Journal of Educational and Behavioral Statistics
http://jebs.aera.net

Statistical Power for Random Assignment Evaluations of Education Programs

Peter Z. Schochet

JOURNAL OF EDUCATIONAL AND BEHAVIORAL STATISTICS 2008 33: 62 originally published online 22 October 2007
DOI: 10.3102/1076998607302714

The online version of this article can be found at:
http://jeb.sagepub.com/content/33/1/62

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

Journal of Educational and Behavioral Statistics
March 2008, Vol. 33, No. 1, pp. 62–87
DOI: 10.3102/1076998607302714

Statistical Power for Random Assignment Evaluations of Education Programs

Peter Z. Schochet
Mathematica Policy Research, Inc.

This article examines theoretical and empirical issues related to the statistical power of impact estimates for experimental evaluations of education programs. The author considers designs where random assignment is conducted at the school, classroom, or student level, and employs a unified analytic framework using statistical methods from the literature. Focusing on standardized test scores of elementary school students, this article discusses appropriate precision standards and, for each design, the required number of schools to achieve those standards using empirical values of intraclass correlations, regression R^2 *values, and other parameters. Clustering effects vary by design but are typically large. Thus, large school samples are required for education trials, and many evaluations will only have sufficient power to detect precise impacts for relatively large subgroups of sites.*

Keywords: *statistical power; experimental designs; evaluations of education programs*

This article examines issues related to the statistical power of impact estimates for experimental evaluations of education programs. I focus on group-based experimental designs because many studies of education programs involve random assignment at the group level (e.g., at the school or classroom level) rather than at the student level. The clustering of students within groups generates design effects that considerably reduce the precision of the impact estimates, because the outcomes of students within the same schools or classrooms tend to be correlated. Thus, statistical power is a concern for these evaluations.

Until recently, evaluations of education programs where the student is the unit of analysis have often ignored design effects due to clustering; thus, many of these studies overestimated the statistical precision of their impact estimates (Hedges, 2004). As a consequence, there is currently much concern among education policy makers about how to interpret impact findings from previous evaluations of education programs and how to properly design future experimental studies to have sufficient statistical power to estimate impacts with the desired level of precision. This is a pressing issue because of provisions in the Education Sciences Reform Act of 2002 specifying, when feasible, the use of experimental designs to provide scientifically based evidence of program effectiveness and substantial taxpayer resources that are currently targeted to large-scale experimental

evaluations of educational interventions by the Institute for Education Sciences (IES) at the U.S. Department of Education (ED).

There is a large literature on appropriate statistical methods under clustered randomized trials. Walsh (1947) showed that if clusters are the unit of random assignment, then conventional analyses will lead to an overstatement about the precision of the results, and the problem becomes more severe as the heterogeneity across clusters increases. Cochran (1963) and Kish (1965) discuss the calculation of design effects under clustered sample designs in terms of the intraclass correlation coefficient (ICC), which is the proportion of variance in the outcome that lies between clusters. In a seminal article, Cornfield (1978) first drew attention in the public health literature to the analytic issues presented by clustered randomized trials. Since that time, there have been extensive methodological developments in adjusting variance estimates for clustered designs (e.g., see Donner & Klar, 2000; Murray, 1998; Raudenbush, 1997). Much of this literature has focused on cluster randomized trials of medical and public health interventions. Despite this literature, however, Varnell, Murray, Janega, and Blitstein (2004) and Ukoumunne, Gulliford, Chinn, Sterne, and Burney (1999) found that only a small percentage of published impact studies that they reviewed in the public health field used appropriate methods to account for clustering.

Less attention has focused specifically on statistical power analyses in the education field. Bryk and Raudenbush (1992), Bloom, Bos, and Lee (1999), and Raudenbush, Spybrook, Liu, and Congdon (2004) discuss appropriate statistical procedures and provide examples but do not systematically consider statistical power issues for specific designs that are typically used to evaluate school interventions and that are based on up-to-date parameter assumptions.

This article builds on the literature in several ways. First, I use a unified hierarchical linear modeling (HLM) framework (Bryk & Raudenbush, 1992) to examine statistical power under commonly used experimental designs in a school setting with a single treatment and control group per site.[1] Although the variance formulas used are known, the way in which they are used and the policy context is new, as is, in particular, my treatment of fixed versus random group effects. Second, I discuss appropriate precision standards for educational evaluations, a topic that has not been systematically addressed in the literature. Third, a list of key parameters (such as ICCs and regression R^2 values) that are required to estimate power levels is compiled. Finally, this is the first article to jointly consider appropriate sample sizes for alternative designs that can serve as a guide for future designs of education-related evaluations.

My empirical analysis focuses on achievement test scores of elementary school and preschool students in low-performing school districts. I focus on test scores due to the accountability provisions of the No Child Left Behind Act of 2001 and the ensuing federal emphasis on testing interventions to improve reading and mathematics scores of young students.

This article is in five sections. First, general issues for a statistical power analysis, including techniques for assessing appropriate precision levels, are discussed. Second, presented is a unified mathematical model to examine sources of variance for school-based evaluations and, third, an application of this framework to specific designs. Fourth, I conduct power calculations for each of the considered designs and, finally, present conclusions.

General Issues for a Statistical Power Analysis

An important part of any evaluation design is the statistical power analysis, which demonstrates how well the design of the study will be able to distinguish real impacts from chance differences. To determine appropriate sample sizes for experimental evaluations, researchers typically calculate minimum detectable impacts, which represent the smallest program impacts—average treatment and control group differences—that can be detected with a high probability. In addition, it is common to standardize minimum detectable impacts into *effect size units*, that is, as a percentage of the standard deviation of the outcome measures (also known as Cohen's *d*), to facilitate the comparison of findings across outcomes that are measured on different scales (Cohen, 1988). Hereafter, I denote minimum detectable impacts in effect size units as MDEs.

Mathematically, the MDE formula can be expressed as follows:

$$MDE = Factor(\alpha, \beta, df) * \sqrt{\text{Var}(impact)}/\sigma, \quad (1)$$

where Var(*impact*) is the variance of the impact estimate, σ is the standard deviation of the outcome measure, and *Factor*(.) is a constant that is a function of the significance level (α), statistical power (β), and the number of degrees of freedom (*df*, which equals the number of groups minus the number of strata minus 1).[2] *Factor*(.) becomes larger as α and *df* decrease and as β increases (see Table 1).

