Abstract Title Page

Title: Detecting Intervention Effects Across Context: An Examination of the Power of Cluster Randomized Trials

Authors and Affiliations:

Jessaca Spybrook
Western Michigan University
Jessaca.spybrook@wmich.edu

Abstract Body

Background:

In the past decade, cluster randomized trials (CRTs) have emerged as a common design in the evaluation of educational interventions. In fact, since 2002, the Institute of Education Sciences (IES) alone has funded over 100 CRT's (http://ies.ed.gov/). The purpose of these studies is to build a base of reliable evidence on which to base education practice and policy (Whitehurst, 2003). In order to yield high-quality and reliable evidence, the studies must be well-designed and implemented. Although there are many facets to a strong design and implementation, I restrict this study to an examination of key components of the research design.

A critical component of the design is adequate statistical power. However, the term statistical power is rather general and thus should be qualified by adequate statistical power to detect what? Much of the work to date has focused on statistical power to detect the main effect of treatment for different types of CRTs (Bloom, 2005; Donner and Klar, 2000; Konstantopolous, 2008; Murray, 1998; Raudenbush,1997; Raudenbush & Liu, 2000; Authors, 2007; Schochet, 2008). Authors (2009) examined the CRTs funded by IES between 2002 and 2006 and found that they were powered to detect a main effect of treatment ranging from 0.18 to 0.40 and 0.20 to 1.0 for studies funded by the National Center for Educational Evaluation and Regional Assistance (NCEE) and the National Center for Educational Research (NCER), respectively. The precision of the studies increased over the 4 year time span, suggesting that researchers were becoming more adept at planning studies to detect a meaningful treatment effect.

However, an important problem facing education researchers is that the main effect of treatment may be of limited utility to a practitioner in a particular school or site if the treatment effects vary substantially from site to site. It is plausible that *context matters* in education (Berliner, 2002; Cohen, Raudenbush, & Ball, 2002). For example, an intervention may be more effective for low-income students than for high-income students or in urban schools compared to rural schools; its effectiveness may depend on the skill and knowledge of the teachers or the resources available to a school. Understanding the context in which an intervention is likely to be effective will make the results more applicable and therefore more useful to different schools, districts, and students. Thus powering a study for the main effect of treatment may not always be sufficient.

Purpose:

The purpose of this paper is to twofold. The first objective is to examine how to calculate power for three types of treatment heterogeneity including 1) the variability in treatment effects across sites, 2) site-specific treatment effects, and 3) moderator effects at the cluster or student level. The second objective is to examine the power to detect each type of treatment effect heterogeneity on a set of funded CRTs. Given the length of this proposal, I primarily focus on the empirical findings from the set of funded CRTs. The power calculations are included in the full paper.

Sample

The sample includes the studies in the first wave of CRTs funded by IES, or those funded between 2002 and 2006 by NCER and NCEE. These studies represent a range of CRTs on various topics and with different research designs and sample sizes. The majority of these studies were not explicitly required to be powered beyond the main effect of treatment so these studies are used simply to demonstrate the power of a set of funded CRTs to detect heterogeneity of treatment effects.

I identified a total of 54 CRTs of educational evaluations in pre-K through grade 12 in the first wave of studies funded by IES. Forty-nine of those studies are included in the current study. The relevant design and sample size information was unavailable for the remaining 5 studies at the time of this paper. Of the 49 studies, 41 were funded by NCER and 8 were funded by NCEE. The majority of the studies targeted pre-K and elementary students (approximately 65 percent). A variety of topic areas were represented including, but not limited, to social and character development, math and science, teacher professional development, and literacy, reading, and writing with the majority of the studies focused on the latter area (approximately 42 percent).

The 49 studies can be represented by four types of CRTs: 2-level CRT, 3-level CRT, 3-level MultisiteCRT (MSCRT), and 4-level MSCRT. Table 1 in Appendix B provides the basic features of each design and the number of studies in each category. From the table, we can see that multisite trials were the most common designs.

