Abstract Title Page

Title: A Framework for Designing Cluster Randomized Trials with Binary Outcomes

Author(s): Jessaca Spybrook, Andres Martinez

Abstract Body

Background / Context / Significance:

Cluster randomized trials (CRTs) have become increasingly common in education as a means for evaluating the effectiveness of educational interventions. In fact, since 2002, more than 80 CRTs have been funded by the Institute of Education Sciences (IES), the research branch of the U.S. Department of Education. The trials examine a broad array of interventions including, but not limited to, reading curricula, math curricula, science curricula, professional development programs, and social and character development programs. The interventions target students as young as pre-K through post-high school (http://ies.ed.gov/).

CRTs have become more widespread in evaluations of the effectiveness of educational programs and policies for two primary reasons. First, when they are feasible and if they are well designed and implemented, randomized trials are the best way to establish causal relationships (Boruch, 1997; Boruch, DeMoya, & Synder, 2002; Cook, 2002). Second, the natural clustering in our education system, students within classrooms within schools within districts, and the fact that educational interventions are typically delivered at the classroom, school, or district level, make CRTs particularly relevant for education studies (Bloom, 2005; Boruch & Foley, 2000; Cook, 2005). The goal is that over time the evidence provided by rigorous evaluations of educational programs or policies, rigorous being defined as experimental or high-quality quasi-experimental studies, will accumulate and transform education into an evidence-based field (Whitehurst, 2003).

However, the sheer presence of CRTs to evaluate educational programs and policies is not enough to transform education into an evidence-based field. As noted above, the trials must be well-designed and implemented in order to generate high-quality evidence of program effectiveness. Although there are many elements involved in the design and implementation of a study, we limit the scope of this paper to the statistical power of the study. We focus on power because underpowered studies represent a serious threat to the success of CRTs (Boruch, 2005; Boruch & Foley, 2000).

The field has made substantial progress in terms of how to calculate statistical power for CRTs for continuous outcomes, such as academic achievement, in the past 15 years. Raudenbush (1997) introduced power calculations for a two-level CRT (2-level CRT) and illustrated two key points in a power analysis for a CRT: 1) the total number of clusters influences the power more than the total number of individuals and 2) the higher the intraclass correlation (ICC), or the variability between clusters relative to the total variability, the lower the statistical power. In 1998, Murray published a book dedicated to the design and analysis of CRTs which was followed by another book by Donner and Klar (2000) on the same topic, though specifically geared towards the health sciences. Since then, numerous others have contributed by extending the work to additional designs including three level designs and blocked designs (Konstantopoulos, 2008; Raudenbush, Spybrook, & Martinez, 2007; Schochet 2008).

The accessibility of planning parameters has also contributed to the improved accuracy of power analyses for CRTs in education. Several studies have provided evidence to suggest that ICCs for academic achievement are likely to be between 0.15 and 0.25 (Bloom, Richburg-Hayes, & Black, 2007; Bloom, Bos, & Lee, 1999; Hedges & Hedberg, 2007; Schochet, 2008). Bloom, Richburg-Hayes, and Black (2007) also illustrated the importance of including covariates and

provided empirical evidence suggesting that for achievement outcomes, pretests could explain between 40 and 80 percent of the variation in the outcome.

However, outcomes of interest are not always continuous in nature. For example, a key outcome in education is graduation status. Ultimately, increasing the number of students who graduate from high school is an important national goal in education with long term implications for the future work force. However, graduation status is not a continuous outcome but rather a binary outcome i.e. a student either graduates or does not graduate. Binary outcomes rely on different assumptions than continuous outcomes hence the power analysis will necessarily be different.

Power analyses for binary outcomes in single level designs has been well documented (Fleiss, 1981; Hsieh, Block, & Larsen, 1998; Diggle, Heagerty, Liang, & Zeger, 2002). Leon (2004) presented power tables for repeated observations of a binary outcome. Murray (1998) extended the work from single level studies to CRTs. The power calculations presented by Murray (1998) use the same parameters as those used for continuous outcomes including sample sizes at all levels, difference in the outcome between clusters, and the ICC. Moerbeek, VanBreukelen, and Berger (2001) examine the optimal level of randomization and the optimal allocation of units when the outcome is binary for two-level CRTs. They use an alternative approach in which the power calculations do not include the ICC, a standardized parameter that is commonly used in power calculations for continuous outcomes. Instead, the power calculations use the unstandardized within cluster variance and between cluster variance. However, an established framework for power analyses for CRTs with binary appears to be much less developed than for continuous outcomes.

