Optimal Sample Allocation in Group-Randomized Mediation Studies with a Group-level Mediator

Abstract

We derive sample allocation formulas that maximize the power of several mediation tests in two-level group-randomized studies under a linear cost structure and fixed budget. The results suggest that the optimal individual sample size is typically smaller than that associated with the detection of a main effect and is frequently less than 10 under parameter values commonly seen in the literature. However, the optimal sample allocation can be heavily influenced by the group-to-individual cost ratio, the ratio of the treatment-mediator to mediator-outcome path coefficients, and the outcome variance structure. We illustrate these findings with a hypothetical group-randomized trial examining a school discipline reform policy and conclude with discussion of results. To encourage utilization of the sample allocation formulas we implement them in the *R* package [masked for blind review] and the [masked for blind review] software.

Keywords: mediation, optimal sample allocation, optimal design, group-randomized studies, multilevel models

Literature has consistently emphasized the advantages of randomized experiments when assessing the effects of treatments, interventions, programs, or services of interest (e.g.,Raudenbush, 1997; Schneider, Carnoy, Kilpatrick, Schmidt, & Shavelson, 2007; Shadish, Cook, & Campbell, 2002). When studies involve populations in a hierarchical structure (e.g., students in schools) the randomization of individuals is sometimes impractical. Researchers can often mitigate concerns and align study design with theory by assigning entire clusters or groups to a treatment condition. Group-randomized studies designed around extant hierarchical structures are often logistically more efficient, ebb ethical concerns, and produce results that better generalize to the policy contexts they are intended to inform (Raudenbush, 1997; Spybrook & Raudenbush, 2009).

Determining power, or the probability of detecting a treatment effect when one truly exists, in a group-randomized experiment is driven by the sample sizes (at each level), effect size, error variance of the treatment effect, outcome variance decomposition, and, if included, variance explained by prognostic covariates (e.g., Spybrook, Shi, & Kelcey, 2016; Kelcey & Shen, 2016). Sample size allocation (i.e., the balance between the number of groups and number of individuals per group) is an important consideration in power analyses for group-randomized studies. Sample size is a primary driver of statistical power and, unlike many of the other factors influencing power, is often within the control of researchers. When power is a concern, group-randomized studies typically privilege the number of groups over the number of individuals per group because it yields greater gains in efficiency. However, this strategy is practically limited by the additional costs incurred by sampling higher level clusters (Raudenbush, 1997; Kelcey & Phelps, 2013; 2014). Optimizing sample allocation balances these concerns by identifying the

most efficient sample of groups and individuals within a limited budget—that is, optimal sample allocation focuses on identifying sampling strategies that yield the maximum power for a given design and budget (e.g., Kelcey, Phelps, Spybrook, Jones, & Zhang, 2017).

Group-randomized studies designed to detect the effect a treatment has on an outcome (i.e., main or total effect) provide valuable evidence concerning 'what works' but there is growing recognition of the need to more comprehensively investigate the theory guiding the intervention by probing the mechanisms through which the treatment is presumed to work (IES, US DoE, NSF, 2013; Raudenbush & Sadoff, 2008; Kelcey, Dong, Spybrook, & Cox, 2017). The prevailing approach to examine these mechanisms is to establish a sequence of structural relationships through a mediation analysis (Baron & Kenny, 1986; Imai, Keele, & Tingley, 2010). Mediation analyses examine the changes in an outcome produced by exposure to a treatment as they operate through an intermediate or mediating variable. Statistically, mediation is assessed using the indirect effect of the treatment on the outcome via a mediator. These analyses enable researchers to investigate, for example, how an intervention worked, if it worked and differentiate between theory and implementation failure, if it failed.

Despite the value of mediation analyses, the utility of group-randomized studies, and the practicality of optimal design, sparse literature is available examining the intersection of these components (VanderWeele, 2015 ; Hox, Moerbeek, Kluytmans, & van de Schoot, 2014; Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017). This limitation constrains the scope, quality, and efficiency of designs available to researchers for studies of multilevel mediation because optimal and general guidelines regarding sample allocation in studies of multilevel mediation are entirely unclear and unavailable. To address this problem, we

derive optimal sample allocation formulas for group-randomized studies designed to detect multilevel mediation. We investigate mediation in two-level group-randomized studies with a group level mediator and an individual level outcome (i.e., 2-2-1 mediation) within a multilevel linear path model. Similar to prior literature, we define optimal sample allocation as the mix of individual- and group-level sample sizes that maximizes power under a fixed budget and cost structure (e.g., Raudenbush, 1997). In what follows, we outline power formulas to detect multilevel mediation effects under three design-centered tests: the Sobel, joint, and Monte Carlo confidence interval tests (MC interval test). Next, we derive formulas for optimal sample allocation under a linear cost structure framework for group-randomized studies of 2-2-1 mediation and probe the roles of the governing parameters. We illustrate the application of the optimal sample allocation formulas with an example involving a school discipline reform policy and conclude with a brief discussion of results and recommendations.

Multilevel Mediation

Our analyses examine two-level group randomized designs in which groups are assigned to one of two treatment conditions (T). The multilevel mediation model captures the indirect effect of the treatment on an individual-level outcome (Y) as it passes through a group-level mediator (M). Later we extend this model to include individual-level covariates, their group-level means, and group-level covariates. We draw on the typical multilevel linear path formulation (Krull & MacKinnon, 2001) such that

$$\textit{Mediator model}\ \text{(Level 2)} \qquad M_j = \pi_0 + aT_j + \varepsilon_j \qquad \varepsilon_j \sim N(0, \sigma^2_{M|}) \tag{1}$$

$$\textit{Outcome model}\ \text{(Level 1)} \qquad Y_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2_Y) \tag{2a}$$

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

The mediation equation (i.e., Equation 1) represents the treatment-mediator relationship and describes variation in the group-level mediator as a function of group-level variables. We use M_j as the mediator for group *j*, T_j as the treatment assignment coded as ±½ with associated coefficient *a*, and ε_j as the error term that is assumed to follow an independent normal distribution with a mean of zero and variance $\sigma^2_{M|}$. The outcome equation (i.e., Equations 2a and 2b) represents the mediator-outcome relationship such that Y_{ij} is the outcome for individual *i* in group *j*. The error term for level-1 is represented by e_{ij} and is assumed to be normally and independently distributed with a mean of zero and variance σ^2_Y. At the group-level, we introduce *b* as the conditional relationship between the mediator and the outcome, $c^{'}$ as the direct effect of the treatment. and u_{0j} as the group-specific random effects that are assumed to follow an independent normal distribution with a mean of zero and variance $\tau^2_{Y|}$. Like prior literature, we assume sequential ignorability and the typical recursive model structure and subsequently extend this model structure to condition on covariates (e.g., Allison, 1995). Further, we assume variables are fully reliable and caution readers that failing to meet this assumption does influence, power, sample size requirements, and accuracy of parameter estimates (Li and Beretvas, 2013).

Test Statistics and Power

Our derivations focus on the indirect effect of the treatment on the outcome through the mediator as captured by the product of the coefficients (ab) method under maximum likelihood estimation (MacKinnon, 2008). In this section, we introduce three tests to assess the statistical

significance of the mediation effect (i.e., ab): the Sobel test, the joint test, and the MC interval test. These tests represent a range of popular methods and each can be employed in the context of study design (i.e., before data have been collected) and analysis (i.e., after data have been collected). We describe each test and methods to calculate power based on the corresponding test statistics.