Statistical Models:

For illustration purposes, I provide the models for a 3-level MSCRT[1] with a brief description of the relevant power analyses. Assume that the level-one units are students, the level-two units are schools, and the level-three units, or the sites, are districts. Using HLM notation (Raudenbush & Bryk, 2002), the student level model is:

$$Y_{ijk} = \pi_{0jk} + e_{ijk} \qquad (1)$$

where y_{ijk} is the outcome for individual $i = \{1,...,n\}$ in school $j = \{1,...,J\}$ in district $k = \{1,...,K\}$; π_{0jk} is the mean for school *j* in district *k*; and $e_{ijk} \sim N(0,\sigma^2)$ is the error associated with each student. The school-level model is:

$$\pi_{0jk} = \beta_{00k} + \beta_{01k}T_{jk} + r_{0jk} \qquad (2)$$

where T_{jk} is an indicator for the treatment or control group, with -½ for control and ½ for treatment; β_{00k} is the mean for district *k*; β_{01k} is the treatment effect for district *k*; and $r_{0jk} \sim N(0,\tau_\pi)$ is the error associated with each school. The district level model is:

$$\beta_{00k} = \gamma_{000} + u_{00k}$$
$$\beta_{01k} = \gamma_{010} + u_{01k} \qquad (3)$$

where γ_{000} is the overall mean; γ_{010} is the overall treatment effect.

We can choose to treat the district effects, u_{00k} and u_{01k} as fixed or random effects, depending on the goal of the study. If the purpose of the study is to generalize to a larger universe of sites, then the sites are treated as random effects. In this case, u_{00k} represents the random effect associated with each site mean and u_{01k} is the random effect associated with each site treatment effect where var(u_{00k})=$\tau_{\beta_{00}}$ and var(u_{01k})=$\tau_{\beta_{11}}$.

However, if the goal is not to generalize to a broader universe of sites, then the sites are treated as fixed effects. More specifically, u_{00k}, for $k \in \{1,2,...,K\}$, are fixed effects associated

[1] The models and power calculations for all designs are provided in the full paper.

with each site mean, constrained to have a mean of zero, and u_{01k} , for $k \in \{1,2,...,K\}$, are fixed effects associated with each site treatment effect, constrained to have a mean of zero.

Variability in Treatment Effects Across Sites

The variability in treatment effects across sites is relevant for studies where the sites are treated as random effects. The treatment effect variability is defined as var(u_{01k})= $\tau_{\beta_{11}}$ in equation 3 when the sites are random effects. The test of the null hypothesis, H_0: $\tau_{\beta_{11}}$ =0 relies on an *F* test as defined by Raudenbush & Liu, 2000; Authors, 2009.

Site-specific Treatment Effects

Estimating site-specific treatment effects is appropriate when sites are treated as fixed effects or the variability across sites is significant. I focus on the case of fixed site effects and the power for the test of the null hypothesis, H_0: u_{01k} =0 (equation 3) is the same as the main effect of treatment in a single site and is largely dependent the number of clusters per site (Kirk, 1982).

Moderator Effects

In statistical terms, moderator effects are interactions between the moderator variable, such as district type, and the treatment effect. I explore power for moderator variables at the individual, cluster, and site levels. The power is primarily influenced by the sample size corresponding to the level of the moderator and the level of the treatment[2].

Findings:

For all analyses, I used empirical estimates of design parameters based on the recent work of Bloom, Richburg-Hayes, & Black, 2007; Flay & Collins, 2005; Hedges & Hedberg, 2007. The design parameters are given in Table 2 in Appendix B.

