The Implications of "Contamination" for Experimental Design in Education

Christopher H. Rhoads

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2011 36: 76 originally published online 30 November 2010

DOI: 10.3102/1076998610379133

Journal of Educational and Behavioral Statistics
February 2011, Vol. 36, No. 1, pp. 76–104
DOI: 10.3102/1076998610379133

The Implications of "Contamination" for Experimental Design in Education

Christopher H. Rhoads
Northwestern University

Experimental designs that randomly assign entire clusters of individuals (e.g., schools and classrooms) to treatments are frequently advocated as a way of guarding against contamination of the estimated average causal effect of treatment. However, in the absence of contamination, experimental designs that randomly assign intact clusters to treatments are less efficient than designs that randomly assign individual units within clusters. The current article considers the case of contamination processes that tend to make experimental and control subjects appear more similar than they truly are. The article demonstrates that, for most parameter values of practical interest, the statistical power of a randomized block (RB) design remains higher than the power of a cluster randomized (CR) design even when contamination causes the effect size to decrease by as much as 10%–60%. Furthermore, from the standpoint of point estimation, RB designs will tend to be preferred when true effect sizes are small and when the number of clusters in the experiment is not too large, but CR designs will tend to be preferred when true effect sizes are large or when the number of clusters in the experiment is large.

Keywords: *contamination; cluster randomized experiments; experimental design; multilevel models; statistical power*

Educational systems have a natural hierarchical structure, with students nested within classrooms, which are themselves nested within schools, which are themselves nested within districts, and so on. Evaluation researchers often use these natural groupings of individuals into clusters when conducting field experiments.

Although many experimental designs have been developed (see Kirk, 1995), the two most widely used experimental designs in education research are variants of one of the two basic designs: the cluster randomized (CR) design and the randomized block (RB) design. The CR design uses clusters as the unit of random assignment. Hence, in the CR design, all individuals within a given cluster receive the same treatment. The RB design instead uses clusters as blocks, so that units are randomly assigned to treatments within clusters. For reasons of statistical efficiency, among other reasons, researchers usually aim for an equal number of units within each block to be allocated to each treatment. I shall use the

term *randomized block* to refer to a design where an equal number of subjects is randomly assigned to each treatment within each cluster.

In some instances, the nature of the particular program being evaluated dictates a CR design. For instance, a whole-school reform logically must be implemented in an entire school. However, in many cases, the nature of the intervention is such that a researcher planning a field trial could contemplate using either a CR design or a RB design.

According to the terminology used in this article, the unit within the educational hierarchy that defines "clusters" and "units within clusters" will depend on the particular educational intervention being evaluated and the statistical model hypothesized by the researcher. For instance, a study of a professional development effort might consider schools as clusters, in which case the design decision would involve deciding between a CR design where all teachers within a given school are assigned to receive or not receive professional development and an RB design where different teachers within each school are assigned to receive professional development. As a different example, an evaluation of an individualized tutoring program at a single school might regard classrooms as clusters and the design decision would involve deciding between randomly assigning all students eligible for tutoring within a given classroom to an experimental or control condition or randomly assigning some students within a given classroom to the experimental group and some to the control group.

One argument frequently advanced in favor of the CR design is its potential to minimize *contamination* effects (Donner & Klar, 2000). The phenomenon of contamination is also variously referred to as *leakage* (Plewis & Hurry, 1998), *spillover effects* (Bloom, 2005), or *treatment diffusion* (Shadish, Cook, & Campbell, 2002). Contamination occurs when interaction between individuals randomly assigned to different treatment conditions causes some individuals to receive features of a treatment to which they were not assigned. For instance, if an intervention that is designed to reduce dropout rates by increasing engagement with school is delivered to some at-risk children but not others within a given school, it is possible that the children not receiving the intervention will also have their engagement with school increased as a result of the increased engagement of their peers.

The ability of CR designs to minimize contamination by ensuring greater physical separation between individuals receiving different treatments is often cited as a reason for preferring a CR design to an RB design (Raudenbush, 1997; Raudenbush, Martinez, & Spybrook, 2007). The literature contains statements such as "the cluster randomized design is the most appropriate design for implementation research as it minimizes the potential for contamination across trial groups" (Campbell, Mollison, & Grimshaw, 2001, p. 396) and "randomizing schools is ... the design of choice for evaluating classroom-level innovations, if they are likely to 'spillover' within schools from treatment classrooms to control classrooms" (Bloom, Richburg-Hayes, & Black, 2007, p. 30). Commenting on published

justifications for adopting a CR design, Donner and Klar (2000) find that a desire to "minimize or remove contamination" is a "particularly common concern" (p. 3).

However, it is well known that the standard error of estimates of treatment effects is larger in experimental designs that randomize clusters than it is in designs that randomize units within clusters (Blitstein, Hannan, Murray, & Shadish, 2005; Bloom et al., 2007; Cornfield, 1978; Konstantopoulos, 2008a, 2008b; Raudenbush et al., 2007; Schochet, 2008). Typically, only a relatively small number of clusters are recruited into an experiment, and the standard error in CR designs depends mainly on the number of clusters in the experiment, rather than the number of individuals. Hence, the researcher attempting to design an effective field trial faces a dilemma. She can randomize by cluster to avoid the potential for contamination, but she risks a less efficient design. She can randomize within cluster to maximize statistical efficiency, but she risks a contaminated estimate of the treatment effect.

Unfortunately, the existing literature provides little guidance regarding how to make this design choice. Much of the literature seems to imply that cluster randomization should be preferred whenever any amount of contamination is expected. Schochet (2008) argues that designs that randomize within cluster are appropriate when spillover effects are "expected to be small" but gives no guidance as to what might constitute a small spillover effect.

The current article assumes that the contamination process will tend to make treatment and control subjects look more similar, on average, than they really are. Thus, if the uncontaminated effect size is a positive number, the contaminated effect size will be a smaller positive number. It seems reasonable to assume that most contamination processes will be of this sort. However, there may be some sorts of contamination that will tend to spuriously increase the observed effect size. For instance, an intervention that has many different ingredients, all of which must be implemented properly to produce positive effects, may well result in negative effects if only a portion of the intervention is received, as would likely occur when the control group is contaminated by the treatment group. An example might be interventions designed to build leadership capacity in schools (Harris & Lambert, 2003). The results given in this article will not apply if contamination is of this sort.

The existing literature on comparing CR and RB designs when there is contamination has occurred almost exclusively in the health sciences field (Borm, Melis, Teerenstra, & Peer, 2005; Slymen & Hovell, 1997; Torgerson, 2001). The statistical model underlying existing work assumes that there are a large number of clusters in the experiment (so, for instance, it can be assumed that the null distribution of a test statistic is approximately distributed as a standard normal). These articles also assume that treatment effects are homogeneous across clusters. Furthermore, existing work has used statistical power as the criterion for evaluating different designs. That is, existing work has assumed that an experimental design should be preferred if it achieves fixed Type I and Type II error rates with a smaller total sample size.

It is unsurprising that existing work has used statistical power as the criterion for comparing designs. Power analysis is assuredly the dominant paradigm for deciding the sample size required for a given research project. Most, if not all, major grant making institutions in education research require a power analysis to determine the appropriate sample size of the studies that they fund (Institute of Education Sciences, 2009; National Science Foundation, 2009; Scheier & Dewey, 2007).

However, much recent work has recognized that researchers make a serious mistake when they summarize the results of their work solely in terms of the results of significance tests (Schmidt & Hunter, 1997; Ziliak & McCloskey, 2004). This work notes that it is crucial to report estimates of the size of treatment effects in addition to information about the "statistical significance" of the effect, and some recent work has explored the idea of sample size planning to achieve accurate parameter estimates rather than to achieve a specified statistical power (Maxwell, Kelley, & Rausch, 2008). Hence, this article compares the RB and CR designs both using statistical power as the criterion for deciding between the designs and using the quality of point estimates of the average treatment effect as the criterion.

The article proceeds as follows. In the section on Homogeneous Treatment Effects, the case where treatment effects are constant across clusters is considered. The CR and RB designs are first compared under the assumption that the design that achieves fixed Type I and Type II error rates with a smaller total sample size should be preferred. Existing results are extended to the situation where there is only a small number of clusters are in the experiment. Finally, the CR and RB designs are compared with regard to which design results in a better estimator of the treatment effect. As is customary, mean squared error (MSE) is used as the criterion for measuring the quality of the estimator, with the design that results in a smaller value of MSE being preferred.

The section on Heterogeneous Treatment Effects develops results for the case where treatment effects vary across clusters. Once again, the CR and RB designs are compared, first using statistical power as the criterion for choosing between the designs and then using MSE as the criterion. Interestingly, regardless of whether treatment effects vary across clusters, the situations where the RB design is preferred to the CR design are not, in general, the same using the power criterion as they are using the MSE criterion.

Homogeneous Treatment Effects

Previous Work

Slymen and Hovell (1997) compare the CR design to a third design, rarely used in education research, where clusters are ignored in the randomization process, so that treatment is assigned at the individual level irrespective of clusters. They compare CR and individually randomized designs in terms of total sample

size required to achieve fixed Type I and Type II error rates. Their calculations assume a large number of clusters in the experiment in that z rather than t quantiles are used in sample size computations. Torgerson (2001) presents similar calculations.

