Teacher's Corner

Journal of Educational and Behavioral Statistics
Vol. XX, No. X, pp. 1–23
DOI: 10.3102/1076998616655442

Power for Detecting Treatment by Moderator Effects in Two- and Three-Level Cluster Randomized Trials

Jessaca Spybrook
Western Michigan University

Benjamin Kelcey
University of Cincinnati

Nianbo Dong
University of Missouri–Columbia

Recently, there has been an increase in the number of cluster randomized trials (CRTs) to evaluate the impact of educational programs and interventions. These studies are often powered for the main effect of treatment to address the "what works" question. However, program effects may vary by individual characteristics or by context, making it important to also consider power to detect moderator effects. This article presents a framework for calculating statistical power for moderator effects at all levels for two- and three-level CRTs. Annotated R code is included to make the calculations accessible to researchers and increase the regularity in which a priori power analyses for moderator effects in CRTs are conducted.

Keywords: *statistical power; cluster randomized trials; moderator effects*

In the past 15 years, there has been a strong shift toward the use of randomized trials (RTs), and specifically cluster RTs (CRTs), to evaluate the impact of educational programs and interventions. In CRTs, intact clusters (e.g., schools) are assigned to treatment conditions rather than individuals (e.g., students). CRTs are frequently an effective way to study interventions because they permit researchers to accommodate existing school structures and interventions that are designed to operate at the school level (Spybrook & Raudenbush, 2009). In order to yield rigorous evidence of whether a program works, however, such studies must be carefully designed. A principal consideration in the design of CRTs is the power or probability with which a study can detect effects if they exist.

The body of literature on statistical power for CRTs has largely focused on detecting average/main effects of treatment. A sizable number of articles and books have been published on this topic (i.e., Bloom, 1995; Donner & Klar,

2000; Hedges & Rhoads, 2009; Konstantopoulos, 2008; Liu, 2014; Murray, 1998; Raudenbush, 1997; Raudenbush & Liu, 2001; Schochet, 2008). The designs covered include two- and three-level CRTs as well as blocked or multisite CRTs (MSCRTs). This body of literature has clearly established that (1) the number of clusters is more influential than the number of individuals per cluster in terms of increasing the power of a study to detect the main effect of treatment of a given magnitude; (2) the more variability in the outcome across clusters, the greater the number of clusters needed; and (3) including a cluster-level covariate that is highly correlated with the outcome is often a cost-effective and efficient strategy for increasing the power.

Conducting the power calculations for the main effect of treatment for CRTs has also become much easier. Standard statistical software programs, for example, SAS Version 9.4, allow users to conduct power calculations for CRTs using procedures for mixed models (Littell, Milliken, Stroup, Wolfinger, & Schabenberger, 2006). In addition, several stand-alone programs for power calculations for CRTs exist including Optimal Design Plus (Raudenbush et al., 2011), CRT Power (Borenstein & Hedges, n.d.), and PowerUp! (Dong & Maynard, 2013).

A priori power calculations for the main effect of a treatment help ensure the study has the capacity to address the "what works" question. However, there is a growing recognition that there are important explanatory questions that need to be addressed if we are to fully understand the validity and value of substantive theories and interventions in education. One critical line of inquiry that is largely missing in conventional study designs concerns treatment effect moderation—or questions examining "for whom and under what circumstances" an intervention works. For example, it may be that an intervention is more effective in urban schools compared to rural schools or for girls compared to boys, such that school or individual characteristics moderate the treatment effect. Understanding the context in which an intervention is likely to be effective is fundamental to understanding the extent to which results are scalable and applicable to a wide range of schools and students.

The importance of studying moderation has gained considerable momentum in the field. For instance, in 2012, the conference theme for the annual meeting of the Society for Research on Educational Effectiveness was *Understanding Variation in Treatment Effects* and highlighted the importance of understanding how to design studies to enable them to better assess heterogeneity of treatment effects. Moderator effects that measure the treatment effect difference between subgroups represent one type of heterogeneous treatment effect. More recently, funders have started to strongly recommend a priori power analyses for tests of moderator effects (Institute of Education Sciences, 2016, p. 60). However, the literature for conducting power analyses for moderator effects in CRTs is less developed than for main effects.

Much like the case of power for the main effect of treatment in CRTs, the classic experimental design literature provides a framework for power

calculations for moderator effects in CRTs. For example, one could consider a two-level CRT with a moderator at the individual level as a split-plot design with treatment as a whole-plot factor and the individual-level moderator as a split-plot factor (Littell et al., 2006). However, as evidenced by the large literature on power for main effects for CRTs, the reformulation of such designs within the familiar purview of hierarchical or multilevel models is prominent in education. This reframing facilitates direct connections among multilevel designs, hypothesis testing, and multilevel models that reduce power calculations to principles that are more concrete and accessible to researchers. Such restructuring has promoted a more informed appreciation for the factors that govern power and led to more reasonable approximations of power in recent CRTs (e.g., Spybrook & Raudenbush, 2009). Hence, it is critical to present power calculations for moderator effects within the context of CRTs and multilevel models and directly connect them to power calculations for the main effect. Only a handful of articles have done this. Raudenbush and Liu (2001) derive power formulas for site-level moderator effects for multisite trials in which individuals are randomly assigned within sites. Bloom (2005) and Jaciw (2014) focus on two-level CRTs with a binary Level 1 or Level 2 moderator. They provide formulas for the minimum detectable effect size difference (MDESD) or the smallest effect size difference that can be detected with power set to 0.80. Spybrook (2014) provides empirical estimates of the power of a set of funded CRTs to detect moderator effects but does not delineate an approach to estimate the power to detect moderation within the context of CRTs.

Statistical software options for calculating power for moderator effects for CRTs are also more limited than for power for the main effect. For example, none of the three most widely used programs for calculating power for CRTs, Optimal Design Plus (Raudenbush et al., 2011), CRT Power (Borenstein & Hedges, n.d.), or PowerUp! (Dong & Maynard, 2013), have specific functionality for calculating power for testing moderation.

The purpose of this article is to extend the literature and the tools available for power analyses for moderator effects in nested CRTs. As mentioned above, Bloom (2005) and Jaciw (2014) present MDESD formulas for the two-level CRT. We extend this work to power calculations for binary moderators at any level in a three-level CRT. We also implement the power formulas for moderator effects for the two-level and three-level CRT through two user-friendly tools to facilitate the use of these power formulas in planning CRTs. The tools include annotated R code and implementation of the formulas in PowerUp![1] (http://www.causalevaluation.org/). We expect these tools will help make this work accessible to education researchers and increase the regularity in which a priori power analyses for moderator effects in CRTs are conducted.