Purpose / Objective / Research Question / Focus of Study:

The purpose of this paper is to provide a framework for approaching a power analysis for a CRT with a binary outcome. We suggest a framework in the context of a simple CRT and then extend it to a blocked design, or a multi-site cluster randomized trial (MSCRT)[2]. The framework is based on proportions, an intuitive parameter when the outcome is binary. In addition, we provide sample power tables to provide readers with some intuition regarding sample sizes for CRTs with binary outcomes.

Statistical Models:

Following the hierarchical linear modeling (HLM) framework (Raudenbush & Bryk, 2002), the level-1 model is comprised of three parts: the sampling model, the link function, and the structural model. The level-1 sampling model defines the probability that the event will occur. Let Y_{ij}=1 if an event (often called a "success") occurs and Y_{ij}=0 if not. The sampling model is:

$$Y_{ij} \mid \phi_{ij} \sim B(m_{ij}, \phi_{ij}) \qquad [1]$$

for $i \in \{1,2,...,n_j\}$ students per school and for $j \in \{1,2,...,J\}$ schools;

where m_{ij} is the number of trials for student i in school j; and

ϕ_{ij} is the probability of success for student i in school j.

[2] Due to the space limitation, we focus on the simple CRT in this proposal. In the full paper, both the CRT and MSCRT will be included.

The expected value and variance of Y_{ij} are:

$$E(Y_{ij} \mid \phi_{ij}) = m_{ij}\phi_{ij} \qquad Var(Y_{ij} \mid \phi_{ij}) = m_{ij}\phi_{ij}(1-\phi_{ij}) \quad [2]$$

Note that in the case of a Bernoulli trial, $m_{ij} = 1$ so the expected value of $Y_{ij} \mid \phi_{ij}$ reduces to ϕ_{ij} and the variance reduces to $\phi_{ij}(1-\phi_{ij})$. A common link function for a binary outcome is the logit link:

$$\eta_{ij} = \log\left(\frac{\phi_{ij}}{1-\phi_{ij}}\right) \quad [3]$$

where η_{ij} is the log odds of success.

The third part of the level-1 model is the structural model:

$$\eta_{ij} = \beta_{0j} \quad [4]$$

where β_{0j} is the average log odds of success per school *j*.

The level-2 model has the same form as the level-2 model for a 2-level CRT with a continuous outcome. However, the interpretation of the parameters differs because of the logit link function:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}W_j + u_{0j}, \qquad u_{0j} \sim N(0,\tau) \quad [5]$$

where γ_{00} is the average log odds of success across schools;

γ_{01} is the treatment effect in log odds;

W_j is ½ for treatment and -½ for control;

u_{0j} is the random effect associated with each school mean; and

τ is the between school variance in log odds.

In combined form, the model is:

$$\eta_{ij} = \gamma_{00} + \gamma_{01}W_j + \mu_{0j} \ . \quad [6]$$

Power Calculations:

We use a first order Taylor series approximation to linearize the model. Under MQL, we linearize ϕ_{ij} around the fixed part of equation 6 (Breslow & Clayton, 1993). After linearization, the hypothesis testing and power calculations are very straightforward. We are interested in testing whether the treatment effect, $\gamma_{01} = 0$. Under the null hypothesis, the test statistic follows a central t-distribution. Under the alternative hypothesis, the test statistic follows a noncentral t-distribution with *J*-2 degrees of freedom and noncentrality parameter λ, where

$$\lambda = \frac{\gamma_{01}}{\sqrt{\frac{4(\tau + \sigma^2/n)}{J}}}. \quad [7]$$

The power for a two-sided test is:

$$\text{Power} = 1 - \varphi(t_{\alpha/2,J-2} - \lambda) + \varphi(-t_{\alpha/2,J-2} - \lambda) \quad [8]$$

where φ is the cumulative distribution function for the t-distribution; and

$t_{\alpha/2,J-2}$ is the critical value under the null hypothesis with *J-2* degrees of freedom.

