Foundational Methods: Power Analysis

Randomized trials (RTs) are the gold standard for education impact studies as a well-designed and implemented RT yields an unbiased estimate of the treatment effect (Hedges, 2017). In an RT, individuals or groups of individuals are randomly assigned to either receive the treatment that is being tested or to a comparison condition, which often utilizes the business as usual (BAU) model or program. The intervention being tested might be a reading improvement program, an online math curriculum, or a professional development program for teachers. A RT is highly regarded in evaluating program impact because random assignment creates two or more equivalent groups based on observed or unobserved characteristics (Shadish, Cook, and Campbell, 2002). For example, the process of randomly assigning students to treatment and comparison groups would theoretically create two groups that have a similar composition of gender, race or ethnicity, and socioeconomic characteristics, or observed characteristics. In addition, random assignment also ensures that the two groups also have similar composition of unobserved characteristics, or characteristics that cannot be directly measured, such as attitude and self-efficacy (U.S. Department of Education, 2003). Creating equivalent treatment and comparison groups is an essential element in designing RTs because it partitions out the preexisting differences between groups to ensure that the observed treatment effect is the result of the intervention.

There are many components that contribute to the validity of the inferences in an RT, including the primary outcome measures and analytic plan (Shadish, Cook, and Campbell, 2002). Although all of these are important, in this chapter we focus specifically on the statistical power to detect an effect of a given magnitude. Statistical power is the probability of detecting an effect of a given magnitude when the true effect is non-zero. The probability of making a Type II error

increases as the statistical power of the study decreases. Low statistical power may lead to invalid inferences about the presence or absence of a treatment effect.

A statistical power analysis is typically carried out to determine the probability of an RT to detect the main treatment effect (Hedges and Rhoads, 2010). The main treatment effect is the average effect of the intervention. It is often quantified in terms of its effect size, defined as the standardized difference in the mean outcome between the treatment and comparison conditions. The sensitivity of an RT to detect the main treatment effect often focuses on three overlapping topics: the standard error associated with the treatment effect, the power to detect the main effect, and the smallest effect size a study is designed to detect (Hedges and Hedberg, 2012). The latter two topics are typically the focus of *a-priori* power analysis, or a power analysis is conducted during the planning stage of an RT.

Demonstrating that a study is sufficiently powered to detect an effect with a certain magnitude of interest is an imperative component of planning an RT. For example, suppose a team of researchers is developing a new online mathematics tutoring program for students and plans to test the impact of the program on mathematics achievement. They design an RT in which students are assigned to either the new online tutoring program or the current program. At the conclusion of the study, the team finds that there is no difference in mean outcomes for the two groups. A key question one might ask is whether there is no difference because the new program is not better than the current program or if the study simply is not sufficiently powered to detect a meaningful treatment effect? That is, was the sample size too small so that the estimated treatment effect was imprecise leading to a failure to reject the null hypothesis? This is not a question that one should ask at the end of a study. An *a-priori* power analysis is critical to

ensure that the study is designed to enable strong evidence of whether an intervention is effective.

The purpose of this chapter is to present an overview of statistical power for RTs seeking to test the impact of an educational intervention. We lay the foundation by beginning with the statistical power to detect the treatment effect for a single-level RT, or a single level study in which students are the unit of random assignment. Then we discuss the statistical power to detect the treatment effect for a cluster randomized trial (CRT) or a study in which clusters, such as schools, are assigned to conditions and outcomes are assessed at the student level. CRTs are common in impact studies in education given the nested structure of schools, students within schools, and the fact that interventions are often implemented at the school level. We end with a discussion of other types of statistical power analyses that are relevant for impact studies of educational interventions. Though our discussion focuses on studies with randomized designs, the models and methods of conducting power analyses also apply to non-randomized designs.

1. Power for Single Level RTs

1.1 Model without covariates

Recall the online mathematics tutoring program example. Students are randomly assigned to either the new program or the current program. Student mathematics achievement is the primary outcome. The level 1, or student-level model is:

$$Y_i = \beta_0 + \beta_1 T_i + e_i \qquad e_i \sim N(0, \sigma^2) \tag{1}$$

Where Y_i is the mathematics score for individual $i = \{1, \ldots, n\}$; β_0 is the overall mean mathematics score; β_1 is the mean difference between the treatment and comparison group or the main treatment effect; T_i is a treatment indicator with -½ for control and ½ for treatment, and e_i is the residual error associated with students with variance σ^2.

