Strategies for Improving Precision in Group-Randomized Experiments

Stephen W. Raudenbush
University of Chicago

Andres Martinez and Jessaca Spybrook
University of Michigan

Interest has rapidly increased in studies that randomly assign classrooms or schools to interventions. When well implemented, such studies eliminate selection bias, providing strong evidence about the impact of the interventions. However, unless expected impacts are large, the number of units to be randomized needs to be quite large to achieve adequate statistical power, making these studies potentially quite expensive. This article considers when and to what extent matching or covariance adjustment can reduce the number of groups needed to achieve adequate power and when these approaches actually reduce power. The presentation is nontechnical.

Keywords: *group-randomized experiments, multilevel research design*

Iɴᴛᴇʀᴇsᴛ has rapidly increased in studies that randomly assign social units to alternative treatment conditions. Such units include whole schools (Borman et al., 2005; Cook, Hunt, & Murphy, 2000; Flay, 2000; Mosteller, Light, & Sachs, 1996; Porter, Blank, Smithson, & Osthoff, 2005), housing projects (Bloom & Riccio, 2005; Sikkema, 2005), neighborhoods or whole communities (Hannan, Murray, Jacobs, & McGovern, 1994; Sherman & Weisburd, 1995; Teruel & Davis, 2000; Weisburd, 2000, 2005), and physician practices (Donner & Klar, 2000; Grimshaw, Eccles, Campbell, & Elbourne, 2005; Leviton & Horbar, 2005; Murray, 1998). During the 1980s, the U.S. Department of Labor began supporting randomized trials as a means for understanding which employment and training programs

effectively increased earnings. Similarly, during the 1990s, the U.S. Department of Health and Human Services started supporting randomized trials to determine which drug prevention programs and other risk reduction programs were effective (Bloom, 2005; Mosteller & Boruch, 2002). In 2002, with the creation of the new Institute for Education Sciences, the U.S. Department of Education established the conduct of randomized trials as a priority (Education Sciences Reform Act, 2002). In fact, most of these studies treat whole classrooms or schools rather than students as the units of randomization.

The decision to assign groups rather than individuals to treatments is essential when interventions are designed to treat entire collectives of persons. This is true, for example, in whole-school

5

reform efforts designed to engage all of the teachers in a school in a joint effort to improve instruction (Borman et al., 2005) or when a violence prevention program is intended to operate by changing the normative climate in an entire school (Flay, 2000). Group-based trials are also often preferred when there is a danger of "spillover effects" or when logistical, ethical, or political considerations discourage person-based randomization (Bloom, 2005). Cook (2005) suggested that by targeting the group level, an intervention may improve group-level processes, resulting in greater long-term impacts on individuals who come into contact with the group.

In most group-based studies, the statistical power to detect treatment effects depends more strongly on the number of groups available than on the number of persons per group (Bloom, Bos, & Lee, 1999; Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). At the same time, the number of groups to be sampled will also typically drive costs. For example, in a school-based intervention, it is far more expensive to recruit a single school and to sustain that school's engagement than it is to sample and assess an additional student once a school has agreed to participate. Thus, the number of groups drives cost and power more than the total sample size of students.

An evaluation of statistical power for a typical group-randomized study can create the impression that the study will be exceedingly expensive. Consider, for example, a hypothetical study in which J groups are to be assigned either to an experimental or to a control condition ($J/2$ groups in each), with $n = 100$ persons per group assessed after the treatment on a continuous outcome variable. Assume that the true mean difference between the two groups is equivalent to 0.25 on the scale of the outcome standard deviation. Such effect sizes (ESs) are widely regarded as nonnegligible in education research focused on academic achievement (Bloom, 2005). Finally, assume that 85% of the variation lies between persons within groups, with 15% of this variation lying between groups within treatment conditions. This variance ratio is common in school-based research, particularly for achievement outcomes in elementary schools (Bloom, Richburg-Hayes, & Black,

2005). The data will be analyzed by means of a simple *t* test using group means as the outcomes. Under these conditions, 82 schools (41 per condition) must be sampled to achieve statistical power of 0.80 to reject the null hypothesis of no program effect at the 5% level of significance (Raudenbush, 1997). In many research settings, the cost of studying 82 groups will be daunting. For example, in a study of whole-school reform, the cost of recruiting 82 schools, implementing the intervention in half of them, sustaining the involvement of those schools in data collection, and dispatching data collectors can be expected to be many millions of dollars. Given current research funding, such per study costs would severely limit the number of such studies and therefore limit the prospects of a plan to rely on school-based randomized studies for information about school improvement.

In the hypothetical example above, no attempt was made to identify or use prior information that might have predicted participant outcomes. Yet it is well known that such information can often be ascertained at comparatively low cost and that when this information is used effectively, the sample-size requirements for experiments can be reduced, sometimes substantially. Given the high cost of sampling groups, it becomes essential to find efficient experimental designs and analytic approaches that capitalize on prior information to increase statistical power.

Early in the past century, the pioneers of experimental design in agriculture learned that pretreatment information can often be exploited to improve statistical power (Cochran, 1957; Fisher, 1926, 1936, 1949). Two key approaches are available:

1. *Prerandomization blocking*: Prior to treatment assignment, experimental units can be classified into subclasses called blocks, such that within each block, the units are likely to be similar on the outcome. Units are then assigned to treatment conditions within blocks. In this way, variation between blocks is eliminated from the assessment of experimental error. If the blocking is highly successful, such that variation between blocks on the outcome far exceeds variation within blocks, the number of experimental units required to achieve a given

level of power can be dramatically smaller than the number of units required without such blocking.

2. *Covariance adjustment*: Once again, prior to randomization, the researcher identifies characteristics of units that strongly predict the outcome. More specifically, the researcher obtains data on a covariate (i.e., a pretreatment variable), typically measured on a continuous scale. After implementing the treatment and assessing outcomes, the researcher specifies a linear statistical model in which the covariate and indicators of treatment group membership are predictors. If the covariate has a strong linear association with the outcome, and if this association is similar within each treatment condition, the analysis can achieve much higher power than an analysis omitting the covariate. As a result, using a well-chosen covariate can greatly reduce the number of units needed to achieve a given power.

In sum, both approaches use prior information to reduce uncertainty about outcomes. Prerandomization blocking builds this information into the design such that randomization occurs within blocks. In contrast, analysis of covariance (ANCOVA) exploits the linear association between a covariate and the outcome in the analysis phase of the study.

The key aim of the current article is to clarify conditions under which the use of blocking and covariance adjustment can reduce the number of groups required to achieve adequate power in group-randomized studies under specific conditions. This effort may be regarded as the extension of classic Fisherian principles of experimental design to settings in which social units ("groups"), rather than individuals, are the units of randomization and treatment. Throughout the discussion, we draw on the key principles at play when persons are the unit of randomization and show how they extend nicely to the case of group-randomized studies, with certain key modifications that we shall highlight.

This is not the first article to consider such issues. Several authors have provided useful accounts of the trade-offs that arise in selecting alternative designs and analytic methods in group-randomized trials (for reviews, see Bloom, 2005; Donner & Klar, 2000; Hughes, 2005; Martin, Diehr, Perrin, & Koepsell, 1993;

Murray, 1998; Raudenbush, 1997). Our aim is to provide a precise specification of how the power of two alternative approaches, blocking and ANCOVA, compares with the power of a standard design that does not exploit pretreatment information. We make comparisons for specific sample sizes and parameter values. We hope to enable readers to use these results and the freely available software we describe to improve the planning of group-randomized studies (Raudenbush, Liu, Spybrook, Martinez, & Congdon, 2006).[1]

For simplicity, we focus on the case in which the aim is to compare two treatments: a novel "experimental" approach and a more standard "control" approach. In this case, blocks each consisting of two units may be formed. These blocks are called *matched pairs*. We consider the question of when and to what extent such matching will boost the statistical power for a given sample size or reduce the sample size needed to achieve a given level of power. We also consider the utility of ANCOVA in such a two-treatment setting.

In this setting, the aim is to provide sharp answers to specific questions about trade-offs in selecting approaches to design and analysis. We ask the following questions:

1. Under what conditions and to what extent will pretreatment matching reduce the number of groups needed to achieve adequate power to detect an ES of interest?

2. Under what conditions and to what extent will covariance adjustment reduce the number of groups needed to achieve adequate power to detect such an effect?

3. Suppose that a continuous covariate is available. Such a covariate might be used either to form matched pairs or as a covariate. Specifically, under what conditions and to what extent will matching approximate the efficiency of ANCOVA?

We restrict our attention to "two-level designs" (e.g., persons within groups, with groups as the unit of randomization) and draw on the literature from "single-level designs" (e.g., persons as units of randomization). We reserve a brief discussion of more complex designs (e.g., students within classrooms within schools that are assigned to
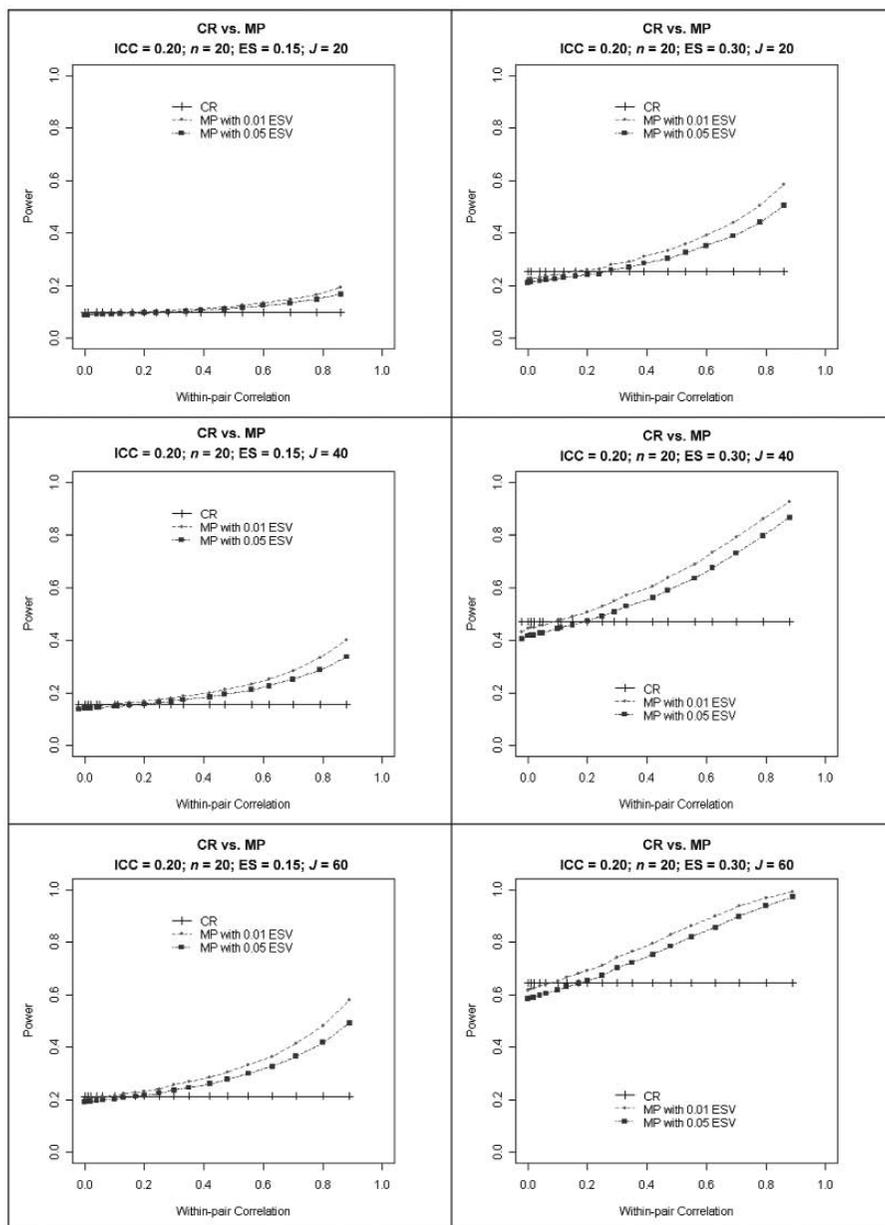
7

FIGURE 1.   *Completely randomized (CR) design versus matched-pairs (MP) design, intracluster correlation (ICC) = .20.*
Note. ES = effect size; ESV = ES variability.

treatments, repeated-measures designs) for a concluding section, noting that the software described can handle a number of more complex designs.