The article is organized as follows. We begin with the model for a two-level CRT and briefly walk through the power calculations for the main effect of treatment. This is for pedagogical purposes, as it allows us to anchor notation

and concepts in the more familiar power analyses for the main effect of treatment and directly transfer these to the less familiar power analyses for moderator effects. Next we provide the model and tests for a cluster-level and individual-level moderator in a two-level CRT and three-level CRT for balanced designs. To make direct connection among the approaches, we purposefully unpack and connect the models, test statistics, and noncentrality parameters. Then we extend to the case of unbalanced designs. Next we present several practical and deliberate examples of how to conduct a power analysis for different moderator effects. In the concluding section, we summarize the key components of the power calculations, explore the results and the implications of powering for moderator effects in the design of two- and three-level CRTs, and discuss future directions for this work.

Two-Level CRTs

Main Effect of Treatment

Suppose a team of researchers are planning a two-level CRT with students nested within schools and treatment assigned at the school level. Mathematics achievement is the outcome of interest. The Level 1 or student-level model is

$$Y_{ij} = \beta_{0j} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2), \tag{1}$$

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

The Level 2 model or cluster-level model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00}), \tag{2}$$

where γ_{00} is the grand mean math achievement; γ_{01} is the mean difference between the treatment and control group or the main effect of treatment; 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}. We assume equal allocation of clusters to treatment and control.

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 (Raudenbush, 1997)

$$\mathrm{Var}(\hat{\gamma}_{01}) = 4(\tau_{00} + \sigma^2/n)/J. \tag{3}$$

Note the variance is a function of the within-cluster variance, σ^2; the between-cluster variance, τ_{00}; the sample size within cluster, n; and the total number of clusters, J. The 4 is a result of $J/2$ clusters per condition, since we are assuming a balanced design.[2]

In this case, we are testing H_0: $\gamma_{01} = 0$. The power for the test is (Kirk, 1982)

$$\begin{aligned} \text{Power} &= \text{Prob}\ (\text{Reject}\ H_0 | H_0\ \text{is false}) \\ &= \text{Prob}\ (F > F_{\alpha;1,J-2}) \\ &= 1 - \text{Prob}\ (F < F_{\alpha;1,J-2}), \end{aligned} \tag{4}$$

where F is the F-statistic $= \frac{\text{MS}_T}{\text{MS}_C}$ from the sample (see Kirk, 1982, for details), MS_T is the mean squares for the treatment, MS_C is the mean squares for the cluster, and $F_{\alpha;1,J-2}$ is the critical value under the null hypothesis with 1 numerator degree of freedom and $J - 2$ denominator degrees of freedom.

If the null hypothesis is true, then the F-statistic follows the central F-distribution. If the null hypothesis is false, then the F-statistic follows the noncentral F-distribution with a noncentrality parameter, λ. The noncentrality parameter is a ratio of the squared main effect of treatment to the variance of the estimated treatment effect, as shown in Equation 5

$$\lambda = \frac{\gamma_{01}^2}{\text{Var}(\hat{\gamma}_{01})} = \frac{\gamma_{01}^2}{4(\tau_{00} + \sigma^2/n)/J}. \tag{5}$$

As the noncentrality parameter increases, the power increases. Thus, for the main effect of treatment in a two-level CRT, increasing the total number of clusters, J, has a greater effect on increasing power than increasing the total number of individuals per cluster, n, holding everything else constant. Note that it is common to standardize the parameters and reexpress λ as

$$\lambda = \frac{\delta^2}{4[\rho + (1-\rho)/n]/J}, \tag{6}$$

where $\rho = \frac{\tau_{00}}{\tau_{00}+\sigma^2}$ is the intraclass correlation (ICC) or percentage of variance between clusters, and $\delta = \frac{\gamma_{01}}{\sqrt{\tau_{00}+\sigma^2}}$ is the standardized effect size.

Before we move to the cluster-level moderator, we outline the extension for the main effect of treatment to the case with a cluster-level covariate. It is common practice to include a cluster-level covariate in the design of a CRT in order to increase the precision of the estimate (Raudenbush, Martinez, & Spybrook, 2007). Although an individual-level covariate may also be included, we focus on the cluster-level covariate because this directly reduces the between-cluster variance and is often more readily available and less expensive to collect than an individual-level covariate (Bloom, Richburg-Hayes, & Rebeck-Black, 2007). In this case, the Level 2 model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \gamma_{02} W_j + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00|W}). \tag{7}$$

The proportion of Level 2 variance explained by the covariate W is $R^2_{|W} = 1 - \frac{\tau_{00|W}}{\tau_{00}}$. Using the standardized parameters, the noncentrality

parameter for the case with one cluster-level covariate and hence $J - 3$ degrees of freedom is

$$\lambda_{|W} = \frac{\delta^2}{4\left[(1 - R^2_{|W})\rho + \left((1 - \rho)/n\right)\right]/J}. \tag{8}$$

Note that the cluster-level covariate cannot reduce the variance at Level 1 since it is the same within clusters.

Cluster-level moderator. Suppose that the pool of schools in the previous study includes different types of schools, such as urban and rural schools. The research team suspects that the treatment effect may differ in urban schools compared to rural schools. Hence, they are interested in whether type of school, urban or rural, moderates the treatment effect. For illustrative purposes, suppose half the schools in the study are urban and half are rural and that they are equally allocated across conditions. The Level 1 model is identical to Equation 1. The Level 2 model or cluster-level model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01} T_j + \gamma_{02} S_j + \gamma_{03}(T_j S_j) + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00|S}), \tag{9}$$

where γ_{00} is the grand mean; γ_{01} is the mean difference between the treatment and control group; T_j is a treatment indicator, with $-½$ for control and ½ for treatment; S_j is a school type indicator, with $-½$ for urban and ½ for rural; γ_{02} is the school type effect, γ_{03} is the Treatment $\times$ School Type interaction; and r_{0j} is the residual error associated with clusters with variance $\tau_{00|S}$. The proportion of Level 2 variance explained by the moderator S and the interaction of S and T is $R^2_{|S} = 1 - \frac{\tau_{00|S}}{\tau_{00}}$.

The moderator effect is estimated by $\hat{\gamma}_{03} = [\bar{Y}^R_E - \bar{Y}^U_E] - [\bar{Y}^R_C - \bar{Y}^U_C]$. The variance of the estimated moderator effect is

$$\mathrm{Var}(\hat{\gamma}_{03}) = 16[(1 - R^2_{|S})\tau_{00} + \sigma^2/n]/J. \tag{10}$$

Note that the 16 in front is a function of the fact that there are now $J/4$ clusters per condition since there are now four conditions, rural experimental, urban experimental, rural comparison, and urban comparison.