Borm, Melis, Teerenstra, and Peer (2005; hereafter BMTP) suggest an interesting compromise between CR and RB designs that they label "pseudo cluster randomization." I do not consider the case of pseudo cluster randomization; however, I do adopt many of the concepts and much of the notation used by BMTP for the current discussion comparing CR and RB designs.

The Model

The statistical model underlying existing work comparing CR and RB designs with contamination is usually left unstated but implicitly this work has assumed that treatment effects are homogeneous across clusters. The implicit statistical model can be described as follows.

Assume that there are two treatments of interest and that $2m$ clusters, each of size n, are available for the experiment. Let Y_{jk}^{E} represent the outcome of the kth individual within the jth cluster who is assigned to the experimental condition and Y_{jk}^{C} represent the outcome of the kth individual within the jth cluster who is assigned to the control condition. Then, the model is

$$Y_{jk}^{E} \sim N\left(\mu^{E}, \sigma_{B}^{2} + \sigma^{2}\right) \tag{1}$$

$$Y_{jk}^{C} \sim N\left(\mu^{C}, \sigma_{B}^{2} + \sigma^{2}\right) \tag{2}$$

$$\mathrm{Cov}\left(Y_{jl}^{C}, Y_{jk}^{E}\right) = \sigma_{B}^{2}; \tag{3}$$

$$\mathrm{Cov}\left(Y_{jl}^{i}, Y_{jk}^{i}\right) = \sigma_{B}^{2}; \quad i = E, C; \quad l \neq k \tag{4}$$

$$\mathrm{Cov}\left(Y_{jl}^{i}, Y_{km}^{i}\right) = 0; \quad i = E, C; \quad j \neq k. \tag{5}$$

Many results will depend on the intracluster correlation coefficient (ICC), defined as $\rho = \sigma_{B}^{2} / \left(\sigma_{B}^{2} + \sigma^{2}\right)$.

Although the notation used in Equations 1–5 is somewhat nonstandard, it has been adopted to have a unified notation for describing both the RB and the CR designs. When the RB design is considered, the model described by Equations 1–5 corresponds to the usual two-way unrestricted analysis of variance (ANOVA) mixed model, with treatment as a fixed factor, clusters as a random factor crossed with treatment, and where the treatment by cluster variance component is set to zero. When the CR design is considered, the model described by Equations 1–5 corresponds to an ANOVA model with treatment as a fixed factor and clusters as a random factor nested within treatment. These ANOVA models

are the usual models used to compute power for the respective designs under consideration (Raudenbush, 1997; Raudenbush & Liu, 2000).

Formalizing Contamination

Let the two treatments under investigation be labeled E and C (for "experimental" and "control"). Contamination is formalized as follows. Let μ_E and μ_C be population mean outcomes under treatments E and C in the absence of contamination and let $d = \mu_E - \mu_C$. The "$\bullet$" notation is used to denote averaging across a given subscript. Then, in the RB design, the expected outcome for those receiving treatment E is

$$E_{RB}\left(Y^E_{\bullet\bullet}\right) = \mu_E - c_E d.$$

So c_E is the population-level average contamination of experimental group subjects by control group subjects, given that half of all subjects are assigned to the experimental condition in each cluster. Similarly, the expected outcome for those assigned to the control condition is

$$E_{RB}\left(Y^C_{\bullet\bullet}\right) = \mu_C - c_c d.$$

So c_C is the population-level average contamination of control group subjects by experimental group subjects, given that half of all subjects are assigned to the experimental condition in each cluster. The results presented will depend only on the total contamination

$$c_T = c_C + c_E. \tag{6}$$

Thus, the expected value of $Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}$ with contamination is given by

$$E_{RB}\left(Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}\right) = d(1 - c_T). \tag{7}$$

As noted above, this article assumes that the contamination process will make subjects "more similar" in the sense that contamination will cause the value of d to decrease in absolute value without changing sign. This implies the following restrictions on the values of c_C and c_E:

$$0 \le c_c; 0 \le c_E; c_E + c_c \le 1. \tag{8}$$

This article will maintain the assumption that cluster randomization successfully ensures an uncontaminated estimate of the treatment effect. Thus, the expected value of the estimated average treatment effect under the CR design is

$$E_{CR}\left(Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}\right) = d. \tag{9}$$

Comparing the Two Designs Using Statistical Power as a Criterion

Comparing the Designs for Large Experiments

The variance of the estimated average treatment effect under the model specified in Equations 1–5 is

$$\mathrm{var}_{CR}\left(Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}\right) = \frac{2\sigma^2}{mn}\frac{(1+(n-1)\rho)}{1-\rho} \tag{10}$$

for the CR design and

$$\mathrm{var}_{RB}\left(Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}\right) = \frac{2\sigma^2}{mn} \tag{11}$$

for the RB design.

As noted by BMTP, if the test statistic used to test the null hypothesis of no treatment effect is assumed to follow a normal distribution (e.g., if there are a large number of clusters in the experiment), then the number of clusters required to achieve fixed Type I and Type II error rates under a given design is proportional to t^{-2}, where

$$t^{-2}_{des} = \frac{\mathrm{var}_{des}\left(Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}\right)}{E^2{}_{des}\left(Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}\right)}.$$

So, for instance, $t^{-2}_{RB} = \left(\frac{2\sigma^2}{mn}\right)\Big/\left(d^2(1-c_T)^2\right)$. Thus, t^{-2} provides a simple criterion for evaluating different experimental designs with regard to the sample size required to achieve fixed Type I and Type II error rates, with smaller values of t^{-2} being preferred.

It follows that the ratio of the sample size required under a CR design to the sample size required under an RB design is given by

$$t^{-2}_{RB}/t^{-2}_{CR} = (1-\rho)\Big/\left((1+(n-1)\rho)(1-c_T)^2\right). \tag{12}$$

To choose between the CR and RB designs, it is necessary only to know whether the ratio given in Equation 12 is greater than 1 or less than 1. Therefore, it may be useful to think of the design problem facing the experimenter in terms of the maximum allowable contamination (MAC) for fixed values of n and ρ, where MAC is defined as the contamination value that satisfies

$$(1-\rho)\Big/(1+(n-1)\rho)(1-c_T)^2 = 1.$$

This results in the following definition of MAC (subscripted by p, to indicate that this is the maximum allowable contamination when the two designs are compared using power as the criterion, and by hom, to indicate that the homogeneous treatment effects assumption is maintained).

$$\mathrm{MAC}_{p,\mathrm{hom}}(n,\rho) = 1 - \sqrt{\frac{1-\rho}{1+(n-1)\rho}}. \tag{13}$$

$\mathrm{MAC}_{p,\mathrm{hom}}$ is the maximum amount of contamination that can be tolerated before the CR design becomes preferable to the RB design. It is useful to explore the properties of MAC for extreme values of n and ρ. Hence,

$$\lim_{\rho \downarrow 0} \mathrm{MAC}_{p,\mathrm{hom}} = 0; \tag{14a}$$

$$\lim_{\rho \uparrow 1} \mathrm{MAC}_{p,\mathrm{hom}} = 1; \tag{14b}$$

$$\lim_{n \to \infty} \mathrm{MAC}_{p,\mathrm{hom}} = 1. \tag{14c}$$

Thus, for very large n and for very large ρ, contamination could remove almost all of the original treatment effect and yet the RB design would still have more power than the CR design to detect non-null treatment effects. Although Equation 14a would seem to imply that for very small ρ the CR design will always be preferred to the RB design, Figure 1 shows that the counteracting influence of Equation 14c implies that for many values of n that are of practical interest, the RB design will be preferred, even for very small values of ρ.

Figure 1 graphs $\mathrm{MAC}_{p,\mathrm{hom}}$ against ρ for fixed values of n. The curves displayed in Figure 1 can be regarded as "indifference curves." For values of (c_T,ρ) on the curve, the experimental planner would be indifferent between a CR design and an RB design. For values of (c_T,ρ) below the curve, she would prefer the RB design and for values above the curve, she would prefer the CR design.

To understand the practical implications of Figure 1, it is necessary to have some sense of what values of ρ can be expected for educational experiments. Recent publications by Hedges and Hedberg (2007), Bloom, Richburg-Hayes, and Black (2007), and Schochet (2008) are instructive in this regard.

Hedges and Hedberg (2007) presented a compendium of intraclass correlations computed from surveys of academic achievement in the United States where clusters are defined by schools. For statistical models that do not control for any individual or school-level covariates (unconditional models), they find values of ρ that are generally in the $.15 < \rho < .25$ range, with higher values associated with lower grade levels. For statistical models that condition on pretest scores and/or demographic variables, values of ρ are generally between .03 and .20.

Bloom et al. (2007) present some values of ρ for standardized reading and math test score outcomes, computed from several large urban school districts in the United States. Estimates of school-level unconditional ICCs, computed within districts, range from .15 to .29.