Variability in Treatment Effects Across Sites

As noted in the models section, the variability in treatment effects across sites is only relevant in the multisite trials. Further, it is only applicable to studies in which sites are treated as random effects and studies that are not matched pairs designs since the matched pairs design confounds the treatment-by-site-variance and the within cluster variance. In total, there are 9 studies that meet the criteria. For each of the 9 studies, let us assume that the main effect of treatment is 0.30. If the effect size variability across sites (esv) is 0.01, the interval around the treatment effect is $0.30 \pm 2\left(\sqrt{0.01}\right) = \left(0.10, 0.50\right)$. In other words, across the sites, the treatment effect may vary from 0.10 to 0.50. These values may be reasonable and although the magnitude of the effect may vary across sites, it is always positive. Now suppose that the esv is 0.03. This creates an interval from $0.30 \pm 2\left(\sqrt{0.03}\right) = \left(-0.046, 0.646\right)$. In some sites, the treatment effect may be 0 or even a small negative value. While we may consider that 0.03 is still a reasonable esv, it is likely that we would want to be able to detect it because it suggests that the treatment effect may be 0 in some sites. Hence from a power perspective, we might want to power a study to detect a minimum detectable esv (mdesv) of 0.03. Table 3 in Appendix B presents the mdesv for the 9 studies. Only 1 study is powered to detect a mdesv in the range of 0.03. One additional study is powered for a mdesv around 0.056. In the remaining 7 studies, the mdesv is greater than 0.09. In these studies, if an esv less than 0.09 is meaningful, it would likely go undetected in these studies.

[2] The noncentrality parameter for the power for the tests of the moderator effects are included in the full paper. The R code is also included in the paper.

Site-specific Treatment Effects

I calculated the mdes for site specific treatments effects for multisite trials that treated the sites as fixed effects. There were 10 studies that met the criteria. Table 4 in Appendix B presents the mdes for the site-specific treatment effects. Half of the studies were powered to detect an mdes between 0.41 and 0.60 whereas the remaining studies were powered to detect an mdes greater than 0.80. Although an mdes greater than 0.80 is outside the range of what is often seen in studies of educational interventions, a range from 0.41 to 0.60 is nearer to what may be deemed reasonable. However, an effect size less than 0.41 is often practically meaningful and would likely be undetected in these studies.

Moderator Effects

Currently I have calculated the power for the moderator effects for all studies at the level of the treatment. For example, for a 2-level CRT, I have power for a cluster level moderator, etc. The full paper will include power for moderator effects at all levels. The mdes for moderator effects at the level of treatment for each study is displayed in Figure 1in Appendix B. Approximately 16 percent of the studies were able to detect treatment-level moderator effects between 0.20 and 0.40 with an additional 25 percent powered to detect treatment-level moderator effects between 0.40 and 0.60 and the remaining 59 percent powered for greater than 0.60.

Conclusions:

For the past 10 years, we have focused on powering studies to detect the main effect of treatment and the result is that it is becoming more common to have studies with solid designs and adequate power to detect the main effect of treatment. However, because of the importance of context in education, it is time to move beyond the main effect of treatment in designing CRTs. The full paper describes three types of treatment effect heterogeneity and examines how to calculate power for each type. In this proposal, I focus on the findings from applying the power calculations to the first wave of CRTs funded by IES.

Overall, the sample of studies had minimal power to detect a reasonable level treatment effect variability across sites. Approximately half of the studies examined had power to detect site-specific treatment effects in the range of 0.40 to 0.60, which is on the high side of what empirical work has revealed to be reasonable effect sizes in achievement studies for example. The mdes for treatment level moderator effects was between 0.20 and 0.60 in almost 40 percent of the cases, which is positive, although the magnitude of typical moderator effects may be smaller than the main effect of treatment and thus should also be considered. A key challenge is that maximizing power to detect the main effect of treatment may not be consistent with for example, maximizing power to detect treatment effect variability across sites, which makes it difficult to meet both objectives given the size of many of the studies in this sample and more broadly many of the CRTs in the field.

For researchers designing future CRTs with the goal of moving beyond the main effect of treatment, we recommend prioritizing the type of heterogeneity of interest and considering the heterogeneity in addition to the main effect of treatment in the planning of the study. However, given the current size and scope of studies, powering for a specific type of treatment effect heterogeneity may not always be feasible. In this case, it is most appropriate to explicitly address the lack of power to detect treatment effect heterogeneity so that these analyses may be noted as exploratory since they are underpowered.

Appendix A. References

Berliner, D. (2002). Educational research: the hardest science of all. *Educational Researcher*, 31(8), 18-21.

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: Russell Sage Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30-59.

Cohen, D.K., Raudenbush, S.W., & Ball, D.L.,(2002). Resources, instruction, and Research. *In F. Mosteller & R. Boruch (Eds.), Evidence matters: Randomized trials in education research*. Washington, DC: Brooking Institution Press.