As an example, consider an experimental design with a single treatment and control group and $\alpha = .05$ and $\beta = .80$. In this case, for a given sample size and design structure, there is an 80% probability that a two-sample *t* test will yield a statistically significant impact estimate at the 5% significance level if the true impact were equal to the MDE value in Equation 1.

Precision Standards

A key issue for any evaluation is the precision standard to adopt for the impact estimates, which determines appropriate sample sizes. There are two key factors that need to be considered. First, the precision standard should depend on what impact is deemed meaningful in terms of future, longer term student outcomes (such as high school graduation, college attendance, earnings, etc.). Second, the precision standard should depend on what intervention effects are realistically attainable.

TABLE 1
Values for Factor*(.) in Equation 2 of Text, by the Number of Degrees of Freedom for One- and Two-Tailed Tests, and at 80% and 85% Power*

Number of Degrees of Freedom	One-Tailed Test		Two-Tailed Test	
	80% Power	85% Power	80% Power	85% Power
2	3.98	4.31	5.36	5.69
3	3.33	3.61	4.16	4.43
4	3.07	3.32	3.72	3.97
5	2.94	3.17	3.49	3.73
6	2.85	3.08	3.35	3.58
7	2.79	3.02	3.26	3.49
8	2.75	2.97	3.20	3.42
9	2.72	2.93	3.15	3.36
10	2.69	2.91	3.11	3.32
11	2.67	2.88	3.08	3.29
12	2.66	2.87	3.05	3.26
13	2.64	2.85	3.03	3.24
14	2.63	2.84	3.01	3.22
15	2.62	2.83	3.00	3.21
20	2.59	2.79	2.95	3.15
30	2.55	2.75	2.90	3.10
40	2.54	2.74	2.87	3.07
50	2.53	2.72	2.86	3.06
60	2.52	2.72	2.85	3.05
70	2.51	2.71	2.84	3.04
80	2.51	2.71	2.84	3.04
90	2.51	2.71	2.83	3.03
100	2.51	2.70	2.83	3.03
Infinity	2.49	2.68	2.80	3.00

Note: All figures assume a 5% significance level.

There is no uniform basis for adopting precision standards in educational research, and this critical issue has not been rigorously addressed in the literature, primarily because it is often difficult to determine what size impacts are meaningful, especially for young children. In this section, we discuss several procedures that can be used in practice.

Examine impact results from previous evaluations. One approach for adopting a precision standard is to use impact results found in previous rigorous evaluations similar to the one under investigation that found beneficial impacts. Another related and widely used approach is to use meta-analysis results from previous impact studies across a broad range of disciplines to examine the magnitude of

impacts that have been achieved (Cohen, 1988; Lipsey & Wilson, 1993). Citing these meta-analysis results, many evaluations of education programs adopt standardized effect sizes of .20, .25, or .33 as the precision standard.

Although these approaches can be used to determine what impacts could be attainable for a particular intervention, they do not necessarily address what impacts are meaningful. As discussed next, I believe that these precision standards are somewhat high for testing the efficacy of education interventions on student standardized test scores.

Adopt a benefit–cost framework. One approach for assessing meaningful standardized effect sizes is to select samples large enough to detect impacts such that program benefits (measured in dollars) would offset program costs. For instance, several studies have indicated that a one-standard-deviation increase in either math or reading test scores for elementary school children is associated with about 8% higher earnings when the students join the labor market (Currie & Thomas, 1999; Murnane, Willet, & Levy, 1995; Neal & Johnson, 1996). Krueger (2000) estimates that the present discounted value of this higher earnings stream over a worker's lifetime due to a one-standard-deviation increase in test scores is about $37,500.[3,4] As a consequence, if an intervention improved test scores by .20 standard deviations, the present value of lifetime earnings would be $7,500, which was roughly the nationwide total expenditures per pupil in 1997–1998. Because most interventions are likely to cost less than $7,500 per pupil, these results suggest that a precision standard of .20 standard deviations might be too large from a benefit-cost standpoint.

Examine the natural progression of students. Another approach is to adopt a precision standard based on the natural growth of student outcomes over time. Kane (2004) found that the test performance of elementary school students in math and reading grows by about .70 standard deviations per grade. I found similar results using scaled SAT-9 test score data from the Longitudinal Evaluation of School Change and Performance (LESCP) in Title I schools.[5] Assuming that test score gains occur uniformly throughout the school year, an average test score gain of about .70 standard deviations suggests that a standardized effect size of .20 corresponds with roughly 3 months of instruction (assuming a regular 10-month school year). This is a large impact given all else that is occurring in students' lives. Thus, according to this metric, it might be appropriate to adopt a smaller, more attainable precision standard.

Examine the distribution of outcomes across schools. Another metric is to assess what an MDE implies about movements in mean student outcomes in a typical school relative to the distribution of outcomes across a broader set of schools. This approach again suggests that effect sizes of .20 to .33 are large.

For instance, I analyzed California Achievement Test (CAT/6) data for third graders using data from the 2004 California Standardized Testing and Reporting (STAR) Program. Consider a school at the 25th percentile of the math or reading

test score distribution. A 33% effect size implies that the intervention would move that school from the 25th to 37th percentile of the score distribution, which is a large increase.[6] A more attainable 10% effect size would move the school from the 25th to 29th percentile. LESCP data for SAT-9 scores of third graders in Title 1 schools yield similar findings.

A related method is to assess the magnitude of MDEs by translating them into nominal impacts for binary outcomes. For example, according to the National Assessment of Educational Progress (NAEP), nearly 70% of fourth graders nationally performed below the Proficient level in reading and math in 2003 (National Center for Education Statistics, 2004). For this binary outcome, effect sizes of .33 and .20 translate into large impacts of about 15.0 and 9.2 percentage points, respectively. A smaller, more realistic effect size of .10 translates into an impact of about 4.6 percentage points.

In sum, there is no standard basis for assessing appropriate precision standards for experimental impact evaluations of education programs. A precision standard of between .20 and .33 of a standard deviation is often used and is justified on the basis of meta-analysis results across a range of fields. This approach also represents a reasonable compromise between evaluation rigor and evaluation cost. However, it must be viewed as somewhat ad hoc. Other methods suggest that smaller effect sizes are meaningful for examining intervention effects on test scores.

Mathematical Framework for the Variance Calculations

The MDE calculations depend critically on the standard errors of the impact estimates. In this section, I use an HLM approach to develop a unified framework for identifying sources of variance under commonly used experimental designs in a school setting with a single treatment and control group per site.