Schochet (2008) reports numerous preschool and elementary school-level ICC values, both unconditional and conditional. The range of ICCs reported is

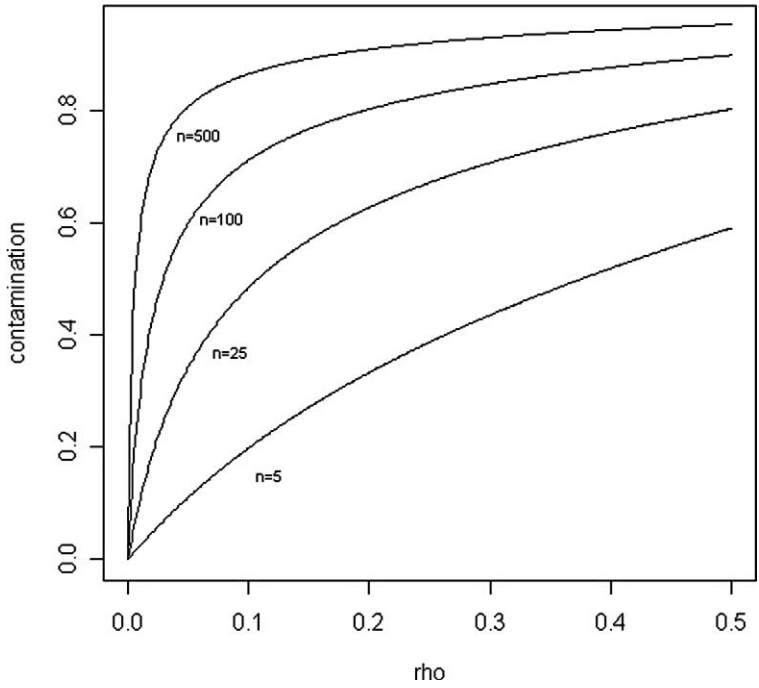

FIGURE 1. *Values of ρ and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for fixed Type I and Type II error rates. Homogeneous treatment effects, infinite degrees of freedom. CR = cluster randomized; RB = randomized block.*

generally consistent with those reported by Hedges and Hedberg, although in one study of 4-year-olds, an unconditional ICC of .38 was obtained. Schochet also gives some limited information about classroom-level ICCs within schools, suggesting that a value of .15 is reasonable. In sum, it seems reasonable to assume that the range of intraclass correlation values of practical interest to education researchers is contained in the interval (0, .5).

In light of the above information about plausible values of ρ, Figure 1 shows that in many design situations of interest, the RB design remains preferable to the CR design even when the amount of contamination is quite large. For example, suppose that researchers will study reading outcomes of fourth grade students within each school in the experimental sample. Assume that the class size will be approximately $n = 25$. Hedges and Hedberg (2007) report a school level intraclass correlation of .242 for fourth grade reading achievement. For these values for n and ρ, the RB design (which assigns students within schools to the intervention) would be preferable to the CR design (which assigns entire schools to the intervention) even when contamination causes a 60% deterioration in the observable treatment effect.

Comparing the Designs in Small Experiments

In fact, the computations in the previous subsection understate the case in favor of the RB design. Cornfield (1978) noted that two penalties are paid when researchers randomize clusters rather than individuals. First, the variance of the estimated average treatment effect increases. Second, the degrees of freedom available to estimate that variance decrease. By assuming a large number of clusters in the experiment, BMTP are able to ignore the second penalty and focus only on the increase in variance. However, many educational experiments are run with only a few clusters participating. In these cases, it is necessary to take account of the different degrees of freedom available to estimate the variance of the estimated average treatment effect under the different designs.

Summary Quantities and Small Sample Test Statistics

To compare the small sample distributions of the test statistics used to test the null hypothesis of no treatment effect in the RB and CR designs, it is necessary to define some summary quantities. Define the following error "sum of squares" quantity relevant to the CR design:

$$\mathrm{SSB}^{\mathrm{CR}} = n\left(\sum_{j=1}^{m}\left(Y_{j\bullet}^{E} - Y_{\bullet\bullet}^{E}\right)^2 + \sum_{j=1}^{m}\left(Y_{j\bullet}^{C} - Y_{\bullet\bullet}^{C}\right)^2\right).$$

The statistic

$$\frac{\sqrt{\frac{mn}{2}}\left(Y_{\bullet\bullet}^{E} - Y_{\bullet\bullet}^{C}\right)}{\sqrt{\frac{\mathrm{SSB}^{\mathrm{CR}}}{2m-2}}} = t_{CR}$$

can be used to test the null hypothesis of no treatment effect in the CR design. This statistic has the t distribution with $2m - 2$ degrees of freedom under the null hypothesis.

Now, define the following error "sum of squares" quantities relevant to the RB design:

$$\mathrm{SSTC}^{\mathrm{RB}} = \frac{n}{2}\sum_{j=1}^{2m}\left(Y_{j\bullet}^{E} - Y_{\bullet\bullet}^{E} - Y_{j\bullet}^{\bullet} - Y_{\bullet\bullet}^{\bullet}\right)^2 + \frac{n}{2}\sum_{j=1}^{2m}\left(Y_{j\bullet}^{C} - Y_{\bullet\bullet}^{C} - Y_{j\bullet}^{\bullet} - Y_{\bullet\bullet}^{\bullet}\right)^2$$

$$\mathrm{SSW}^{\mathrm{RB}} = \sum_{j=1}^{2m}\sum_{k=1}^{n/2}\left(Y_{jk}^{E} - Y_{j\bullet}^{E}\right)^2 + \sum_{j=1}^{2m}\sum_{k=1}^{n/2}\left(Y_{jk}^{C} - Y_{j\bullet}^{C}\right)^2.$$

The statistic

$$\frac{\sqrt{\frac{mn}{2}}\left(Y^{E}_{\bullet\bullet} - Y^{C}_{\bullet\bullet}\right)}{\sqrt{\frac{\mathrm{SSW}^{\mathrm{RB}} + \mathrm{SSTC}^{\mathrm{RB}}}{2mn - 2m - 1}}} = t_{RB}$$

may be used to test the null hypothesis of no treatment effect in the RB design. It has a t distribution with $2mn - 2m - 1$ degrees of freedom under the null hypothesis.

Because all experiments in fact involve a finite number of clusters, the analysis in the current section is technically more correct than the comparison of the CR and RB designs presented above, which assumes a null z distribution for the relevant test statistic. However, the presentation in terms of the z distribution did involve some considerable simplifications relative to the current presentation. In particular, to compare the sample size requirements of the RB and CR designs for experiments with a small number of clusters, it is necessary to specify the desired Type I error rate, the desired statistical power to detect a true uncontaminated treatment effect of a given size, and the minimum standardized treatment effect that the study should be able to detect with the specified power (MDES). When a z null distribution is assumed, the relative sample size of the two designs is invariant to these three parameters.

In experiments involving only a few moderately sized clusters having $2mn - 2m - 1$ degrees of freedom instead of $2m - 2$ degrees of freedom to estimate, the variance of the estimated average treatment effect can provide a substantial additional power advantage to the RB design. However, once m is above 10–15 or so, the advantage of having additional degrees of freedom is not large. In such cases, it is generally sufficient to act as though the true null distribution of the test statistics under consideration is standard normal.

Define

$$\delta_T = \frac{d}{\sqrt{\sigma_B^2 + \sigma^2},} \tag{15}$$

so that δ_T is the uncontaminated treatment effect standardized by the total standard deviation of an individual observation. Figure 2 reproduces the "indifference curve" presentation of Figure 1, with indifference curves for the z distribution in solid lines and indifference curves for the t distribution in broken lines. Figure 2 assumes two-sided Type I error rate = .05, desired power = .80, and $\delta_T = .2$. Figure 3 is equivalent to Figure 2 except for the effect size is changed to $\delta_T = .4$.

A Type I error rate of .05 and a desired power of .80 are considered standard numbers to use for computing the necessary sample size in an educational intervention (Cohen, 1977). The values of δ_T were chosen to cover the range of standardized treatment effects generally considered practically meaningful in

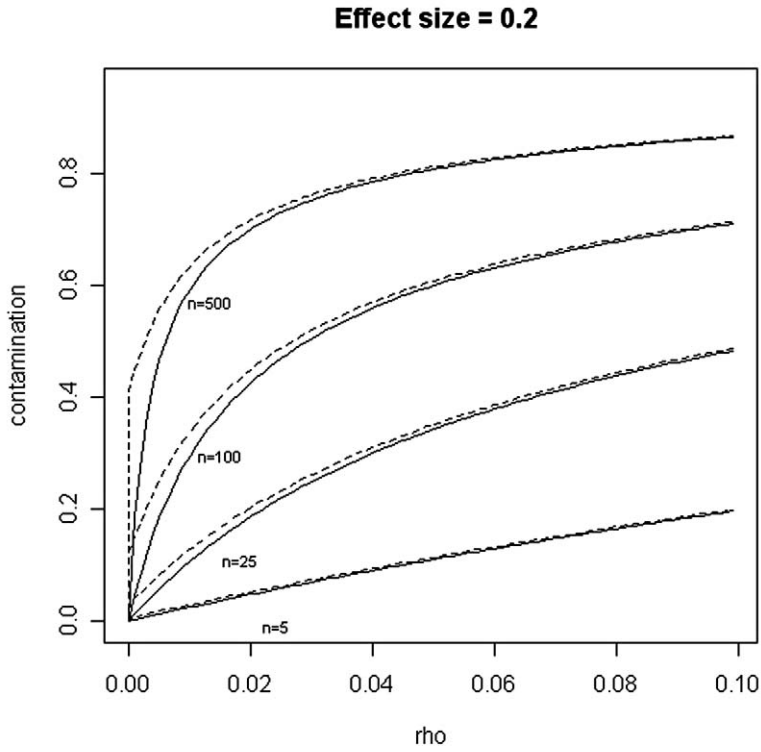

FIGURE 2. *Values of* ρ *and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for two-sided level of significance = .05, power = .80, and* $\delta_T = .2$*. Homogeneous treatment effects, comparison of infinite and finite degrees of freedom. CR = cluster randomized; RB = randomized block.*

education. For instance, Schochet (2008) has noted that the values of .2, .25, and .33 have often been chosen as desired minimum detectable effect sizes for educational evaluations. In fact, Schochet argues that effect sizes less than .20 may well have practically meaningful consequences. Bloom et al. (2007) make a similar claim, stating that "program effect sizes for student achievement of as little as .10–.20 might be policy relevant." However, given the expense associated with conducting studies large enough to detect standardized effect sizes less than .20, it seems unlikely that the educational evaluation community will find it possible to consistently design studies capable of detecting standardized effect sizes of less than .20 with high power.