The treatment effect is estimated by $\hat{\beta}_1 = \bar{Y}_E - \bar{Y}_C$, where $\bar{Y}_E$ is the mean for the treatment group and $\bar{Y}_C$ is the mean for the control group. The variance of the estimated treatment effect, which describes uncertainty of the estimate effect, is:

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{P(1-P)n}. \quad (2)$$

Note the variance of the treatment effect is a function of the total sample size (n), the proportion of students assigned to the treatment condition (P), and the between-student variance within a treatment condition (σ^2). The associated hypothesis test is:

$$H_0: \beta_1 = 0 \quad (3)$$

$$H_1: \beta_1 \neq 0$$

This test is based on the F statistic, which is defined as:

$$F_{statistic} = \frac{(MS_{treatment})}{(MS_{error})} \quad (4)$$

which denotes the relationship between treatment variance and error variance.

The power for F-test is (Kirk, 1982):

$$\begin{aligned} Power &= Prob\ (Reject\ H_0 | H_0\ is\ false) \quad (5) \\ &= Prob\left(F > F_{\alpha;1,J-2}\right) \\ &= 1 - Prob\left(F < F_{\alpha;1,J-2}\right) \end{aligned}$$

where $F_{\alpha;1,J-2}$ is the critical value under the null hypothesis with 1 numerator degree of freedom and $n - 2$ denominator degrees of freedom. Under the null hypothesis, the F-statistic follow the central F-distribution. Under the alternative hypothesis, the F-statistic follows the noncentral F-distribution, represented by a non-centrality parameter λ. As Equation 6 shows, the non-

centrality parameter is a function of the true treatment effect and variance of the estimated treatment effect.

$$\lambda = \frac{\beta_1^2}{Var(\hat{\beta}_1)} = \frac{\beta_1^2}{\sigma^2/[P(1-P)n]} \quad (6)$$

It is common to standardize the parameters. The standardized treatment effect can be expressed as $\delta = \beta_1/\sqrt{\sigma^2}$. Set $\sigma^2 = 1$ and substituting δ into Equation 6, we get the standardized version as:

$$\lambda = \frac{\delta^2}{1/[P(1-P)n]} \quad (7)$$

The non-centrality parameter is related to the power of the test. That is, the power increases as the λ increases. Equation 7 suggests λ is a function of the sample size (n), the standardized treatment effect (δ), and the proportion of individuals in the treatment condition (P). As n increases, the non-centrality parameter increases, thus the power also increases. Applying this concept to the example of designing a study to test an online mathematics tutoring program, the researcher could improve the power to detect the treatment effect by increasing the number of students in the sample. This approach is often adopted by researchers since sample size typically under the control of the researcher. The power also increases as δ increases. However, the magnitude of the treatment effect is a function of the intervention and not something that is typically under the control of the researcher.

1.2 Model with a covariate

Researchers often include covariates to increase the precision of the estimated treatment effect. A pre-treatment measure, such as the students' mathematics achievement prior to

receiving the intervention, is often incorporated in the model. In this case, the level-1 or the student level model is an extension of Equation 1:

$$Y_i = \beta_0 + \beta_1 T_i + \beta_2 X_i + e_i \qquad e_i \sim N\left(0, \sigma^2_{|x}\right) \tag{8}$$

where X_i is the pre-treatment covariate and β_2 is the effect associated with the covariate. The proportion of variance explained by the covariate is $R^2_{|x} = 1 - \frac{\sigma^2_{|x}}{\sigma^2}$ where σ^2 is the between-person variance within a treatment condition and $\sigma^2_{|x}$ is the variance conditional on the covariate

Under the alternative hypothesis, the *F*-statistic follows a non-central *F*-distribution with 1 numerator and $n - 3$ denominator degrees of freedom. Note that the inclusion of a covariate reduces the degrees of freedom by 1. The standardized version of the non-centrality parameter is:

$$\lambda = \frac{n\delta^2}{1/[P(1-P)](1-R^2_{|x})} \tag{9}$$

As Equation 9 suggests, λ is a function of the sample size (n), the percent of variance explained by the covariate ($R^2_{|x}$), the treatment effect which was standardized by between-person variance within a treatment condition (δ), and the proportion of individuals assigned to the treatment condition (P). As $R^2_{|x}$ increases, the non-centrality parameter increases, thus the power also increases.

Continuing with the example of designing a study to test an online mathematics tutoring program, the researcher could include a pre-intervention measure that is highly correlated with the mathematics achievement outcome, such as the student's mathematics achievement outcome from the previous year, to increase the power of test for a given sample size and treatment effect.