I consider the following four-level model corresponding to students in Level 1 (indexed by h), classrooms in Level 2 (indexed by i), schools in Level 3 (indexed by j), and school districts in Level 4 (indexed by k):

$$
\begin{aligned}
&\textit{Level } 1 : \textit{Students} : && Y_{hijk} = \alpha_{0ijk} + \beta_{0ijk} R_1 T_{hijk} + e_{hijk} \\
&\textit{Level } 2 : \textit{Classrooms} : && \alpha_{0ijk} = \alpha_{00jk} + \beta_{00jk} R_2 T_{0ijk} + u_{0ijk}^{random_2} \\
& && \beta_{0ijk} = \beta_{00jk} + \eta_{0ijk}^{random_2} \\
&\textit{Level } 3 : \textit{Schools} : && \alpha_{00jk} = \alpha_{000k} + \beta_{000k} R_3 T_{00jk} + r_{00jk}^{random_3} \\
& && \beta_{00jk} = \beta_{000k} + \theta_{00jk}^{random_3} \\
&\textit{Level } 4 : \textit{Districts} : && \alpha_{000k} = \lambda_0 + \lambda_1 R_4 T_{000k} + \tau_{000k}^{random_4} \\
& && \beta_{000k} = \lambda_1 + \pi_{000k}^{random_4}.
\end{aligned}
\tag{2}
$$

In the model, Y_{hijk} is a continuous outcome measure for a student; $R_q (q = 1, 2, 3, 4)$ is an indicator variable equal to 1 if the point of random assignment is at Level q, and 0 otherwise ($\Sigma R_q = 1$); $T$is an indicator variable equal to 1 for treatment group units and 0 for controls; and the superscript, $random_q$, equals 1 if the group effect in Level q is treated as a random error term, and 0 if it is treated

as a fixed parameter (a topic that I discuss in more detail in the next section). The α terms represent level-specific intercepts, and the β terms represent level-specific treatment effects that are applicable depending on the point of random assignment. The λ_1 parameter represents the average impact estimate (the difference in mean outcomes between treatment and control group students), and the λ_0 parameter represents the control group mean. The random student-level errors, e_{hijk}, are assumed to be *iid* $N(0, \sigma_e^2)$.

Clustering arises due to the random error terms in the Level 2 through 4 equations. For HLM level q, group effects are considered to be random (that is, $random_q = 1$) if either (a) random *assignment* of units is conducted at level q, or (b) units at level q are considered to be randomly *sampled* from a broader universe of units before or after random assignment takes place. When included, I assume that u^1_{0ijk} are *iid* $N(0, \sigma_u^2)$ classroom-specific random error terms that capture the correlation between the outcomes of students in the same classroom; r^1_{00jk} are *iid* $N(0, \sigma_r^2)$ school-specific error terms that capture the correlation between the outcomes of students in the same school; τ^1_{000k} are *iid* $N(0, \sigma_\tau^2)$ district-specific error terms; and η^1_{0ijk}, θ^1_{00jk}, and π^1_{000k} are *iid* $N(0, \sigma_\eta^2)$, *iid* $N(0, \sigma_\theta^2)$, and *iid* $N(0, \sigma_\pi^2)$ random error terms, respectively, that represent the extent to which *treatment effects* (that is, average impacts) vary across classrooms, schools, and districts, respectively. The random error terms across equations are assumed to be distributed independently of each other.[7]

To more concisely examine sources of variance for the considered designs, I use a single-equation version of the HLM framework (e.g., Murray, 1998) by recursively inserting the Level 2 through 4 equations into the Level 1 equation, which yields the following expression:

$$Y_{hijk} = \lambda_0 + \lambda_1 T_h + \{L_1 e_{hijk} + L_2[u^{random_2}_{0ijk} + R_1 T_h \eta^{random_2}_{0ijk}] + L_3[r^{random_3}_{00jk} + (R_1 + R_2) T_h \theta^{random_3}_{00jk}] + L_4[\tau^{random_4}_{000k} + (R_1 + R_2 + R_3) T_h \pi^{random_4}_{000k}]\}, \quad (3)$$

where, for simplicity, I subscript T by h only, and where, for future reference, I include L_q indicator variables ($q = 1, 2, 3, 4$), which equal 1 if Level q is included in the considered design, and 0 otherwise (in which case, I omit subscripts corresponding to Level q). In all designs, $L_1 = 1$.

Equation 3 yields a simple differences-in-means estimator. As discussed later, the model can be generalized to include level-specific covariates measured at baseline that can improve the precision of the impact estimates.

Variance Calculations for Commonly Used Designs

In this section, I use Equation 3 to examine sources of variance for designs where the following units are randomly assigned to a research status: (a) students within sites (schools or districts); (b) classrooms within schools; and (c) schools within districts. Each of these designs is nested within the four-level model discussed above.

TABLE 2
Summary of Alternative Designs

Design #, Unit of Random Assignment in Equation 3, and Level With $R_q = 1$	HLM Levels Included in Equation 3 (Levels With $L_q = 1$[a])	Sources of Clustering in Equation 3 (Levels With $random_q = 1$[b])	Equation # for Variance Formulas
I: Students within sites (schools or districts): $R_1 = 1$	3 or 4	none	6
II: Students within sites: $R_1 = 1$	3 or 4	3 or 4	7
III: Students within schools: $R_1 = 1$	2 and 3	2 and 3	8
IV: Classrooms within schools[c]: $R_2 = 1$	2 and 3	2	9
V: Classrooms within schools (at least two classrooms per school): $R_2 = 1$	2 and 3	2 and 3	10
VI: Classrooms within schools (only two classrooms per school): $R_2 = 1$	3	3	10 with $\sigma_u^2 = 0$
VII: Schools within districts: $R_3 = 1$	3 and 4	3	11
VIII: Schools within districts: $R_3 = 1$	2, 3, and 4	2 and 3	12

Note: Subscripts of 2, 3, and 4 denote HLM levels corresponding to classrooms, schools, and districts, respectively. HLM = hierarchical linear model.
a. Level 1 (students) is included in all models (that is, $L_1 = 1$). All other levels not listed (and corresponding subscripts) are excluded from Equation 3.
b. All other level-specific group effects included in Equation 3 are assumed to be fixed (that is, $random_q = 0$).
c. This design is pertinent if (a) there are at least two treatment and control classrooms per school and school fixed effects are included in the analysis, or (b) there is only one classroom per condition per school, but school fixed effects are not included in the analysis.