Looking at Figures 2 and 3, it is clear that accounting for the exact small sample distribution of the test statistics will make the most difference when the value of δ_T is relatively large, when the value of ρ is small, and when the within-cluster sample

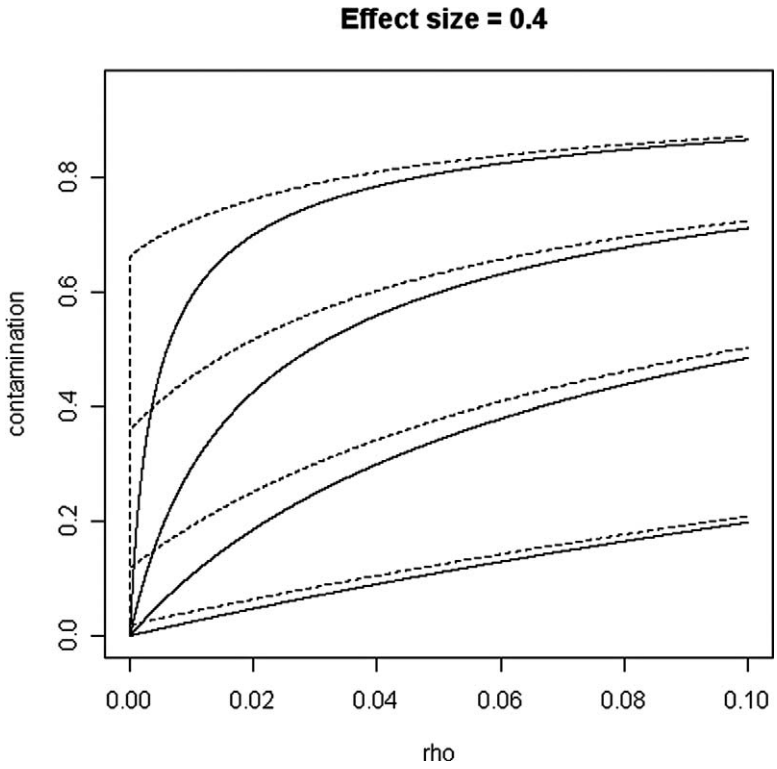

FIGURE 3. *Values of ρ and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for two-sided level of significance = .05, power = .80, and* $\delta_T = .4$*. Homogeneous treatment effects, comparison of infinite, and finite degrees of freedom. CR = cluster randomized; RB = randomized block.*

size, n, is relatively large. The reason is that under these conditions, .80 power can be achieved with a fairly small sample of clusters, which implies that there will be very few degrees of freedom available to estimate the variance of the estimated average treatment effect in the CR design. For instance, when $n = 500$, $\rho = .003$, and $\delta_T = .4$, .80 power can be achieved in the CR design with only $m = 2$ clusters assigned to each treatment group and so the test statistic used in the CR design would have only two degrees of freedom. With the parameter values $n = 500$, $\rho = .003$, and $\delta_T = .4$, the large sample presentation given in Figure 3 would imply an $MAC_{p,\text{hom}}$ value of .368. However, correctly accounting for the small sample distributions of the test statistics implies that contamination as high as .687 can be tolerated before the CR design is preferred to the RB design.

In contrast, consider the situation where $n = 100$, $\rho = .1$, and $\delta_T = .2$. In this case, $m = 44$ clusters are required to achieve .8 power in the CR design. The large

sample presentation implies an $MAC_{p,\text{hom}}$ value of .713. Correctly accounting for the degrees of freedom gives a value of $MAC_{p,\text{hom}} = .716$.

It may seem unnecessary to some to bother with the analysis in this section, given that ICC values in educational experiments are generally expected to be in the $.15 < \rho < .25$ range and that it is usually desirable to detect small effect sizes. However, considerably smaller ICCs and somewhat larger minimum detectable effects will be obtained if the analysis conditions on covariates to remove residual variance. Hedges and Hedberg (2007) found conditional ICCs as low as .03 in surveys measuring academic achievement. In one sample of schools from Louisville, Kentucky, Gargani and Cook (2005) found that an unconditional ICC of .11 could be converted to a conditional ICC of .0175 if prior year pretest scores were included in the model as a covariate. Additionally, between community ICCs in the .001 to .01 range, in conjunction with cluster sizes of $n = 500$ or above are commonly observed in community intervention trials conducted in the epidemiology field (see, for instance, Hannan, Murray, Jacobs, & McGovern, 1994).

Comparing the Two Designs Using Accuracy of Estimation as a Criterion

Although most studies are planned to achieve fixed Type I and Type II error rates, it is also important that studies be designed to achieve good estimates of the size of the treatment effect. The standard criterion that is used to measure the quality of estimators is the MSE. The MSE of an estimator, $\hat{\theta}$, of a population parameter, θ, is defined as

$$\text{MSE} = \text{var}(\hat{\theta}) + (\text{Bias}(\hat{\theta}))^2. \tag{16}$$

The *bias* of an estimator is defined as the difference between the expected value of the estimator and the parameter that it is meant to estimate, thus,

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta} - \theta). \tag{17}$$

Before determining the MSE of the estimators that result from the CR and RB designs, it is necessary to determine which is the main quantity of interest, the unstandardized average treatment effect, d, or the standardized average treatment effect, δ_T. The reasons to focus interest on either d or δ_T have been discussed at length in the literature and are not repeated here. The interested reader may see Baguley (2009).

If the factor used to standardize the treatment effect, $\sigma_T = \sqrt{\sigma_B^2 + \sigma^2}$, is known, then the decision between d and δ_T as the primary quantity of interest is immaterial for the current purposes. In that case, the MSE of an estimator of δ_T may be found by simply dividing the MSE of the corresponding estimator of d by σ_T. If σ_T is not known, then MSE calculations to evaluate estimators of δ_T are considerably more complicated. For simplicity, the current article focuses attention on the MSE of the usual estimator of δ_T when σ_T is known.

As noted above, the results presented may also be regarded as applying to the MSE of estimators of d, where MSE is reported in the units of the total standard deviation. Even when primary interest is in δ_T with σ_T unknown, the results presented here may be regarded as approximately correct, because the additional factors appearing in the MSE computations are quite similar for estimators resulting from the RB design and estimators resulting from the CR design.

Computing MSE and Comparing the Two Designs With Respect to MSE

Assume that the population average (standardized) treatment effect will be estimated by the quantity $\hat{\delta}_T = \frac{Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}}{\sigma_T}$. Then, it follows immediately from the results given in Equations 7, 9, 10, and 11 that

$$\mathrm{MSE}_{\mathrm{RB,hom}} = \frac{2}{mn}(1-\rho) + c_T^2\delta_T^2, \quad \text{and} \tag{18}$$

$$\mathrm{MSE}_{\mathrm{CR,hom}} = \frac{2}{mn}(1-(n-1)\rho). \tag{19}$$

A researcher choosing between the RB and CR designs will be concerned with situations where the relation $\mathrm{MSE}_{\mathrm{RB,hom}} < \mathrm{MSE}_{\mathrm{CR,hom}}$ holds. Thus, as was the case when comparing the CR and RB designs with respect to power, we are led to define the maximum allowable contamination by equating Equations 18 and 19 and solving for c_T. We obtain

$$\mathrm{MAC}_{\mathrm{MSE,hom}} = \sqrt{\frac{2\rho}{m}}\,\frac{1}{\delta_T}. \tag{20}$$

We subscript MAC in this case by MSE, to indicate that the designs are being compared with respect to their MSE, and by hom, to indicate that the homogeneous treatment effects assumption is maintained.

Examining Equation 13, we see that $\mathrm{MAC}_{p,\mathrm{hom}}$ does not depend on the number of clusters in the experiment, $2m$, but does depend on n, the size of the clusters. However, $\mathrm{MAC}_{\mathrm{MSE,hom}}$ does not depend on n but does depend on m. Equation 20 makes clear that, from an estimation standpoint, when either m or δ_T is very large the CR design will almost always be preferred. Unlike Equation 13, it is possible for Equation 20 to be greater than 1, the logical upper bound for c_T. In this case, regardless of the amount of contamination, the RB design will be preferred from the standpoint of MSE.

Table 1 presents values of $\mathrm{MAC}_{\mathrm{MSE,hom}}$ for various values of δ_T, ρ, and m. Values of $\mathrm{MAC}_{\mathrm{MSE,hom}}$ greater than 1 are reported as 1. For comparison purposes, the first three rows of Table 1 present values of $\mathrm{MAC}_{p,\mathrm{hom}}$ for $n = 5$, 25, and 100, respectively.