Sobel Test

The first test is the asymptotic Sobel test. This often-employed conventional test of mediation effects compares the ratio of the estimated mediation effect to its estimated standard error. When defining the mediation effect as the product of the *a* and *b* path coefficients, we form the Sobel test statistic as

$$z_{ab}^{Sobel} = ab / \sqrt{\sigma_{ab}^2} \quad . \tag{3}$$

Here, σ_{ab}^2 is the error variance associated with the mediation effect as estimated by the product of the *a* and *b* paths. Prior literature has shown that a good estimate of this error variance can be obtained as a function of the individual paths and their individual error variances

$$\sigma_{ab}^2 = b^2\sigma_a^2 + a^2\sigma_b^2 + \sigma_a^2\sigma_b^2 \quad . \tag{4}$$

In the literature, the final term that captures the product of the error variance is often dropped because it has been shown to be small (i.e., $\sigma_a^2\sigma_b^2 \approx 0$). Thus, the Sobel test statistic for the mediation effect is (Sobel, 1982)

$$z_{ab}^{Sobel} = ab / \sqrt{b^2\sigma_a^2 + a^2\sigma_b^2} \quad . \tag{5}$$

When using maximum likelihood to estimate the parameters, the Sobel test statistic has an asymptotic normal distribution. Therefore, the test statistic is compared to a standard normal

distribution to determine the statistical significance of the mediation effect. Comparing these test statistics to a standard normal distribution allows for inferential testing of the mediated effects. Assuming the alternative hypothesis is true, these test statistics follow a non-central distribution with the ratios as the non-centrality parameter. The power of a two-sided test to detect the mediation effect is then

$$P(|z^{Sobel}|>z_{critical}) = 1-\Phi(z_{critical}-z^{Sobel})+\Phi(-z_{critical}-z^{Sobel}) \quad (6)$$

where Φ is the normal distribution with $z_{critical}$ as the chosen critical value (e.g., 1.96) corresponding to a nominal type one error rate and z^{Sobel} is the test statistic that compares the mediation effect to its respective standard errors.

Research indicates that the Sobel test and other applications that use the delta method produce accurate variance estimates for mediated effects (Oehlert, 1992). However, the use of a standard normal distribution to draw inferences about the Sobel test statistic can affect the accuracy of type one and type two error rates. It is well-established that the sampling distribution of the mediation effect is asymptotically normal but tends to be skewed and kurtotic in small sample sizes (Bollen & Stine, 1990; MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). In these cases, referring the Sobel test statistic against a standard normal distribution can result in conservative inferences and low power (Hayes & Scharkow, 2013). We include the Sobel test in our analyses despite these issues because of its historical popularity and ease of implementation. Next we extend, the joint test of significance and the Monte Carlo confidence interval test which address the shortcomings of the Sobel test by avoiding its distributional assumptions (MacKinnon et al., 2002; MacKinnon, Lockwood, & Williams, 2004).

Joint Test

The aptly named joint test is constructed by combining separate tests of the individual *a* and *b* paths (Hayes & Scharkow, 2013). Using the maximum likelihood estimation for the path coefficients outlined above, the statistical test of the treatment-mediator association (i.e., *a* path) is

$$z_a = a / \sigma_a \tag{7}$$

and the test of the mediator-outcome association (i.e., *b* path) is

$$z_b = b / \sigma_b \tag{8}$$

The joint test determines the significance of the mediation effect using a simultaneous evaluation of inferences from the z_a and z_b results. When both z_a and z_b are statistically significant the mediation effect is considered statistically significant. Hayes and Scharkow (2013) found the joint test performed similarly to bootstrap methods and had a good balance of type one error rates and power. The joint test does not involve an estimate of the mediation effect or an associated standard error and therefore does not directly produce effect sizes or confidence intervals (MacKinnon et al., 2002).

Under the joint test, power to detect a mediated effect is the product of the power to detect the *a* path and the power to detect the *b* path. We operationalized the joint test using normal distributions and formulate the joint test power functions as follows

$$P(|z_a| > z_{critical} \;\&\; |z_b| > z_{critical}) = (1-\Phi(z_{critical}-z_a)+\Phi(-z_{critical}-z_a))*(1-\Phi(z_{critical}-z_b)+\Phi(-z_{critical}-z_b)) \tag{9}$$

with $\Phi()$ as the normal cumulative density function and $z_{critical}$ as the corresponding critical value.

Monte Carlo Interval Test

The MC interval test is a resampling based method that draws random samples of the principal paths that constitute the mediation effect with the sampling variability equal to the error variance of the path approximations (Preacher and Selig, 2012). Typically, the MC interval test assumes the maximum likelihood estimates of the path coefficients have a multivariate normal sampling distribution with means, variances, and covariances based on the maximum likelihood estimates (Preacher & Selig, 2012). Given our multilevel linear path model, the MC interval test would estimate the mediated effects using

$$\begin{pmatrix} a^* \\ b^* \end{pmatrix} \sim MVN\left(\begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}, \begin{pmatrix} \hat{\sigma}^2_{\hat{a}} & \hat{\sigma}_{\hat{a},\hat{b}} \\ \hat{\sigma}_{\hat{a},\hat{b}} & \hat{\sigma}^2_{\hat{b}} \end{pmatrix}\right). \tag{10}$$

Here, the product of random draws of a^* and b^* are used to create a sampling distribution. This sampling distribution approximates the sampling distribution of the mediation effects. Asymmetric confidence intervals are constructed using predetermined values. For example, selecting values for the mediation effect associated with the 2.5th and 97.5th percentiles create a 95% confidence interval for the estimation of the mediation effect. When these confidence intervals exclude the null value (e.g., zero), the mediation effect is considered significant.

The MC interval test offers several advantages over other tests of mediation effects. Recent studies have shown its performance is similar to bootstrap based methods (Hayes & Scharkow, 2013; Preacher & Selig, 2012) but it is much less computationally intensive because it does not require resampling from complete data sets. Additionally, the MC interval test does not make distributional assumptions about the product of path coefficients which allow researchers

to avoid the type one error rate issues associated with the Sobel test (Preacher & Selig, 2012). Power for the MC interval test is determined by the proportion of asymmetric confidence intervals that exclude the null value (Preacher & Selig, 2012).

Error Variance

Evident from the aforementioned tests, in order to identify designs that maximize power while considering cost, we must track the error variance associated with the mediation effects. Our development of the error variances begins with an outline of the error variance associated with each path coefficient. Assuming a balanced random assignment of groups to treatment conditions the error variance of the *a* path coefficient associated with the maximum likelihood estimator can be reduced to

$$\sigma_a^2 = \frac{4\sigma_{M|}^2}{n_2} \quad . \tag{11}$$

For a two-level random intercept model as presented in the outcome equation (i.e., Equation 2ab), the error variance associated with the maximum likelihood estimator of the group-level path coefficient *b* can be reduced to

$$\sigma_b^2 = \frac{\tau_{Y|}^2 + \sigma_Y^2 / n_1}{n_2 \sigma_{M|}^2} \tag{12}$$

where $\tau_{Y|}^2$ is the conditional group variance and σ_Y^2 is individual outcome variance, n_1 and n_2 represent the number of individuals per group and the number of groups, and $\sigma_{M|}^2$ is the variance of the mediator conditional on the treatment (Kelcey, Dong, Spybrook, & Cox, 2017; Kelcey, Dong, Spybrook, & Shen, 2017).

We adapt the restructured formulas of the error variance of the paths to capture the error variance associated with the Sobel test. For a two-level group-randomized study using the multilevel linear path formulation from Equations 1 and 2ab, the error variance under the Sobel test for the mediation effect is

$$\sigma^2_{Sobel} = a^2\left(\frac{\tau^2_{Y|} + \sigma^2_Y / n_1)}{n_2\sigma^2_{M|}}\right) + b^2\left(\frac{4\sigma^2_{M|}}{n_2}\right). \quad (13)$$

The Sobel test statistic under these conditions is then formed by substituting this equation into Equation 3.