The power for the cluster-level moderator effect, γ_{03}, is an extension of the power for the main effect of treatment. The hypothesis of interest in this case is H_0: $\gamma_{03} = 0$, the F-statistic is a ratio of $MS_{T:S}$, which is the mean squares for the interaction, to the MS_C, the degrees of freedom for the test are $J - 4$, and the noncentrality parameter, $\lambda_{|S}$ is the ratio of the squared treatment effect to the variance of the estimated moderator effect

$$\lambda_{|S} = \frac{\gamma^2_{03}}{16[(1 - R^2_{|S})\tau_{00} + \sigma^2/n]/J}. \tag{11}$$

For consistency with main effect calculations, we standardize by setting $\tau_{00} + \sigma^2 = 1$. Hence, the noncentrality parameter using standardized notation is

$$\lambda_{|S} = \frac{\delta^2_{\text{CLmod}}}{16[(1 - R^2_{|S})\rho + (1 - \rho)/n]/J}, \tag{12}$$

where $\delta_{\text{CLmod}} = \frac{\gamma_{03}}{\sqrt{\tau_{00}+\sigma^2}}$.

As discussed above, it is common to include a cluster-level covariate to increase precision, thus the new Level 2 model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}T_j + \gamma_{02}S_j + \gamma_{03}(T_jS_j) + \gamma_{04}W_j + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00|WS}). \tag{13}$$

The addition of the cluster-level covariate further reduces the Level 2 variance where $R^2_{|WS} = 1 - \frac{\tau_{00|WS}}{\tau_{00}}$ is the proportion of Level 2 variance explained by the covariate W, the moderator S, and the interaction of S and T. The standardized noncentrality parameter, with $J - 5$ degrees of freedom, is

$$\lambda_{|WS} = \frac{\delta^2_{\text{Clmod}}}{16\left[\left(1 - R^2_{|\text{WS}}\right)\rho + \left((1 - \rho)/n\right)\right]/J}. \tag{14}$$

Similar to the main effect of treatment, it is clear that the total number of clusters is the key sample size for increasing the power to detect cluster-level moderator effects. However, there are also important differences in the noncentrality parameters in Equations 7 and 14. First, the multiplier in the variance of the estimated treatment effect is 4 times larger for the moderator effect. This is a result of having $J/4$ clusters per condition rather than $J/2$ clusters per condition. Second, the set of covariates being conditioned on differs. For main effects, we condition on cluster-level covariate(s), whereas for moderator effects, we condition on the cluster-level covariate(s) and the moderator. Third, the numerator is a standardized differential treatment effect rather than a main effect. We briefly consider the role of these three factors before we move to the individual-level moderator.

First, consider the case in which, $R^2_{|WS}$ is equal to $R^2_{|W}$. In this case, the moderator is not explaining any additional variance. Hence, the variance of the estimated treatment effect for the cluster-level moderator is 4 times greater than the main effect of treatment. If we set the magnitude of the treatment effect and moderator effect to be the same, then it is clear that more clusters would be needed to achieve the same level of power for the treatment effect and cluster-level moderator effect. However, the situation is actually more challenging because it seems likely that the magnitude of the cluster-level moderator effect will be smaller than the main effect of treatment in many practical settings because it is the difference in the treatment effect for two groups (Aguinis, Beaty, Boik, & Pierce, 2005). The smaller size of the cluster-level moderator combined

with the larger variance means that many more clusters would be needed to achieve a given level of power for the moderator effect than are typically needed for the main effect of treatment.

Next, suppose that $R^2_{|WS}$ is greater than $R^2_{|W}$. Mathematically it is clear that this will help reduce the variance of the moderator effect compared to the main effect of treatment helping to improve the power for the moderator effect. However, in education studies, binary moderators generally explain little additional variance beyond that explained by commonly used cluster-level covariates. This is primarily because the most common cluster-level covariate in education studies is a pretest. Studies have shown that cluster-level pretests are likely to explain 60% to 80% of the variation in the outcome (see Spybrook, 2013, for a review of empirical studies). Further, school characteristics have not been shown to explain much additional variation beyond the pretest (Bloom et al., 2007). This means that although $R^2_{|WS}$ may be greater than $R^2_{|W}$, the difference is not likely to be large. Combined with the fact that the magnitude of the moderator effect will likely be smaller than the main effect of treatment, it will be challenging to achieve adequate power to detect a cluster-level moderator effect.

Individual-level moderator. We might be interested in whether gender moderates the treatment effect. The Level 1 model or student-level model is now

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{ij} + e_{ij} \qquad e_{ij} \sim N(0, \sigma^2_{|x}), \tag{15}$$

where Y_{ij} is the math achievement for individual $i = \{1, \ldots, n\}$ in school $j = \{1, \ldots, J\}$; β_{0j} is the mean achievement in school j; X_{ij} is an indicator for gender, with $-½$ for boys and ½ for girls; β_{1j} is the gender gap in school j; and e_{ij} is the residual error associated with students. Note that gender explains the variation at Level 1 such that $R^2_{|X} = 1 - \frac{\sigma^2_{|X}}{\sigma^2}$. We hold the gender effect constant within clusters. For pedagogical reasons, we assume that the means of gender are same among clusters and do not include the aggregated version of gender in the Level 2 model. However, this could easily be included in the models below. The school-level model is

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}T_j + r_{0j} \qquad r_{0j} \sim N(0, \tau_{00}), \\ \beta_{1j} &= \gamma_{10} + \gamma_{11}T_j \end{aligned} \tag{16}$$

where γ_{00} is the mean achievement across schools; T_j is a treatment indicator, with $-½$ for control and ½ for treatment; γ_{01} is the average treatment effect; γ_{10} is the gender gap; γ_{11} is the Treatment × Gender interaction; and r_{0j} is the error associated with mean achievement across schools with variance τ_{00}. Note we do not allow the gender gap to vary randomly across schools, although the model could be modified to reflect a random gender gap.

The individual-level moderator effect is estimated by $\hat{\gamma}_{11} = [\bar{Y}_E^G - \bar{Y}_E^B] - [\bar{Y}_C^G - \bar{Y}_C^B]$. The variance of the estimated moderator effect is

$$\text{Var}(\hat{\gamma}_{11}) = 16[(1 - R^2_{|x})\sigma^2]/nJ. \quad (17)$$

Similar to the case of a cluster-level moderator, the 16 in front of the variance is a function of the four groups, girls in treatment, boys in treatment, girls in comparison, and boys in comparison. However, unlike the variance for the cluster-level moderator effect, the between-school variance, τ_{00}, does not contribute to the variance of the estimated moderator effect. This is because the differences in boys and girls are within schools and hence school effects cancel out.

The hypothesis of interest in this case is H_0: $\gamma_{11} = 0$, the F-statistic is a ratio of $MS_{T:X}$, which is the mean squares for the interaction effect, to the MS_C with degrees of freedom $n \times J - J - 2$. Given that the noncentrality parameter is a ratio of the squared treatment effect to the variance of the estimated treatment effect, it can be expressed as

$$\lambda_{|x} = \frac{\gamma_{11}^2}{[16(1 - R^2_{|x})\sigma^2]nJ}. \quad (18)$$

In order to be able to compare results with the power for the main effect of treatment and cluster-level moderators, we standardize the same way as above

$$\lambda_x = \frac{\delta^2_{\text{INDmod}}}{[16(1 - R^2_{|x})(1 - \rho)]/nJ}, \quad (19)$$

where $\delta_{\text{INDmod}} = \frac{\gamma_{11}}{\sqrt{\tau_{00}+\sigma^2}}$.