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

Flay, B. R., & Collins, L. (2005). Historical review of school-based randomized trials for evaluation problem behavior prevention programs. *The Annals of the American Academy of Political and Social Science, 599*, 115-146.

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

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

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

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Raudenbush, S.W. & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5(2),* 199-213.

Schochet, P. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(*1*), 62-87.

Institute of Education Sciences. Retrieved August 1, 2011, from http://ies.ed.gov/.

Authors (2007).

Authors (2009).

Authors (2009).

Appendix B. Tables and Figures

Table 1. The basic design features of the CRTs identified in the study proposals

	Two-Level Cluster Randomized Trial	Three-Level Cluster Randomized Trial	Three-Level Multisite Cluster Randomized Trial	Four-Level Multisite Cluster Randomized Trial
Level of Randomization	2	3	2	3
Blocking	No	No	Yes	Yes
Number of Studies	8	5	30	6
Example of Nesting	Students, Schools	Students, Classrooms, Schools	Students, Classrooms, Schools	Students, Classroom, Schools, Districts

Table 2. Design parameters for calculating power for heterogeneity of treatment effects.

	ICC Level 2	ICC Level 3	R2 Level 2	R2 Level 3
2-level CRT	0.15,0.0.2	NA	0.6,0.6	NA
3-level CRT	0.07,0.02	0.15,0.05	NA	0.6,0.6
MSCRT TRMT L2	0.15,0.02	NA	0.6,0.6	NA
MSCRT TRMT L3	0.07,0.02	0.15,0.05	NA	0.6,0.6

Note. The first number was used for academic outcomes. The second number was for non-academic outcomes.

Table 3. Minimum detectable effect size variability for power = 0.80.

Minimum detectable ESV	Frequency
0.0 - 0.030	1
0.031 - 0.060	1
0.061 - 0.090	0
0.091 - 0.120	1
0.121 - 0.150	1
0.151 - 0.180	0
0.181 - 0.210	0
Greater than 0.211	5

Table 4. Minimum detectable effect size for site-specific treatment effects.

Minimum detectable ES	Frequency
0.00 - 0.20	0
0.21 - 0.40	0
0.41 - 0.60	5
0.61 - 0.80	0
0.81 - 1.00	3
Greater than 1.00	2

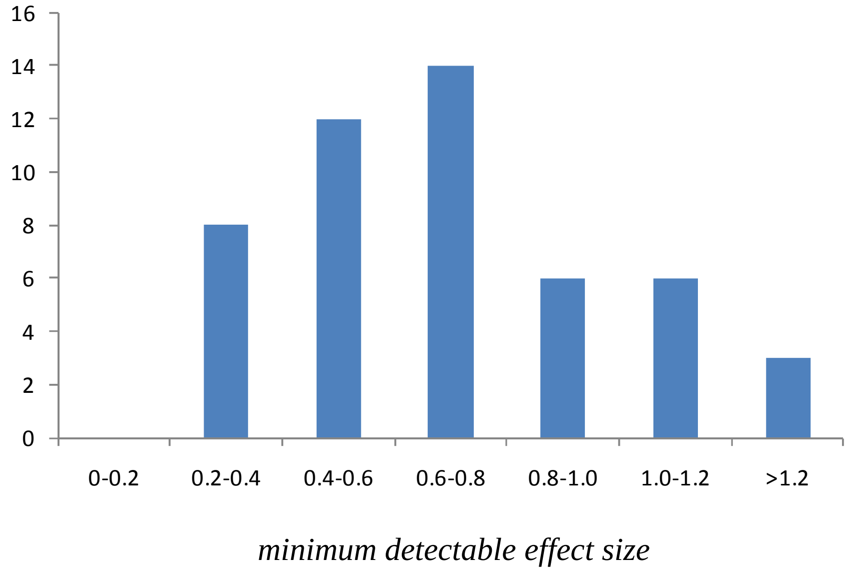

Figure 1. Minimum detectable effect size for moderator effects at the level of the treatment.