Equations 11 and 12 represent the error variance associated with the *a* path and *b* path separately and therefore the error variance of each component of the joint test. For a two-level group-randomized study using the multilevel linear path formulation from Equations 1 and 2, the test statistic using the joint test for the *a* path is formed by substituting Equation 11 into Equation 7 and for the *b* path substituting Equation 12 into Equation 8.

The MC interval test uses a distribution of products (i.e., *a*b**) to estimate the mediation effect so the error variance will be asymptotically equal to that of the Sobel test although in finite sample sizes it can deviate from the asymptotic approximation. While power of the MC interval test is only moderately approximated by the Sobel test, for the recursive models applied here prior research has indicated that the error variance of the product ($\sigma^2_{ab} = b^2\sigma^2_a + a^2\sigma^2_b + \sigma^2_a\sigma^2_b$) is a good approximation of the error variance of the MC interval test (e.g., Kelcey, Dong, Spybrook, & Shen, 2017).

Optimal Sample Allocation

Studies of Main Effects

In group randomized studies the power to detect a main effect increases with greater differences between treatment conditions on the outcome variable or through decreased variance in the treatment effect. Examining the variance of the treatment effect helps illuminate the relationship between individual sample size (n_1), group sample size (n_2) and power (Raudenbush, 1997; Spybrook, Raudenbush, Liu, Congdon, & Martínez, 2011). Consider an unadjusted multilevel outcome model for the main effect

$$\textit{Outcome model}\ \text{(Level 1)} \qquad Y_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma_Y^2) \tag{14a}$$

$$\text{(Level 2)} \qquad \beta_{0j} = \gamma_{00} + cT_j + u_{0j} \qquad u_{0j} \sim N(0, \tau_{Y|T}^2) \tag{14b}$$

with the coefficient *c* as the main effect, $\tau_{Y|T}^2$ as the group-level outcome variance solely conditional upon the treatment (T_j). Other terms retain a similar meaning from the outcome equation presented in Equation 2ab. Based on this formulation, the variance of the treatment effect in a two-level group-randomized study of main effects can be expressed as

$$Var(c) = \frac{4(\tau_{Y|T}^2 + \sigma_Y^2 / n_1)}{n_2} \tag{15}$$

Researchers typically do not have control over the magnitude of the treatment effect, individual-level outcome variance, or group-level outcome variance so manipulating these parameters is not a viable means to increase study power. However, researchers can control sample allocation or include prognostic covariates to improve study power. Increasing n_1 or n_2 increases power by reducing the *Var(c)* but with varying effectiveness. Increasing individual level sample size provides diminishing returns in reducing the *Var(c)* while increasing group sample size reduces the *Var(c)* until the power rate approaches 1. Sampling a large number of

groups is a simple and effective means to achieve adequate power rates but this recommendation may be problematic as sampling additional groups can be costly and funding is often limited.

Application of optimal design principles can identify the number of individuals per group and the number of groups that produce the minimum error variance or approximately the maximum power within budgetary constraints. Our analyses consider the conventional linear cost formulation (Raudenbush, 1997) such that

$$T = c_2 n_2 + c_1 n_2 n_1 \tag{16}$$

T is the total funds available for a study, c_1 is the sampling cost for each individual, c_2 is sampling cost for each group, and each is typically measured in monetary units (e.g., dollars). We can now express the variance of the treatment effect with the cost function in Equation 17.

$$Var(c) = \frac{4(\tau^2_{Y|T} + \sigma^2_Y / n_1) * (c_2 + c1 n_1)}{T} \tag{17}$$

Minimizing Equation 17 in terms of n_1 results in an optimal number of individuals to sample ($n_1^{opt.me}$) which is expressed as

$$n_1^{opt.me} = \frac{\sigma_Y}{\tau_{Y|T}} \times \sqrt{c_2 / c_1}\ . \tag{18}$$

Sampling $n_1^{opt.me}$ individuals per group minimizes the variance of the treatment effect (i.e., main effect) therefore maximizing power under a fixed cost. This is the most efficient allocation of resources—designs that sample more or less individuals in favor of less or more groups will tend to be less efficient and yield lower levels of power under the same total cost. For two-level designs, to find the number of groups to sample one substitutes the optimal n_1 value into $n_2 = T / (c_2 + c_1 n_1)$.

Studies of 2-2-1 Mediation

In studies of 2-2-1 mediation the power associated with detecting mediation effects also depends on individual and group sample size. Evident from Equations 11 and 12, group sample size has a greater influence on power than individual sample size. The same linear cost formulation from Equation 16 is applicable to studies of 2-2-1 mediation (Raudenbush, 1997) along with the formulation for group sample size. Below we derive the optimal sample allocation formulas for two-level group-randomized studies that track effects from a group-level treatment to an individual-level outcome as they pass through a group-level mediator (i.e., 2-2-1 mediation) using the Sobel test and joint test. We subsequently evaluate the extent to which the optimal sample allocation under the Sobel or joint test serves as an accurate approximation of optimal sample allocation when using the MC interval test. We begin with optimal sample allocation formulas when the multilevel path model has no covariates and then extend the formulas to accommodate covariates.

Optimal Sample Allocation without Covariates

Sobel test

Despite the known shortcomings of the Sobel test, it is instructive to examine this test because it provides closed-form expressions that are simple to maximize/minimize and outlines the asymptotic behavior of the estimated mediation effect. Similar to prior optimal design frameworks, we derive optimal sample allocation formulas for the Sobel test by minimizing the error variance (Raudenbush, 1997).

Substituting the linear cost formulation of n_2 into Equation 13, we re-express the error variance of the Sobel test statistic as

$$\sigma^2_{Sobel} = a^2\left(\frac{(\tau^2_{Y|} + \sigma^2_Y / n_1)}{(T/(c_2 + c_1 n_1))\sigma^2_{M|}}\right) + b^2\left(\frac{4\sigma^2_{M|}}{(T/(c_2 + c_1 n_1))}\right) \quad (19)$$

Minimizing the error variance of the Sobel test with respect to n_1 by taking the first derivative yields

$$\frac{\partial}{\partial n_1}\sigma^2_{Sobel} = \frac{4b^2 c_1 \sigma^2_{M|}}{T} + \frac{a^2 c_1(\tau^2_{Y|} + \sigma^2_Y / n_1)}{T\sigma^2_{M|}} - \frac{a^2\sigma^2_Y(c_2 + c_1 n_1)}{T\sigma^2_{M|}n_1^2} \quad (20)$$

Solving for n_1 provides the optimal sample allocation $\left(n_1^{opt}\right)$ for a two-level group-randomized study designed to detect 2-2-1 mediation under the Sobel test

$$n_1^{opt} = \sqrt{\left(\frac{c_2}{c_1}\right)\frac{a^2\sigma^2_Y}{4b^2\sigma^4_{M|} + a^2\tau^2_{Y|}}} \quad (21)$$

In Equation 21, we formulated the optimal individual sample size for the Sobel test using group- and individual-level variance terms for the mediator and outcome. It is also possible to express these variance terms as a function of the path coefficients (i.e., *a, b, c'*). When the mediator and outcome are standardized to have variances of one, we can express the individual-level outcome variance as $\sigma^2_Y = 1 - \rho$ where ρ is the intraclass correlation coefficient of the outcome, the conditional group-level variance of the mediator as (Kelcey, Dong, Spybrook, & Cox, 2017)

$$\sigma^2_{M|} = 1 - \frac{a^2}{4} \quad (22)$$

and the conditional group-level variance of the outcome as

$$\tau^2_{Y|} = \rho\left(1 - \left(\frac{(ab + c')^2}{4\rho} + \frac{b^2}{\rho}\left(1 - \frac{a^2}{4}\right)\right)\right) \quad (23)$$

Replacing the variance terms, we can now express the optimal individual sample size for the Sobel test in terms of path coefficients as

$$n_1^{opt} = \sqrt{(\frac{c_2}{c_1})\frac{a^2(1-\rho)}{4b^2(1-\frac{a^2}{4})^2 + a^2(\rho - \frac{(ab+c')^2}{4} - b^2(1-\frac{a^2}{4}))}}. \quad (24)$$

Substituting the n_1^{opt} produced by Equation 21 or 24 into the formulation for group sample size provides the optimal number of groups to sample. Utilizing this sample allocation provides the researcher with an optimally designed study under the Sobel test.