1.3 Power calculation for individual level RCTs

There are two main approaches to conducting a power analysis: the sample size approach and the effect size approach. The sample size approach calculates the number of individual level units necessary for a study to detect a standardized main treatment effect of a given magnitude with a specified level of power, often power = 0.80, which is generally the acceptable level of power for designing educational studies. Software programs, like Optimal Design Plus (Raudenbush et al., 2011), CRT Power (Borenstein and Hedges, n.d.), and PowerUp! (Dong and Maynard, 2013) are freely available for conducting power analysis for individual-level and multi-level RTs. These software programs can also be used to conduct power analysis for studies with non-randomized designs.

Suppose the researchers designing the online mathematics tutoring program are interested in determining the number of students they need to detect a treatment effect of 0.20 standard deviations (SD). Figure 1 shows the study needs at least 790 students to detect a treatment effect of 0.20 at power = 0.80 if using a model without covariates (solid line). However, as the dashed line in Figure 1 demonstrates, the study only needs to recruit 288 students if using a model with a pretest covariate that explains 64% of the between-student variance within a treatment condition ($R^2_{|x}$ = 0.64). This is an example where choosing appropriate covariate(s) in the model could significantly reduce the cost associated with the study. Note that we set the target treatment effect size as 0.20 SD in this example. This is based on the effect sizes of 0.20 – 0.30 SD commonly found in education interventions (Hill et al., 2009)

[Insert Figure 1 here]

The effect size approach determines the minimum detectable effect size (MDES) a study can detect with a given number of individual-level units at a given level of power, typically

power = 0.80 (Bloom, 1995). Continuing with the example of an online mathematics tutoring program, suppose the research team is limited to a sample of 500 students. They want to know the MDES assuming power = 0.80. Figure 2 demonstrates that the MDES is 0.25 based on the model without covariates (solid line). However, the dashed line in Figure 2 shows that the MDES is 0.15 when considering a pretest covariate that explains 64% of the between-student variance within a treatment condition. In other words, the study could detect a smaller treatment effect with power = 0.80 when the model includes a covariate, while other parameters remain constant.

[Insert Figure 2 here]

2. Power for Two-Level CRTs

In a cluster randomized trial (CRT), groups of individuals, or clusters are randomly assigned to treatment and comparison conditions. This type of design is prevalent in education because individual units are naturally nested within clusters. For instance, students are naturally nested in schools, where students within one school experience the same type of school climate and administrative support. Because of this nesting structure, interventions are often assigned to schools and students within a school receive the same treatment.

2.1 Model without Covariates

Suppose a research team is interested in assessing the effectiveness of a supplemental reading program and plans to assign schools to either implement the program or carry on with the program that is already in place. The team seeks to determine the impact of the program on student reading achievement. This is a typical two-level CRT with students nested within schools, schools randomly assigned to treatment and comparison conditions and outcome measured at the student level.

The level-1, or the student level model is:

$$Y_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2) \tag{10}$$

where Y_{ij} is the reading score for individual $i = \{1, \ldots, n\}$; β_{0j} is the mean reading score for school j; e_{ij} is the residual error associated with students with variance σ^2.

The level 2, or the school-level model is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00}) \tag{11}$$

where γ_{00} is the grand mean reading achievement; γ_{01} is the mean difference between the treatment and control groups or the main treatment effect; T_j is a treatment indicator with -½ for control and ½ for treatment, and r_{0j} is the residual error associated with schools with variance τ_{00}.

The treatment effect is estimated by $\hat{\gamma}_{01} = \bar{Y}_E - \bar{Y}_C$ where $\bar{Y}_E$ is the mean for the treatment group and $\bar{Y}_C$ is the mean for the control group. The variance of the estimated treatment effect is:

$$Var(\hat{\gamma}_{01}) = \frac{(\tau_{00} + \sigma^2/n)}{JP(1-P)} \tag{12}$$

Equation 12 shows that the variance of the estimated treatment effect is a function of the between-cluster variance within a treatment condition (τ_{00}), within-cluster or between-student variance (σ^2), the sample size within a cluster (n), and the total number of clusters (J). The associated hypothesis is:

$$H_0: \gamma_{01} = 0$$

$$H_1: \gamma_{01} \neq 0 \tag{13}$$

and the power for the test is:

$$Power = Prob\ (Reject\ H_0 | H_0\ is\ false) \tag{14}$$

$$= Prob\left(F > F_{\alpha;1,J-2}\right)$$

$$= 1 - Prob\left(F < F_{\alpha;1,J-2}\right)$$

In the case of 2-Level CRT, the F-statistic = $\frac{(MS_{Treatment})}{(MS_{Cluster})}$. Under the null hypothesis, the F statistic follows a central F distribution with 1 numerator degree of freedom and $J - 2$ denominator degrees of freedom. Under the alternative hypothesis, the F statistic follows a non-central F distribution represented by the non-centrality parameter λ.