There are important differences in the noncentrality parameter for the individual-level moderator compared to the noncentrality parameters for main effect of treatment and the cluster-level moderator. The key difference is that the between-cluster variance is not a part of the denominator for individual-level moderator effects. As a result, the number of individuals per cluster becomes as important as the total number of clusters. This differs from the main of treatment and the cluster-level moderator effect where the number of individuals per cluster was less critical and the number of clusters was the key sample size.

Before we move to the three-level case, we briefly summarize the key findings from the two-level case. From a sample size perspective, the total number of clusters is the most influential sample size for increasing the power to detect the main effect of treatment and a cluster-level moderator effect. However, the variance of the cluster-level moderator effect can be up to 4 times as large as the main effect of treatment which means many more clusters are necessary in order to detect a treatment effect of the same magnitude. Given that moderator effects tend to be smaller than main effects and education CRTs are typically

designed to detect the main effect, the potential to design education CRTs with the capacity to detect cluster-level moderator effects of a reasonable magnitude may be limited. The situation is much more optimistic for designing two-level CRTs to detect individual-level moderator effects. This is a result of two factors: the between-school variance does not impact the power calculations and the number of individuals per cluster is equally as important as the total number of clusters. Hence, for a fixed total number of clusters, while increasing the number of individuals per cluster will not yield measurable gains for the power to detect the main effect of treatment or the cluster-level moderator effect, it has the potential to yield important gains in the power for the individual-level moderator effect.

Three-Level CRT

Main Effect of Treatment

Next we extend the work to the case of a three-level CRT. Suppose the same team of researchers are considering including a middle level in the study, teachers, so that they have a three-level CRT with students nested within teachers nested within schools. Math achievement remains the outcome of interest. The Level 1 or student-level model is

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2), \tag{20}$$

where Y_{ijk} is the math achievement for individual $i = \{1, \ldots, n\}$ in teacher $j = \{1, \ldots, J\}$ in school $k = \{1, \ldots, K\}$; π_{0jk} is the mean math achievement for teacher j in school k; and e_{ijk} is the error associated with students with variance σ^2. The Level 2 or teacher-level model is

$$\pi_{0jk} = \beta_{00k} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_\pi), \tag{21}$$

where β_{00k} is the mean math achievement for school k and r_{0jk} is the error associated with teachers with variance τ_π. The Level 3, or school-level model is

$$\beta_{00k} = \gamma_{000} + \gamma_{001} T_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}}), \tag{22}$$

where γ_{000} is the grand mean math achievement; γ_{001} is the mean difference between the treatment and control group or the main effect of treatment; T_k is a treatment indicator, with $-½$ for control and ½ for treatment; and u_{00k} is the error associated with schools with variance $\tau_{\beta 00}$. We assume equal allocation of clusters to treatment and control.

The treatment effect is estimated by $\hat{\gamma}_{001} = \bar{Y}_E - \bar{Y}_C$. The variance of the estimated treatment effect is

$$\mathrm{Var}(\hat{\gamma}_{001}) = \frac{4[\tau_{\beta_{00}} + (\tau_\pi + \sigma^2/n)J]}{K}. \tag{23}$$

The hypothesis of interest is H_0: $\gamma_{001} = 0$, the F-statistic is a ratio MS_T to the MS_C, the degrees of freedom for the test are $K - 2$, and the noncentrality parameter is

$$\lambda = \frac{\gamma_{001}^2}{4[\tau_{\beta_{00}} + (\tau_\pi + \sigma^2/n)/J]/K}. \quad (24)$$

Like in the case of the two-level CRT, it is common to standardize the parameters such that

$$\lambda = \frac{\delta^2}{4\{\rho_\beta + [\rho_\pi + (1 - \rho_\beta - \rho_\pi)/n]/J\}/K}, \quad (25)$$

where $\delta = \frac{\gamma_{001}}{\sqrt{\tau_{\beta_{00}} + \tau_\pi + \sigma^2}}$ is the standardized effect size; $\rho_\beta = \frac{\tau_{\beta_{00}}}{\tau_{\beta_{00}} + \tau_\pi + \sigma^2}$ is the ICC at Level 3, or the percentage of the total variance at Level 3; and $\rho_\pi = \frac{\tau_\pi}{\tau_{\beta_{00}} + \tau_\pi + \sigma^2}$ is the ICC at Level 2, or the percentage of the total variance at Level 2.

It is also typical to include a school-level covariate to increase the power of the study. Assuming a school-level covariate, such as school-level pretest, the Level 3 variance will be reduced by $R^2_{|W} = 1 - \frac{\tau_{\beta_{00}|W}}{\tau_{\beta_{00}}}$. In this case, the noncentrality parameter is

$$\lambda_W = \frac{\delta^2}{4\{(1 - R^2_{|W})\rho_\beta + [\rho_\pi + (1 - \rho_\beta - \rho_\pi)/n]/J\}/K}. \quad (26)$$

School-level moderator. Again we assume that half the schools in the study are urban and half are rural and that they are equally allocated across conditions. The Level 1 and Level 2 models are identical to Equations 20 and 21. The new Level 3 model is

$$\beta_{00k} = \gamma_{000} + \gamma_{001}T_k + \gamma_{002}S_k + \gamma_{003}(T_kS_k) + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}|S}), \quad (27)$$

where γ_{000} is the grand mean math achievement; γ_{001} is the mean difference between the treatment and control group; T_k is a treatment indicator, with $-½$ for control and ½ for treatment; S_k is a school type indicator, with $-½$ for urban and ½ for rural; γ_{002} is the school type effect; γ_{003} is the Treatment × School Type interaction or the school-level moderator effect; and u_{00k} is the residual error associated with schools. Note that where $R^2_{|S} = 1 - \frac{\tau_{\beta_{00}|S}}{\tau_{\beta_{00}}}$ is the proportion of Level 3 variance explained by the moderator and the interaction of the moderator and treatment.