The optimal individual sample size formulas for the Sobel test indicate cost structure, variance structure, path coefficients, and intraclass correlation influence the size of n_1^{opt}. First, higher group to individual cost ratios inflate optimal individual sample size. As the cost ratio increases (i.e., when groups are much more expensive than individuals) it becomes more efficient to increase the number of individuals sampled within groups. This influential relationship is also observable in two-level group randomized studies of main effects.

Second, higher group-level conditional variance in the outcome (i.e., $\tau_{Y|}^2$) reduces the optimal individual sample size while individual-level variance inflates it. This relationship is also true for optimal individual sample size in studies of main effects. When there is greater variance at level-one more individuals are sampled and when there is greater variance at level-two more groups are sampled. Lastly, the influence of specific path coefficients and intraclass correlation coefficient can be seen in Equation 21. Increasing the *a* coefficient, the group-level relationship, results in greater n_1^{opt} values while increases in the *b* coefficient result in smaller n_1^{opt} values. Stronger relationships at level-two (e.g., *a*) require fewer groups to detect an effect

while stronger relationships at level-one (e.g., b) requires fewer individuals to detect an effect. Increasing intraclass correlation coefficients degrade the value of individuals sampled within a group so higher intraclass correlation leads to reductions in the optimal individual sample size.

As we shall see with the example cost structures below, a rough and fallible rule of thumb regarding the optimal individual-level sample size under the Sobel test in group-randomized studies of 2-2-1 mediation is that it will typically be less than 10 with small a/b path coefficient ratios (i.e., $a/b < 2$) and small cost ratios (i.e., $c_2/c_1 < 100$). As a/b or c_2/c_1 increases, the optimal individual sample size can exceed 10 but almost never exceeds 30.

Joint Test

Next, we considered optimal sample allocation when using the joint test in a group-randomized study of 2-2-1 mediation. Because the joint test is comprised of two separate sub-tests concerning the *a* and *b* paths, we cannot indirectly maximize power by maximizing the product of the test statistics or minimizing the error variances. As a result, we must directly maximize power to identify optimal sample allocations.

Using the linear cost formulation of n_2 we can express the tests of the *a* and *b* paths as

$$z_a = a / \sqrt{\frac{4\sigma^2_{M|}}{T/(c_2 + c_1 n_1)}} \tag{25}$$

and

$$z_b = b / \sqrt{\frac{\tau^2_{Y|} + \sigma^2_Y / n_1}{\sigma^2_{M|}(T/(c_2 + c_1 n_1))}}. \tag{26}$$

Substituting these terms into the joint test power function (i.e., Equation 9) and taking its derivative relative to n_1 gives

$$\frac{\partial}{\partial n_1} = \left(-\frac{\frac{b}{5}\exp\left[-\frac{1}{2}\left(-z_{critical}+z_b\right)^2\right]\left(\frac{c_1\sigma_b^2}{c_2+c_1n_1}-\frac{\sigma_Y^2(c_2+c_1n_1)}{T\sigma_{M|}^2n_1^2}\right)}{\sigma_b^3}+\frac{\frac{b}{5}\exp\left[-\frac{1}{2}\left(z_{critical}+z_b\right)^2\right]\left(\frac{c_1\sigma_b^2}{c_2+c_1n_1}-\frac{\sigma_Y^2(c_2+c_1n_1)}{T\sigma_{M|}^2n_1^2}\right)}{\sigma_b^3}\right)$$
$$\left(1-\frac{1}{2}erfc\left(\frac{-z_{critical}+z_a}{\sqrt{2}}\right)+\frac{1}{2}erfc\left(\frac{z_{critical}+z_a}{\sqrt{2}}\right)+\left(-\frac{\frac{a}{10}c_1\sigma_{M|}^2\exp\left[-\frac{1}{2}\left(-z_{critical}+z_a\right)^2\right]}{T\sigma_a^3/8}+\frac{\frac{a}{10}c_1\sigma_{M|}^2\exp\left[-\frac{1}{2}\left(z_{critical}+z_a\right)^2\right]}{T\sigma_a^3/8}\right)\right. \quad (27)$$
$$\left.\left(1-\frac{1}{2}erfc\left(\frac{-z_{critical}+z_b}{\sqrt{2}}\right)+\frac{1}{2}erfc\left(\frac{z_{critical}+z_b}{\sqrt{2}}\right)\right)\right)$$

where *erfc()* is the complementary error function. While there is no simple closed-form analytic solution that identifies the optimal sample allocation as in the case of the Sobel test, we can identify the optimal sample allocation for a given cost structure by finding the root of this derivative using numerical methods. Although the resulting derivative and its solution appears cumbersome, finding its root and the optimal sample size is quite straightforward when implemented in software.

Just like with the Sobel test, it is possible to restructure the joint test optimal sample allocation equations in terms of the path coefficients (i.e., *a, b, c'*). For the joint test, we found

$$z_a = a \Big/ \sqrt{\frac{4(1-(a^2+4))}{T/(c_2+c_1n_1)}} \quad (28)$$

for the *a* path and for the test of the *b* path we found

$$z_b = b \Big/ \sqrt{\frac{(\rho-\frac{(ab+c')^2}{4}-b^2(1-\frac{a^2}{4}))+(1-\rho)/n_1}{(1-\frac{a^2}{4})(T/(c_2+c_1n_1))}}\,. \quad (29)$$

The solution to the optimal sample allocation in terms of the path coefficients is then possible by substituting the path formulations of the variance terms into the n_1^{opt} formula for the joint test found in Equation 27.

The formula for n_1^{opt} under the joint test has two general components that balance maximizing z_a and z_b. Each component is an addend made up of the product of parts representing the power functions for the *a* and *b* paths. The complementary error function ($erfc$) is utilized in the second part of each addend to determine probabilities related to statistics (e.g., z_a and z_b) falling inside and outside of critical regions (e.g., $z_{critical}$). In the first component, the parts representing the power function of the *b* path is divided by the variance of the *b* path and multiplied by parts representing the power function of the *a* path. Conversely, the second component consists of parts representing the power function of the *a* path divided by the variance of the *a* path multiplied by the parts representing the power function of the *b* path. The simultaneous consideration of *a* and *b* power and variance components demonstrates the formula balancing each test when determining the optimal individual sample size. However, the *b* path component is weighted more heavily (i.e., factor of *b/5* verse *a/10*) indicating increases in the *b* path inflate optimal individual sample size for the joint test more quickly than increases in the *a* path.

The factors that influence optimal individual sample size under the joint test are similar to the factors that influence optimal individual sample size under the Sobel test. In the formulation above for determining the optimal individual sample size under the joint test, we see familiar factors such as a, b, c_1, c_2, $\tau^2_{Y|}$, σ^2_Y, and $\sigma^2_{M|}$. Under most conditions, the roles of these factors in shaping the optimal individual sample size for the joint test parallel their roles for the Sobel test and MC interval test. We detail exceptions in a subsequent section. Our general guidelines regarding the size of the optimal individual sample under the Sobel test also apply to individual sample size for the joint test with the exception that at larger cost ratios (e.g., $c_2/c_1 \approx 1000$) and

when $a/b \approx 1$, the optimal individual sample size for the joint test tends to be somewhat smaller than optimal individual sample size under the Sobel test. Again, we caution readers with regard to over applying these fallible rules of thumb.