This article is organized as follows. In the next section, we consider matching in the context of group-randomized designs, addressing Question 1 above. In the following section, we turn to covariance adjustment (Question 2). Next, we compare matching and covariance

adjustment (Question 3). A concluding section summarizes the results, provides guidance for planning new studies, and briefly explores more complex designs.

Matching Versus Not Matching

The potential benefits associated with matching in cases in which individual participants are

8

randomized to treatment groups have been a topic of methodological research for many decades (Cochran, 1957; Federer, 1955; Fisher, 1926, 1936, 1949; Kempthorne, 1952). Excellent summaries can be found in classic texts on experimental design (see, e.g., Kirk, 1982). The basic principle of blocking (of which matching is a special case) was elucidated by Fisher (1926) in the context of agricultural research. He criticized a simple experiment in which plots of land were assigned at random to receive one of five varieties of seed to assess the causal effect of such varieties on plant growth:

> On most land, however, we shall obtain a smaller standard error, and consequently a more valuable experiment, if we proceed otherwise. The land is divided first into seven blocks, which, for the present purpose, should be as compact as possible; each of these blocks is divided into five plots, and these are assigned in this case to five varieties, independently, and wholly at random. If this is done, those components of soil heterogeneity which produce differences in fertility between plots of the same block will be completely randomized, while those components which produce differences in fertility between blocks will be completely eliminated. (p. 509)

Blocking is not always helpful, however. If little variation lies between blocks, blocking can actually reduce power because it entails a loss of degrees of freedom. Even when blocking does little to boost power, it can be useful because of a potential increase in the "face validity" of an experiment. We discuss these issues below in the context of person-randomized trials before turning to the main focus of this section: matching in group-randomized trials.

Loss of Degrees of Freedom

In a *completely randomized* (CR) design, individual participants are randomized to treatment groups, and the treatment group means are compared on an outcome variable. There is no attempt to use prior information to improve the efficiency of the design. Because only two quantities (the two group means) need to be computed to obtain an ES, the available degrees of freedom are $M - 2$, where M is the total number of participants.

In contrast, suppose that prior to randomization, the individuals are rank ordered on the basis of some information X collected beforehand. Next, the experimenters match the individuals such that the first two in the ranking constitute Pair 1, the next two constitute Pair 2, and so on. Within each pair, individuals are then assigned at random to treatment and control conditions, the treatment conditions are implemented, and an estimate of the treatment effect is obtained from each pair. In this case, prior information on the individuals is explicitly embedded into the design. The general name for such a design is a *randomized block design*, and in the special case in which each block contains two units, it is called a *matched-pairs* (MP) design. The number of quantities computed here is $(M/2) + 1$ (a treatment effect for each pair and an average treatment effect), and the available degrees of freedom are then $M - [(M/2) + 1] = (M/2) - 1$, half those in the CR design. When the overall sample size M is small, losing degrees of freedom can substantially increase the critical value of the t test. So if matching is ineffective, this loss of degrees of freedom will reduce power. This penalty becomes negligible as M increases.

Improving Face Validity

The foregoing text may suggest that if the sample size is small, matching is not a good strategy. However, this is not necessarily true. A benefit of matching other than its potential contribution to power involves *face validity*. For example, an experimenter may find pairs of individuals who have the same gender and ethnicity and who are also close on X. By matching also on gender and ethnicity, the experimenter ensures that after randomization, the two treatment groups will be identical with respect to their gender and ethnic compositions. In contrast, the CR design allows chance differences between the two treatment groups on these socially and perhaps politically salient variables. Such chance differences are unlikely to be large if M is large, but when M is modest or small, embarrassing chance differences in gender and ethnic composition between experimental participants and controls might arise, producing a comparison between groups that look different on salient variables.
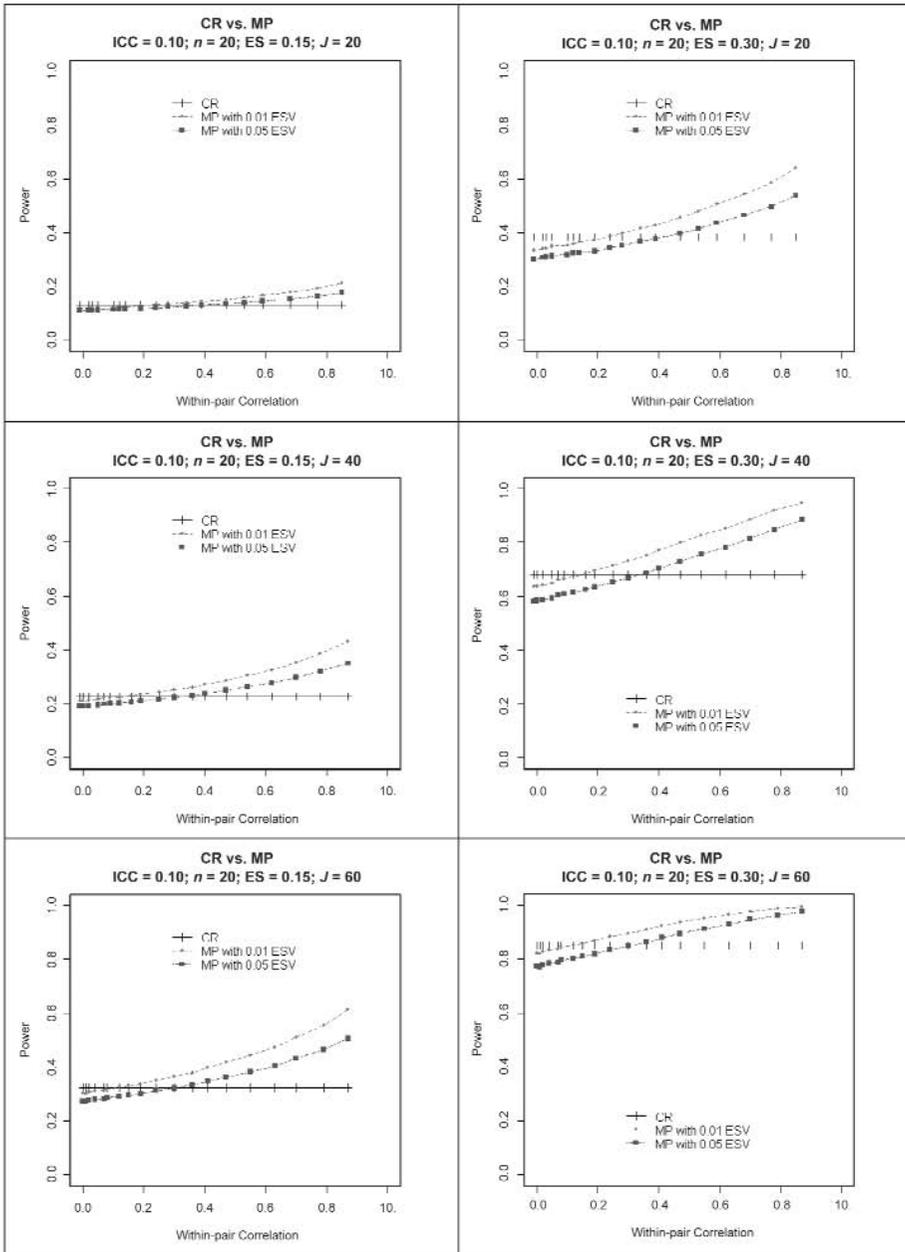
9

FIGURE 2. *Completely randomized (CR) design versus matched-pairs (MP) design, intracluster correlation (ICC) = .10.*
Note. ES = effect size; ESV = ES variability.

Matching in Group-Randomized Studies

Many of the key findings in the classic literature on person-randomized studies can be extended to designs in which whole groups (e.g., classrooms) are the unit of randomization. In the context of group-randomized studies, the unit of the outcome variable (the individual) is nested within the unit of randomization (the group). Because of this nesting, the calculation of the power to detect treatment effects is slightly more complicated. To clarify how the MP design compares with the CR design in the context of group-randomized studies,
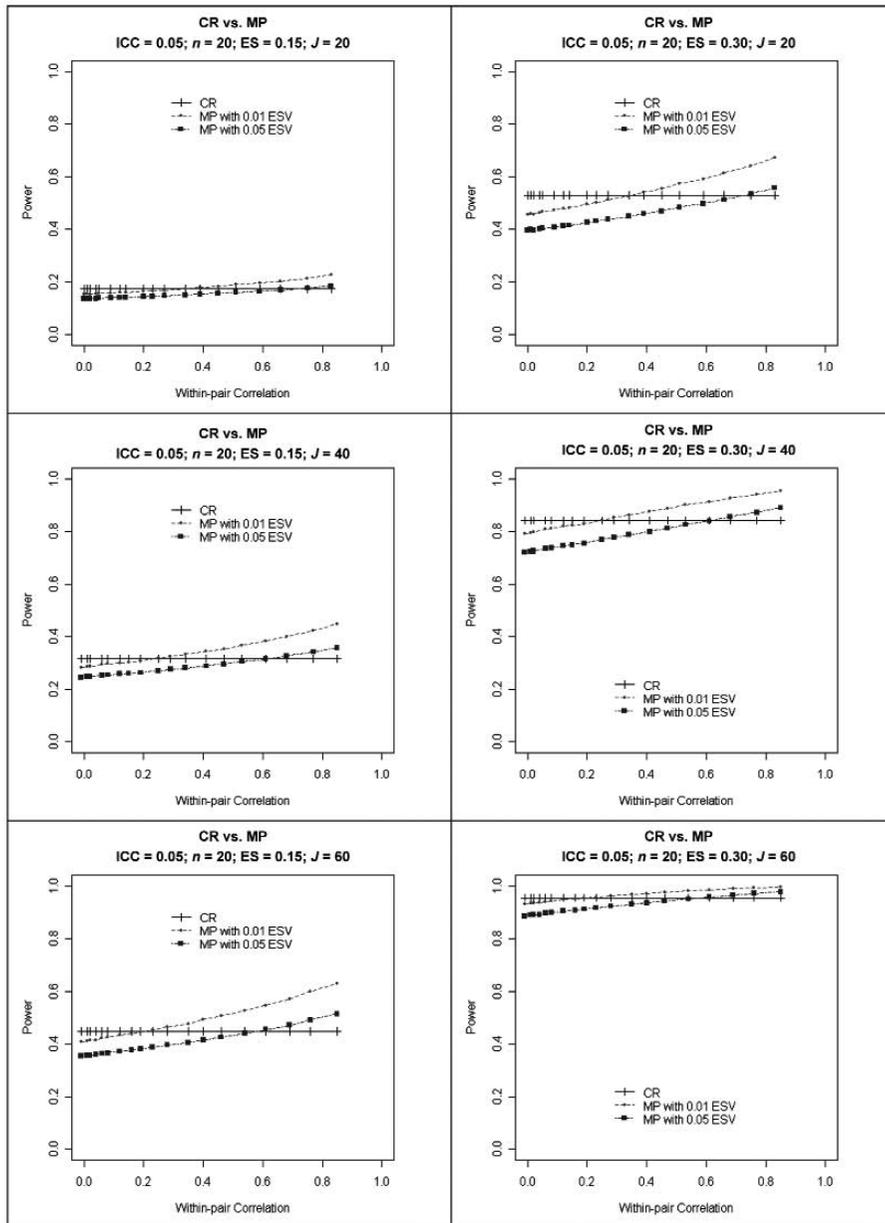
10

FIGURE 3.   *Completely randomized (CR) design versus matched-pairs (MP) design, intracluster correlation (ICC) = .05.*
Note. ES = effect size; ESV = ES variability.

consider a simple setting with two treatment conditions.