$$\lambda = \frac{\gamma_{01}^2}{Var(\hat{\gamma}_{01})} = \frac{JP(1-P)\gamma_{01}^2}{(\tau_{00}+\sigma^2/n)} \tag{15}$$

As in the case of the single-level RT, it is common to standardize the parameters, Hence we can re-express λ as:

$$\lambda = \frac{JP(1-P)\delta^2}{[\rho+(1-\rho)/n]} \tag{16}$$

where δ is the standardized treatment effect $\delta\ \delta = \frac{\gamma_{01}}{\sqrt{\tau_{00}+\sigma^2}}$ and ρ is the intraclass correlation (ICC) $\rho = \frac{\tau_{00}}{\tau_{00}+\sigma^2}$, or the proportion of between-cluster variance within a treatment condition relative to the total variance within a treatment condition. .

Recall that the power increases as the noncentrality parameter increases. Equation 16 demonstrates that power increases as the treatment effect (δ) increases, the sample size increases and specifically the total number of clusters (J) and the ICC (ρ) decreases. As noted earlier, the magnitude of the treatment effect is a function of the intervention and hence not typically under the control of the researcher. The ICC is a function of the clusters. That is, the more homogenous

the clusters, schools for example, the smaller the ICC. Therefore, the researcher may have some control in reducing the ICC during the sampling stage of the study. However, the sample size is still the parameters that the researcher can influence the most. Though it is important to note that increasing the total number of clusters (*J*) has a much stronger effect on the power to detect the main treatment effect than increasing the number of individuals per cluster (*n*).

2.2 Model with a cluster-level covariate

Similar to the case of an individual-level RT, it is common to include covariates to increase the precision of the estimated treatment effect (Bloom, Richburg-Hayes, and Black, 2007). In a two-level CRT, cluster-level covariates, such as school mean achievement scores are frequently incorporated in impact analyses because they are more readily available and less costly to collect compared to the data for individual-level covariates. In addition, they explain the between-cluster variance and thus drive down the ICC, which increases the power to detect the treatment effect, holding everything else constant. Hence, we limit our discussion to the case of cluster-level covariates.

The level-1 model is the same as the level-1 model without covariates. The level-2, or the cluster-level model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}W_j + r_{0j} \qquad r_{0j} \sim N\left(0, \tau_{00|W}\right) \tag{17}$$

where W_j is the cluster-level covariate and γ_{02} is the effect associated with the covariate. The proportion of between-cluster variance within a treatment condition explained by the covariate is $R^2_{|W} = 1 - \frac{\tau_{00|W}}{\tau_{00}}$.

Under the alternative hypothesis, the *F*-statistic follows a non-central *F*-distribution with 1 numerator and $J - 3$ denominator degrees of freedom. Note that the inclusion of a covariate reduces the degrees of freedom by 1. The standardized version of the non-centrality parameter is:

$$\lambda = \frac{JP(1-P)\delta^2}{\left[(1-R^2_{|W})\rho+(1-\rho)/n\right]} \quad (18)$$

Note that the difference between Equation 18 and the non-centrality parameter for the model without covariates (Equation 16) is the factor $\left(1 - R^2_{|W}\right)$, which reduces the between-cluster variance within a treatment condition. However, a cluster-level covariate does not reduce the within-cluster variance. The more variance explained by the cluster-level covariate, the greater the power, holding everything else constant.