Table 2 summarizes the various designs that I consider with a single treatment and control group per site. For each design, it displays values for R_q, L_q, and $random_q$ in Equation 3 and displays equation numbers in the text for the variance formulas. Power analyses for these designs can be conducted using the Optimal Design software program (Raudenbush et al., 2004).

Random Assignment of Students Within Sites: Fixed-Effects Case

In some designs, students in purposively selected schools or districts are randomly assigned directly to the treatment and control groups without regard to the classrooms or schools that the students attend, for example, the 21st Century

Community Learning Centers Program (Dynarski et al., 2004) and the Impact Evaluation of Charter Schools Strategies (Gleason & Olsen, 2004). It is clear that these designs are not appropriate for testing classroom-based interventions where random assignment at the classroom or teacher level is required. Furthermore, these designs are appropriate only if potential "spillover" effects are expected to be small, so that students in the control group are expected to "receive" little of the intervention through their contact with students in the treatment group.

A central issue in the variance calculations for this design is whether site effects should be treated as fixed or random. Under the fixed-effects case, the impact estimates are viewed as generalizing to the study sites only, whereas under the random effects case, the impact estimates are viewed as generalizing to a broader population of sites similar to the study sites, which introduces design effects.

Although this issue needs to be addressed for each study, I believe that the fixed-effects case is usually more realistic in evaluations of education interventions. For most evaluations, sites are purposively selected for the study for a variety of reasons (such as the site's willingness to participate, whether the site has a sufficient number of potential program participants to accommodate a control group, and so on). Furthermore, most evaluations are efficacy trials where a relatively small number of purposively selected sites is included in the study. Thus, in many instances, it is untenable to assume that the study sites are representative of a broader, well-defined population. Furthermore, inflating the standard errors to incorporate between-site effects will slant the study in favor of finding internally valid impact estimates that are not statistically significant, thereby providing less information to policy makers on potentially promising interventions. Instead, I believe, in general, that it is preferable to treat site effects as fixed and to assess the generalizability of study findings by examining the pattern of the impact estimates across sites. This approach is likely to yield credible information on the extent to which specific interventions could be effective and whether larger scale studies are warranted to examine whether they truly are effective.

Using Equation 3, the impact estimate for site g under the fixed-effects design is $(\hat{\lambda}_1 + \hat{\theta}^0_{0.g.})$ if sites are schools, and $(\hat{\lambda}_1 + \hat{\pi}^0_{0..g})$ if sites are districts. The variance of these site-specific impact estimates can be expressed as follows (see Design I in Table 2):

$$\mathrm{Var}(\hat{\lambda}_{1g}) = \frac{\sigma_e^2}{m_g p_{1g}(1-p_{1g})}, \tag{4}$$

where m_g is the total sample size of treatment and control students in site g, and p_{1g} is the proportion of students assigned to the treatment group in the site.

Pooled impact estimates across sites are calculated as a weighted average of the impact estimates in each site, and the associated variances are obtained by aggregating the site-specific variances in Equation 4 as follows:

$$\text{Var}(\textit{pooled impact}) = \sum_{g=1}^{s} w_g^2 \frac{\sigma_e^2}{m_g p_{1g}(1-p_{1g})}, \tag{5}$$

where s is the number of sites and w_g is the weight associated with site g, where the weights sum to unity. Each site could be given equal weight in the analysis, or weights could be constructed to be inversely proportional to site-specific variances if they are allowed to vary (Fleiss, 1986).

To reduce notation and to facilitate comparisons with the other designs discussed below, I use the following simplified (approximate) version of Equation 5:

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_e^2}{s[mp_1(1-p_1)]}, \tag{6}$$

where m is the average sample size per site and p_1 is the average sampling rate to the treatment group per site.

I note that estimating pooled impacts is appropriate in many site-based evaluations, because even in cases where the tested interventions differ somewhat across sites and serve different populations, the interventions are usually within the same general category (such as a reading or math curriculum, an after-school program, a technology, or a teacher preparation or mentoring model) and often share common features and a common funding source. Thus, it is typically of policy interest to examine the overall efficacy of promising interventions within a general class of treatments, even though the results must be interpreted carefully.

Random Assignment of Students Within Sites: Random-Effects Case

In some designs, site effects are treated as random. This can occur in two ways. First, purposively selected sites could be considered representative of a broader population of similar sites. Second, in some evaluations, sites are actually randomly sampled from a larger pool of sites. This type of design is typically employed in large-scale studies of a well-established program that require externally valid impact estimates and where the burden of evidence of program effectiveness is set high, for example, the National Evaluation of Upward Bound (Myers & Schirm, 1999) and the National Job Corps Study (Schochet, Burghardt, & Glazerman, 2001).

In these random-effects designs, study results are generalized more broadly than in the fixed-effects designs but involve a cost in terms of precision levels. Consider a two-level model corresponding to students and schools (see Design

II in Table 2). For this design, Equation 3 yields the following variance formula for the pooled impact estimate ($\hat{\lambda}_1$):

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_\theta^2}{s} + \frac{\sigma_e^2}{s[mp_1(1-p_1)]}, \tag{7}$$

where σ_θ^2 is the variance of the impacts across schools, and where other parameters are defined as above. Design effects occur under this design because of the first variance term, which is deflated by the number of schools (not students). If sites are considered to be districts rather than schools, then, in Equation 7, σ_θ^2 is replaced by σ_π^2 and s is replaced by d, the number of districts.

The variance formulas presented above can be easily generalized to account also for additional levels of clustering within sites. For instance, if classrooms in study schools were considered to be representative of a broader population of classrooms in these schools, then Equation 3 yields the following variance formula (see Design III in Table 2):

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_\theta^2}{s} + \frac{\sigma_\eta^2}{sc} + \frac{\sigma_e^2}{sc[np_1(1-p_1)]}, \tag{8}$$

where c is the average number of classrooms per school and n is the average number of students per classroom. Additional variance terms could also be included to account for potential "treatment-induced" correlations between the outcomes of treatments if the intervention is administered in small groups, thereby creating potential correlations between the outcomes of treatments within each small group (Murray, Varnell, & Blitstein, 2004; Raudenbush, 1997).

Random Assignment of Classrooms Within Schools

A design that is commonly used in evaluations of school interventions is when classrooms or teachers within study schools are randomly assigned to the treatment or control groups, for example, the Evaluation of the Effectiveness of Educational Technology Interventions (Dynarski et al., 2004). This type of design is appropriate for interventions that are administered at the classroom or teacher level and where potential spillover effects across classrooms are deemed to be small.