Example

Suppose a researcher is planning to assess the effectiveness of a new schoolwide literacy program (the treatment condition) to be implemented in a number of schools in a given large city. A single classroom containing 20 students will provide data from each school. The schools under consideration vary considerably in mean achievement and are quite segregated ethnically: about two thirds of the schools serve predominately minority students, while the remaining schools enroll mainly non-Hispanic White children. The researcher wonders how many schools will be needed to achieve power to detect a given

11

standardized ES and whether matching schools prior to randomization might help achieve this goal.

Under the CR design, intact schools, rather than isolated students, are assigned at random to treatment and control conditions. The literacy programs are administered, and students are then assessed on an outcome variable Y, the score on a standardized reading test. The comparison of outcomes needs to take into account the nesting of the students within the schools.

Similarly, under the MP group-randomized design, researchers first locate pairs of schools, rather than of individuals, that are likely to be similar on their mean reading outcomes. Researchers then rank order the schools on the basis of expected achievement and create the pairs: the first two schools constitute Pair 1, the next two schools constitute Pair 2, and so on. Within each pair, one school is assigned at random to the new literacy program condition while the other is kept as a control. The programs are then implemented, and students are assessed on Y. Again, the comparison of outcomes must take into account the nesting of the individuals within the groups and of the groups within the pairs.

The question that naturally arises is whether the MP design has substantially more power than the CR design. Intuitively, the answer will be yes if X (the matching variable) strongly predicts Y (reading outcomes). If so, schools within any given pair, which are by design very similar on X, will also tend to be very similar on Y in the absence of the treatment effect. In the extreme case in which matches on X within pairs are perfect and in which X perfectly predicts Y in the absence of treatment, the difference between pair members on Y will perfectly reflect the impact of the new literacy program within that pair, and the average of these differences will give an excellent estimate of the population average effect of the new program.

Generalization

The power associated with the CR design depends on five factors:
1. the size of the true average ES, in units of the outcome standard deviation;
2. the fraction of the variance in the outcome that lies between schools, equivalent to the intracluster correlation (ICC);

3. the number of schools studied, denoted J, with $J/2$ schools assigned at random to the experimental condition and $J/2$ schools assigned at random to the control condition;
4. the number of students studied per school, denoted by n; and
5. the adopted level of statistical significance, α.

The power associated with the MP design depends on these five factors plus two additional factors:

6. the correlation between latent true outcome means of schools within pairs, ρ_{pairs}; and
7. the magnitude of variation in the treatment effect across pairs.

Factor 6, the within-pair correlation, is also equivalent to the proportion of variance in the latent mean outcomes that lies between pairs. The fact that a correlation is equivalent to a variance ratio may seem odd, but that is easily shown to be the case. Factor 7 is referred to as the ES variability (ESV). The ESV quantifies how much the treatment effect varies across pairs. For example, the school differences in treatment implementation may result in school differences in the intervention effect. This variation in the treatment effect is captured in the design by the ESV.

Within the past 15 years, a number of authors have compared the power of the CR design with that of the MP design in the context of group-randomized studies (Freedman, Green, & Byar, 1990; Gail, Mark, Carroll, Green, & Pee, 1996; Martin et al., 1993). Freedman et al. (1990) examined how the MP design increased the precision of the estimate of the treatment effect relative to the CR design, restricting their attention to a comparison of the standard errors of the treatment effect (as noted by Martin et al., 1993). This approach, although useful, does not incorporate the effect on power of the loss of degrees of freedom using the MP design. Martin et al. (1993) then considered the power of CR and MP designs for small group-randomized studies, in which the loss of degrees of freedom would likely matter. They concluded that if there were fewer than 20 groups, matching should be used only if the correlation between the matching variable and the outcome is greater than about .45. Otherwise,

12

the loss of degrees of freedom in the MP design would result in a less powerful design than the CR design (see also Hughes, 2005). Bloom (2005) provided calculations showing the required predictive power of matching to increase the power of the test relative to a CR design for varying numbers of groups in the context of social studies.

Making the comparison precise

Figure 1 compares the power of CR and MP designs. In both cases, ES = {0.15, 0.30}, J = {20, 40, 60}, and ICC = .20, with n = 20 and α = .05. This ICC is commonly regarded as close to the higher end of what is typically found for achievement outcomes (Shochet, 2005).

Power in the MP design will also depend on how much the ES varies from pair to pair. Such ESV is allowed to be 0.01 and 0.05. To clarify, suppose ES = 0.15 and ESV = 0.01. If the pair-specific ESs vary randomly across pairs according to the normal distribution, one would expect 95% of the pairs to yield ESs in the range of $0.15 \pm 1.96\sqrt{0.01}$ ES units, that is, over the interval (–0.05, 0.35). An ESV larger than 0.01 would yield quite large plausible value intervals, perhaps larger than is realistic, although 0.05 is considered an upper bound.

Looking at the upper left panel of Figure 1, when J = 20 (10 matched pairs) and ES = 0.15, power for the CR design is extremely low, at about 0.10. Power for the MP design is not much better, even when matching explains 90% of the outcome variation. When ES = 0.30 (upper right panel), things are not much better. The CR design yields about 0.25 power, and the MP design does not help much unless it explains a large fraction of the variation. Still, even if 90% of the variation is explained, power is unacceptable, at about 0.60. It is clear that the ES will need to be quite large if such a small study (J = 20) is to yield adequate power. Notice that the utility of matching also depends on the ESV, but this dependence is weak. Indeed, the curves for ESV = 0.01 and ESV = 0.05 are virtually indistinguishable despite the fact that ESV = 0.05 is very large indeed. Across all the scenarios shown in the figure, the smaller ESV of 0.01 gives slightly greater power than the larger ESV of 0.05.

Increasing the number of schools will naturally add power to the study. Consider the center right panel of Figure 1, in which ES = 0.30 and J = 40 (20 matched pairs). Power for the CR design is about 0.47. The utility of matching depends on the percentage of outcome variation explained. Until about 25% of the outcome variation is explained, matching actually makes things slightly worse. But matching increasingly helps as the percentage of variance explained increases toward 1.0 and can even lead to a well-powered study if it successfully explains about 80% of the variation in the outcome. Looking now at the bottom pair of panels, the benefit of increasing J to 60, especially for the larger ES, can be seen. The benefits of matching are again clear. When ES = 0.30, power increases beyond 0.80 if about 60% of the variance in school means is explained.

Recall that Figure 1 portrays a "bad case" in that the ICC is .20. When the ICC is large, there is considerable variation between schools, meaning that the CR design requires many schools to achieve adequate power for a given ES. Matching reduces the need to have so many schools by accounting for some of the large variation between schools. However, matching may actually hurt unless ρ_{pairs}, the percentage of variance explained between schools, achieves a minimum level. The minimum level is about 25% when J = 20 and declines to about 15% when J = 60.

Figure 2 gives a somewhat more optimistic scenario because now the ICC is .10 rather than .20. This means that about 10% of the overall variation in outcomes lies between schools. Power is uniformly higher than in Figure 1, holding constant J and ES. However, the benefit from matching is smaller in Figure 2 than in Figure 1. Now, ρ_{pairs}, the percentage of variance explained by matching, must be higher before matching begins to help. And the benefit of matching after reaching "the threshold" is smaller than in Figure 1.

These principles are even clearer when Figure 3 is examined, in which an ICC of only .05 is assumed. Now, the CR design achieves nearly respectable power when ES = 0.30 and J = 40 and quite respectable power when J = 60. Notice that now, there is little to gain from matching because little of the variation lies between schools, and matching schools can
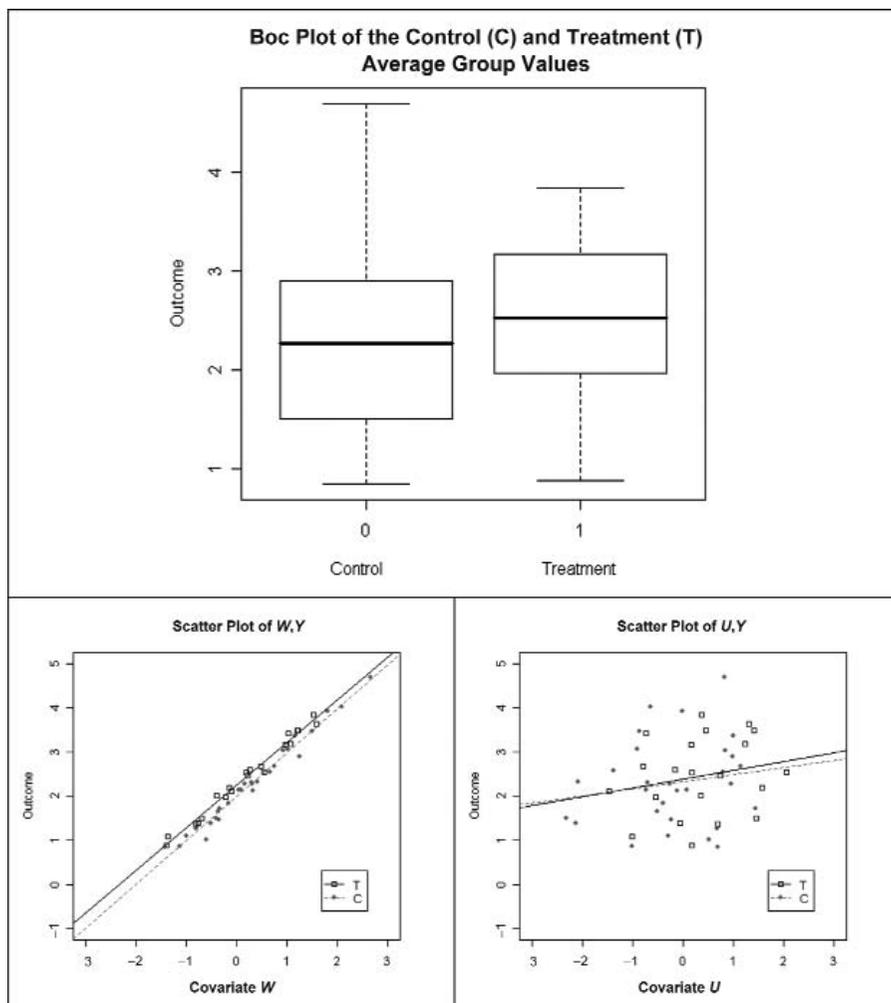
13

FIGURE 4.   *Analysis of covariance example for two-level design.*

help only in reducing that small fraction of variation further.