Monte Carlo Interval Test

Because the MC interval test is a resampling-based test that cannot be easily captured through closed-form expressions, we draw on closed-form asymptotic approximations of the sampling variance of the mediation effect to track its optimal sampling strategies. As previously noted, asymptotic approximations to the error variance (but not the sampling distribution) of the product of two random variables performs quite well even in moderate sample sizes (e.g., (Kelcey, Dong, Spybrook, & Cox, 2017). For this reason, the MC interval test and the Sobel test should have asymptotically equivalent optimal sample allocations. For example, when T=1000, $c_1 = 1$, $c_2 = 10$, a=.5, b=.3, $c' = .1$, and $\rho = .2$, the Sobel test $n_1^{opt} \approx 2.4$ and so we expect the corresponding optimal n_1 for the MC interval test to be about 2.4. The Sobel error variance and the Monte Carlo error variance under this example are plotted in Figure 1. Evident from this figure, the empirical optimal n_1 for the MC interval test aligned well with the Sobel optimal n_1.

To assess this correspondence more generally, we conducted a simulation study assessing the quality of the Sobel optimal sample allocation formula in approximating the optimal sample allocation for the MC interval test. Our simulation study probes the magnitude of the differences between optimal sample allocation based on the Sobel test formulas and the true optimal sample allocation for the MC interval test. Before considering the full simulation and results, Figure 2 presents MC interval test power as a function of individual sample size. The n_1 value that aligns with the maximum possible power using the MC interval test (i.e., true optimal sample

allocation) and the n_1^{opt} suggested by the Sobel test formulas are marked to provide an illustrated example of the differences under investigation by the simulation. Figure 2 clearly shows true and recommended optimal individual sample sizes diverge as the cost ratio increases indicating some inaccuracies in the optimal sample allocation for the MC interval test if using the Sobel formulas.

However, it is important to consider the consequences to power rates not just the differences between true and recommended optimal individual sample sizes. For example, the first curve (a) in Figure 2 indicates no difference in the individual sample size that aligns with maximum power for the MC interval test and the individual sample size suggested by the Sobel test optimal sample allocation formulas. In this case, there is no loss in power if we use the Sobel formulas to determine optimal sample allocation for the MC interval test. The second curve (b) indicates a small difference between recommended individual sample size and true optimal sample size but only a minor loss of power. The final two curves (c and d) indicate greater differences between the recommended and true optimal individual sample sizes but the power curves are relatively flat resulting in only a minor loss in power when using the Sobel formulations.

The purpose of our simulation was to identify the empirical individual sample size that maximized power under the MC interval test and compare it to the optimal individual sample size suggested by the Sobel formulations. We aimed to substantiate our theoretical derivations and provide an initial assessment of the precision of this theory under practical sample sizes and common parameter specifications. Because our Sobel-based formulas for optimal sampling indicated that the magnitude of the *a* path, *b* path, and their ratio influenced optimal sample size,

we considered three different values for a/b to cover a range of mediation effects: when the a path was greater than, less than and equal to the b path. Cost ratio can vary greatly depending on the substantive context of the study. For example, compared to collecting group-level data, collecting individual-level data may require time intensive processes that are expensive. In our simulations, we held individual costs at $c_1 = 1$ and varied such that $c_2 = 5, 10, 100, 1000$. This range of c_2/c_1 ratios represents a range of substantive studies in which the group and individual costs vary widely. Total cost was set to 100 times the cost of each group ($T = 100c_2$). Finally, we set $c' = 0.1$ to represent a small direct effect and indicate partial mediation and the unconditional intraclass correlation coefficient to $\rho = 0.2$ based on previous investigations (Hedges and Hedberg, 2007). For each condition, the MC interval test was repeated 100 times using 1000 draws per test (see Table 1).

We found some (ultimately) minor discrepancies between the Sobel formula implied optimal individual sample size and individual sample size that aligned with greatest achievable power for the MC interval test. These differences occurred at higher cost ratios and were more pronounced when the path coefficient ratio was small. For example, in the conditions presented in the final panel of Figure 2 maximum power with MC interval test was achieved using $n_1 \approx 21$ but the optimal sample allocation suggested by the Sobel formulas was $n_1^{opt} = 7$. Fortunately, the loss in power corresponding to differences in recommended individual sample sizes was consistently negligible (i.e., $\leq 3\%$) even when recommended and true optimal sample sizes differed substantially. Differences in the power rates did increase as the cost ratio increased and tended to be greatest when $a/b \approx 1$. Even under these conditions, differences between

maximum power and power using the optimal sample allocation from the Sobel formulas remained around 3%.

This evidence suggests that the Sobel formulas for optimal sample allocation often provided a reasonably good approximation for sample allocation that maximizes power under the MC interval test in group-randomized studies of 2-2-1 mediation. We conclude that it will typically be appropriate for researchers to use the same formulas and process described for the Sobel test to approximate the optimal sample allocation for a group-randomized study of 2-2-1 mediation when using the MC interval test. However, the consistent quality of this approximation needs more future assessment. Because we apply the optimal sample allocation formulas for the Sobel test to the MC interval test, the guidelines regarding typical optimal individual sample sizes and principal path formulations are the same for the Sobel test and MC interval test.

Influences and Implications

Cost Ratio

We next examined the absolute and relative influence of the cost ratio on optimal individual sample size in two-level group-randomized studies probing main and mediation effects under a variety of conditions. Subsequently, we considered the consequences to power as study designs deviated from the optimal individual sample size. These investigations allowed us to better understand the behavior of optimal sampling because they outline how the optimal individual sample size changes as a function of the cost ratio and the rates with which sub-optimal sampling undermines statistical power.

In an absolute sense, the results indicated that optimal individual sample size for studies of main and mediation universally increased as the cost ratio increased. The results in Table 2 provide examples of this relationship across a variety of conditions. For instance, when *a*=.5, *b*=.3, $c' = .1$, and $\rho = .3$, the MC interval test and Sobel test have an optimal individual sample size of $2, 7, \text{and } 22$ when the cost ratio is $5, 100 \text{ and } 1000$ respectively.

In a relative sense, the results suggested that regardless of cost structure, the optimal individual-level sample size for mediation effects will typically be less than that of main effects. However, under the Sobel or MC interval test, increases in the cost ratio result in increases in the optimal individual-level sample size at roughly the same rate for main and mediation effects. For example, with an intraclass correlation coefficient of 0.1, going from a cost ratio of 5/1 to 100/1 increases the optimal individual sample size by a factor of about four for the main effect (i.e., n_1^{opt} goes from 7 to 30) and about five for the mediation effects (i.e., n_1^{opt} goes from 3 to 15; Table 2). This multiplicative similarity does not appear to apply as much to the joint test. Under the same example, the optimal individual sample size for the joint test increases only by a factor of about two (i.e., n_1^{opt} goes from 3 to 7).

The results also suggested that the influence of the cost ratio on optimal sample allocation was moderated by the intraclass correlation coefficient—the optimal number of individuals was reduced as it increased. The major exception to this pattern occurred under the joint test with a high cost ratio. Under these conditions increasing intraclass correlation led to increases in the optimal individual sample size.

Comparison with main effects. The optimal individual sample size for studies of main effects tended to be greater than the optimal individual sample size for comparable studies of

mediation effects. This was true across cost ratios but was influenced by path coefficient values. With large path coefficient ratios (i.e., a/b) the differences between optimal individual sample size for studies of main and mediated effects was minimal. As the path coefficient ratio approached one optimal individual sample sizes in studies of mediation effects decreased causing greater differences between studies of main and mediated effects.