If clusters are represented by schools and we assume that most experiments in the education field will involve somewhere between 10 and 100 schools, an

TABLE 1

$MAC_{MSE,hom}$ Values: Homogeneous Treatment Effects and MSE as Evaluation Criterion

	ρ	0.001	0.01	0.02	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4
$MAC_{p,hom}$	$n = 5$	0.002	0.024	0.047	0.110	0.198	0.271	0.333	0.388	0.436	0.480	0.520
	$n = 25$	0.012	0.106	0.186	0.343	0.486	0.570	0.629	0.673	0.708	0.737	0.762
	$n = 100$	0.047	0.295	0.427	0.600	0.713	0.768	0.804	0.829	0.849	0.865	0.878
δ_T	M											
0.2	2	0.16	0.50	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	5	0.10	0.32	0.45	0.71	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	10	0.07	0.22	0.32	0.50	0.71	0.87	1.00	1.00	1.00	1.00	1.00
0.2	20	0.05	0.16	0.22	0.35	0.50	0.61	0.71	0.79	0.87	0.94	1.00
0.2	30	0.04	0.13	0.18	0.29	0.41	0.50	0.58	0.65	0.71	0.76	0.82
0.2	40	0.04	0.11	0.16	0.25	0.35	0.43	0.50	0.56	0.61	0.66	0.71
0.2	50	0.03	0.10	0.14	0.22	0.32	0.39	0.45	0.50	0.55	0.59	0.63
0.5	2	0.06	0.20	0.28	0.45	0.63	0.77	0.89	1.00	1.00	1.00	1.00
0.5	5	0.04	0.13	0.18	0.28	0.40	0.49	0.57	0.63	0.69	0.75	0.80
0.5	10	0.03	0.09	0.13	0.20	0.28	0.35	0.40	0.45	0.49	0.53	0.57
0.5	20	0.02	0.06	0.09	0.14	0.20	0.24	0.28	0.32	0.35	0.37	0.40
0.5	30	0.02	0.05	0.07	0.12	0.16	0.20	0.23	0.26	0.28	0.31	0.33
0.5	40	0.01	0.04	0.06	0.10	0.14	0.17	0.20	0.22	0.24	0.26	0.28
0.5	50	0.01	0.04	0.06	0.09	0.13	0.15	0.18	0.20	0.22	0.24	0.25
0.8	2	0.04	0.13	0.18	0.28	0.40	0.48	0.56	0.63	0.68	0.74	0.79
0.8	5	0.03	0.08	0.11	0.18	0.25	0.31	0.35	0.40	0.43	0.47	0.50
0.8	10	0.02	0.06	0.08	0.13	0.18	0.22	0.25	0.28	0.31	0.33	0.35
0.8	20	0.01	0.04	0.06	0.09	0.13	0.15	0.18	0.20	0.22	0.23	0.25
0.8	30	0.01	0.03	0.05	0.07	0.10	0.13	0.14	0.16	0.18	0.19	0.20
0.8	40	0.01	0.03	0.04	0.06	0.09	0.11	0.13	0.14	0.15	0.17	0.18
0.8	50	0.01	0.03	0.04	0.06	0.08	0.10	0.11	0.13	0.14	0.15	0.16

Note: MAC = maximum allowable contamination; MSE = mean squared error. Shaded values indicate that any amount of contamination can be tolerated and the randomized block design is still preferred.

examination of Table 1 reveals that, when the true effect size is small ($\delta_T = .2$) and values of the ICC are in the .1–.3 range commonly reported for academic achievement outcomes, $MAC_{MSE,hom}$ will range from about 0.30 to 1.00. That is, contamination can cause a deterioration of anywhere from 30% to 100% of the observable treatment effect and the RB design will still have smaller MSE than the CR design. Suppose instead that the effect size is $\delta_T = .5$ (which would be considered large for experiments studying academic outcomes), that one hundred schools participate in the study ($m = 50$), and that the ICC is a relatively small .1. This configuration of parameters is relatively unfavorable for the RB design, yet even in this case, contamination of as much as 13% can be tolerated before the CR design is preferred to the RB design.

Table 1 makes clear that the situations where a particular design may be preferred from the standpoint of power are not necessarily the same as the situations where that design would be preferred from an estimation perspective. For instance, suppose that a researcher has at her disposal $2m = 60$ schools and she will sample $n = 5$ students within each school to study. She expects a relatively small ICC of $\rho = .05$ (perhaps because she will condition on covariates), a somewhat small effect size of $\delta_T = .2$ and she suspects that contamination may be as high as 20%. Under this scenario, $MAC_{MSE,hom} = .29$ and $MAC_{p,hom} = .11$. Thus, if primary interest is in power to reject the null hypothesis of no treatment effect, the CR design should be chosen; however, if primary interest is in obtaining a good estimate of the size of the treatment effect, then the RB design should be chosen.

Heterogeneous Treatment Effects

The assumption that treatment effects are constant across clusters appears to be standard in the health sciences literature; however, it is not recommended in educational experiments where it is possible, if not likely, that the effect of treatment may vary across different educational settings. Thus, I now consider the problem of choosing between CR and RB designs under a model that allows the effect of treatment to vary across clusters. To specify this model, we decompose the between-cluster variance component from the homogeneous treatment effects model, σ_B^2, into two parts, σ_{Bl}^2 and σ_{TC}^2, so that

$$\sigma_B^2 = \sigma_{Bl}^2 + \sigma_{TC}^2. \tag{21}$$

The σ_{Bl}^2 component of variance represents the portion of the total between-cluster variance that is "removed" from the variance of the estimated treatment effect in a blocked design. The σ_{TC}^2 component of variance is the portion of the total between-cluster variance which cannot be removed by blocking and can be interpreted as one half of the total variance in cluster-specific treatment effects. With these parameter definitions, the model for the data when treatment effects vary across clusters can be written as:

$$Y_{jk}^{E} \sim N\left(\mu^{E}, \sigma_{Bl}^{2} + \sigma_{TC}^{2} + \sigma^{2}\right) \tag{22}$$

$$Y_{jk}^{C} \sim N\left(\mu^{C}, \sigma_{Bl}^{2} + \sigma_{TC}^{2} + \sigma^{2}\right) \tag{23}$$

$$\mathrm{Cov}\left(Y_{jl}^{C}, Y_{jk}^{E}\right) = \sigma_{Bl}^{2}; \tag{24}$$

$$\mathrm{Cov}\left(Y_{jl}^{i}, Y_{jk}^{i}\right) = \sigma_{Bl}^{2} + \sigma_{TC}^{2}; \quad i = E, C; \quad l \neq k \tag{25}$$

$$\mathrm{Cov}\left(Y_{jl}^{i}, Y_{km}^{i}\right) = 0; \quad i = E, C; \quad j \neq k. \tag{26}$$

Notice that the model given by Equations 22–26 is basically the homogeneous treatment effects model given in Equations 1–5 with σ_B^2 rewritten as $\sigma_B^2 = \sigma_{Bl}^2 + \sigma_{TC}^2$. The only exception is in the representation of $\mathrm{Cov}\left(Y_{jl}^{C}, Y_{jk}^{E}\right)$, which was denoted σ_B^2 in the previous model but is now represented as σ_{Bl}^2. Because $\sigma_B^2 = \sigma_{Bl}^2$ when treatment effects do not vary across clusters, the two models agree in the homogeneous treatment effects case. As noted above for the homogeneous treatment effects case, when the RB design is considered, the model described by Equations 22–26 corresponds to the usual unrestricted two-way ANOVA mixed model, with treatment as a fixed factor and clusters as a random factor crossed with treatment. When the CR design is considered, the model described by Equations 22–26 corresponds to an ANOVA model with treatment as a fixed factor and clusters as a random factor nested within treatment. These ANOVA models are the usual models used to compute power for the respective designs under consideration (see, for instance, Raudenbush, 1997; Raudenbush & Liu, 2000). Of course, because all subjects in a given cluster receive the same treatment, it is impossible to obtain separate estimates of σ_{Bl}^2 and σ_{TC}^2 in the CR design. Instead, this design only allows estimation of the parameter $\sigma_B^2 = \sigma_{Bl}^2 + \sigma_{TC}^2$.

Deciding between the CR and the RB designs under the model given by Equations 22–26 will require making assumptions about the parameter $\omega = \sigma_{TC}^2 / \sigma_B^2$, which represents the proportion of between-cluster variation that cannot be removed by blocking. By definition, $0 \leq \omega \leq 1$. However, it is very difficult to obtain empirical estimates of where in the [0,1] interval the ω parameter is likely to be located. This is because necessarily the variance of treatment effects across clusters will depend on the particular intervention being studied, which has generally not been well studied prior to the planned experiment (hence, the rationale for the planned experiment).

However, there are some limited suggestions in the literature about reasonable values of this parameter. Schochet (2008) defines $\rho_\theta = \omega\rho$ and reports values of ρ_θ between .04 and .08 for an evaluation of the 21st Century Community

Learning Centers program, where clusters are schools and outcomes are reading and math test scores. Estimates for the school-level ICC from this study were between .09 and .24, thus the ρ_θ estimates given imply values of ω between .16 and .89, depending on the grade level.