The moderator effect is estimated by $\hat{\gamma}_{003} = \left[\bar{Y}^R_E - \bar{Y}^U_E\right] - \left[\bar{Y}^R_C - \bar{Y}^U_C\right]$. The variance of the estimated moderator effect is

$$\mathrm{Var}(\hat{\gamma}_{003}) = 16[(1 - R^2_{|S})\tau_{\beta_{00}} + (\tau_\pi + \sigma^2/n)/J]/K. \quad (28)$$

The hypothesis of interest in this case is H_0: $\gamma_{003} = 0$, the F-statistic is a ratio of $MS_{T:S}$ to the MS_C, the degrees of freedom for the test are $K - 4$, and the noncentrality parameter, $\lambda_{|S}$, is

$$\lambda_{|S} = \frac{\gamma^2_{003}}{16[(1 - R^2_{|S})\tau_{\beta_{00}} + (\tau_\pi + \sigma^2/n)/J]/K}. \tag{29}$$

Hence, the noncentrality parameter using standardized notation is

$$\lambda_{|S} = \frac{\delta^2_{\text{SCHmod}}}{16\{(1 - R^2_{|S})\rho_\beta + [\rho_\pi + (1 - \rho_\beta - \rho_\pi)/n]/J\}/K}, \tag{30}$$

where $\delta_{\text{SCHmod}} = \frac{\gamma_{003}}{\sqrt{\tau_{\beta_{00}} + \tau_\pi + \sigma^2}}$ and all other terms were defined previously.

Note that we could also include a cluster-level moderator, W, where $R^2_{|SW} = 1 - \frac{\tau_{\beta_{00}|SW}}{\tau_{\beta_{00}}}$ is the proportion of Level 3 variance explained by the covariate, the moderator term, and the interaction. Like the two-level CRT, the noncentrality parameter for the school-level moderator with a school-level covariate looks very similar to the noncentrality parameter with a school-level covariate in Equation 30 except that it is conditioned on a different set of variables and the multiplier is a 16 rather than a 4. As in the case of the two-level CRT, it is unlikely that the moderator will explain a large proportion of the variance beyond that explained by the common school-level covariate, the school-level pretest.

Teacher-level moderator. Given the three levels, we can also test for moderator effects at the teacher level. For example, teacher experience may moderate the effect of the intervention. Assume that teacher experience is quantified as new teacher (teaching 0–5 years) or veteran teacher (teaching more than 5 years). The Level 1 model remains the same as in Equation 20. The new Level 2 or teacher-level model is

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}M_{jk} + r_{0jk} \qquad r_{0jk} \sim N(0, \tau_{\pi|M}), \tag{31}$$

where β_{00k} is the mean math achievement for school k; M_{jk} is an indicator for teacher experience, with $-½$ for 0–5 years, or new teacher, and ½ for more than 5, or veteran teacher; β_{01k} is the teacher experience gap in school k; and r_{0jk} is the residual error associated with teachers. Note that we assume that percentage of variance explained by teacher variance is $R^2_{|M} = 1 - \frac{\tau_{\pi|M}}{\tau_\pi}$. Although it is common for Level 2 variables to be aggregated up to Level 3 and included in the model, for ease of interpretation, we do not do this. The Level 3 or school-level model is

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + \gamma_{001}T_k + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta 00}), \\ \beta_{01k} &= \gamma_{010} + \gamma_{011}T_k \end{aligned} \tag{32}$$

where γ_{000} is the grand mean math achievement; γ_{001} is the mean difference between the treatment and control group; T_k is a treatment indicator, with $-½$ for control and ½ for treatment; γ_{010} is the teacher experience gap; γ_{011} is the Treatment × Teacher Experience interaction; and u_{00k} is the residual error associated with schools with variance $\tau_{\beta 00}$. Note that we do not allow the experience gap to vary randomly across schools, although this assumption could be relaxed.

The parameter of interest is γ_{011}. The moderator effect is estimated by $\hat{\gamma}_{011} = [\bar{Y}_E^{\text{New}} - \bar{Y}_E^{\text{Exp}}] - [\bar{Y}_C^{\text{New}} - \bar{Y}_C^{\text{Exp}}]$. The variance of the estimated moderator effect is

$$\text{Var}(\hat{\gamma}_{011}) = 16\left[\left(\left(1 - R^2_{|M}\right)\tau_\pi + \sigma^2/n\right)/J\right]/K. \tag{33}$$

The F-statistic in this case though is a ratio $\text{MS}_{T:M}$, which is the mean squares for the interaction to the MS_C with $J \times K - J - 2$ degrees of freedom. The noncentrality parameter is defined as

$$\begin{aligned} \lambda_{|M} &= \frac{\gamma^2_{011}}{16\left[\left(\left(1 - R^2_{|M}\right)\tau_\pi + \sigma^2/n\right)/J\right]/K} \quad \text{or} \\ \lambda_{|M} &= \frac{\delta^2_{\text{TCH mod}}}{16\left[\left(\left(1 - R^2_{|M}\right)\rho_\pi + \left(1 - \rho_\beta - \rho_\pi\right)/n\right)/J\right]/K}, \end{aligned} \tag{34}$$

where $\delta_{\text{TCHmod}} = \frac{\gamma_{011}}{\sqrt{\tau_\beta + \tau_\pi + \sigma^2}}$.

Note that the between-school variance does not enter the calculations for the power of the teacher-level moderator, as the difference between new and experienced teachers is within schools. The teacher-level variance and student-level variance are the only two variance components that affect the power. Hence, the number of teachers per school becomes a much more critical sample size in the power calculations.

Individual-level moderator. We might also be interested in whether gender moderates the treatment effect in a three-level CRT. The Level 1 model or student-level model is now

$$Y_{ijk} = \pi_{0jk} + \pi_{1jk}X_{ijk} + e_{ijk} \qquad e_{ijk} \sim N(0, \sigma^2_{|x}), \tag{35}$$

where Y_{ijk} is the math achievement for individual $i = \{1, \ldots, n\}$ in teacher $j = \{1, \ldots, J\}$ in school $k = \{1, \ldots, K\}$; π_{0jk} is the mean achievement for teacher j in school k; X_{ijk} is an indicator for gender, with $-½$ for boys and ½ for girls; π_{1jk} is the gender gap for teacher j in school k; and e_{ij} is the residual error associated with students. The percentage of variance explained by gender is

$R^2_{|x} = 1 - \frac{\sigma^2_{|x}}{\sigma^2}$. For simplicity of interpretation, we do not aggregate gender to the next levels. The teacher-level model is

$$\begin{aligned} \pi_{0jk} &= \beta_{00k} + r_{00k} \qquad r_{00k} \sim N(0, \tau_\pi), \\ \pi_{1jk} &= \beta_{10k} \end{aligned} \tag{36}$$

where β_{00k} is the mean achievement across schools; r_{00k} is the error associated with mean achievement across teachers in schools with variance τ_π. Note we do not allow the gender gap to vary randomly across teachers in schools. The new Level 3 model is

$$\begin{aligned} \beta_{00k} &= \gamma_{000} + \gamma_{001} T_K + u_{00k} \qquad u_{00k} \sim N(0, \tau_{\beta_{00}}), \\ \beta_{10K} &= \gamma_{100} + \gamma_{101} T_K \end{aligned} \tag{37}$$

where γ_{000} is the grand mean achievement; T_k is a treatment indicator, with $-½$ for control and ½ for treatment; γ_{001} is the overall treatment effect; γ_{100} is the gender gap; γ_{101} is the Treatment × Gender interaction; and r_{00k} is the error associated with schools with variance $\tau_{\beta_{00}}$.