Differences across tests. The relationship between optimal individual sample size and the cost ratio was more complex when comparing the mediation tests. Optimal individual sample sizes were similar across mediation tests at lower cost ratios. However, at larger cost ratios optimal individual sample sizes differed among mediation tests depending on the path coefficient ratio. Compared to the Sobel and MC interval tests, the joint test had larger optimal individual sample sizes when the path coefficient ratio was large and smaller optimal individual sample sizes when the path coefficient ratio was small. It should be noted that increasing intraclass correlation values reduced this disparity.

Deviating from Optimal Design

Next, we considered the consequences of deviating from the optimal individual sample size in two-level group-randomized studies of 2-2-1 mediation and similar studies of main effects. Lower cost ratios increased the rate at which power depreciated as sample allocation deviated from the optimal individual sample size. This is true for studies of main and mediated effects but the influence of cost ratio on power loss was much more pronounced in studies of mediation effects. Figure 3 shows power rates by individual sample size for a study of main effects and a study of mediation effects using the Sobel, joint, and MC interval test at different cost ratios. In studies of mediation effects with large cost ratios power depreciated slowly as

sample allocation deviated from optimal values. This resulted in large but practically insignificant differences in optimal individual sample sizes between mediation tests when the cost ratio was large. For example, when $c_2/c_1 = 100$, the optimal sample of individuals for the Sobel test was eight but only 3 for the joint test. However, under those same conditions power rates were nearly identical for each test using individual sample sizes anywhere between 3 and 30. Under a large cost ratio, the importance of optimal individual sample size is greatly reduced because deviations have little influence on power. The converse is true under a small cost ratio as slight deviations from the optimal individual sample size greatly reduce power rates in studies of mediation effects. The range of individual sample sizes that had little influence on power under the large cost ratio (i.e., individual sample sizes from 3 to 30) cut power rates in half under a small cost ratio. For example, with a small cost ratio and similar conditions as the above example power rates were ≈ .6 when $n_1 \approx 3$ but when $n_1 \approx 30$ power rates were ≈ .3.

Now consider the study of main effects in the middle example of Figure 3 (i.e., cost ratio of 10/1), power rates were ≈ .5 when n_1 ranged from 5 to 20 reflecting only a minor loss of power as sample allocation deviated from optimal. For a similar study of 2-2-1 mediation using the joint test or MC interval test, power rates ranged from ≈ .6. to ≈ .3 across an equivalent range of individual sample sizes reflecting the greater consequences of deviating from optimal sample allocation in studies of mediation effects.

The Sobel test, joint test, and MC interval test vary in sensitivity to detect the mediation effect and therefore power rates varied (as expected based on previous literature) but power loss as study designs moved away from optimal sample allocation was fairly uniform across the tests. For example, in Figure 3 the joint test and MC interval test maintain their power advantage over

the Sobel test as study designs moved away from optimal sample allocation although the advantage does diminish at lower cost ratios.

A hypothetical study with a small cost ratio best demonstrates the consequences of deviating from optimal sample allocation. Figure 4 presents power rates by individual sample size with a cost ratio of $2/1$. Under these conditions optimal individual sample size was small. For the study of main effects (i.e., $n_1^{opt} = 3$) and for the study of mediation effects using the Sobel, joint, or MC interval test (i.e., $n_1^{opt} = 2$). Power rates $\geq 80\%$ were achieved across all the studies and using any mediation test under an optimal design. The joint test and MC interval test were especially well powered (i.e., power rate $\approx 98\%$) but increasing individual sample sizes from 2 to anything > 10 resulted in an underpowered study. In group randomized studies the group often includes many individuals (e.g., students in classrooms, teachers in schools) so researchers can be tempted to sample more individuals per group than necessary under the false notion it will increase power. This reasoning is dangerous to study sensitivity but especially so when the cost ratio is small. Under the example conditions described here, a large individual sample size (i.e., $n_1 \approx 25$) resulted in power rates under 50% for each mediation test.

To summarize, high cost ratios caused divergent optimal individual sample sizes across mediation tests but these differences were not practically significant. At small cost ratios, there were smaller differences in optimal individual sample sizes across mediation tests but these differences had a great deal of practical significance. Deviating slightly from the optimal individual sample size when the cost ratio was small resulted in a significant loss of power. It is crucial to utilize optimal sample allocation in studies of 2-2-1 mediation with small cost ratios to prevent significant power loss but deviating from optimal sample allocation with a higher cost

ratio is less detrimental to power. This was also true for studies of main effects but the influence of cost ratio on power depreciation as sample allocation deviated from optimal values was less pronounced.

Optimal Sample Allocation with Covariates

The inclusion of covariates has been shown to be an excellent strategy to increase the precision of group-randomized studies of main effects (e.g., Bloom, Richburg-Hayes, & Black, 2007) and mediated effects (e.g., Kelcey, Dong, Spybrook, & Shen, 2017). Covariates can explain variation in the outcome and/or mediator and potentially increase the precision of estimates (Kelcey, Dong, Spybrook, & Cox, 2017). For example, variables representing past academic achievement when the outcome is academic achievement serve as popular prognostic covariates in educational research (e.g., Hedges and Hedburg, 2007). Further, in the context of mediation, random assignment of the treatment but not the mediator introduces the possibility of confounding in the mediator-outcome relationship (i.e., violations of the sequential ignorability assumption). For this reason, inclusion of covariates in the outcome model will almost always be required to obtain an unbiased estimate of the mediator-outcome relationship. The formulations below include the same set of covariates in both the mediator and outcome model to best address the sequential ignorability assumption and improve the precision of both the mediator and outcome estimates but matching sets of covariates are not required. To allow for these advantages, we extended our optimal sample allocation formulas to designs that include covariates. Below we augment our models to include individual-level covariates (X), their group-level means ($\bar{X}$), and group-level covariates (W) in our multilevel linear path model from Equations 1 and 2ab such that

Mediator model (Level 2) $$M_j = \pi_0 + aT_j + \sum_{l=1}^{L} \pi_l W_{lj} + \sum_{k=1}^{K} \alpha_k \bar{X}_{kj} + \varepsilon_j \qquad \varepsilon_j \sim N(0, \sigma^2_{M|}) \tag{30}$$

Outcome model (Level 1) $$Y_{ij} = \beta_{0j} + \sum_{k=1}^{K} \beta_k (X_{kij} - \bar{X}_{kj}) + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2_{Y|}) \tag{31a}$$

(Level 2) $$\beta_{0j} = \gamma_{00} + bM_j + c^{'} T_j + \sum_{l=1}^{L} \gamma_l W_{lj} + \sum_{k=1}^{K} \lambda_k \bar{X}_{kj} + u_{0j} \qquad u_{0j} \sim N(0, \tau^2_{Y|}) \tag{31b}$$

In the mediation equation (i.e., Equation 30) we add W_{lj} as group-level covariates, with π_l as the corresponding path coefficients, $\bar{X}_{kj}$ as group-level aggregates of individual-level covariates, with α_k as the corresponding path coefficients, and ε_j as the error term that is assumed to be normally distributed with a mean of zero and variance $\sigma^2_{M|}$. In the outcome equation (i.e., Equation 31ab) we add X_{kij} for individual-level covariates with coefficients β_k. Here we use group-mean centering for the individual-level covariates because it is common in the literature and can aid in the interpretability of effects (e.g., Pituch & Stapleton, 2012; Zhang, Zyphur, & Preacher, 2009). We would arrive at equivalent results if using grand-mean centering or not centering at all with our random intercept models (Enders & Tofighi, 2007; Kreft, de Leeuw, & Aiken, 1995). The individual-level error term is represented by e_{ij} and is assumed to be normally distributed with a mean of zero and variance $\sigma^2_{Y|}$. At the group-level, we introduce γ. and λ as the respective path coefficients for group-level covariates and individual aggregated covariates, and u_{0j} as the group-specific random effects that are assumed to be normally distributed with a mean of zero and variance $\tau^2_{Y|}$. Under this new formulation $\sigma^2_{M|}$ is conditional on T_j and now W_{lj} and $\bar{X}_{kj}$, $\sigma^2_{Y|}$ is conditional on X_{kij} and $\tau^2_{Y|}$ is conditional on M_j and T_j and now

W_{lj} and $\bar{X}_{kj}$. We employ cluster means rather than a latent mean model for group-level covariates (e.g., $\bar{X}_{kj}$) because of the larger sample sizes typically required for latent mean models. However, the use of cluster means can introduce unreliability into the mediator and outcome models. Adding unreliability into these models can bias sample size and power recommendations.