The implications of this exercise are clear. Matching helps most when between-school variation is large and therefore a big problem for the CR design. The CR design then requires a very large number of schools to "cut down the noise" caused by random variation in school means within treatments. Matching helps by explaining some or most of this variation, so that fewer schools are required under matching than under the CR design, as long as the percentage of variation in school means explained by matching achieves a threshold. This threshold is higher when few schools are sampled and goes up further as the ICC declines. When little

variation lies between schools in the absence of matching, matching will not help much, because there is little variation for matching to explain. Indeed, matching is more likely to hurt when the ICC is small than when the ICC is large.

Using a Covariate Versus Not Using a Covariate

Matching prior to randomization is one way to increase power. A second major strategy for increasing the power of an experiment is the use of ANCOVA. The efficacy of ANCOVA in the context of group-randomized studies has been discussed by several authors (see, e.g., Bloom,
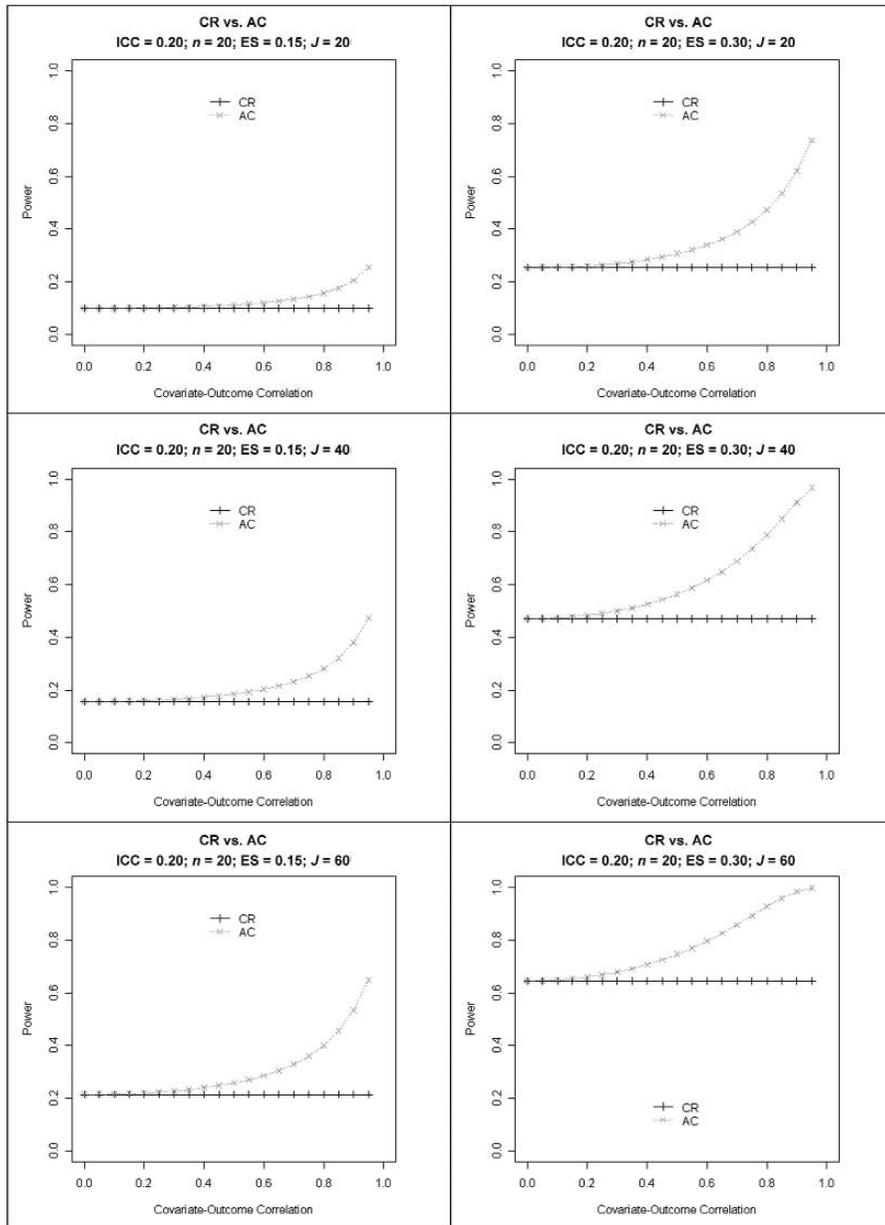
14

FIGURE 5.   *Completely randomized (CR) design versus analysis of covariance (AC), intracluster correlation (ICC) = .20.*
Note. ES = effect size.

2005). A question arising in this literature is whether the covariate should be a person-level characteristic or a group characteristic. According to Bloom (2005), correlations at the group level in the social sciences are typically higher than correlations at the individual level. In addition, group-level data are often more accessible and may be less expensive to acquire. For example,

consider a study designed to test the effect of a reading intervention in schools in which reading achievement is the outcome and all third grade classrooms within a school are assigned to the same treatment condition. A useful covariate would be last year's third grade scores on the reading test. Finding last year's average third grade scores for each school will typically be
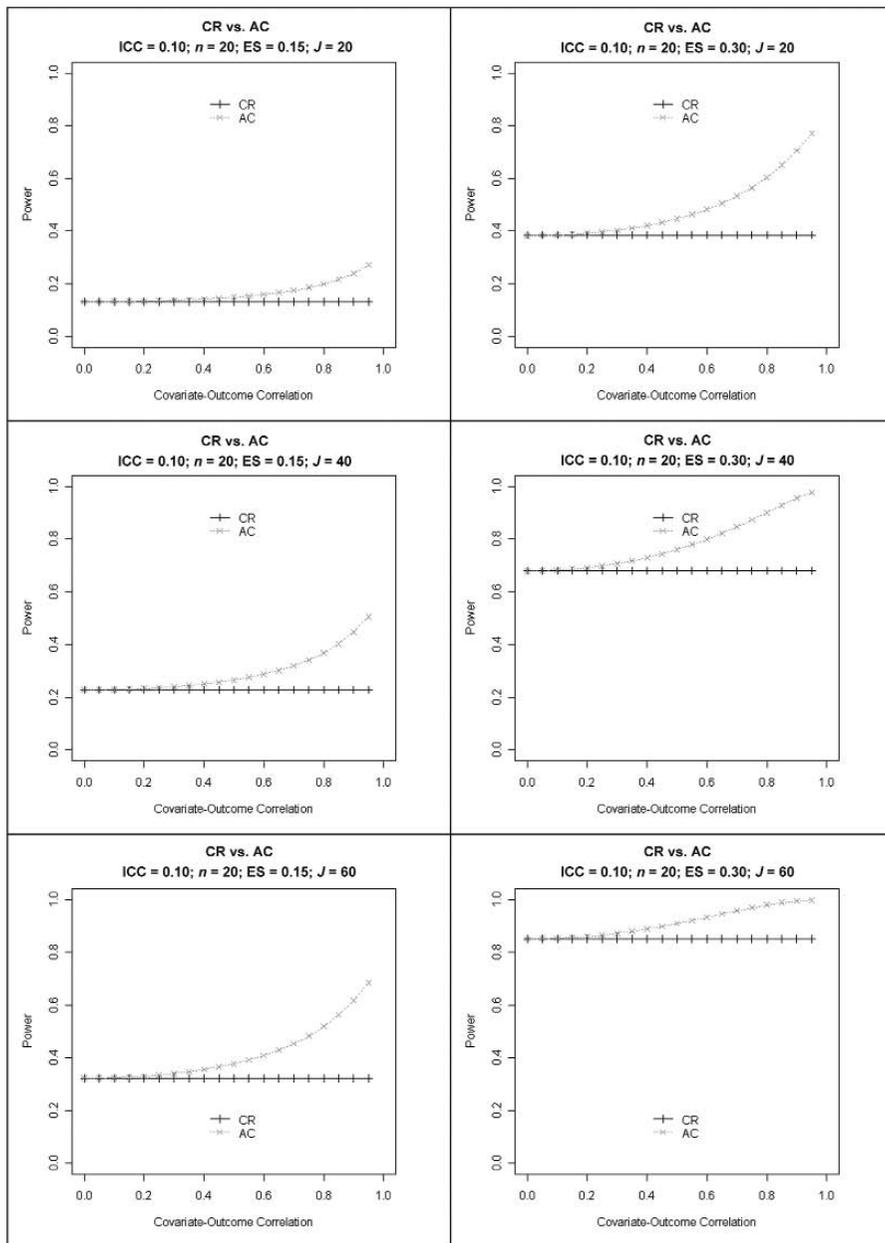
15

FIGURE 6. *Completely randomized (CR) design versus analysis of covariance (AC), intracluster correlation (ICC) = .10.*
Note. ES = effect size.

inexpensive. Examples of this type arise frequently, so we focus here on covariance adjustment for group-level covariates.

Example

Consider once again a study in which schools are to be assigned at random either to an experimental group that receives a new literacy program or to a control group that does not. On completion of the program, the two groups are compared on their average values of *Y*, the school-level sample mean reading test score. Let μ denote the "true" school mean, of which *Y* is an estimate.