The proportion of variance between-clusters within a treatment condition (ICC) and the proportion of variance explained by a cluster-level covariate ($R^2_{|W}$) are referred to as the design parameters for conducting power analysis. These design parameters are context-specific and depend on the specific outcome, sample, nesting structure, etc. Empirical estimates of these design parameters for some outcomes, such as student achievement outcomes, are widely available in the literature (Hedges and Hedberg, 2007; Hedges and Hedberg, 2012; Kelcey & Shen, 2016; Spybrook, Westin, and Taylor, 2016). The literature generally suggests the school-level ICCs, such that involved in the power calculations for two-level CRTs where schools are at level-2, are in the range of 0.14 – 0.26 for student mathematics and reading achievement outcomes (Hedges and Hedberg, 2007) and 0.17 – 0.31 for student science achievement outcomes (Spybrook, Westin, and Taylor, 2016). Similarly, studies of teacher development that draw on two-level CRTs where teachers are nested within schools 2 suggest that ICCs are in the range of 0.16 – 0.35 for mathematics knowledge for teaching and 0.08 to 0.24 for reading

knowledge for teaching (Kelcey & Phelps, 2013a; 2013b). The school-level R^2 values associated with achievement pretest covariates are 0.71 – 0.80 for student mathematics, reading, and science achievement outcomes (Hedges and Hedberg, 2007; Spybrook, Westin, and Taylor, 2016). These design parameters are pertinent in the accuracy of power calculation for CRTs. In the event the empirical estimates are not available for a particular outcome or context, a small pilot study in similar schools with a similar outcome may be useful for generating estimates of the design parameters.

2.3 Power calculation for two-level CRTs

The two approaches for conducting power calculation for individual-level RTs also apply to CRTs. Suppose a research team assigns the supplemental reading program to schools, such that half of the schools receive the intervention and the other half continues with the current program. The team is interested in examining the program's impact on students' reading achievement. Using the sample size approach, the team wants to determine the number of schools the study needs to detect a treatment effect of 0.20 with 50 students per school at power = 0.80. The team assumes that 15% of the variance is between the clusters (i.e., ICC = 0.15). Figure 3 shows the study needs at least 132 schools based on the model without any covariates (solid line). Suppose the team plans to include the school average reading score from a year prior to implementing the intervention as a level-2 covariate. Further, this covariate explains 80% of the between-cluster variance within a treatment condition. Figure 3 shows the study only needs approximately 40 schools to achieve a power of 0.80 (dashed line), which is approximately 1/3 of those needed for the model without any covariate.

[Insert Figure 3 here]

Similarly, the research team could use the effect size approach for power analysis. Suppose the same the research team has a limited budget and they could only recruit 50 schools and 50 students per school. Figure 4 demonstrates that the MDES is 0.33 based on the model without any covariates (solid line), which is larger than the 0.20 – 0.30 effect size that studies in education typically are designed to detect. However, when the school average pretest covariate is included as a cluster-level covariate that explains 80% of the between-cluster variance within a treatment condition, the MDES is 0.18 (dashed line in Figure 4). This new MDES is below the range of 0.20 – 0.30, which suggests that the study is sufficiently powered to detect a meaningful treatment effect. This example highlights the importance of cluster-level covariates in reducing the sample sizes needed to sufficiently power a study.

[Insert Figure 4 here]

3. Future Directions

As the field is accumulating more knowledge on the main treatment effects of educational interventions, which answers the "*what works?*" question, researchers are also interested in the capacity of RTs to answer more questions. For example, "*for whom?*", "*under what conditions?*", and "*how?*" questions are quintessential for understanding the applicability of interventions across educational contexts and the underlying mechanisms the interventions work to improve student achievement. Researchers often rely on moderator and mediator analyses for answering these questions. As such, more researchers are planning studies and conducting power analyses for main, moderator, and mediator effects (Spybrook et al., 2020).

Moderator effects represent differential treatment effects. For example, one may ask, does the treatment have a differential effect based on gender or pretest score? From a modeling

perspective, they are represented by an interaction term between the moderator and the treatment. In an individual-level RT, moderators are the individual-level. In a two-level CRT, moderators may be at the individual-level or cluster-level. The literature on calculating power for moderator effects in CRTs is growing (e.g. Bloom, 2005; Jaciw, Lin, and Ma, 2016; Spybrook, Kelcey, and Dong, 2016; Dong et al., in press; Dong, Kelcey, and Spybrook, in press) and software is also available for conducting these analyses (PowerUp!-Moderator).

Mediator effects are key to understanding the mechanisms through which an intervention comes to be effective. For example, the researcher may want to test the extent to which a new science curriculum operates by first increasing students' interest in science such that the increased interest leads to greater science achievement. Interest in science represents the mediator in this example. Literature for power calculations for mediator effects is also growing, particularly for CRTs (e.g. Kelcey et al., 2020; Kelcey, Spybrook, and Dong, 2019) and software is also available for conducting these analyses (PowerUp!-Mediator).