As the noncentrality parameter increases, the power increases. Although the power calculations appear quite straightforward, you may recall that the estimates necessary for the noncentrality parameter are in log odds, which is not a readily usable metric. For example, if a researcher is designing a study in which graduation status is the primary outcome of interest, he is more likely to think about the proportion of students graduating than the log-odds of student graduating. Also, in terms of the variability across schools, he is more likely to think about the variability in graduation rates across schools, not the variability in log-odds of graduation rates across schools. Because a power analysis is only as good as the parameters that are used, we propose to use more intuitive parameters to guide the power analysis.

Usefulness / Applicability of Method:

The phrase binary outcome immediately conjures up the term proportions. Thus we use the proportions to guide the power analyses. We conduct the power analysis from estimates of four parameters including the proportion of successes in the treatment group, ϕ_E, the proportion of successes in the control group, ϕ_C, and a lower and upper bound on the proportion of successes in the control group, $\phi_{C_{LB}}$ and $\phi_{C_{UB}}$. Next we describe how the proportions are translated into the noncentrality parameter necessary for the power calculations.

We begin by examining the numerator of the noncentrality parameter, or the difference in the treatment and control group. The proportion of successes in the treatment and control group can easily be converted to log odds using the following $\eta_E = \log\left(\frac{\phi_E}{1-\phi_{iE}}\right)$ and $\eta_C = \log\left(\frac{\phi_C}{1-\phi_C}\right)$, such that the difference between η_E and η_C is now the estimate of the difference between the treatment and control group in log odds. The denominator of the noncentrality parameter includes two variance components, σ^2 and τ. Because the outcome is binary, the within school variance is a function of the proportion of successes in the treatment and control group and is easily calculated from estimates of the proportions. The between cluster variance, τ, in the context of a binary outcome or in terms of log odds is not an obvious or intuitive parameter for study planners. However, it is more likely that a researcher can estimate the lower bound, $\phi_{C_{LB}}$, and upper bound, $\phi_{C_{UB}}$, of a 95 percent plausible value range for the proportion of successes among the control schools. Converting these bounds to log odds, $\eta_{C_{LB}} = \log\left(\frac{\phi_{C_{LB}}}{1-\phi_{C_{LB}}}\right)$ and $\eta_{C_{UB}} = \log\left(\frac{\phi_{C_{UB}}}{1-\phi_{C_{UB}}}\right)$, we can now assume that the log-odds follow an approximately normal distribution. The midpoint of the interval is $\eta_{C_M} = \left(\frac{\phi_{C_{UB}} + \phi_{C_{LB}}}{2}\right)$. A 95 percent plausible value interval around η_{C_M} is $\eta_{C_M} \pm 1.96(\sqrt{\text{var}(\eta_{C_M})}$. The term $\text{var}(\eta_{C_M})$ represents τ, the between cluster variation among the control schools. Algebraic manipulation of the plausible interval reveals that $\tau = \left(\frac{\eta_{C_M} - \eta_{C_L}}{1.96}\right)^2$. In other words, if the

researcher can estimate an upper and lower bound of successes across control schools, τ can easily be calculated. Hence γ_{01}, σ^2, and τ, the three parameters required for the power calculations in equation 8 can be calculated from the four proportions, all of which are intuitive for researchers designing studies with binary outcomes.

Example:

Suppose that a team of researchers are interested in testing the effectiveness of a new stay-in-school campaign. They select a sample of 30 schools to participate in the study. The outcome is whether or not a student in 12^{th} grade graduates. On average there are 150 12^{th} graders per school. Based on school history, they expect that the graduation rate across schools is about 70 percent, with a range from 55 to 90 percent. They believe that the treatment, participation in the stay-in-school campaign, will boost graduation rates by 9 percentage points.