One way to interpret this design is that a "miniexperiment" is being conducted in each school. If school effects are treated as fixed, pooled impact estimates across schools are calculated by averaging the impact estimates from each miniexperiment. The variance formula for these pooled impacts can be expressed as follows (see Design IV in Table 2):

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_u^2}{s[cp_2(1-p_2)]} + \frac{\sigma_e^2}{s[cp_2(1-p_2)]n}, \tag{9}$$

where p_2 is the average proportion of classrooms assigned to the treatment group per school.

Several important features of this variance formula are worth mentioning. First, in some evaluations, all children in the study classrooms are included in the study. Under the fixed-effects scenario, one could then argue that student effects should not be included in the variance calculations. However, it is customary to include these student-level terms, because it is usually the case that some children will not provide follow-up data due to study nonconsent, attrition, and interview nonresponse. Thus, students in the follow-up sample are often considered to be representative of a larger pool of students in the study schools. An alternative approach is to apply a finite population correction to the student-level variance term (Schochet, 2005).

Second, in some evaluations, students within each of the participating schools and grades are randomly assigned to classrooms at the start of the school year, for example, the Teach For America (TFA) Evaluation (Decker & Glazerman, 2004). This design ensures that the average baseline characteristics of students in the treatment and control group classrooms are similar. Although this design reduces classroom effects, it does not remove them. This is because classroom effects arise from two sources: (a) differences in the quality of teachers within schools, and (b) systematic differences in the types of children who are assigned to different classrooms. The random assignment of children to classrooms reduces the second source of variance but not the first source.

Third, if school effects are treated as random, then Equation 3 yields the following variance expression (see Design V in Table 2):

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_\theta^2}{s} + \frac{\sigma_u^2}{s[cp_2(1-p_2)]} + \frac{\sigma_e^2}{s[cp_2(1-p_2)]n}. \qquad (10)$$

Finally, due to limitations in the number of available classrooms, it is often the case that only one treatment and control classroom can be randomly assigned per school. In this case, there are not enough degrees of freedom to estimate σ_u^2, which represents the extent to which classroom-level outcomes vary within schools and treatment condition (Murray, 1998). One approach is to set σ_u^2 to 0 in Equation 10 and to use the resulting variance formula for either the random or fixed-effects specifications (see Design VI in Table 2). Another possibility for the fixed-effects specification is to use Equation 9 and ignore the school fixed effects in the analysis, that is, the stratification by school. A final approach is to combine similar schools into larger strata, thereby making it possible to estimate between-classroom effects.

Random Assignment of Schools

In some designs, schools within districts are randomly selected to the treatment and control groups, for example, the Social and Character Development

Research Program (Schochet, James-Burdumy, & Kisker, 2004). These designs are necessary for testing interventions that are (a) school based or (b) classroom based, but where potential spillover effects across classrooms are deemed to be serious. I focus on designs where school districts volunteer for the study and, thus, where district effects are treated as fixed.

Clustering at the school level only. For a school-based evaluation, one design option is not to sample classrooms within the study schools. For this option, either all relevant classrooms in the selected schools are included in the research sample or students are sampled directly to the research sample without regard to the classrooms that they are in.

Under this design and using Equation 3, the impact estimate in district g is $(\hat{\lambda}_1 + \hat{\pi}^0_{0.0g})$. The pooled impact estimate across districts is then calculated as a simple or weighted average of the district-specific impact estimates, with the following variance:

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_r^2}{dsp_3(1-p_3)} + \frac{\sigma_e^2}{[dsp_3(1-p_3)]cn}, \tag{11}$$

where d is the number of school districts, s is now the average number of schools per district, and p_3 is the average proportion of schools per district assigned to the treatment group (see Design VII in Table 2).

Clustering at the school and classroom levels. For a school-based design, there could also be clustering at the classroom level. This would occur if, to conserve project resources, classrooms were sampled within the study schools or if the full set of classrooms in the study schools were considered to be representative of a larger population of classrooms. In this case, the variance formula for the pooled impact estimate can be expressed as follows (see Design VIII in Table 2):

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_r^2}{dsp_3(1-p_3)} + \frac{\sigma_u^2}{[dsp_3(1-p_3)]c^*} + \frac{\sigma_e^2}{[dsp_3(1-p_3)]c^*n}, \tag{12}$$

where c^* is the number of sampled classrooms (which is assumed to be the same in the treatment and control schools).

Estimating Intraclass Correlation Coefficients

A critical issue for the MDE calculations is what estimates to use for the error variances in the above variance formulas. In this section, I present empirical values for these error variances based on those found in the literature and from new tabulations from several recent large evaluations conducted by Mathematica Policy Research Inc. I report ICCs (denoted by ρ) for standardized test scores of elementary school and preschool students in low-performing schools. An ICC represents the proportion of variance in the outcome that lies between

clusters and is the relevant parameter for the power calculations because MDEs are in standard deviation units.

Table 3 displays empirical values for ρ_r. The table shows that estimates of ρ_r vary somewhat by data source and grade level and typically become smaller when adjusted for district fixed effects. Thus, although applicable values will depend on the study context, the examined data sources suggest that ρ_r values often range from .10 to .20 for standardized test scores. Thus, in the illustrative power calculations below, I use the midpoint, .15.

There is less evidence on plausible values for the ICC at the classroom level, ρ_u, because there are fewer data sources that have student-level data on multiple classrooms within schools. LESCP and TFA data suggest a ρ_u value of about .16, which is similar to ρ_r. Thus, mean student test scores tend to differ as much across classrooms within schools as they do across schools. In the power calculations, I assume the same .15 value for ρ_u and ρ_r.

A plausible value for the impact-related ICC, ρ_θ, will depend on the relative effectiveness of the tested interventions across schools. Data indicate, however, that this ICC is typically small. For instance, in the evaluation of the 21st Century Program, the value of ρ_θ for math and reading test scores was about .04 across elementary schools and .08 across middle schools. Similarly, in the evaluation of the School Dropout Demonstration Assistance Program, the value of ρ_θ for student grades was .06. Finally, in the Early Head Start evaluation, ρ_θ was about .06 for Bayley scores and .08 for the MacArthur Communicative Development Inventories. However, because of the uncertainty of this ICC, I assume a conservative ρ_θ value of .15.