Comparing the Two Designs for Large Experiments

Expected value calculations are unchanged when the heterogeneous treatment effects model is maintained and the variance of the estimated treatment difference under the CR design remains

$$\text{var}_{\text{het,CR}} = \left(\frac{2\sigma^2}{mn}\right)\left(\frac{1+(n-1)\rho}{1-\rho}\right). \tag{27}$$

However, the variance of the estimated treatment difference under the RB design is

$$\text{var}_{\text{het,RB}} = \left(\frac{2\sigma^2}{mn}\right)\left(\frac{1+(n\omega/2-1)\rho}{1-\rho}\right). \tag{28}$$

Thus, the ratio of the sample size required under a CR design to the sample size required under an RB design is given by

$$t_{RB}^{-2}/t_{CR}^{-2} = [1+(n\omega/2-1)\rho]/[1+(n-1)\rho](1-c_T)^2. \tag{29}$$

It is again instructive to define the maximum allowable contamination MAC by setting the right-hand side of Equation 29 equal to 1 and solving for c_T to give

$$\text{MAC}_{p,\text{het}}(n,\rho,\omega) = 1 - \sqrt{\frac{1+(n\omega/2-1)\rho}{1+(n-1)\rho}}. \tag{30}$$

It is useful to examine some of the properties of the function given by Equation 30. First, clearly $\text{MAC}_{p,\text{het}}$ is a decreasing function of ω for fixed values of n and ρ.

Looking at the limit of $\text{MAC}_{p,\text{het}}(n,\rho,\omega)$ as different parameters approach their extremes for fixed values of the other parameters reveals:

$$\lim_{\rho\downarrow 0} \text{MAC}_{p,\text{het}} = 0; \tag{31}$$

$$\lim_{\rho\uparrow 1} \text{MAC}_{p,\text{het}} = 1 - \sqrt{\omega/2}; \tag{32}$$

$$\lim_{n\to\infty} \text{MAC}_{p,\text{het}} = 1 - \sqrt{\omega/2}; \tag{33}$$

$$\lim_{\omega\uparrow 1} \text{MAC}_{p,\text{het}} = 1 - \sqrt{\frac{1+(n/2-1)\rho}{1+(n-1)\rho}}; \tag{34}$$

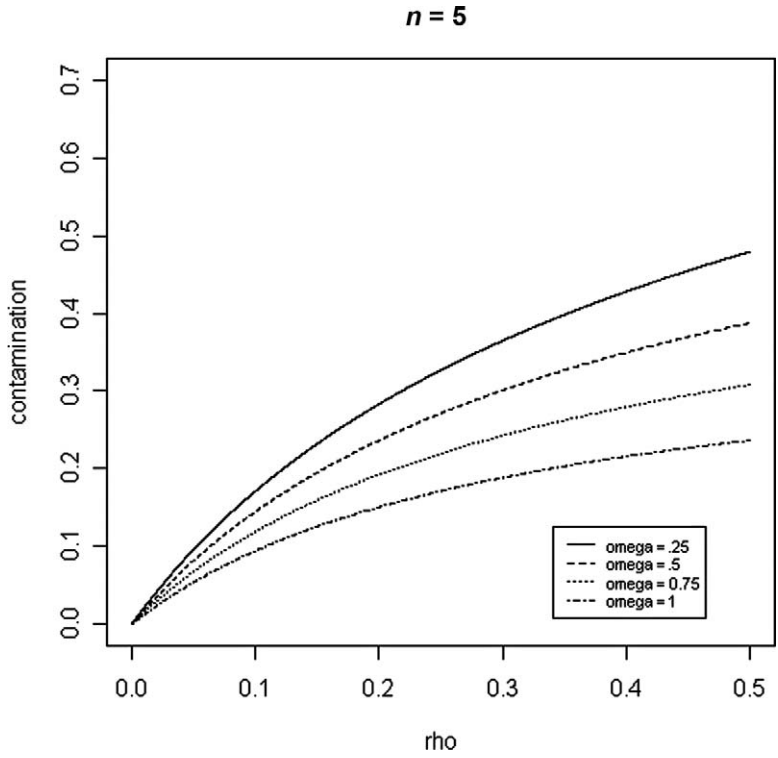

FIGURE 4. *Values of* ρ *and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for fixed Type I and Type II error rates. Heterogeneous treatment effects, small* n*. CR = cluster randomized; RB = randomized block.*

$$\lim_{\omega \downarrow 0} \mathrm{MAC}_{p,\mathrm{het}} = \mathrm{MAC}_{p,\mathrm{hom}} = 1 - \sqrt{\frac{1-\rho}{1+(n-1)\rho}}. \tag{35}$$

Given the results presented in Equations 32 and 33, it is of interest to know how the quantity $1 - \sqrt{\omega/2}$ varies with ω. Some evaluations at representative values of ω are .646 at $\omega = .25$, .592 at $\omega = .5$, .388 at $\omega = .75$, and .293 at $\omega = 1$. Comparing Equations 32 and 33 with the analogous equations for the homogeneous treatment effects case demonstrates that as n or ρ get large the RB design is no longer necessarily preferred. However, because it seems reasonable to assume that contamination will rarely exceed .29, and $1 - \sqrt{\omega/2}$ will never be less than .29 when n or ρ is very large, the general rule that large values of n and ρ tend to favor the RB design still holds.

Figures 4–7 graph $\mathrm{MAC}_{p,\mathrm{het}}$ for increasingly larger values of n. In my discussion, I assume that education researchers will typically be interested in values of

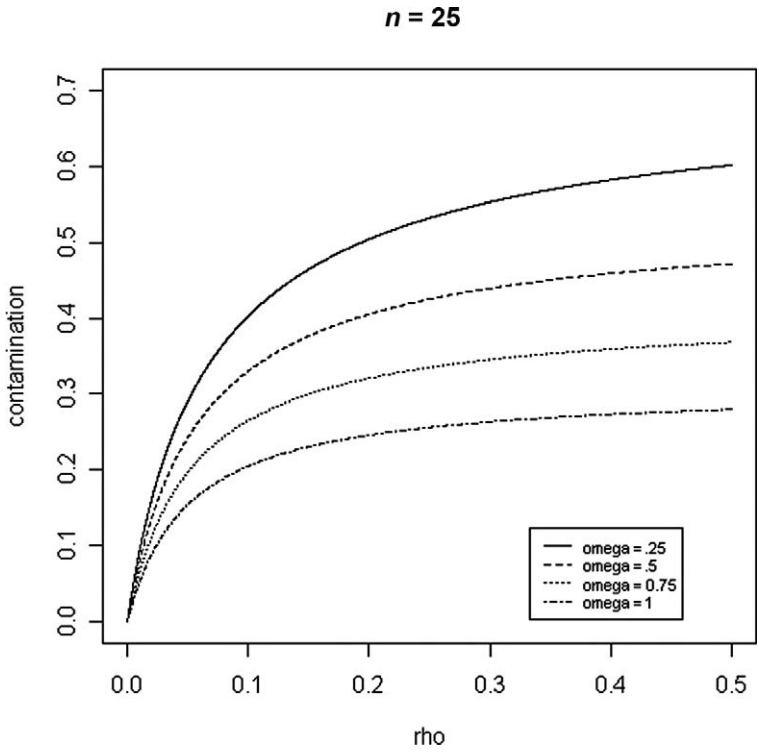

FIGURE 5. *Values of* ρ *and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for fixed Type I and Type II error rates. Heterogeneous treatment effects, moderate* n. *CR = cluster randomized; RB = randomized block.*

$MAC_{p,het}$ when ρ is approximately .2. Assuming this value for the ICC, Figure 4 shows that when $n = 5$, a 10% contamination of the treatment effect can be tolerated before the CR design is preferred, even when the treatment effect heterogeneity is at its maximum possible value. If more modest levels of heterogeneity are assumed, contamination as high as 25% can be tolerated. Figures 5–7 show how much more contamination can be tolerated when the size of the clusters randomized increases. At an ICC of .2, Figure 5 shows that when $n = 25$, at least a 25% contamination of the treatment effect can be tolerated, and if treatment heterogeneity is modest, contamination of 50% or more can be tolerated. Recall that when $\omega = 1$, the maximum amount of contamination that can be tolerated under any configuration of the other design parameters of interest is 29.3%. From Figure 5, we see that when $n = 25$, we are already quite close to this value for an ICC of .2. Hence, further increases in the within-cluster sample size n will not change $MAC_{p,het}$ very much. Figures 6 and 7 illustrate this point visually. However, when ω is at a less extreme value, increases in n may still have an