Following, Kelcey, Dong, Spybrook, and Shen (2017), and assuming the outcome and mediator are standardized to have a variance of one, the new conditional variance terms can be expressed as

$$\tau^2_{Y|} = \rho(1 - R^2_{Y\bar{z}}) - \frac{(ab + c')^2}{\tau} - b^2(1 - R^2_{\bar{z}}) - \frac{a^2}{\tau} \quad (32)$$

for the group-level variance term associated with the outcome (e.g., Y) where $R^2_{Y\bar{z}}$ represents the variance explained in the outcome at the group level by covariates in the vector $\vec{Z}$ (e.g, W and $\bar{X}$) and $R^2_{M\bar{z}}$ represents the variance explained in the mediator at the group level by covariates in the vector $\vec{Z}$ (e.g, W and $\bar{X}$). For the individual-level variance term associated with the outcome,

$$\sigma^2_{Y|} = (1 - R^2_{YL1})(1 - \rho) \quad (33)$$

where R^2_{YL1} represents the variance explained in the outcome at the individual-level by covariates in the vector $\vec{Z}$ (e.g, X). Finally, when the mediator is standardized to have an unconditional variance of one ($\sigma^2_M = 1$), the variance term associated with the mediator can be expressed as

$$\sigma^2_{M|} = 1 - (R^2_{M\bar{z}} + a^2/4). \quad (34)$$

To extend the optimal sample allocation formulas to group-randomized studies of 2-2-1 mediation that include covariates we simply substitute the new conditional variance terms in the previously presented n_1^{opt} formulas.

Illustration

Consider a hypothetical group-randomized trial in which entire schools adopt a new reform based student discipline policy aimed at reducing student discipline referrals (e.g. Osher, Bear, Sprague & Doyle, 2010). General reductions in student discipline may have unintended negative consequences (e.g., Eden, 2017) so it is important to understand the mechanisms through which the discipline policy is working. There are strong connections between school climate and student behavior (Hoffman, Hutchinson, and Reiss; 2009) so school climate serves as an important mediator between the discipline policy and student discipline outcomes. This is an example of 2-2-1 mediation with the school-level treatment, school discipline policy, influencing the school-level mediator, school climate, which in turn affects the student-level outcome, student discipline referrals. While we employ a hypothetical example to easily highlight and manipulate key factors in optimal design for 2-2-1 mediation, group-randomized designs aimed at 2-2-1 mediation have been shown to be plausible for educational research (Kelcey, Dong, Spybrook, & Shen, 2017). These studies have included the examination of teacher professional development, implementation, and student science outcomes (e.g., Desimone & Hill, 2017), and teacher professional development, classroom quality, and student literacy outcomes (e.g., Yoshikawa et al., 2015).

We assume a common cost structure (e.g., Raudenbush, 1997) with a simple total budget of T=5000 monetary units and an initial ratio of cost per school to cost per student of $c_2/c_1 =$

15. We assume a partially mediated relationship between the student discipline policy, school climate, and student discipline outcomes such that $c' = 0.1$, because few studies have shown complete mediation. We further assume that student discipline outcomes within a school are correlated around $\rho = 0.1$ and that researchers anticipate that the policy change will have a strong influence on school climate but a smaller effect on the more distal student outcome such that $a = 0.8$ and $b = 0.1$. We caution that the parameter values used here are meant to be illustrative only. Applications of the optimal design framework should consult previous empirical studies specific to the substantive area under investigation to outline plausible values. Figure 5 presents power rates for each test and a matching study of the main effect as a function of individual sample size.

To determine the optimal design of this study we first find the optimal number of students per school to sample. Under the Sobel test and MC interval test the optimal number of students per school to sample is $n_1^{opt} \approx 10$ (see Equation 24); using the joint test results in $n_1^{opt} \approx 13$ (see Equation 27). Using the optimal individual sample sizes (i.e., $n_1^{opt} = 10$ for the Sobel and MC interval test and $n_1^{opt} = 13$ for the joint test), we determine the optimal number of schools to sample (i.e., 198 schools under the Sobel and MC interval test and 181 schools under the joint test). The sample of schools was rounded down to ensure we stayed within budget. These results indicate that under the Sobel or MC interval test, sampling 10 students per school and 198 total schools is the most efficient design given our budget constraints while under the joint test the most efficient sample allocation is 13 students and 181 schools. Under these same conditions a group-randomized study of the main effect between school discipline policy and

student discipline outcomes has an optimal sample allocation of 12 students per school and 188 schools.

In this example, there are only minor differences between the mediation tests in terms of optimal sample allocation but significant differences in power. The differences in the heights of the power curves in Figure 5 illustrates these well-established differences in test sensitivity (e.g., MacKinnon et al., 2002; Preacher & Selig, 2012). The power rate for the Sobel test was $\approx .79$ while the joint and MC interval tests had power rates of $\approx .88$. Prior research suggests that the MC interval test or the joint test should be preferred because they avoid the distributional issues of the Sobel test (Hayes & Scharkow, 2013). The corresponding study of the main effect had a similar optimal sample allocation as the study focused on the mediated effect and a power rate of $\approx .82$. This suggests a similar sampling allocation in our hypothetical example provides an efficient and adequately powered study of both the main and mediated effects. Adequate power rates were achieved under the MC interval test with an optimal sample of 10 students in 198 schools and under the joint test with an optimal sample of 13 students in 181 schools. As a comparison, a researcher using approximately the same total sample size but a convenient sample of 100 students and 20 schools would be well within budget but severely underpowered using any of the three tests; the power rate for the Sobel test would be $\approx .20$, the MC interval test would achieve power rates of $\approx .11$, and power rates for the joint would be $\approx .13$. Changes in design, data collection, or data accessibility can change the group to individual cost ratio and this can have a large influence on optimal sample allocation. Consider the example study with a new group to individual cost ratio of 10, it is still best if we utilize the MC interval test or joint test but adequate power is now achieved using the Sobel test. The optimal design when using the

MC interval test includes a sample of eight students per school and 272 schools while under the joint test the optimal sample allocation is 10 students in 245 schools. Under these conditions, results indicate that we can still achieve power rates above .80 when sampling more than 30 students per school. Under the new cost ratio n_1^{opt} values decrease while n_2^{opt} increases. Power decreases at a greater rate as we deviate from optimal sample allocation but with a fixed budget of *T=5000* monetary units we achieve high power rates. In other words, we have more excess power to lose before it becomes practically significant.

Discussion

There has been a growing recognition of the need to supplement main effects investigations by probing the mechanisms through which interventions are presumed to work. However, unlike the literature base outlining design strategies for planning efficient and effective studies of main effects, literature on strategies for planning studies of mediation is very limited. To address one aspect of this gap, we extended optimal sample allocation to group-randomized studies of 2-2-1 mediation. Using a multilevel path formulation, we derived the optimal sample allocation formulas for the Sobel and joint test and demonstrated that the optimal sample allocation formulas for the Sobel test approximate optimal sample allocation under the MC interval test. Probing these formulas revealed the roles of governing parameters, typical sizes for optimal individual sample size, and the consequences of deviating from the optimal design.