The moderator effect of interest is γ_{101}, which is estimated by $\hat{\gamma}_{101} = [\bar{Y}^G_E - \bar{Y}^B_E] - [\bar{Y}^G_C - \bar{Y}^B_C]$. The variance of the estimated moderator effect is

$$\text{Var}(\hat{\gamma}_{101}) = [16(1 - R^2_{|x})\sigma^2]nJK. \tag{38}$$

The F-statistic in this case though is a ratio $\text{MS}_{T:X}$ to the MS_C with $n \times J \times K - J \times K - K - 2$ degrees of freedom and the noncentrality parameter is defined as

$$\lambda_{|x} = \frac{\gamma^2_{101}}{[16(1 - R^2_{|x})\sigma^2]nJK} \quad \text{or} \quad \lambda_{|x} = \frac{\delta^2_{\text{INDmod}}}{[16(1 - R^2_{|x})(1 - \rho_\beta - \rho_\pi)]/nJK}, \tag{39}$$

where $\delta_{\text{INDmod}} = \frac{\gamma_{101}}{\sqrt{\tau_\beta + \tau_\pi + \sigma^2}}$.

Note that because the moderator is at Level 1 and hence the difference between boys and girls is within teacher, the between-teacher and between-school variance components are removed from the variance of the moderator effect. Hence, the number of individuals per cluster is a critical sample size in power calculations for the individual-level moderator effect in a three-level CRT.

The three-level CRT is a natural extension of the two-level CRT and findings are similar. That is, for the main effect and the school-level moderator, the power is most heavily influenced by the total number of clusters. In addition, designing a study to detect a school-level moderator will require many more clusters than designing a study to detect the main effect of treatment since the moderator effect will likely be smaller and the variance of the estimated moderator effect may be up to 4 times that of the main effect. As we consider lower level moderators, the sample size at the level of the moderator becomes more important. That is, for a teacher-level moderator, the number of teachers is as important as the total

number of schools, given that the between-school variance is removed. Further, for a student-level moderator, the number of students is as influential as the number of teachers per school and the total number of schools because the between-teacher and between-school variance components are removed from the moderator effect.

Unbalanced Designs

Thus far, we have assumed perfectly balanced designs. For example, in a two-level CRT with 40 total schools we assumed the ideal case, 20 schools in treatment and 20 schools in control, and 10 rural and 10 urban school in each condition. Given this structure, we maximize the power for both the test of the main effect and the cluster-level moderator.

However, in practice, it may not always be feasible to achieve a perfectly balanced design. For example, suppose that in order to increase the likelihood of schools participating in a study, the researchers plan to assign 28 schools to the treatment condition and 12 to the control condition. We can use the same formulas described above for the test of the main effect of treatment by replacing the total number of clusters with the effective sample size for the calculations. In this case, the effective sample size is 2 times the harmonic mean (HM). The HM of the treatment and control conditions is $\mathrm{HM} = \frac{2}{\frac{1}{J_T}+\frac{1}{J_C}}$ or $\mathrm{HM} = \frac{2}{\frac{1}{28}+\frac{1}{12}} \sim 16.8$. Hence, the effective sample size for the calculations is 16.8 clusters per condition for a total number of 33.6 clusters.[3] Given that the power is strongly influenced by the total number of clusters, the power to detect an effect of a given magnitude will be less for designs that are not balanced.

The same process can be used for the moderator power calculations. However, now we need the HM of the four groups: rural treatment, urban treatment, rural control, and urban control. Suppose that of those 28 treatment clusters, 14 are rural schools and 14 are urban schools, and of the 12 control clusters, 6 are rural schools and 6 are urban schools. The HM of the four groups is $\mathrm{HM} = \frac{4}{\frac{1}{J_{\mathrm{TR}}}+\frac{1}{J_{\mathrm{TU}}}+\frac{1}{J_{\mathrm{CR}}}+\frac{1}{J_{\mathrm{CU}}}}$ or $\mathrm{HM} = \frac{4}{\frac{1}{14}+\frac{1}{14}+\frac{1}{6}+\frac{1}{6}} \sim 8.4$. Hence, the effective sample size is 8.4 clusters per condition for a total number of 33.6 clusters. In essence, the total number of clusters used for the calculations is 33.6 rather than 40, which will reduce the power to detect a cluster-level moderator effect of a given magnitude.

The logic of this example is applicable for any of the power calculations discussed in this article and can be applied as follows: First, identify the estimator for the effect of interest. Second, identify the sample size for each of the groups included in the estimator. If they are not equal, calculate the HM for each group. For power calculations, for the main effect of treatment, double the HM to calculate the total effective sample size that will be used for the calculations. For power calculations, for the moderator effects, calculate the HM for each of

the four groups and multiply it by four for the effective sample size that can be used for the calculations. The effective total sample size for the power calculations may be different for the main effect and the moderator effects depending on the allocation of clusters and individuals.

It is important to note that in practice the sample sizes for the different moderator variables may be beyond the control of the researcher. For example, there may not be an equal number of boys and girls in a class, or new and experienced teachers within a school, or rural and urban schools in the sample of schools willing to participate in the study. As the imbalance increases, the power to detect an effect of a given magnitude will decrease. Hence, to the extent possible, it is important to identify moderators of interest prior to recruiting for a study and to consider these variables during the recruitment process.

Examples

We begin with an example of a two-level CRT. Continuing with the idea of mathematics achievement as the primary outcome, suppose that based on past studies of the intervention, a team wants to design a study to detect a main effect of treatment that has a standardized effect size of 0.20. They plan to test the intervention in urban and rural schools and are interested in whether the treatment effect is moderated by school type. Recognizing that the moderator effect will be smaller than the main effect of treatment, they are interested in detecting a cluster-level moderator effect of 0.10. Suppose the team is limited to a total of 40 schools, 20 urban and 20 rural, with 100 students per school. Assume that they assign 20 schools to each condition and that the number of urban and rural schools in each condition is balanced. Note that this assumption could be relaxed to allow for imbalance in groups in which case the HM calculations described above would apply. Based on the literature (Hedges & Hedberg, 2009, 2014), they estimate an ICC of 0.23. They have access to a school-level covariate, last year's scores, and assume an $R^2_{|W} = 0.66$. They estimate that school type will explain additional variance at the school level and thus $R^2_{|SW} = 0.75$. Note that the R code provided in Appendix A, available in the online version of the journal, was used for the power calculations in the examples.