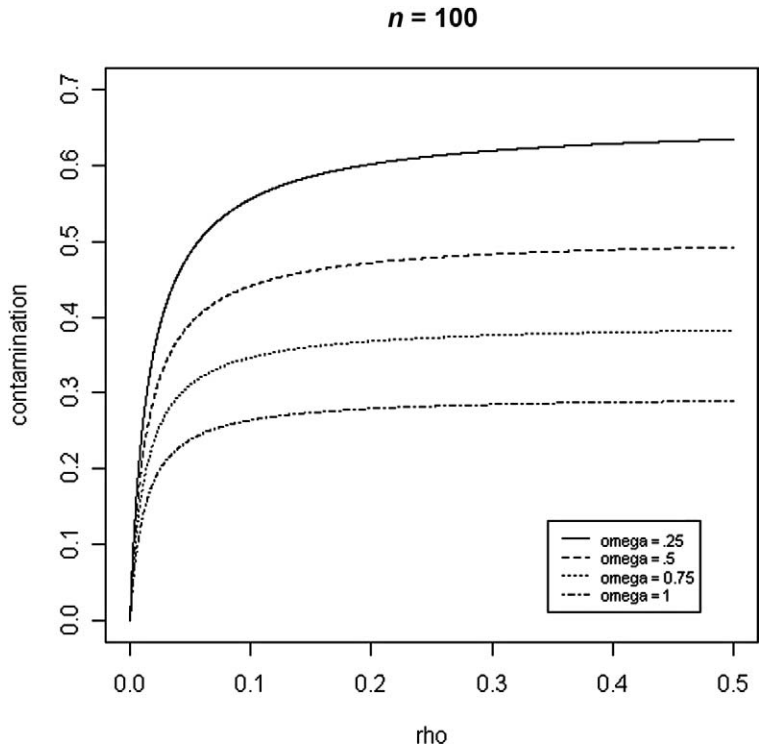

FIGURE 6. *Values of* ρ *and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for fixed Type I and Type II error rates. Heterogeneous treatment effects, large* n*. CR = cluster randomized; RB = randomized block.*

appreciable impact on the maximum allowable contamination. When the ICC is .2 and $\omega = .25$, MAC changes from about 50% at $n = 25$ to over 60% at $n = 500$.

When there is more variation in treatment effects across clusters, less contamination can be tolerated before the CR design is preferred to the RB design. However, even when there is substantial heterogeneity of treatment effects, it is still striking just how much contamination can be tolerated before the CR design is preferable to the RB design. For instance, across all values of n and ω displayed in Figures 4–7, when $\rho = .2$ contamination of anywhere between .1 and .65 can be tolerated. If n is at least 25, contamination of at least 20% can be tolerated even when ω is at its maximum value.

Comparing the Two Designs in Small Experiments

Unlike the homogeneous treatment effects model, the degrees of freedom available to estimate the variances in Equations 27 and 28 are quite similar for

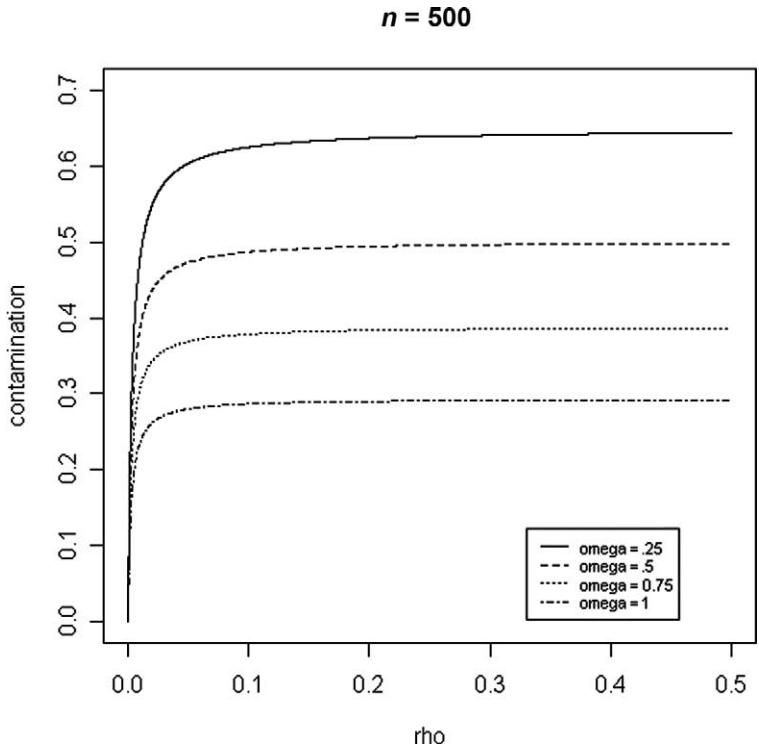

FIGURE 7. *Values of* ρ *and* c_T*, which result in equivalent sample size requirements for the RB and CR designs for fixed Type I and Type II error rates. Heterogeneous treatment effects, very large* n*. CR = cluster randomized; RB = randomized block.*

the heterogeneous treatment effects model. They remain $2m - 2$ under the CR design but are reduced from $2mn - 2m - 1$ to $2m - 1$ under the RB design. Given the small difference in degrees of freedom, $\mathrm{MAC}_{p,\mathrm{het}}$ values computed from Equation 30 should be very accurate approximations to the true, small sample amount of tolerable contamination. Thus, I do not present exact comparisons of sample size requirements based on the t distribution for this model.

Comparing the Two Designs Using Accuracy of Estimation as a Criterion

As was the case for the homogeneous treatment effects model, again assume that the population average (standardized) treatment effect will be estimated by the quantity $\hat{\delta}_T = \frac{Y^E_{\bullet\bullet} - Y^C_{\bullet\bullet}}{\sigma_T}$, where $\sigma_T = \sqrt{\sigma_B^2 + \sigma^2}$. It follows immediately from Equations 7, 9, 27, and 28 that

$$\mathrm{MSE}_{\mathrm{RB,het}} = \frac{2}{mn}(1 + (n\omega/2 - 1)\rho) + c_T^2\delta_T^2, \quad \text{and} \tag{36}$$

$$\mathrm{MSE}_{\mathrm{CR,het}} = \frac{2}{mn}(1 + (n-1)\rho). \tag{37}$$

The situations where the RB design is preferred to the CR design are where $\mathrm{MSE}_{\mathrm{RB,het}} < \mathrm{MSE}_{\mathrm{CR,het}}$. Thus, by setting $\mathrm{MSE}_{\mathrm{RB,het}} = \mathrm{MSE}_{\mathrm{CR,het}}$ are solving for c_T we obtain

$$\mathrm{MAC}_{\mathrm{MSE,het}} = \sqrt{\frac{2\rho}{m}}\frac{1}{\delta_T}\sqrt{1 - \frac{\omega}{2}}. \tag{38}$$

Comparing $\mathrm{MAC}_{\mathrm{MSE,het}}$ to $\mathrm{MAC}_{\mathrm{MSE,hom}}$ as defined in Equation 20, it is clear that the formulas for MAC differ only by a factor of $\sqrt{1 - \frac{\omega}{2}}$. Because this factor is never greater than 1, the relation $\mathrm{MAC}_{\mathrm{MSE,het}} \leq \mathrm{MAC}_{\mathrm{MSE,hom}}$ holds. When ω is at its minimum value, $\omega = 0$, then $\mathrm{MAC}_{\mathrm{MSE,het}} = \mathrm{MAC}_{\mathrm{MSE,hom}}$. When ω is at its maximum value, $\omega = 1$, then $\mathrm{MAC}_{\mathrm{MSE,het}} = \sqrt{\frac{\rho}{m}}\frac{1}{\delta_T}$, so that $\mathrm{MAC}_{\mathrm{MSE,het}}$ differs from $\mathrm{MAC}_{\mathrm{MSE,hom}}$ by at most a multiplicative factor of $1/\sqrt{2}$, or about .71. $\mathrm{MAC}_{\mathrm{MSE,het}}$ depends on the number of clusters in the experiment, $2m$, but not on the cluster size, n.

Table 2 presents values of $\mathrm{MAC}_{\mathrm{MSE,het}}$ for various values of δ_T, ρ, ω, and m. Values of $\mathrm{MAC}_{\mathrm{MSE,het}}$ greater than 1 are reported as 1. For comparison purposes, the first three rows of Table 2 present values of $\mathrm{MAC}_{p,\mathrm{het}}$ for $n = 5$, 25, and 100, respectively.

For the purposes of discussion, I focus attention on cases where $\rho = .1$ or .2, where $\delta_T = .2$ or .5, and where the number of clusters in the experiment is fairly large ($m = 40$ or 50). In these situations, when the effect size is small ($\delta_T = .2$), a substantial amount of contamination (anywhere from 22% to 47%) can be tolerated before the CR design is preferred to the RB design. When the effect size is larger ($\delta_T = .5$), less contamination can be tolerated. However, even in this case, at least 10% contamination can generally be tolerated, even when treatment effects vary substantially across clusters.