We have implemented the optimal sample allocation formulas in the freely available *R* package [masked for blind review] and the [masked for blind review] software available at www.[masked for blind review] along with supplementary documents to encourage use among applied researchers. A substantial body of research has been developed for optimal sample

allocation in group-randomized studies of main effects (e.g., Hedges & Borenstein, 2014; Konstantopoulos, 2009; Raudenbush, 1997). Our extension of this work for studies of mediation provides applied researchers the methodological tools to incorporate the practical concern of cost into their study design. Closing this gap improves the quality and efficiency of designs available to researchers for the study of multilevel mediation.

Results from probing the formulas indicated that the optimal schemes for 2-2-1 mediation are governed by five primary parameters: (a) the intraclass correlation coefficient of the outcome, (b) the magnitude of the treatment-mediator path coefficient (*a*), (c) the magnitude of the mediator-outcome path coefficient (*b*), (d) the magnitude of the direct treatment-outcome path coefficient (*c'*), (e) and the sampling cost structure (c_1 and c_2). Both the intraclass correlation coefficient and sampling cost structure are also key parameters for determining optimal sample allocation for studies of main effects (Raudenbush, 1997). Here, the increased complexity of determining optimal sample allocation for mediated effects becomes abundantly clear. The magnitude of the *a*, *b*, and *c'* coefficients are now necessary for determining the optimal sample allocation. Furthermore, applied researchers must select among a variety of tests to determine the significance of the mediated effect and this selection not only influences the optimal sample allocation but the relationship between other parameters and the optimal sample allocation.

For the applied researcher, it becomes more difficult to accurately determine optimal sample allocation due to the involvement of these additional influences. For these study designs it is imperative to have a robust theoretical and empirical literature base to provide a foundation for the parameter estimates used in determining optimal sample allocation. The additional

complexity and difficulty of determining optimal sample allocation in group-randomized studies of 2-2-1 mediation must not deter researchers from taking on the task as deviations from optimal design in these studies are more detrimental to study efficiency than similar studies of main effects, especially with small group to individual cost ratios. While determining optimal sample allocation in these designs is complex it is also crucial to planning an appropriately powered and efficient study.

Results from the application of the formulas suggested that the optimal individual sample size in these types of studies tends to be smaller than that of main effects—less than 10 under many common design parameter values and typically less than 30 under less common values. We caution readers that these rules of thumb are fallible and optimal sampling is influenced greatly by the factors detailed above (e.g., cost ratio, intraclass correlation, and path coefficients). We again emphasize the need for readers to carefully identify appropriate study-specific design parameter values and thoughtfully consider how they may influence optimal individual sample size.

Based on these rules of thumb and other results there are likely to be difference between optimal sampling for the main and mediated effects. Investigations of multilevel mediation are often conducted in conjunction with studies of main effects leading to the possibility of conflicting optimal sample allocations within a single design. In these cases, applied researchers must balance the considerations from optimal sample allocation for main and mediated effects. While these decisions may often involve study specific considerations, our results provide some guidance on balancing conflicting optimal sample allocations by outlining the potential loss of power for mediation effects associated with sub-optimal sampling. Future research, however,

should more explicitly consider the balance between optimal sampling for the main and mediation effects as well as identifying sampling strategies that are jointly or globally optimal.

Our investigation also provides an initial assessment regarding the relative efficiency of sub-optimal designs and their robustness to inaccurate parameter estimates. Even in well researched substantive areas, applied researchers are unlikely to perfectly specify the parameter values necessary for optimal design of a group-randomized study of 2-2-1 mediation (e.g., cost ratio, intraclass correlation, and path coefficients). For these reasons, it is important that future work further examines the relative efficiency of sub-optimal designs and the sensitivity of optimal sampling plans to incorrect initial parameter estimates (e.g., Korendijk, Moerbeek, & Maas, 2010). For example,

One practical limitation of our results is that with some parameter combinations the optimal individual sample size can be quite small—for instance, several example scenarios recommended optimal sampling schemes of only one, two, or three individuals per group (see Table 2). In these situations, researchers should use caution because missing data could quickly undermine design efficacy. It may be prudent to consider sampling additional individuals when attrition is likely or the parameter estimation technique warrants.

Here and in general, optimal sampling strategies are intended as a theoretical guide to sampling—researchers must balance theoretical optimums with practical constraints. For example, we again caution that our assumption of fully reliable measures of the outcome and mediator may be untenable in some educational research contexts and that failing to meet this assumption influences, power, sample size requirements, and the accuracy of parameter

estimates. The resulting choice of individual sample size for a given study should be a study-specific balance of these considerations.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

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

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: Classical and bootstrap estimates of variability. *Sociological Methodology*, *20*, 115–140.

Desimone, L. M., & Hill, K. L. (2017). Inside the black box: Examining mediators and moderators of a middle school science intervention. *Educational Evaluation and Policy Analysis*, *39*, 511-536.

Eden, M. (2017). School discipline reform and disorder: Evidence from New York City Public Schools. *The Education Digest*, *83*, 22-28.

Enders, C., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*, 121–138.

Fritz, M. S., & Mackinnon, D. P. (2007). Required sample size to detect the mediated effect. *Psychological Science*, *18*, 233–239.

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. V., & Borenstein, M. (2014). Conditional optimal design in three- and four-level experiments. *Journal of Educational and Behavioral Statistics*, *39*, 257-281.

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

Hox, J. J., Moerbeek, M., Kluytmans, A., & van de Schoot, R. (2014). Analyzing indirect effects in cluster randomized trials: The effect of estimation method, number of groups and group sizes on accuracy and power. *Frontiers in Psychology*, *5(78)*, 1–7.

Imai, K., Keele, L., & Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, *15*, 309–334.

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

Kelcey B., Dong, N., Spybrook, J., & Shen, Z. (2017). Statistical power for causally-defined mediation in group-randomized studies. *Multivariate Behavioral Research*, Advance online publication.

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., (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., 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.

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

Konstantopoulos, S. (2009). Incorporating cost in power analysis for three-level cluster-randomized designs. *Evaluation Review*, *33*, 335-357.

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

Krull, J. L., & MacKinnon, D. P. (2001). Multilevel modeling of individual and group level mediated effects. *Multivariate Behavioral Research*, *36*, 249–277.

Li, X., & Beretvas, S. N. (2013). Sample size limits for estimating upper level mediation models using multilevel SEM. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*, 241–264.

MacKinnon, D. P. (2008). *Introduction to statistical mediation analysis*. Mahwah, NJ: Lawrence Erlbaum.

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*, 83–104.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 37–67.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*, 27–29.

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., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods*, 40(3), 879–891.

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., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*, 199–213.

Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness*, *1*, 138–154.

Schneider, B., Carnoy, M., Kilpatrick, J., Schmidt, W.H., & Shavelson, R.J. (2007). Estimating causal effects using experimental and observational designs (report from the Governing Board of the American Educational Research Association Grants Program). Washington, DC: American Educational Research Association.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston: Houghton Mifflin.

Spybrook, J., Shi, R., & Kelcey, B. (2016). Progress in the past decade: An examination of the precision of cluster randomized trials funded by the U.S. Institute of Education Sciences. *International Journal of Research and Method in Education. 39,* 255–267.

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. (2015). *Explanation in causal inference: Methods for mediation and interaction*. New York, NY: Oxford University Press.

Yoshikawa, H., Leyva, D., Snow, C. E., Treviño, E., Barata, M. C., Weiland, C., Gomez, C. J., Moreno, L., Rolla, A., D'Sa, N., & Arbour, M. C. (2015). Experimental impacts of a

teacher professional development program in Chile on preschool classroom quality and child outcomes. *Developmental Psychology*, *51*, 309-322.

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.