However, there are many challenges ahead. One challenge being the moderator and mediator effect sizes studies should aim to detect. Education researchers typically plan CRTs to detect the main treatment effect of 0.20 – 0.30 SDs (Hill et al., 2008). The same reference effect sizes for planning studies to detect a moderator or mediator effects are in need of further exploration. As more CRTs are incorporating these types of analyses in their designs, we hope to see more evidence of moderation and mediation effects accumulate in the field.

References

Bloom, H.S. (1995). Minimum detectable effect: a simple way to report the statistical power of experimental designs. *Evaluation Review*, 19, 547-556.

Bloom, H. S. (2005). *Randomizing groups to evaluate place-based programs*. In H. S. Bloom (Ed.), Learning more from social experiments: evolving analytic approaches (pp. 115-172). New York, NY: Russel Sage.

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(1), 30-59.

Borenstein, M., & Hedges, L. V. (n.d.). *CRT-power*. Englewood, NJ: Biostat.

Dong, N. and Maynard, R. A. (2013). *PowerUp!*: A tool for calculating minimum detectable effect sizes and sample size requirements for experimental and quasi-experimental designs. *Journal of Research on Educational Effectiveness*, *6*(1), 24-67.

Dong., N., Kelcey, B., Spybrook, J., & Bulus, M. (in press). Power analyses for moderator effects with (non)randomly varying slopes in cluster randomized trials. *Methodology*.

Dong, N., Kelcey, B., & Spybrook, J. (in press). Design considerations in multisite randomized trials to probe moderated treatment effects. *Journal of Educational and Behavioral Statistics*.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29(1), 60-87.

Hedges, L. V., Hedberg, E. C., & Kuyper, A. M. (2012). The variance of intraclass correlation in three- and four-level models. *Educational and Psychological Measurement*, 72(6), 893-909.

Hedges, L.V., & Rhodes, C. (2010). *Statistical power analysis in education research.* U.S. Department of Education. Washington, DC: National Center for Special Education Research, Institute of Education Sciences.

Hedges, L. V. (2017). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, 11(1), 1-21.

Hill, C. J.; Bloom, H. S.; Black, A. R.; & Lipsey, M. (2008). empirical benchmark for interpreting effect sizes in research. *Child Development Perspectives*, 2(3), 172-177.

Jaciw, A. P., Lin, L., & Ma, B. (2016). An empirical study of design parameters for assessing differential impacts for students in group randomized trials. *Evaluation Review*, 40(5), 410-443.

Kelcey, B., Dong, N., Spybrook, J., & Shen, Z. (2017) Experimental Power for Indirect Effects in Group-randomized Studies with Group-level Mediators. *Multivariate Behavioral Research*, 52(6), 699-719.

Kelcey, B., Spybrook, J., & Dong, N. (2019) Sample size planning for cluster-randomized interventions probing multilevel mediation. *Prevention Science*, 20(3), 407-418.

Kelcey, B., Xie, Y., Spybrook, J., & Dong, N. (2020). Power and sample size determination for multilevel mediation in three-level cluster randomized trials. *Multivariate Behavior Research*.

Kirk, R. E. (1982). *Experimental design*. New York, NY: John Wiley.

Raudenbush, S. W., Spybrook, J., Congdon, R., Liu, X. F., Martinez, A., & Bloom, H. (2011). *Optimal design software for multi-level and longitudinal research (Version 3.01)* [Software].

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-experimental Designs*. Belmont, CA: Wadsworth Cengage Learning.

Spybrook, J., Kelcey, B., & Dong, N. (2016). Power for detecting treatment by moderator effect in two and three-level cluster randomized trials. *Journal of Educational and Behavior Statistics,* 41(6), 605-627.

Spybrook, J., Westine, C. D., & Taylor, J. A. (2016). Design parameters for impact research in science education: a multistate analysis. *AERAOpen*, 2(1), 1-15.

Spybrook, J., Zhang, Q., Kelcey, B., & Dong, N. (2020). Learning from Cluster Randomized Trials in Education: an assessment of the capacity of studies to determine what works, for whom, and under what condition. *Educational Evaluation and Policy Analysis*, 42(3), 354-374.

U.S. Department of Education. (2003). *Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide*. Washington, DC: Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

U.S. Department of Education. (2012). *National Board for Education Science (NBES) briefing material for board members.* Washington DC: Institute of Education Science.

U.S. Department of Education. (2010). *Request for Applications: Education Research Grants.* Washington, DC: Institute of Education Sciences.

U.S. Department of Education. (2018). *Request for Applications: Education Research Grants.* Washington, DC: Institute of Education Sciences.