ANCOVA modifies the script in that the investigators decide to exploit the availability of information on *W*, the prior mean achievement of
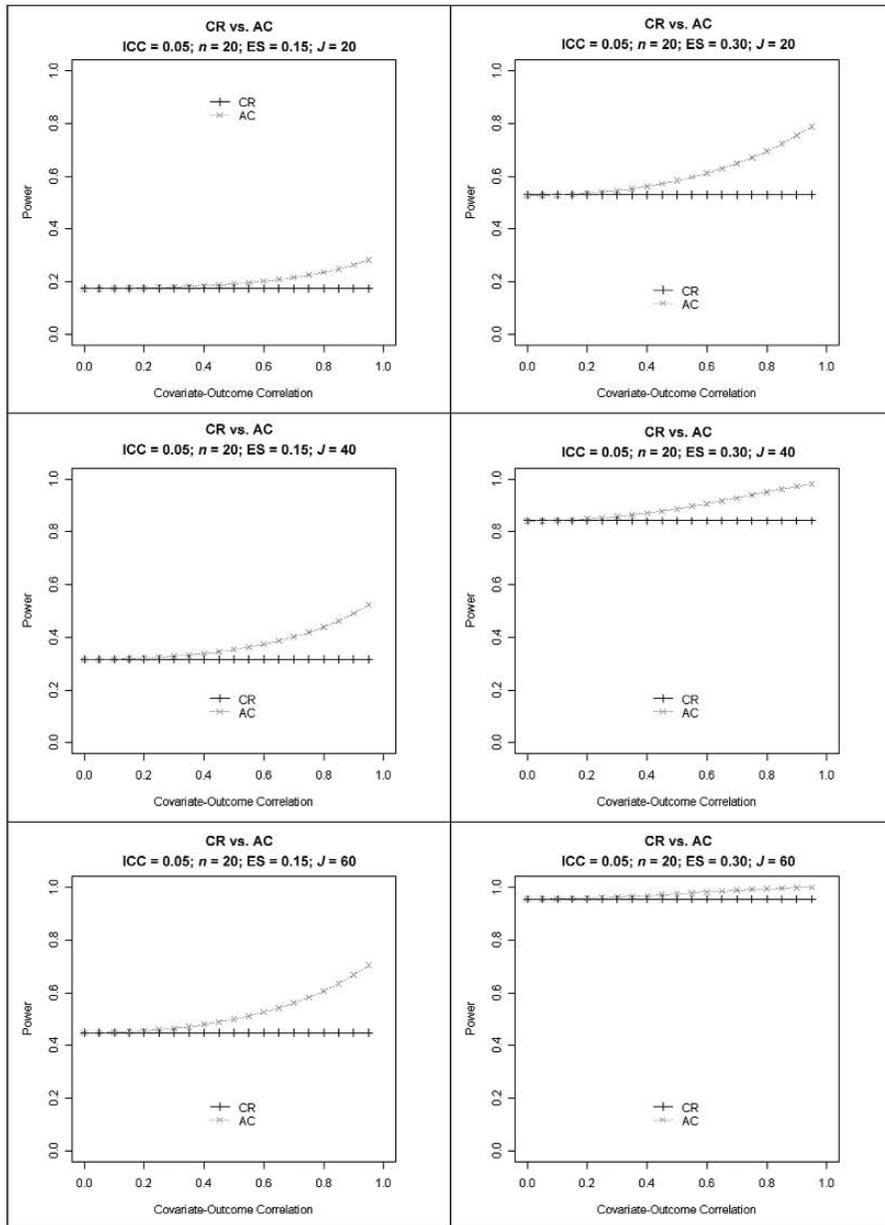
16

FIGURE 7. *Completely randomized (CR) design versus analysis of covariance (AC), intracluster correlation (ICC) = .05.*
Note. ES = effect size.

school J. Many studies have found a strong linear association between school-level mean achievement and school-level mean outcome, so such a W is an excellent candidate as a covariate.

As in matching, the experimenter uses ANCOVA to exploit the existence of prior information about schools, information that hopefully is helpful in predicting school-level mean outcomes. One key difference is that although matching builds this prior information into the design of the study (by first matching and then randomizing within pairs), ANCOVA uses a statistical model to "hold constant" the prior variables when evaluating the impact of the treatment on the outcome. In fact, operationally, the ANCOVA design is identical to the CR

17

design: Persons are simply randomly assigned to either the experimental or the control group with no regard for any prior information. However, the analytic model is different. Specifically, the outcome, Y, is regarded as a linear function of two things: the covariate, W, and treatment group membership, indicated by a dummy variable, Z. The aim is to predict how two schools with the same value of W would differ if one received the experimental treatment ($Z = 1$) and one received the control treatment ($Z = 0$). Because treatments are assigned at random, only the treatment effect plus chance differences (other than the covariate) contribute to the predicted difference between two such schools.

Thus, in the scenario described above, using linear regression, the experimenter uses W and treatment group membership Z to predict Y. In this scenario, the expected mean difference between the two groups, holding constant W, is the average impact of the treatment in the population from which the schools were sampled.

Intuitively, ANCOVA will boost power when W (the prior mean achievement of school J) has a strong linear association with Y (the school-level mean reading test score). If so, two schools that have very similar W values but experience different treatments will have very similar predicted Y values in the absence of a treatment effect. Therefore, if, on average, schools that are similar on W but vary with respect to treatment also vary systematically on Y, one will tend to find evidence of a treatment impact.

To illustrate this idea, data were generated in which W strongly predicts Y. To illustrate the logic of ANCOVA, assume an unrealistically high correlation of $\rho_{w\mu} = .95$ between W and the true school mean, μ. For ease of interpretation, test scores are measured in standard deviation units. Schools receiving the new program receive a boost of 0.25 standard deviation units in average scores. Under these assumptions, W and Y are bivariate normal in distribution, each with a variance of 1.0.[2] The sample size per school, n, is set at 100; the number of schools, J, is set at 50; and the ICC is set at .10. Figure 4 shows box plots of the experimental and control groups' mean Y values. The median Y is a bit higher in the experimental group ($Z = 1$) than in the control group ($Z = 0$), but the two groups' distributions overlap considerably. Under the

CR analysis, the t test of the mean difference between groups on mean Y taking the nesting into account results in $t = 0.60$, far from being significant at the 5% level.

Figure 4 gives a scatterplot of Y against W for the two treatment groups. The plot shows an extremely strong, positive, linear association between W and Y (note the sample prediction line within each of the two groups). Note that at nearly all values of W, the experimental group Y values tend to be elevated above those of the control group. Thus, when attention is restricted to schools that have similar values of W, it can be readily discerned that the treatment and control group means are different. The estimate of the average impact of the treatment is the vertical distance between the two nearly parallel lines in the bottom left panel of Figure 4. The test of the average treatment effect under the nested ANCOVA is positive and statistically highly significant ($t = 7.01$, $p < .001$). In effect, controlling for the covariate W has dramatically reduced the "noise" in Y, revealing the "signal," that is, the impact of the treatment.

Generalization

As in the case of matching, controlling for a useless covariate is of no help at all; in fact, it actually hurts, again because of the loss of degrees of freedom. Recall that the degrees of freedom for the CR analysis equals $J - 2$. ANCOVA requires the computation of one additional quantity: the association between the covariate and the outcome. So in this case, $df = J - 3$. The loss of one extra degree of freedom will have a negligible effect on power unless the sample size is very small. Thus, ANCOVA exacts a smaller penalty than the MP design for using useless information on W.

To see how the data look when a useless covariate is used, refer to the bottom right panel of Figure 4. Values of this covariate, labeled U, increase along the horizontal axis, with Y on the vertical axis. Confining attention to cases with similar values of U in no way reduces uncertainty about Y and therefore is of no help in discerning the treatment effect. In this case, ANCOVA, like the CR analysis, gives a nonsignificant test of the impact of the treatment ($t = 0.25$).
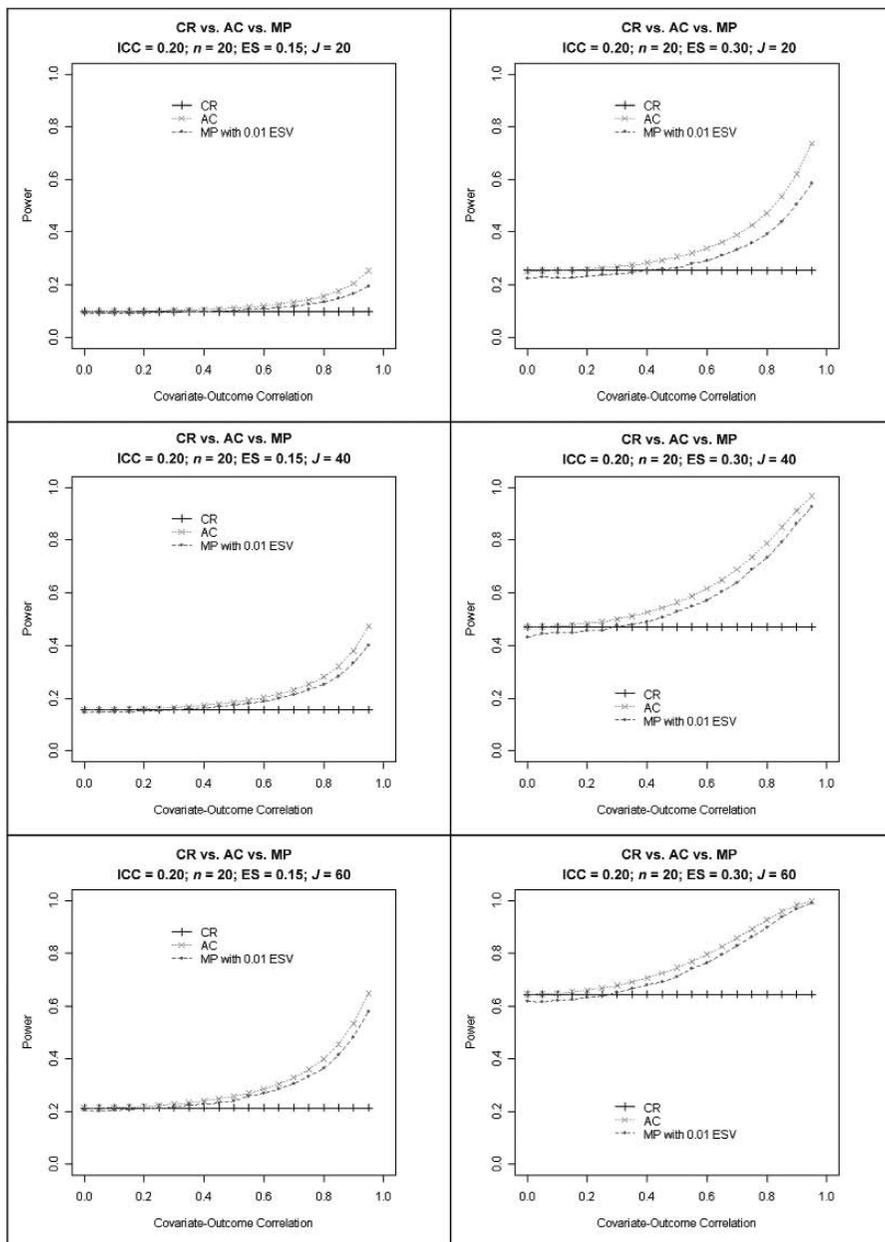
18

FIGURE 8. *Analysis of covariance (AC) versus the matched-pairs (MP) design, intracluster correlation (ICC) = .20.*

Note. CR = completely randomized; ES = effect size; ESV = ES variability.

The key assumptions underlying ANCOVA are that

1. power can increase only to the extent that *X* and *Y* are linearly associated,
2. the magnitude of this association must be the same in the experimental and control groups, and

3. the residuals (the errors of prediction) must be normally distributed with constant variance.

Assumptions 1 and 2 are not needed in using the MP design. Thus, the assumptions for ANCOVA are more restrictive than those needed for the MP analysis, making the latter more flexible.

19

As noted in the matching discussion, the power associated with the CR design depends on five factors:

1. the true ES;
2. the ICC;
3. the total number of schools, *J*;
4. the number of students studied per school, *n*; and
5. the adopted level of statistical significance, α.

The power in the ANCOVA design depends on one additional factor:

6. $\rho_{w\mu}$, the correlation between the school-level covariate and the true school-level mean outcome.