Figure 1 shows the power to detect the main effect of treatment of 0.20 and the cluster-level moderator of 0.10. As expected, the power for the main effect of treatment is always greater than the power for the cluster-level moderator. Assuming 40 total clusters, the power to detect the main effect of treatment in this case is 0.56. In order to reach the acceptable level of 0.80, they would need an additional 30 clusters for a total of 70. The power to detect the cluster-level moderator effect of 0.10 for 40 total clusters is only 0.09. Assuming 70 total clusters, the number needed to reach adequate power for the main effect of treatment, the power is still only 0.13 for a cluster-level moderator of 0.10. It is clear that the number of clusters to achieve adequate power for the cluster-level moderator will be outside a reasonable range.

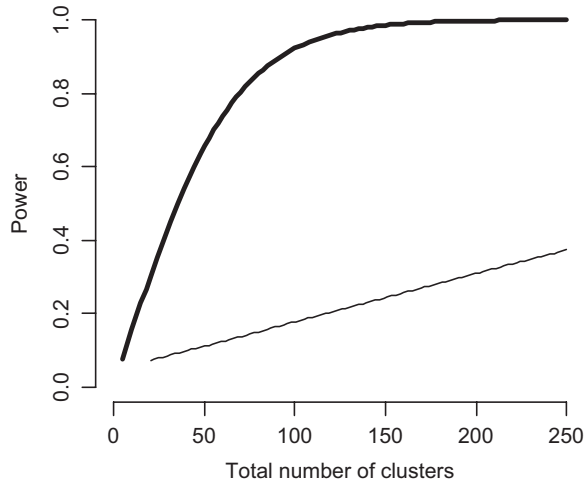

FIGURE 1. *Power curves for main effect of treatment and cluster-level moderator.*

Suppose that a different team of researchers are also designing a study of the same intervention. They were concerned that the treatment effect may have a differential effect on boys and girls, hence they are interested in power the study to detect an individual-level moderator effect in addition to the main effect of treatment. They seek to detect an individual-level moderator effect of 0.10. Assuming the same design parameters as above, a total of 40 schools, 100 students per school (assuming half girls and half boys), an ICC = 0.23, and an $R^2_{|W} = 0.66$, we know the power for the main effect of treatment of 0.20 is 0.56 and that 70 total schools are needed for the power of 0.80. Based on Bloom, Richburg-Hayes, and Rebeck-Black (2007), we assume that gender explains approximately 10% of the variance in achievement within schools, or $R^2_{|x} = 0.10$. The power to detect an individual-level moderator is 0.48 and 0.71, with 40 and 70 schools, respectively. Note the power is much higher than in the cluster-level moderator case because the number of individuals per school is a key sample size. In fact, assuming all parameters are held constant, the power to detect an individual-level moderator of 0.10 with only 40 schools increases to 0.80 if approximately 115 more students are included in each school. Although this would increase the total number of students in the study from 4,000 to 8,600, if the costs associated with adding individuals is small because, for example, a state test is the outcome of interest for the study and all students take the test, it may be very feasible to include more students and have an adequately powered study to detect an individual-level moderator effect.

The three-level CRT follows the same pattern as the two-level CRT. That is, the power for the main effect of treatment and the school-level moderator is

TABLE 1.
Power for Main Effect, Cluster-Level Moderator, Teacher-Level Moderator, and Individual-Level Moderator Effects

	Power: Main Effect	Power: Cluster-Level Moderator	Power: Teacher-Level Moderator	Power: Individual-Level Moderator
$J = 5$				
$n = 10$	0.65	0.10	0.15	0.27
$n = 30$	0.72	0.10	0.20	0.64
$J = 30$				
$n = 10$	0.85	0.13	0.61	0.91
$n = 30$	0.86	0.13	0.79	0.99

Note. The main effect of treatment = 0.20, school-level moderator = 0.10, teacher-level moderator = 0.10, student-level moderator = 0.10, 40 total schools, 5 or 30 teachers per schools, 10 or 30 kids per teacher, an intraclass correlation (ICC) at the school level of 0.15, an ICC at the teacher level of 0.08, $R^2_{|W} = 0.75$, $R^2_{|SW} = 0.80$, $R^2_{|M} = 0.10$, and $R^2_{|X} = 0.10$, and equal allocation of clusters across condition.

driven by the total number of clusters or schools. The power for the lower level moderator effects is also strongly influenced by the sample size of the moderator of interest. For example, Table 1 displays the power to detect the main effect of treatment, cluster-level moderator, teacher-level moderator, and individual-level moderator under the following assumptions: main effect of treatment of 0.20, school-level moderator effect of 0.10, teacher-level moderator effect of 0.10, student-level moderator effect of 0.10, a total of 40 schools, either 5 or 30 teachers per schools, either 10 or 30 kids per teacher, an ICC at the school level of 0.15, an ICC at the teacher-level of 0.08, $R^2_{|W} = 0.75$, $R^2_{|SW} = 0.80$, $R^2_{|M} = 0.10$, and $R^2_{|X} = 0.10$.

The table illustrates the effects of the sample sizes at different levels on different effects. For the main effect of treatment and the cluster-level moderator, the power is not strongly influenced by increases in the number of individuals per teacher or the total number of teachers per school. However, increasing the total number of teachers per school increases the power to detect a teacher-level moderator. The challenge with this is that, in many cases, the total number of teachers per school may be small, particularly if only one grade level is represented. For the individual-level moderator, it is clear that increasing for 10 to 30 students per teachers has a strong effect on the power, particularly in the case with a smaller total number of schools and number of teachers. For most elementary and middle/high schools, a per class sample size of 25 to 30 is quite common. The smaller sample sizes are more prevalent in pre-K studies that may make individual-level moderator effects more difficult to detect in these studies.

Discussion

The capacity of CRTs to provide rigorous evidence of the main effect of treatment has improved in the past decade. That is, more recent CRTs are being designed with adequate power to detect a meaningful main effect of treatment that past CRTs (Spybrook & Raudenbush, 2009). As we start to design studies that enable us to determine whether or not an intervention has an overall effect, we also begin to ask other important question regarding whether the effect is the same across different kinds of schools, teachers, and students. Hence, it becomes important to think about whether we can power CRTs to detect not only the main effect of treatment but also important moderator effects.

Some general patterns emerge from the findings related to the different types of moderator effects. For the purpose of this discussion, consider a two-level CRT with students nested within schools and a three-level CRT with students nested within teachers nested within schools. In both cases, powering for the school-level moderator effect will be challenging, given the current size of CRTs in the field. As illustrated by the formulas and in the examples, the power for a cluster-level moderator tends to be much smaller than the power for the main effect. Given that it is often challenging for teams to afford enough clusters to power for the main effect of treatment, the number of schools required to power for a cluster-level moderator is likely to be outside the budgetary constraints. This suggests that the analysis of school-level moderator effects may require a more meta-analytic approach involving combining across studies.