Finally, I have less information on plausible values for ρ_η, which require data on multiple treatment and control classrooms within schools. Test score data from the TFA evaluation suggest a value of about .15 for ρ_η. However, I assume a more conservative value of .20 in the power calculations to reflect the uncertainty in this parameter.

Incorporating Baseline Covariates

For a given sample design, the most effective strategy for improving precision levels for clustered designs is to use regression models to estimate program impacts (Raudenbush, 1997). The inclusion of relevant baseline student-, classroom-, and school-level explanatory variables in the regression models in Equation 3 can increase power by explaining some of the variance in mean outcomes across schools and across classrooms within schools, that is, by increasing regression R^2 values.

To demonstrate how to incorporate the use of regression models into the variance formulas, I consider the design where the school is the unit of random assignment and generalize Equation 12 as follows:

$$\text{Var}(\textit{pooled impact}) = \frac{\sigma_r^2(1-R_{BS}^2)}{dsp_3(1-p_3)} + \frac{\sigma_u^2(1-R_{BC}^2)}{[dsp_3(1-p_3)]c^*} + \frac{\sigma_e^2(1-R_W^2)}{[dsp_3(1-p_3)]c^*n}.^{8} \quad (13)$$

(*text continues on p. 80*)

TABLE 3
Intraclass Correlation Estimates for Standardized Test Scores Across Elementary Schools and Preschools, by Data Source

Data Source	Description of Data	Standardized Test Measure	Grade and Year	ICC Estimate
		Elementary schools		
Longitudinal Evaluation of School Change and Performance (LESCP)	71 Title I schools in 18 school districts in 7 states	Stanford 9	3rd in 1997; 4th in 1998; 5th in 1999	**Unadjusted** 3rd: Math: .13 3rd: Reading: .13 4th: Math: .24 4th: Reading: .19 5th: Math: .18 5th: Reading: .21 **Adjusted for district effects** 3rd: Math: .08 3rd: Reading: .06 4th: Math: .07 4th: Reading: .07 5th: Math: .11 5th: Reading: .11
Prospects Study: Figures reported in Hedberg, Santana, & Hedges (2004)	372 Title I schools in 120 school districts	Comprehensive Test of Basic Skills (CTBS)	3rd in 1991	**Unadjusted** Math: .23 Reading: .20 **Adjusted**[a] Math: .16 Reading: .18

(continued)

TABLE 3 (continued)

Data Source	Description of Data	Standardized Test Measure	Grade and Year	ICC Estimate
Teach for America Evaluation	17 schools in six cities (Baltimore, Chicago, Los Angeles, Houston, New Orleans, and the Mississippi Delta)	Iowa Test of Basic Skills (ITBS)	2nd to 4th in 2003	**Unadjusted** 2nd: Math: .10 2nd: Reading: .23 3rd: Math: .03 3rd: Reading: .05 4th: Math: .16 4th: Reading: .16
21st Century Community Learning Centers Program	30 schools in 12 school districts	Stanford Achievement Test-9 (SAT-9)	1st, 3rd, and 5th in 2002	**Unadjusted** 1st: Math: .17 1st: Reading: .19 3rd: Math: .19 3rd: Reading: .24 5th: Math: .17 5th: Reading: .09
Data from Rochester: Figures calculated from MDEs reported in Bloom, Bos, & Lee (1999)	25 elementary schools	Pupil Evaluation Program (PEP) Test	3rd and 6th in 1992	**Unadjusted** 3rd: Math: .19 3rd: Reading: .18 6th: Math: .19 6th: Reading: .14
Data from Louisville: Figures reported in Gargani & Cook (2005)	22 schools	Kentucky Core Content Test (KCCT) developed for Kentucky students	grade not reported: 2003	Reading: .11

(continued)

TABLE 3 (continued)

Data Source	Description of Data	Standardized Test Measure	Grade and Year	ICC Estimate
		Preschools		
Early Reading First Evaluation	162 preschools in 68 sites	Expressive One Word Picture Vocabulary (EOW) Test; Preschool Language Scale Auditory Comprehension Subscale (PLS Auditory)	4-year-olds in 2004	**Unadjusted** PLS: .18 EOW: .14 **Adjusted for district effects** PLS: .08 EOW: .08
FACES 2000	219 centers in 43 Head Start programs	Peabody Picture Vocabulary Test (PPVT); Woodcock Johnson Applied Problems (WJMATH); Woodcock Johnson Letter-Word Identification (WJWORD)	4-year-olds in fall 2000	**Unadjusted** PPVT: .38 WJMATH: .13 WJWORD: .16 **Adjusted for district effects** PPVT: .11 WJMATH: .06 WJWORD: .03
Early Head Start Evaluation	families in 17 Early Head Start programs	Bayley Mental Development Index (Bayley MDI); PPVT	3-year-olds between 1996 and 1999	**Unadjusted** Bayley: .19 PPVT: .18
Preschool Curriculum Evaluation (PCER)	113 preschools across 7 PCER grantees	PPVT	4-year-olds in 2004	**Unadjusted** PPVT: .20
Early Childhood Longitudinal Study: Figures reported in Hedberg et al. (2004)	1,000 public and private kindergartens	Early Childhood Longitudinal Study, Kindergarten Class (ECLS-K)	kindergarteners in spring 1999	**Adjusted**[a] Math: .17 Reading: .23

(continued)

TABLE 3 (continued)

Data Source	Description of Data	Standardized Test Measure	Grade and Year	ICC Estimate
National Education Longitudinal Study (NELS): Figures reported in Hedberg et al. (2004)	1,052 schools	NELS: 88 test battery	8th in 1988	**Unadjusted** Math: .24 Reading: .17 **Adjusted**[a] Math: .12 Reading: .08

Note: Tabulations were conducted using SAS PROC MIXED. ICC = intraclass correlation coefficient; MDEs = minimum detectable impacts in effect size units.
a. Adjusted for socioeconomic status, race, and gender.

In this expression, R^2_{BS}, R^2_{BC}, and R^2_W are, respectively, the proportion of the between-school, between-classroom, and within-classroom variances that are explained by the regression model.[9] The most effective explanatory variables are likely to be preintervention measures of the outcome variables.