Making the Comparison Precise

Suppose an experimenter wonders how many schools will be needed to achieve power to detect a predetermined ES and whether ANCOVA might help achieve this goal. The results are graphed in Figures 5–7, following the same logic of the graphs presented in the previous section on matching. Thus, Figure 5 gives results when the ICC (unadjusted for the covariate) is .20, Figure 6 gives results for ICC = .10, and Figure 7 gives results for ICC = .05. In each case, results are seen for a "small" ES of 0.15 and for a larger ES of 0.30 using *J* = 20, *J* = 40, and *J* = 60 and holding *n* constant at 20 and α at .05. In every case, the results are reproduced for the CR design (i.e., the design that does not use the covariate). Power is shown as a function of the magnitude of the correlation between the school-level covariate and the school mean.

In many ways, the results parallel those of the preceding section. Thus, a smaller benefit of using the covariate is seen when the ICC is .10 than when it is .20, and even less benefit is seen for ICC = .05. Once again, the school-level covariate can explain only the variation between schools, and when the ICC is very small, there is little between-school variation to be explained.

One obvious and important difference involves the potential negative effect of using the covariate. Recall that matching can actually undermine power unless the percentage of variance explained by matching reaches a threshold value. In contrast, Figures 5–7 suggest that the penalty for using a useless covariate is negligible in all cases. The reason is clear: Only one degree of freedom is sacrificed in using the covariate, compared with *K* – 1 (where *K* is the number of pairs) in the MP group-randomized design.

Matching Versus ANCOVA

We have explored how two approaches to using prior information can improve power compared with an approach that does not make use of such prior information. As long as there is substantial variation between schools to explain, matching will substantially improve power as the correlation within pairs on the outcome increases. And ANCOVA will substantially improve power when the correlation between the covariate and the outcome increases. The question then naturally arises: Which of these two approaches is most effective in increasing power?

A comparison between ANCOVA and matching makes sense only when ANCOVA is a reasonable approach, that is, when the assumptions required for ANCOVA are at least approximately correct. Yet if the ANCOVA assumptions do hold, ANCOVA will be optimal. However, there are benefits of matching. Recall that matching can ensure balance between treatments on key salient variables, increasing face validity. Moreover, the weaker assumptions required for matching make it appealing.

The question therefore arises as to whether matching might be nearly as good as ANCOVA when the ANCOVA assumptions hold. If so, one might argue that matching ought to be adopted to purchase its unique benefits. The available literature does not provide a conclusive answer to this question.

To date, researchers evaluating matching techniques have tended to overestimate the utility of matching (Hughes, 2005; Klar & Donner, 1998). In this literature, it is common to postulate the existence of a continuous pretreatment variable, say *W*, having a correlation, say $\rho_{w\mu}$, with the group mean outcome μ. One might then assume that matching on *W* will reduce the unexplained variance in the outcome by a factor of $\rho_{w\mu}^2$. Such
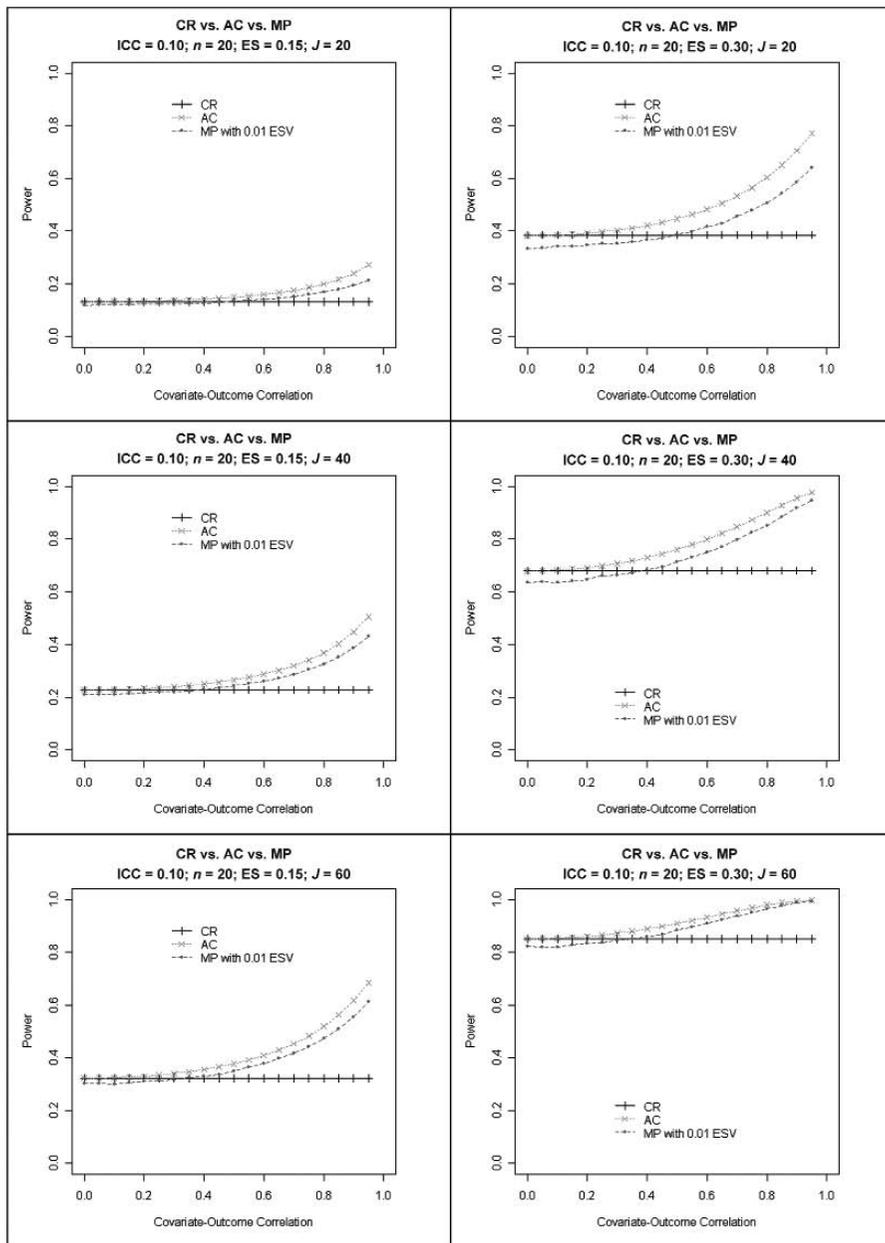
FIGURE 9. *Analysis of covariance (AC) versus the matched-pairs (MP) design, intracluster correlation (ICC) = .10.*

Note. CR = completely randomized; ES = effect size; ESV = ES variability.

an approach overestimates the benefit of matching by overestimating the similarity of units within pairs on *W*, in effect assuming that pair members will be perfectly matched on *W*. Such an assumption is unrealistic in practice, yet there is no closed-form mathematical expression for determining the variance explained by matching on *W* when *W* randomly varies within pairs. We

address this problem through simulation, allowing for a more rigorous comparison between matching and ANCOVA.

First, a group-level covariate *W* and, for each group, a latent "true" group mean outcome, μ, were generated as bivariate normal in distribution, each with variance 1.0 and correlation $\rho_{w\mu}$. Then, cases were rank ordered on *W*, and the
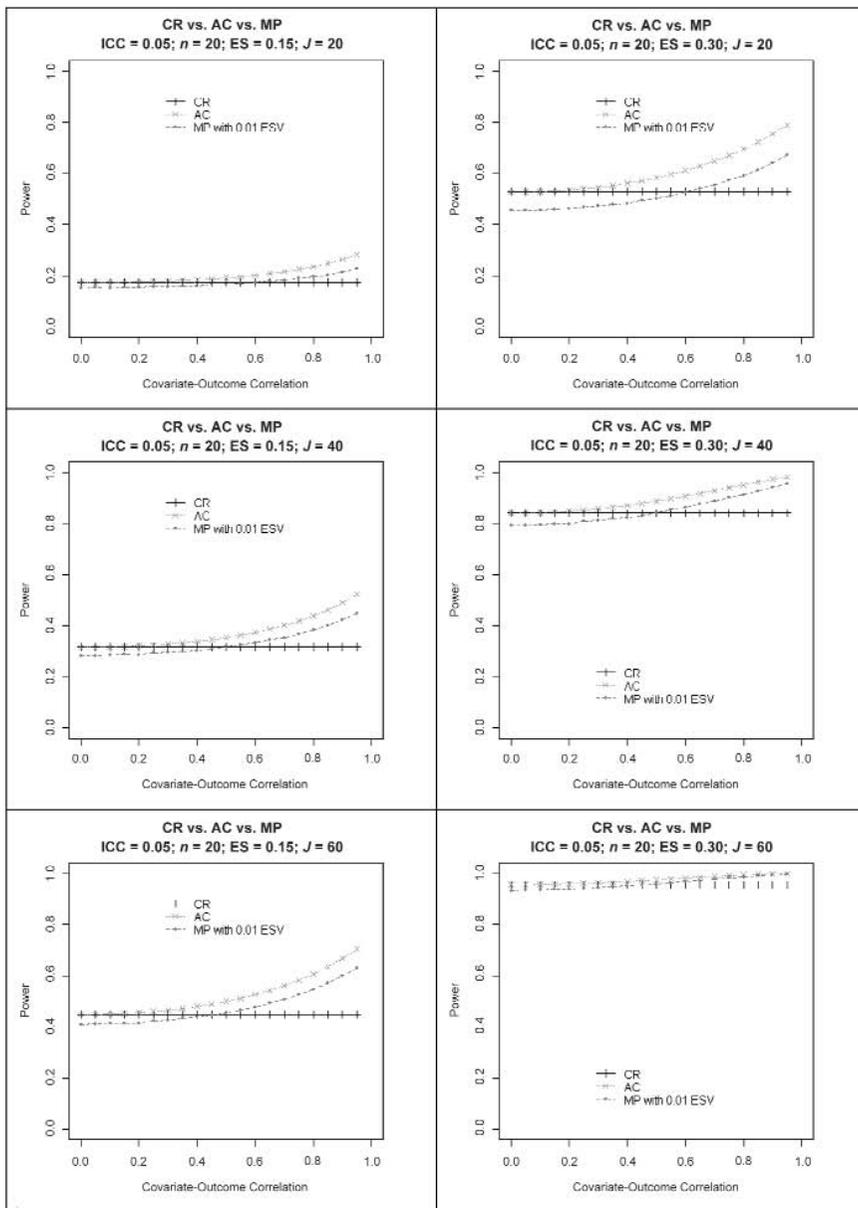
21

FIGURE 10. *Analysis of covariance (AC) versus the matched-pairs (MP) design, intracluster correlation (ICC) = .05.*

Note. CR = completely randomized; ES = effect size; ESV = ES variability.

sample of groups was divided into pairs such that the first- and second-ranked groups constituted Pair 1, the third and fourth constituted Pair 2, and so on. The variances within and between pairs on μ were then computed, enabling the deduction of the correlation between pair members, ρ_{pairs}. To ensure the stability of this estimate, the process was repeated 1,000 times for each

sample size, and the average of the estimates was used as the final value of ρ_{pairs}. Thus, for every sample size, it was possible to regard ρ_{pairs} as a function of $\rho_{w\mu}$, enabling the computation of power for the ANCOVA and the MP analysis at each value of $\rho_{w\mu}$. Power for the MP design can readily be shown to be a function of the number of groups, J; the sample size per group, n; the

22

ICC (the percentage of variance lying between groups in the absence of matching); and ρ_{pairs}. Here, $J = 2K$, where K is the number of pairs.