However, lower level moderator effects hold much more promise from a power perspective. In a three-level CRT with students nested in teachers nested in schools, designing studies to detect teacher-level moderators may be possible. That is, the power for teacher-level moderators is driven more by the number of teachers per schools as shown in the formulas and examples. For studies examining the effect of a whole school intervention in which randomization takes place at the school level and all teachers in the school are involved, it may be reasonable to design the study with adequate power to detect teacher-level moderators because the number of teachers per school may be large. This presents an important opportunity for researchers to be able to answer critical questions about teacher moderator effects that may help improve the likelihood an intervention is effective. For example, if the researchers determine that an intervention is more effective with experienced teachers rather than novice teachers, they may be able to put in additional professional development opportunities for less experienced teachers to help them overcome challenges.

The greatest potential for detecting moderator effects in CRTs lies in individual-level moderator effects. As we saw in both the two- and three-level CRT, the power for the individual-level moderator depends much more heavily on the number of individuals per cluster. In many CRTs in education, all of the students in a school or all of the students in a grade will participate in a study.

This means that there are often large numbers of individuals per cluster. Taking advantage of the number of individuals per cluster and hence asking a priori questions about individual-level moderators can help researchers better understand for whom a program is effective. This is a critical step toward designing, developing, and implementing programs that meet the needs of all students. For example, suppose that a group of researchers testing a math curriculum are concerned that the program is less effective for English-language learners (ELL). In their sample, about half of the students are ELLs and hence they test whether ELL moderates the treatment effect. If findings suggest that there is a differential effect, the team can then explore how to modify the treatment, so that ELL and non-ELLs benefit from the program.

For many other K–12 studies, powering to detect an individual-level moderator of a reasonable magnitude may be a very realistic goal for the study. The exception to this case is when there are only a small number of individuals per cluster. For example, in pre–K classrooms, the number of students per class may be less than seven which would make powering the study to detect individual-level moderators very challenging.

Future Directions

In this article, we focused on clustered designs. Extending the work to MCRTs is critical, as multisite studies are quite common in evaluations of educational interventions. Furthermore, we discussed binary moderators in this article. Moderators can also be continuous in nature, for example, whether the program's effect is moderated by school quality, and extending the work to continuous moderators is another critical step. We also fixed the moderator effects at lower levels. It may not always be the case that moderator effects are fixed and thus allowing the moderator effects to vary randomly is another area for future research. In addition, understanding more about the magnitude of moderator effects is a critical step toward planning studies appropriately. For example, how different is the effect on boys and girls? Or for urban schools versus rural schools? As we begin to develop empirical estimates of the magnitude of moderator effects, we can start to use these effect sizes to guide the power analyses.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) declared the following financial support for the research, authorship, and/or publication of this article: This project has been funded by the National Science Foundation [DGE-1437679, DGE-1437692, DGE-1437745]. The opinions expressed herein are those of the authors and not the funding agency.

Notes

1. For this article, we use the annotated R code for the examples. PowerUp! could also be downloaded at http://www.causalevaluation.org/ and used for the examples.
2. Note we could also express this formula as $\mathrm{Var}(\hat{\gamma}_{01}) = \frac{(\tau_{00}+\sigma^2/n)}{P(1-P)J}$ where P is the proportion of clusters assigned to the treatment condition. Assuming a balanced design, the expressions are equivalent. We assume a balanced design throughout, hence we adopt the notation set forth in Equation 3.
3. It is important to note that using the harmonic mean (HM) is equivalent to Bloom's (2005) approach for the two-level CRT, which asks for the proportion of clusters assigned to each condition. His formula includes an additional term of $\frac{1}{P(1-P)}$ where P is the proportion of clusters assigned to treatment. In a balanced case, this yields the 4 in the variance term. In an unbalanced case, it is equivalent to substituting twice the HM for the total number of clusters.

References

Aguinis, H., Beaty, J. C., Boik, R. J., & Pierce, C. A. (2005). Effect size and power in assessing moderating effects of categorical variables using multiple regression: A 30-year review. *Journal of Applied Psychology*, *90*, 94–107.

Bloom, H. S. (1995). Minimum detectable effects: 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: Russell 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*, 30–59. doi:10.3102/0162373707299550

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

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, *6*, 24–67. doi:10.1080/19345747.2012.673143

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London, England: Arnold.

Hedges, L. V., & Hedberg, E. C. (2014). Intraclass correlations and covariate outcome correlations for planning two- and three-level cluster-randomized experiments in education. *Evaluation Review*, *37*, 445–489. doi:10.1177/0193841X14529126

Hedges, L. V., & Rhoads, C. (2009). *Statistical power analysis in education research* (NCSER 2010-3006). Washington, DC: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education.

Institute of Education Sciences. (2016). *Research grants request for applications for awards beginning in fiscal year 2017: CFDA Number 84.305A*. Washington, DC: U.S. Department of Education.

Jaciw, A. (2014). An empirical study of design parameters for assessing the differential impacts for students in group randomized trials. Working paper. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2516005

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

Konstantopoulos, S. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 66–88.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). *SAS for mixed models* (2nd ed.). Cary, NC: SAS Institute.

Liu, X. (2014). *Statistical power analysis for the social and behavioral sciences*. New York, NY: Taylor and Francis Group.

Murray, D. M. (1998). *Design and analysis of group randomized trials*. New York, NY: Oxford University Press.

Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, *2*, 173–185.

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*, 387.

Raudenbush, S. W., Martinez, A., & Spybrook, J. (2007). Strategies for improving precision in group-randomized experiments. *Educational Evaluation and Policy Analysis*, *29*, 5–29. doi:10.3102/0162373707299460

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].

Schochet, P. Z. (2008). Strategies for power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics*, *33*, 62–87.

Spybrook, J. (2013). Introduction to a special issue on design parameters for cluster randomized trials in education. *Evaluation Review*, *37*, 435–444.

Spybrook, J. (2014). Detecting intervention effects across context: An examination of the precision of cluster randomized trials. *The Journal of Experimental Education*, *82*, 334–357. doi:10.1080/00220973.2013.813364

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. doi:10.3102/0162373709339524

Authors

JESSACA SPYBROOK is an associate professor in the Evaluation, Measurement, and Research Program at Western Michigan University, 1903 W. Michigan Avenue, Kalamazoo, MI 49008, e-mail: jessaca.spybrook@wmich.edu. Her research interests are the design of evaluations of educational interventions, multilevel models, and statistical power.

BENJAMIN KELCEY is an assistant professor in the Quantitative Research Methodologies Program at the University of Cincinnati, Teachers/Dyer Hall, Cincinnati, OH

45221, e-mail: benjamin.kelcey@uc.edu. His research interests are causal inference and measurement methods within the context of multilevel and multidimensional settings such as classrooms and schools.

NIANBO DONG is an assistant professor in the Department of Educational, School, and Counseling Psychology at the University of Missouri, 14 Hill Hall, Columbia, MO 65211, e-mail: dongn@missouri.edu. His research interests are statistical power analysis, causal inference, and the design and analysis of randomized experiments and quasi-experiments.

Manuscript received July 13, 2015
Revision received May 17, 2016
Accepted May 21, 2016