We use Optimal Design V2.0 to produce the power curves. Optimal Design calculates the power based on the four proportions as well as the two sample sizes. Figure 1 displays the power curve for the example. As you can see, the power increases as the number of schools increases. For example, approximately 44 schools (rounded up from 43 assuming equal allocation) would be required to achieve power of 0.80.

General Intuitions:

The example above introduced the intuitive parameters guiding the power analysis and showed the power for one specific case. Given that these parameters are likely more accessible and intuitive for researchers, we examine how the sample size, probabilities of success in the treatment and control conditions, and range of the plausible intervals affect the power. Table 1 provides the power for a fixed total of 20 clusters and 50 individuals per cluster. We vary the probability of success in the treatment and control condition as well as the plausible interval. In the full paper, we provide several tables which vary different parameters and examine the patterns in statistical power.

Conclusions:

Binary outcomes, such as graduation status or retention status, play an important role in studies of educational interventions. Designing studies with binary outcomes requires a shift from traditional parameters we use in the design of studies with continuous outcomes to parameters that are intuitive when dealing with binary outcomes. We propose a framework for conducting power analyses based on proportions. We contend that basing power calculations on intuitive parameters will strengthen the quality and accuracy of power analyses for CRTs with binary outcomes.

Appendices

Appendix A. References

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., Bos, J.M., & Lee, S.W. (1999). Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for the Evaluation of Education Programs. *Evaluation Review*, 23(4), 445-469.

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.

Boruch, R.F (1997). *Randomized experiments for planning and evaluation: A practical guide*. Thousand Oaks, CA: Sage Publications.

Boruch, R.F. (2005). Better Evaluation for Evidence Based Policy: Place Randomized Trials in Education, Criminology, Welfare, and Health. *The Annals of American Academy of Political and Social Science,* 599.

Boruch, R. F., & Foley, E.(2000). The Honestly Experimental Society. In *Validity and Social Experiments: Donald Campbell's Legacy*. Edited by L. Bickman. Thousand Oaks, CA: Sage Publications, 193-239.

Boruch, R.F., DeMoya, D., & Snyder, B. (2002). The Importance of Randomized Field Trials in Education and Related Areas. In *Evidence Matters: Randomized Field Trials in Education Research*. Edited by F. Mosteller & R. Boruch. Washington, D.C: Brookings Institution Press, 50-79.

Breslow, N.E., & Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models, *Journal of the American Statistical Association*, 88(421), 9-25.

Cook, T.D. (2002). Randomized Experiments in Educational Policy Research: A Critical Examination of the Reasons the Educational Evaluation Community has Offered for not Doing Them. *Educational Evaluation and Policy Analysis*, 24(3), 175-199.

Cook, T.D. (2005). Emergent Principles for the Design, Implementation, and Analysis of Cluster-based Experiments in Social Science. *The Annals of American Academy of Political and Social Science,* 599.

Diggle, P.J., Heagerty, P., Liang, K.-Y., & Zeger, S.L. (2002) *Analysis of longitudinal data, 2nd edition*. Oxford: Oxford University Press.

Donner, A. & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold Publishers.

Fleiss, J.L., Levin, B., Paik, M.C. (1981). *Statistical methods for rates and proportions*. New York: Wiley Press.

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.

Hsieh, F.Y., Block, D.A., &Larsen, M.D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17, 1623-1634.

Leon, A.C., (2004). Sample size requirements for comparisons of two gropus on repeated observations of a binary outcome. *Evaluation and the Health Professions*, 27(1), 34-44.

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.

Moerbeek, M., VanBreukelen, G., & Berger, M. (2001). Optimal experimental designs for multilevel logistic models. *Journal of the Royal Statistical Society,* 50(1), 17-30.

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, 2nd edition*. Thousand Oaks: Sage Publications.

Raudenbush, S.W., Martinez, A., & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis*, 29(1), 5-29.

Schochet, P. (2008). Statistical Power for Random Assignment Evaluations of Education Programs. *Journal of Educational and Behavioral Statistics*, 33(1), 62-87.

Whitehurst, G.R. (2003). *New Wine, NewBottles*. http://www.ed.gov/rschstat/research/pubs/ies.html.