The results are graphed in Figures 8–10. These are the same results shown in Figures 5–7, except that the power for the MP design at each value of $\rho_{w\mu}$ has been added. For simplicity, only the case in which ESV = 0.01, a plausible value of ESV, it was reasoned, was included.

As expected, the ANCOVA approach works best. In every case, power associated with matching increases at a rate nearly equal to that of ANCOVA, but always at a lower level. The two are most similar when J is large and the ICC is large. A key difference is that when $\rho_{w\mu}$ is small, matching can do worse than the CR design, whereas the penalty for a small correlation with ANCOVA is negligible. As seen earlier, there is virtually no penalty for using ANCOVA when X is a useless covariate, unless J is very small, smaller than studied here. In contrast, matching can reduce power, particularly when ineffective, when the number of pairs is small and when the ICC is small.

Discussion

The logic of social experimentation often requires that groups, rather than individuals, be the unit of assignment and of treatment. Assignment by randomization confers the same advantages in group-based studies as in person-based studies: A well-implemented randomized study eliminates bias by statistically balancing treatment conditions on all prior characteristics. In this case, the only differences between treatments on prior characteristics are chance differences, and standard significance tests and confidence intervals correctly quantify uncertainty about the existence and magnitude of causal effects.

However, if no effort is made to identify and control prior group characteristics, group-randomized studies will tend to lack statistical power, unless many groups are recruited. This need for many groups arises because the effects of short-term interventions typically implemented in social settings will tend to be modest. The number of groups required to achieve adequate power depends also on the intraclass correlation, also interpretable as the fraction of variation in group-mean outcomes that lies

between groups. Even when this fraction is modest (e.g., less than 10%), the number of groups required to achieve adequate power can be daunting if the ES is modest.

In this article, we have evaluated two alternative approaches to identifying and controlling prior group characteristics: matching and ANCOVA. Under certain circumstances, such uses of prior information can substantially increase power given the number of groups or, equivalently, reduce the number of groups needed given a desired level of power.

In this discussion, we first briefly summarize the key findings. Second, we provide advice on how to use existing data to assess the likely contribution of matching or ANCOVA to improve power. Third, we briefly explore other more complex designs and how available information can be used to increase the power to detect treatment effects.

Key Findings

Matching

It has been demonstrated that matching will tend to enhance statistical power when groups within pairs are well matched on prior characteristics and when those characteristics strongly predict outcomes. In this setting, groups within pairs will tend then to be similar on outcomes, except for the fact that one group experiences the treatment and one does not. Such similarity is measured by the correlation between pair members, ρ_{pairs}, which is equivalent to the correlation between group means on the outcome within pairs and the fraction of variation that lies between pairs on the true mean outcomes. The variation in the true ES across pairs also affects power; power will be greater when this variation is small.

A key factor to consider, however, is the fraction of variation that lies between groups in the absence of matching. This factor is indexed by the ICC. Matching is most helpful when the ICC is comparatively large. If the ICC is tiny, there is little variation between groups to be removed through matching. Thus, the threshold value of ρ_{pairs} decreases with the number of pairs and also with the ICC. Thus, matching will help least, and may in fact hurt, when the number of pairs is small and when the ICC is small. We found, in fact, that in quite plausible

23

circumstances, matching actually hurt when the ICC was .05, even if ρ_{pairs} was moderate to large.

ANCOVA

In group-randomized studies, the most promising covariates are often group-level covariates, *W* (Bloom, 2005). They are typically cheap to measure and often very strongly correlated with the group mean outcome, labeled μ in this study. As one might expect, power tends to increase with the covariate-outcome correlation $\rho_{w\mu}$. However, the benefit of using the covariate depends on the ICC. When the ICC is large, the use of a group-level covariate can substantially increase power. However, for a small ICC, the utility of the covariate will be small. Only if the number of groups is exceedingly small, however, does one pay a non-negligible penalty for using a poorly chosen covariate. However, the required assumptions are more stringent than when matching on pretreatment characteristics.

Matching versus ANCOVA

Its flexibility gives matching certain advantages: There is no requirement that the pretreatment characteristic has a linear association with the outcome to be used. Moreover, as noted, matching on one or more salient variables can enhance face validity. However, when the assumptions of ANCOVA hold and when the correlation is large, it is difficult to beat ANCOVA for boosting power. A question that naturally arises is whether matching on the covariate *X* can provide nearly as much power as ANCOVA when the assumptions of ANCOVA hold. If so, one might opt for matching to enjoy its other benefits.

We were surprised to find little precise guidance in the literature on this question. On reflection, this finding is understandable in that precise mathematical comparisons are difficult if possible at all. We therefore conducted a simulation study. In the simulation study, matching approximated the power of ANCOVA as the number of groups increased and as the ICC increased. However, for many plausible cases, ANCOVA did significantly better than matching, because in these scenarios, either the number of groups or the ICC was too small to enable

matching to approximate the power of ANCOVA (see Figures 8–10).

Match and covary? Given the trade-offs, is it possible to benefit from matching and use *X* as a covariate? Although the precise study of such an option goes beyond the scope of this article, the results are suggestive. On one hand, adding a useless covariate to an MP study will cause little harm unless the number of groups is very small. On the other hand, if the covariate is powerfully related to the outcome, the benefit may be great, particularly if matching was not effective. Thus, it is quite plausible to envision a scenario in which matching is desirable to enhance face validity but adds little power, or even reduces it. Adding a strong covariate may then help. On the other hand, if a good covariate is already available, matching in addition to ANCOVA would seem unpromising in most cases of interest for the purpose of increasing power. In studies with small numbers of groups, matching in addition to ANCOVA may hurt because the loss of degrees of freedom associated with matching increases the critical value of the test statistic for the treatment effect.

Using These Ideas to Plan Group-Randomized Studies

Our hope is that this article will improve the planning of group-randomized studies. Good planning typically requires reasonable estimates of quantities that can be known precisely only after a study has been conducted. In many cases, pilot data or data from archives can be analyzed to provide good estimates of these quantities. If those estimates become available, the reader might use the information presented in the figures to guide planning. These figures cover only a small fraction of the cases that arise in practice, however. Details of how to replicate the results in the figures using the Optimal Design software package are in Appendix A. We now consider how to use the software (Raudenbush et al., 2006) to plan studies in cases not represented in those figures. The discussion is restricted to the case of continuous outcomes. However, the software documentation shows how to use the software

for dichotomous outcomes and presents all equations needed to derive the results in the figures.

CR Design

In person-level studies, one may begin with an assumed standardized ES, that is, the mean difference between experimental subjects and controls in the scale of the outcome standard deviation. This is typically the "minimum ES worth detecting." One then simply calculates the sample size, M, needed to ensure a given level of power at a given level of significance. Alternatively, one may fix the ES and determine the "minimum detectable effect," that is, the smallest ES that can be detected at the given power and significance level. In some cases, prior data will be needed to get a good approximation to the outcome standard deviation. The expected mean difference in outcomes divided by this quantity then constitutes the standardized ES.

Planning group-level CR studies is a little more complicated. To obtain power, one needs not only a hypothesized ES but also an estimate of the ICC, the fraction of variation lying between groups. Prior data will be useful in this regard (see Bloom, 2005; Shochet, 2005). Power is then a function of both the number of groups, J, and the sample size per group, n.

MP Design

For person-randomized studies, the calculation of power requires the same quantity as needed in CR: the standardized ES. In addition, one needs an estimate of ρ_{pairs}, the correlation between outcomes of pair members in the absence of treatment (or, equivalently, the fraction of variance in the outcome that lies between pairs). This quantity can easily be estimated from pilot data if the matching variables and a good substitute for the outcome are available. To estimate the correlation between pairs, one simply matches cases on the matching variables and then computes a one-way analysis of variance on the outcome. A good estimate of ρ_{pairs} is then half the difference between the between- and within-pair mean squares.

In the group-randomized case, planning for MP is again a bit more complicated. One needs the same quantities as required in the CR case (the standardized ES and the ICC) plus an estimate of ρ_{pairs}. One might proceed in a way that is entirely analogous to the procedure described above for the person-level case: match groups, then compute a one-way analysis of variance using the group sample mean outcomes, computing ρ_{pairs} in the same way. However, this procedure is not correct. Such an estimate, call it $\rho_{pairs\ naive}$, will be attenuated as the sample size per group diminishes. What is needed is an estimate of ρ_{pairs} defined as the fraction of variance in "true" group means that is explained by matching (see Appendix B for details).

Analysis of Covariance

In the case of person-level studies, the computation of power will require the standardized ES and a reasonable estimate of ρ_{xy}, the covariate-outcome association. In the group-randomized case, one again needs the standardized ES, but also the ICC and an estimate of $\rho_{w\mu}$, the correlation between the group-level covariate W and the group mean outcome μ. We emphasize that $\rho_{w\mu}$ is not equivalent to the correlation between W and the sample mean outcome (see Appendix B for details).

Other (More Complex) Designs

The discussion in this article has been limited to two-level designs. However, many studies often involve additional levels of clustering. Although precise study of these other designs goes beyond the scope of this article, the findings described in this article provide some guidance on how to increase power in such cases.

One such design is the three-level group-randomized trial with treatment at level 3. In this case, there is an additional layer of clustering between the unit of analysis and the unit of randomization. Suppose a schoolwide study involving several schools is designed to determine the effects of a whole-school reform on students' academic achievement. The reform is

25

implemented at the school level, and the outcomes of interest are at the individual level. However, the effectiveness of the reform may depend on the quality of the teachers or on other classroom-level characteristics. Thus, taking into account the clustering at the level of the classroom is essential.

To increase the power of such a study, an option might be to block the schools by district prior to randomization. This blocking can reduce the between-district variation and thus potentially increase the power. Again, whether the power increases will depend on the strength of districts as a blocking variable and on the number of schools in the study, among other variables. This type of blocking, however, may be desirable for face-validity purposes. Alternatively, a school-level covariate such as school mean prior achievement might be included in the analysis. With three levels in the design, any level covariate could be used; however, the school-level covariate in this case is the most logical for reasons outlined above in the discussion of covariance adjustment.

Another (three-level) design not discussed here is the cluster randomized trial with repeated measures at the individual level. Such a design involves repeated measures nested within individuals nested within groups (Raudenbush & Liu, 2001). Yet again, the main findings about the power-increasing strategies discussed in this article can once again be extended.

All the designs illustrated in the figures plus the three-level group-randomized trial with treatment at all levels and the cluster-randomized trial with repeated measures at the individual level are included in the Optimal Design software and its documentation, available from the William T. Grant Foundation's Web site or from the authors. We encourage researchers to use these resources in planning their studies of group-based initiatives. See Appendix A for details on how to use the software.

Appendix A

All the results in the figures included in this article may be replicated using the Optimal Design for Group Randomized Trials mode of the Optimal Design software. We include here a few examples.