To obtain benchmark regression R^2 values, I examined the fit of models using baseline and follow-up test score data on elementary school students from various data sources: (a) the LESCP, (b) the national evaluation of the 21st Century program, and (c) the TFA evaluation. My analysis indicated that R^2_{BS} and R^2_W values were at least .50 in regression models that included student-level baseline test scores as explanatory variables. Gargani and Cook (2005) and Bloom et al. (1999) found similar values using test score data from Louisville, Kentucky, and Rochester, New York, respectively. However, in the absence of these preintervention measures of the outcome variables, R^2 values were closer to .20.

Illustrative Precision Calculations

In this section, I collate formulas and results from above and calculate illustrative MDE calculations for standardized test scores for the most common designs that I have considered.

Presentation and Assumptions

Tables 4 and 5 display, under various assumptions and for each of the considered designs, the total number of schools that are required to achieve precision targets of .10, .20, .25, and .33 of a standard deviation, respectively. Because the amount and quality of baseline data vary across evaluations, I conduct power calculations assuming conservative R^2 values of 0, .20, and .50 at each group level. My estimates assume a two-tailed test; a balanced allocation (Table 4) or 2:1 allocation (Table 5) across the treatment and control groups; and the ICC values discussed above. Finally, all calculations were conducted under the following assumptions: (a) 80% power, (b) a 5% significance level, (c) the intervention is being tested in a single grade with an average of three classrooms per school per grade and an average of 23 students per classroom (with no subsampling of students), and (d) 80% of students in the sample will provide follow-up (posttest) data. Schochet (2005) presents a more complete set of precision calculations under a wider range of plausible parameter values.

Results

The main results can be summarized as follows:

Clustering matters in school-based evaluations that aim to improve test scores. Precision levels decrease substantially as clustering effects increase. For instance, assuming a zero R^2 value, 14 schools under Design I (student-level random assignment) are required to detect an effect size of .20 standard deviations,

TABLE 4

Required School Sample Sizes to Detect Target Effect Sizes, by Design: Assumes a Two-Tailed Test, a Value of .15 for the Intraclass Correlations, and a Balanced Allocation of the Research Groups

Unit of Random Assignment: Sources of Clustering	Number of Schools Required to Detect an Impact in *SD* Units of:			
	.10	.20	.25	.33
I: Students within schools: No clustering				
$R^2 = 0$	57	14	9	5
$R^2 = .2$	45	11	7	4
$R^2 = .5$	28	7	5	3
II: Students within schools: School-level clustering				
$R^2 = 0$	166	44	29	18
$R^2 = .2$	133	36	24	15
$R^2 = .5$	86	23	16	9
III: Students within schools: School- and classroom-level clustering				
$R^2 = 0$	197	51	34	21
$R^2 = .2$	157	41	28	17
$R^2 = .5$	100	27	18	11
IV: Classrooms within schools: Classroom-level clustering (ignoring school fixed effects)				
$R^2 = 0$	205	51	33	19
$R^2 = .2$	164	41	27	16
$R^2 = .5$	103	26	17	10
VI: Classrooms within schools: School-level clustering				
$R^2 = 0$	213	55	36	22
$R^2 = .2$	170	45	30	18
$R^2 = .5$	106	29	20	12
VII: Schools within districts: School-level clustering				
$R^2 = 0$	519	130	86	50
$R^2 = .2$	415	104	68	40
$R^2 = .5$	259	67	44	26
VIII: Schools within districts: School- and classroom-level clustering				
$R^2 = 0$	667	167	107	63
$R^2 = .2$	534	133	88	51
$R^2 = .5$	333	86	55	33

Note: See the text for formulas and other assumptions underlying the calculations.

TABLE 5
Required School Sample Sizes to Detect Target Effect Sizes, by Design: Assumes a Two-Tailed Test, a Value of .15 for the Intraclass Correlations, and a 2:1 Split of the Research Groups

Unit of Random Assignment: Sources of Clustering	Number of Schools Required to Detect an Impact in *SD* Units of:			
	.10	.20	.25	.33
I: Students within schools: No clustering				
$R^2 = 0$	64	16	10	6
$R^2 = .2$	51	13	8	5
$R^2 = .5$	32	8	5	3
II: Students within schools: School-level clustering				
$R^2 = 0$	172	45	30	18
$R^2 = .2$	138	37	25	15
$R^2 = .5$	89	24	16	10
III: Students within schools: School- and classroom-level clustering				
$R^2 = 0$	207	54	35	22
$R^2 = .2$	166	43	29	18
$R^2 = .5$	103	28	19	12
IV: Classrooms within schools: Classroom-level clustering (ignoring school fixed effects)				
$R^2 = 0$	232	58	37	22
$R^2 = .2$	186	46	31	18
$R^2 = .5$	116	30	19	11
VI: Classrooms within schools: School-level clustering				
$R^2 = 0$	219	57	37	23
$R^2 = .2$	175	46	31	19
$R^2 = .5$	110	30	20	12
VII: Schools within districts: School-level clustering				
$R^2 = 0$	586	147	96	56
$R^2 = .2$	469	117	77	45
$R^2 = .5$	293	76	49	29
VIII: Schools within districts: School- and classroom-level clustering				
$R^2 = 0$	754	189	121	71
$R^2 = .2$	603	151	98	57
$R^2 = .5$	377	96	62	37

Note: See the text for formulas and other assumptions underlying the calculations.

compared with 55 schools for Design VI (classroom-based random assignment) and 130 schools for Design VII (school-level random assignment; see Table 4).

Relatively large school sample sizes are required for the most commonly used designs. For classroom-level random assignment, about 30 to 50 schools are required to detect an effect size of .20 (depending on the R^2 assumption), and the required number of schools is about 65 to 130 under school-level random assignment (see Table 4). As a consequence, because of resource constraints, many evaluations will only have sufficient power to detect precise impacts for relatively large subgroups of sites. Thus, the chosen point of random assignment and the treatment of group effects as random or fixed have important implications for evaluation costs and the range of study questions that can be addressed.

Achieving effect sizes of .10 may not be attainable in many evaluations. As discussed, relatively small test score gains might be meaningful from a benefit-cost standpoint and realistic in terms of the natural progression of students over a school year. However, results suggest that very large sample sizes are required to detect relatively small test score gains. For instance, even with an R^2 value of .50, to detect an effect size of .10, Design VI requires 106 schools and Design VII requires 259 schools (see Table 4). As a consequence, because of cost constraints, some interventions should be tested only if they can be expected to have a relatively large effect on student outcomes.