It remains the case that the situations where a particular design would be preferred from the standpoint of $\mathrm{MAC}_{p,\mathrm{het}}$ are not the same as the situations where that design would be preferred from the standpoint of $\mathrm{MAC}_{\mathrm{MSE,het}}$. For instance, suppose that a researcher has at her disposal $2m = 60$ schools and she will sample $n = 100$ students within each school to study. She expects a moderate-to-small ICC value of $\rho = .1$, a somewhat small effect size of $\delta_T = .2$, fairly substantial variation in treatment effects across clusters ($\omega = .5$), and she suspects that contamination could be quite large, as much as 35%, but no larger than 40%. Under this scenario, $\mathrm{MAC}_{p,\mathrm{het}} = .44$ and $\mathrm{MAC}_{\mathrm{MSE,het}} = .35$. Thus, if primary interest is in statistical power to reject a null hypothesis of no treatment effect,

TABLE 2
$MAC_{MSE,het}$ Values: Heterogeneous Treatment Effects and MSE as Evaluation Criterion

	ρ	0.001			0.05			0.1			0.2			0.4		
	ω	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1
$MAC_{p,het}$	$n = 5$	0.002	0.002	0.001	0.096	0.081	0.054	0.171	0.144	0.094	0.283	0.236	0.150	0.428	0.350	0.216
	$n = 25$	0.011	0.009	0.006	0.291	0.242	0.154	0.403	0.330	0.205	0.504	0.405	0.246	0.582	0.459	0.273
	$n = 100$	0.041	0.035	0.023	0.486	0.392	0.239	0.556	0.441	0.264	0.602	0.472	0.279	0.629	0.489	0.288
δ_T	m															
0.2	2	0.15	0.14	0.11	1.00	0.97	0.79	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
0.2	5	0.09	0.09	0.07	0.66	0.61	0.50	0.94	0.87	0.71	1.00	1.00	1.00	1.00	1.00	1.00
0.2	10	0.07	0.06	0.05	0.47	0.43	0.35	0.66	0.61	0.50	0.94	0.87	0.71	1.00	1.00	1.00
0.2	20	0.05	0.04	0.04	0.33	0.31	0.25	0.47	0.43	0.35	0.66	0.61	0.50	0.94	0.87	0.71
0.2	30	0.04	0.04	0.03	0.27	0.25	0.20	0.38	0.35	0.29	0.54	0.50	0.41	0.76	0.71	0.58
0.2	40	0.03	0.03	0.03	0.23	0.22	0.18	0.33	0.31	0.25	0.47	0.43	0.35	0.66	0.61	0.50
0.2	50	0.03	0.03	0.02	0.21	0.19	0.16	0.30	0.27	0.22	0.42	0.39	0.32	0.59	0.55	0.45
0.5	2	0.06	0.05	0.04	0.42	0.39	0.32	0.59	0.55	0.45	0.84	0.77	0.63	1.00	1.00	0.89
0.5	5	0.04	0.03	0.03	0.26	0.24	0.20	0.37	0.35	0.28	0.53	0.49	0.40	0.75	0.69	0.57
0.5	10	0.03	0.02	0.02	0.19	0.17	0.14	0.26	0.24	0.20	0.37	0.35	0.28	0.53	0.49	0.40
0.5	20	0.02	0.02	0.01	0.13	0.12	0.10	0.19	0.17	0.14	0.26	0.24	0.20	0.37	0.35	0.28
0.5	30	0.02	0.01	0.01	0.11	0.10	0.08	0.15	0.14	0.12	0.22	0.20	0.16	0.31	0.28	0.23
0.5	40	0.01	0.01	0.01	0.09	0.09	0.07	0.13	0.12	0.10	0.19	0.17	0.14	0.26	0.24	0.20
0.5	50	0.01	0.01	0.01	0.08	0.08	0.06	0.12	0.11	0.09	0.17	0.15	0.13	0.24	0.22	0.18

(continued)

TABLE 2 *(continued)*

	ρ	0.001			0.05			0.1			0.2			0.4		
	ω	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1	0.25	0.5	1
0.8	2	0.04	0.03	0.03	0.26	0.24	0.20	0.37	0.34	0.28	0.52	0.48	0.40	0.74	0.68	0.56
0.8	5	0.02	0.02	0.02	0.17	0.15	0.13	0.23	0.22	0.18	0.33	0.31	0.25	0.47	0.43	0.35
0.8	10	0.02	0.02	0.01	0.12	0.11	0.09	0.17	0.15	0.13	0.23	0.22	0.18	0.33	0.31	0.25
0.8	20	0.01	0.01	0.01	0.08	0.08	0.06	0.12	0.11	0.09	0.17	0.15	0.13	0.23	0.22	0.18
0.8	30	0.01	0.01	0.01	0.07	0.06	0.05	0.10	0.09	0.07	0.14	0.13	0.10	0.19	0.18	0.14
0.8	40	0.01	0.01	0.01	0.06	0.05	0.04	0.08	0.08	0.06	0.12	0.11	0.09	0.17	0.15	0.13
0.8	50	0.01	0.01	0.01	0.05	0.05	0.04	0.07	0.07	0.06	0.10	0.10	0.08	0.15	0.14	0.11

Note: MAC = maximum allowable contamination; MSE = mean squared error. Shaded values indicate that any amount of contamination can be tolerated and the randomized block design is still preferred.

the RB design should be chosen. However, if primary interest is in good MSE properties for estimates of treatment effects, then a CR design should be chosen.

Conclusion

The threat of *contamination* or *treatment spillage* is often cited as a rationale for adopting a CR experimental design in field trials of educational interventions. However, the variance of estimates of the average impact of treatment is greatly increased in CR designs relative to designs that randomize individuals within clusters, and hence, in the absence of contamination, statistical power is much less in designs that randomize clusters as compared to designs that block on clusters. The current article has considered situations where the contamination process is expected to decrease the observable average treatment effect without causing the observable average treatment effect to change sign (i.e., contamination cannot change a positive treatment effect into a negative one). This assumption is not trivial. For instance, in a capacity building intervention, it may be that exposure to the "full" treatment produces positive effects, but exposure to only a portion of the treatment (as might happen with control subjects who have been contaminated by exposure to treatment subjects) produces negative effects. In this case, the contaminated treatment effect estimate will be larger than the uncontaminated estimate and the arguments presented in the current article would not hold.

However, many contamination processes can be expected to obey the assumption described above. When contamination does obey this assumption, when there are not too many clusters in the experiment (say less than 100) and when the standardized effect size is not too large (say less than .5), then the RB design will generally result in more precise estimates of the average impact of treatment and more powerful statistical tests of the null hypothesis of no treatment effect than the CR design, even with fairly large amounts of contamination. For values of the intraclass correlation coefficient generally considered to be of interest to education researchers, contamination can deteriorate the observable average treatment effect by between 10% and 60% before the CR design becomes preferable to the RB design.

References

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617.

Blitstein, J. L., Hannan, P. J., Murray, D. M., & Shadish, W. R. (2005). Increasing degrees of freedom in existing group randomized trials through the use of external estimates of intraclass correlation: The df* approach. *Evaluation Review*, *29*, 241–267.

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 Foundation.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision: Empirical guidelines for studies that randomize schools to measure the

impacts of educational interventions. *Educational Evaluation and Policy Analysis*, *29*, 30–59.

Borm, G. F., Melis, R. J. F., Teerenstra, S., & Peer, P. G. (2005). Pseudo cluster randomization: A treatment allocation method to minimize contamination and selection bias. *Statistics in Medicine*, *24*, 3535–3547.

Campbell, M. K., Mollison, J., & Grimshaw, J. M. (2001). Cluster trials in implementation research: Estimation of intracluster correlation coefficients and sample size. *Statistics in Medicine*, *20*, 391–399.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology*, *108*, 100–102.

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research.* New York, NY: Oxford University Press.

Gargani, J., & Cook, T. D. (2005). *How many schools? Limits of the conventional wisdom about sample size requirements for cluster randomized trials* (Working paper). Evanston, IL: Northwestern University.

Hannan, P. J., Murray, D. M., Jacobs, D. R., & McGovern, P. G. (1994). Parameters to aid in the design and analysis of community trials: Intraclass correlations from the Minnesota Heart Health Program. *Epidemiology*, *5*, 88–94.

Harris, A., & Lambert, L. (2003). *Building leadership capacity for school improvement.* New York, NY: McGraw Hill/Open University Press.

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

Institute of Education Sciences. (2009). *Request for applications: Education research grants FY 2010.* Washington, DC: U.S. Department of Education.

Kirk, R. (1995). *Experimental design.* Belmont, CA: Brooks Cole.

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

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 265–288.

Maxwell, S., Kelley, K., & Rausch, J. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, *59*, 537–563.

National Science Foundation. (2009). *Discovery research K–12 program solicitation NSF 09–602.* Washington, DC: Government Printing Office.

Plewis, I., & Hurry, J. (1998). A multilevel perspective on the design and analysis of intervention studies. *Educational Research and Evaluation*, *4*, 13–26.

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

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

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

Scheier, W. M., & Dewey, W. L. (Eds.). (2007). *The complete writing guide to NIH behavioral science grants.* Oxford, NY: Oxford University Press.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. A. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Lawrence Erlbaum.

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

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

Slymen, D. J., & Hovell, M. F. (1997). Cluster versus individual randomization in adolescent tobacco and alcohol studies: Illustrations for design decisions. *International Journal of Epidemiology*, *26*, 765–771.

Torgerson, D. J. (2001). Contamination in trials: Is cluster randomization the answer? *British Medical Journal*, *322*, 355–357.

Ziliak, S. T., & McCloskey, D. N. (2004). Size matters: The standard error of regressions in the American Economic Review. *The Journal of Socio-Economics*, *33*, 527–546.

Author

CHRISTOPHER H. RHOADS is Institute of Education Sciences postdoctoral fellow in policy research at the Institute of Policy Research at Northwestern University, 2040 Sheridan Road, Evanston, IL 60208; e-mail: christopherrhoads2008@u.northwestern.edu. His research interests include the design and analysis of cluster randomized experiments, hierarchical modeling, causal inference, regression-discontinuity designs and meta-analysis.

Manuscript received June 25, 2009
Revision received May 11, 2010
Accepted June 27, 2010