Appendix B. Tables and Figures

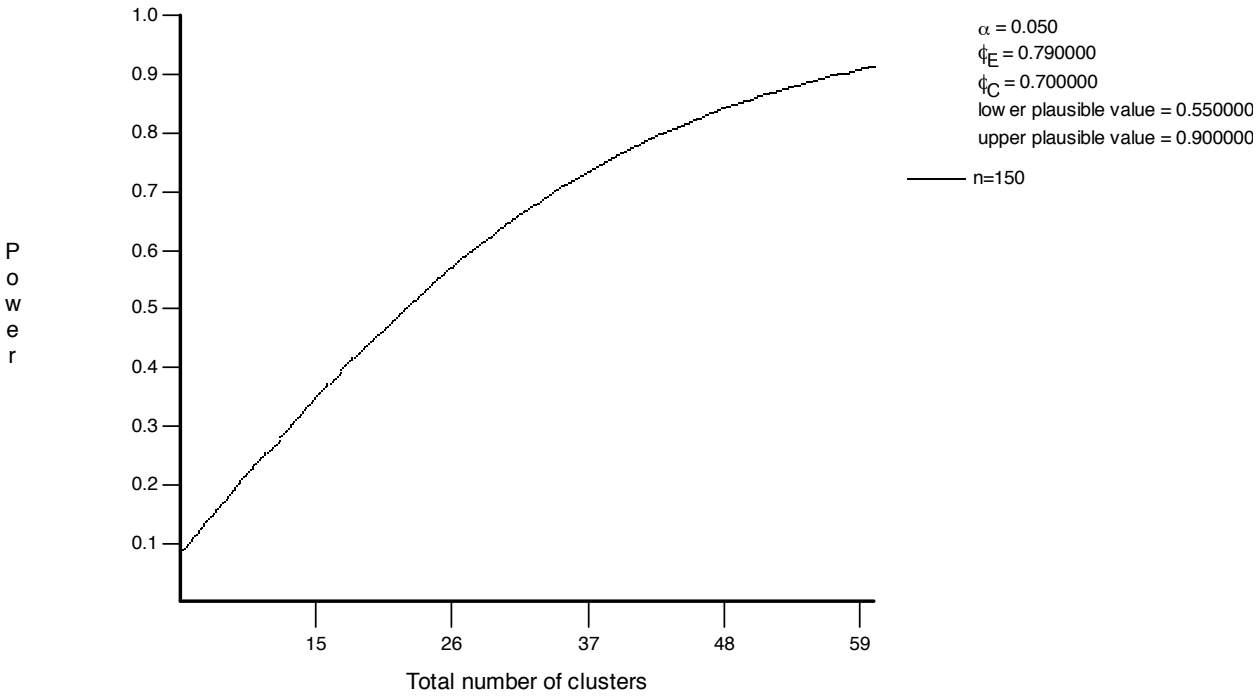

Figure 1. Power curve for 2-level CRT with binary outcome.

Table 1. The power to detect the main effect of treatment given 20 clusters and 50 persons per cluster.

Phi E	Phi C	PI (0.1,0.9)	PI (0.2,0.8)	PI (0.3,0.7)
0.1	0.2	0.30	0.55	
	0.3	0.37	0.94	0.99
	0.4	0.89	0.99	0.99
	0.5	0.97	0.99	
0.2	0.3	0.16	0.31	0.55
	0.4	0.43	0.76	0.97
	0.5	0.71	0.97	0.99
	0.6	0.91	0.99	0.99
0.3	0.4	0.13	0.23	0.43
	0.5	0.34	0.65	0.93
	0.6	0.63	0.93	0.99
	0.7	0.87	0.99	0.99
0.4	0.5	0.12	0.20	0.38
	0.6	0.32	0.61	0.91
	0.7	0.63	0.93	0.99
	0.8	0.91	0.99	
0.5	0.6	0.12	0.20	0.38
	0.7	0.34	0.65	0.93
	0.8	0.71	0.97	
	0.9	0.97		
0.6	0.7	0.13	0.23	0.43
	0.8	0.43	0.76	0.97
	0.9	0.89		
0.7	0.8	0.16	0.31	0.55
	0.9	0.67		
0.8	0.9	0.30		