (i.e., most plausible) parameter values to conduct a second power analysis. This second power analysis serves as the initial or incorrect study power (i.e., numerator in Equation 16). Results from such an undertaking provide researchers with a range of possible sample allocations, power under each sample allocation, and at least some notion of the consequences to study power when employing an optimal sampling scheme based on misspecified values. Researchers conducting their own simulation ensure maximum conformity to their main study and allow interpretation of results in a study-specific context. For example, simulation results indicating minor power loss from employing an inefficient sample allocation is of little concern if the study design is well powered (e.g., > 90%) but the same power loss in a design with barely adequate power (e.g., ≈ 80%) can be practically significant.

Effective studies utilize sample sizes that ensure adequate power and efficient studies make proper use of limited resources. The optimal sample allocation framework is an excellent means to investigate study power and maximize available resources. That said, we caution readers to examine their study-specific factors when applying our results or conducting their own simulation and remind them that optimal sampling strategies provide a theoretical guide rather than a strict set of rules.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at https://doi.org/10.1027/1614-2241/a000169

ESM 1. Text, Tables, Figures (.docx)

This document contains supplemental derivations, tables of complete results, and additional figures.

References **[Author: Please provide DOIs for journal references]**

Authors. (A) **[Author: Please provide full reference details].**

Authors. (B) **[Author: Please provide full reference details].**

Authors. (C) **[Author: Please provide full reference details].**

Authors. (D) **[Author: Please provide full reference details].**

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis, 24*, 81–112.

Early, D. M., Berg, J. K., Alicea, S., Si, Y., Aber, J. L., Ryan, R. M., & Deci, E. L. (2016). The impact of every classroom, every day on high school student achievement: Results from a school-randomized trial. *Journal of Research on Educational Effectiveness, 9*, 3–29.
Gottfredson, D. C., Cook, T. D., Gardner, F. E. M., 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*, 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*, 1918–1927.
Hedges, L., & Borenstein, M. (2014). Conditional optimal design in three- and four-level experiments. *Journal of Educational and Behavioral Statistics, 39*, 1–25.
Hedges, L., & Hedberg, E. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60–87.
Institute of Education Sciences, US Department of Education, & National Science Foundation. (2013). *Common Guidelines for Education Research and Development (NSF 13–126)*. Retrieved from http://ies.ed.gov/pdf/CommonGuidelines.pdf
Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017). Statistical power for causally-defined mediation in group-randomized studies. *Multivariate Behavioral Research*. Advance online publication. **[Author: please update reference]**
Kelcey, B., & Phelps, G. (2013). Considerations for designing group randomized trials of professional development with teacher knowledge outcomes. *Educational Evaluation and Policy Analysis, 35*, 370–390.
Kelcey, B., & Phelps, G. (2014). Strategies for improving power in school randomized studies of professional development. *Evaluation Review, 37*, 520–554.
Kelcey, B., Phelps, G., Spybrook, J., Jones, N., & Zhang, J. (2017). Designing large-scale multisite and cluster-randomized studies of professional development. *Journal of Experimental Education, 85*, 389–410.
Korendijk, E. J. H., Moerbeek, M., & Maas, C. J. M. (2010). The robustness of designs for trials with nested data against incorrect initial intracluster correlation coefficient estimates. *Journal of Educational and Behavioral Statistics, 35*, 566–585.
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*, 77–98.
Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173–185.
Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.
Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological Methodology, 13*, 290–312.
Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the institute of education sciences. *Educational Evaluation and Policy Analysis, 31*, 298–318.
VanderWeele, T. J. (2010). Direct and indirect effects for neighborhood-based clustered and longitudinal data. *Sociological Methods & Research, 38*, 515–544.
VanderWeele, T. J., & Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and Its Interface, 2*, 457–468.
Zhang, Z., Zyphur, M. J., & Preacher, K. J. (2009). Testing multilevel mediation using hierarchical linear models: Problems and solutions. *Organizational Research Methods, 12*, 695–719.

History
Received May 3, 2018
Revision received February 13, 2019
Accepted April 15, 2019
Published online XX, 2019

Funding
This article is based on work funded by the National Science Foundation [#1552535] to Benjamin Kelcey. The opinions expressed herein are those of the authors and not the funding agency.

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

Kyle Cox
Quantitative and Mixed Methods Research Methodologies
University of Cincinnati
Teachers/Dyer Hall
Cincinnati, OH 45221
USA
coxk5@mail.uc.edu

Kyle Cox is a doctoral student in the Quantitative Research Methodologies Program at the University of Cincinnati. His research interests include experimental and quasi-experimental design, multilevel mediation, and their application in educational settings.

Benjamin Kelcey is an associate professor in the Quantitative Research Methodologies Program at the University of Cincinnati. His research interests are causal inference and measurement methods within the context of multilevel and multidimensional settings such as classrooms and schools.