For a more in depth discussion, refer to the software documentation (http://www.wtgrantfoundation.org/ or http://sitemaker.umich.edu/group-based).

Example 1: Replicating the Completely Randomized (CR) Results for Intracluster Correlation (ICC) = .20, n = 20, Effect Size (ES) = 0.15, and J = 20 (Figure 1, Top Left Panel)

Select Cluster Randomized Trial → Power for the main effect of treatment (continuous outcome) → Power vs. effect size → $\rho = 0.20$, $J = 20$, $n = 20$. Clicking on the figure for an ES of 0.15 yields power of about 0.10, which is the same as the power for the CR in the top left panel in Figure 1.

Example 2: Replicating the Matched-Pairs (MP) Results for ICC = .20, n = 20, ES = 0.15, J = 60, and ESV = 0.01 (Figure 1, bottom left graph)

Select Multi-site CRT → Power for treatment effect → Power vs. effect size → $n = 20$, $J = 2$, $K = 30$, $\rho = 0.20$, $\sigma_{\delta}^2 = 0.01$, $B = 0.8$. Clicking on the plot for an effect size of 0.15 yields power of about 0.49, which approximates the power for the MP with 0.01ESV in the bottom left graph in Figure 1.

Example 3: Replicating the ANCOVA Results for ICC = .20, n = 20, ES = 0.15, and J = 60 (Figure 5, Bottom Left Panel)

Select Cluster Randomized Trial → Power for the main effect of treatment (continuous outcome) → Power vs. effect size → $\rho = 0.20$, $J = 60$, $n = 20$, $R_{L2}^2 = .64$. Clicking on the figure for an ES of 0.15 yields power of about 0.40, which is the same as the power for the ANCOVA for a correlation of .80 ($.80^2 = .64$) in the bottom left graph in Figure 5. Note that $R_{L2}^2 = .64$ is entered because it asks for the percentage of variance explained. Figure 5 is based on the correlation.

Appendix B

Estimating ρ_{pairs} for MP Designs for Group-Randomized Studies

Suppose that the planner has obtained pilot data on *J* schools, each with sample size *n*. First,

26

the user computes a one-way random-effects analysis of variance, obtaining an estimate of the between-school variance, τ^2, the within-school variance, σ^2, and, from these, $ICC = \tau^2/(\tau^2 + \sigma^2)$ and $\lambda = \tau^2/(\tau^2 + \sigma^2/n)$.

The next step is to construct $J/2$ matched pairs just as they will be constructed in the study. One then proceeds as follows:

1. Compute the sample means for each of the J schools. This yields a data set with $J/2$ pairs of sample means.

2. Now compute a second one-way analysis of variance with two level-1 units per each of the $J/2$ Level 2 units. This will yield a between-pair variance of τ^{*2}, a within-pair variance of σ^{*2}, and, from these, $ICC^* = \tau^{*2}/(\tau^{*2} + \sigma^{*2})$.

3. Compute $\rho_{pairs} = ICC^*/\lambda$.

Estimating $\rho_{w\mu}$ in the ANCOVA Design for Group-Randomized Studies

The correlation between the group-level covariate (denoted W) and the sample group mean outcome (denoted m), ρ_{wm}, will be an attenuated estimate of $\rho_{w\mu}$ as the sample size used to compute m diminishes. So it is not appropriate simply to compute the sample correlation between W and m. Instead, one can use the relationship $\rho_{w\mu} = \rho_{wm}/\sqrt{\lambda}$, where λ is the reliability of m.

Now suppose that W itself is also the sample mean of a person-level covariate and is therefore a fallible estimate from the pilot sample of the "true mean W." Then let $\rho_{w\mu \text{ naive}}$ be the correlation between the sample mean of W and the sample mean of the outcome. Then the relation $\rho_{w\mu} = \rho_{w\mu \text{ naive}}/\sqrt{(\lambda_m \times \lambda_w)}$ can be used. Thus, the naive estimate is divided by the square root of the product of the two reliabilities, λ_m and λ_w, the reliabilities of the sample mean outcome and the sample mean covariates, respectively.

Notes

[1]The Optimal Design software is freely available from the Web site of the William T. Grant Foundation (http://www.wtgrantfoundation.org). It can be used to estimate the power of individual- and group-randomized studies following the designs discussed in this article, among others. Software documentation containing all the formulas for the power calculations is also accessible on the Web site.

[2]The model generating the data was $y_{ij} = 2 + 0.95W_j + 0.25Z_j + r_j + e_{ij}$, where y_{ij} represents the outcome for student $i = \{1, \ldots, n\}$ in school $j = \{1, \ldots, J\}$; W_j represents the school-level covariate, assumed normally distributed with mean 0 and variance 1; Z_j represents a school-level indicator (1 for treatment, 0 for control); $r_j \sim N(0, \tau)$ represents a random error associated with each school; and $e_{ij} \sim N(0, \sigma^2)$ represents a random error associated with each student. The intraclass correlation is set at .10, n at 100, and J at 50.

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.

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., & Riccio, J. A. (2005). Using place-based random assignment and comparative interrupted time-series analysis to evaluate the jobs-plus employment program for public housing residents. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 19–51). Thousand Oaks, CA: Sage.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2005). *Using covariates to improve precision: Empirical guidance for studies that randomize schools to measure the impacts of educational interventions*. New York: MDRC.

Borman, G. D., Slavin, R. E., Cheung, A., Chamberlain, A., Madden, N., & Chambers, B. (2005). Success for all: First-year results from the National Randomized Field Trial. *Educational Evaluation and Policy Analysis*, *27*, 1–22.

Cochran, W. G. (1957). Analysis of covariance: Its nature and uses. *Biometrics*, *13*(3), 261–281.

Cook, T. D. (2005). Emergent principles for the design, implementation and analysis of cluster-based experiments in social science. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 176–198). Thousand Oaks, CA: Sage.

Cook, T. D., Hunt, H. D., & Murphy, R. F. (2000). Comer's School Development Program in Chicago: A theory-based evaluation. *American Educational Research Journal*, *37*, 535–597.

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

Education Sciences Reform Act, 108 Cong. 2nd Sess. 48 (2002).

Federer, W. T. (1955). *Experimental design*. New York: Macmillan.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture*, *33*, 503–513.

Fisher, R. A. (1936). *Statistical methods for research workers*. Edinburgh, UK: Oliver & Boyd.

Fisher, R. A. (1949). *The design of experiments*. Edinburgh, UK: Oliver & Boyd.

Flay, B. R. (2000). Approaches to substance use prevention utilizing school curriculum plus social environmental change. *Addictive Behaviors*, *25*(6), 861–885.

Freedman, L. S., Green, S. B., & Byar, D. P. (1990). Assessing the gain in efficiency due to matching in a community intervention study. *Statistics in Medicine*, *9*(8), 943–952.

Gail, M. H., Mark, S. D., Carroll, R., Green, S. B., & Pee, D. (1996). On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, *15*(11), 1069–1092.

Grimshaw, J., Eccles, M., Campbell, M., & Elbourne, D. (2005). Cluster randomized trials of professional and organizational behavior change interventions in health care settings. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 71–93). Thousand Oaks, CA: Sage.

Hannan, P. J., Murray, D. M., Jacobs, D. J., & McGovern, P. (1994). Parameters to aid in the design and analysis of community trials: Intraclass correlations from the Minnesota Heart Health Program. *Epidemiology*, *5*(1), 88–95.

Hughes, J. P. (2005). Using baseline data to design a group randomized trial. *Statistics in Medicine*, *24*(13), 1983–1994.

Kempthorne, O. (1952). *The design and analysis of experiments*. New York: John Wiley.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Klar, N., & Donner, A. (1998). The merits of matching in community intervention trials: A cautionary tale. *Statistics in Medicine*, *16*(15), 1753–1764.

Leviton, L. C., & Horbar, J. D. (2005). Cluster randomized trials for the evaluation of strategies designed to promote evidence-based practice in perinatal and neonatal medicine. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 94–114). Thousand Oaks, CA: Sage.

Martin, D. C., Diehr, P., Perrin, E. B., & Koepsell, T. D. (1993). The effect of matching on the power of randomized community intervention studies. *Statistics in Medicine*, *12*(3–4), 329–338.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education*. Washington, DC: Brookings Institution.

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. *Harvard Educational Review*, *66*(4), 797–842.

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

Porter, A. C., Blank, R. K., Smithson, J. L., & Osthoff, E. (2005). Place-based randomized trials to test the effects on instruction practices of a mathematics/science professional development program for teachers. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 147–175). Thousand Oaks, CA: Sage.

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

Raudenbush, S. W., & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401.

Raudenbush, S. W., Liu, X.-F., Spybrook, J., Martinez, A., & Congdon, R. (2006). Optimal Design software for multi-level and longitudinal research (Version 1.77) [Computer software]. Available at http://sitemaker.umich.edu/group-based

Sherman, L. W., & Weisburd, D. (1995). General deterrent effects of police patrol in crime "hot spots": A randomized, controlled trial. *Justice Quarterly*, *12*(4), 625–648.

Shochet, P. A. (2005). *Statistical power for random assignment evaluations of education programs*. Princeton, NJ: Mathematica Policy Research.

Sikkema, K. J. (2005). HIV prevention among women in low-income housing developments: Issues and interventions outcomes in a place-based randomized controlled trial. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 52–70). Thousand Oaks, CA: Sage.

Teruel, G. M., & Davis, B. (2000). *Final report: An evaluation of the impact of PROGRESA cash payments on private inter-household transfers*. Washington, DC: International Food Policy

Research Institute.

Weisburd, D. (2000). Randomized experiments in criminal justice policy: Prospects and problems. *Crime and Delinquency*, *46*(2), 181–193.

Weisburd, D. (2005). Hot spots policing experiments and criminal justice research: Lessons from the field. In R. Boruch (Ed.), *Place randomized trials: Experimental tests of public policy* (pp. 220–245). Thousand Oaks, CA: Sage.

Authors

STEPHEN W. RAUDENBUSH is a professor in the Department of Sociology at the University of Chicago, 1126 E. 59th Street, Chicago, IL 60637; sraudenb@uchicago.edu. He is best known for his expertise in quantitative methodology using the advanced research technique of hierarchical linear models, which allows researchers to accurately evaluate data from school performance. His research pursues the development, testing, refinement, and application of statistical methods for individual change. He also researches the effects of social settings, such as schools and neighborhoods.

ANDRES MARTINEZ is a doctoral student in the combined program in Education and Statistics at the University of Michigan and a visiting research scholar at the University of Chicago, Social Science Research Building, Room 417, 1126 E. 59th Street, Chicago, IL 60637; amzzz@umich.edu. His current research interests include causal inference in hierarchical settings, the measurement of educational settings, and the effectiveness of conditional cash transfer programs in education.

JESSACA SPYBROOK is a doctoral candidate in the combined program in Education and Statistics at the University of Michigan, Institute for Social Research #2050, 426 Thompson Street, Ann Arbor, MI 48106-1248; jessacah@umich.edu. She has been a part of the "Building Capacity for Evaluating Group-Level Interventions" project sponsored by the William T. Grant Foundation since January 2004. She co-authored documentation that accompanies the Optimal Design Software and has been a part of the consulting team that assists researchers in the design and analysis of group-randomized trials.