R^2 *values matter.* An effective strategy for improving precision levels for group-based experimental designs is to use regression models to estimate program impacts. For example, under Design VII, 86 schools are required to achieve an effect size of .25 for an R^2 value of 0, compared with only 44 schools for an R^2 value of .50 (see Table 4). Thus, the availability of detailed baseline data at the aggregate school or individual student level (and in particular, data on preintervention measures of the outcome variables) can substantially improve statistical power under clustered designs.

A 2:1 split of the research groups does not materially reduce precision levels relative to a balanced allocation. The required school sample sizes are only slightly larger under a design with twice as many treatments as controls (or vice versa) than under a design with equal sample sizes across the research groups (Tables 4 and 5).[10] This is an important finding because an unbalanced design might be preferred in some evaluations for operational reasons (e.g., to minimize the number of controls).

Finally, I note that for each of the considered designs, the optimal sample size of districts, schools, classrooms, and students can be calculated by minimizing the MDE subject to a cost function or minimizing costs subject to a fixed MDE. The cost function will depend on study sample sizes as well as unit costs (from recruiting districts, schools, and teachers, obtaining parental consent, conducting interviews and assessments, and so on). The precision calculations presented above

are based on designs that are likely to be relatively cost-effective for attaining a given MDE level, because they assume no subsampling of students. Clustered designs can usually attain the same MDE level with a smaller total student sample if the sample contains more schools and fewer students per school. However, these designs could cost more due to the significant expense of recruiting additional districts and schools, but this will depend on unit costs specific to each project.

Summary and Conclusions

In this article, I have examined theoretical and empirical issues related to the statistical power of impact estimates under commonly used experimental designs for evaluations of education interventions that seek to improve students' standardized test scores. My main conclusion is that clustering effects cannot be ignored when groups such as schools or classrooms are the unit of random assignment. I find that relatively large school sample sizes are required to achieve targeted precision levels under these designs. Furthermore, the required sample sizes of schools and classrooms increase substantially as clustering effects increase and as precision standards are made more stringent. Design effects due to clustering cannot be ignored because of the relatively large ICCs for standardized test scores at the school and classroom levels.

The implication of these findings is that because of study resource constraints, many impact evaluations of education interventions will only have sufficient statistical power to detect impacts at the pooled level and for relatively large subgroups of sites or schools, but not for smaller subgroups. Furthermore, it might not always be feasible from a power standpoint to randomly assign multiple treatments to units. In addition, some evaluations may not have sufficient power to obtain precise estimates for other types of analyses that are often conducted in impact evaluations (such as mediated analyses, latent variable analyses, and so on). As a consequence, we must recognize that many impact evaluations of education programs can be expected to rigorously address broad research questions only and, hence, should be structured to focus on a narrow set of issues. Results from more disaggregated analyses, although important, must be deemed heuristic.

This discussion has stressed that education researchers who conduct power analyses—using, for example, the Optimal Design software program (Raudenbush et al., 2004)—must carefully specify the sources of clustering under their designs and the assumptions underlying them. In particular, the treatment of group effects as fixed or random is an important issue that has major implications for sample size requirements and has often been ignored in the literature. Finally, I emphasize that the collection of detailed baseline data is an important way to reduce clustering effects.

Notes

1. Under designs with multiple interventions (treatment groups), multiple comparison issues must be considered, which are beyond the scope of this article.

2. Specifically, *Factor*(.) can be expressed as $[T^{-1}(1-\alpha)+T^{-1}(\beta)]$ for a one-tailed test and $[T^{-1}(1-\{\alpha/2\})+T^{-1}(\beta)]$ for a two-tailed test, where $T^{-1}(.)$ is the inverse of the student's t distribution function with *df* degrees of freedom (see Murray, 1998, and Bloom, 2004, for derivations of these formulas). Equation 1 ignores the estimation error in the standard deviation.

3. This figure was calculated (a) using the age-earnings profile in the March 1999 Current Population Survey, (b) a 4% discount rate, (c) assuming workers begin earning wages at age 18 and retire at age 65, and (d) a productivity (wage) growth rate of 1% per year.

4. Kane and Staiger (2002) find similar results.

5. Kane's results are based on separate cross-sections of students (which could be affected by cohort effects), whereas the Longitudinal Evaluation of School Change and Performance results are based on the same students over time.

6. The 25th percentile of the school distribution is 605 for reading and 602 for math, and the standard deviation of CAT/6 scores is about 20 scale points.

7. My framework is a variant of a General Linear Mixed Model where it is assumed that random effects are independent and normally distributed (e.g., see Murray, 1998; Searle, Casella, & McCulloch, 1992). I do not consider models with a more general covariance structure for the random effects.

8. As shown in Raudenbush (1997), a small correction factor needs to be applied to the variance formulas when group-level covariates are included in the regression model.

9. It is usually the case that R^2_{BS} and R^2_{BC} are positive, although the school and classroom variance components could increase with regression adjustment, in which case, Equation 14 does not apply (Murray, 1998, provides a more general formulation).

10. Bloom (2004) also documents this finding.

References

Bloom, H. (2004). *Randomizing groups to evaluate place-based programs*. New York: MDRC.

Bloom, H., Bos, J., & Lee, S. (1999). Using cluster random assignment to measure program impacts: Statistical implications for evaluation of education programs. *Evaluation Review*, *23*(4), 445–469.

Bryk, A., & Raudenbush, S. (1992). *Hierarchical linear models for social and behavioral research. Applications and data analysis methods*. Newbury Park, CA: Sage.

Cochran, W. (1963). *Sampling techniques*. New York: John Wiley.

Cohen, J. (1988). *Statistical power analysis for behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Cornfield, J. (1978). Randomization by group: A formal analysis. *American Journal of Epidemiology*, *108*(2), 100–102.

Currie, J., & Thomas, D. (1999). *Early test scores, socioeconomic status and future outcomes* (NBER Working Paper No. 6943). Cambridge, MA: National Bureau of Economic Research.

Decker, P., & Glazerman, S. (2004). *The effects of Teach For America on students: Findings from a national evaluation* (Final report submitted to the U.S. Department of Education). Princeton, NJ: Mathematica Policy Research.

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

Dynarski, M., Moore, M., Rosenberg, L., James-Burdumy, S., Deke, J., & Mansfield, W. (2004). *When schools stay open late: The national evaluation of the 21st Century Community Learning Centers Program, new findings*. Princeton, NJ: Mathematica Policy Research.

Fleiss, J. L. (1986). *The design and analysis of clinical experiments*. New York: John Wiley.

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

Gleason, P., & Olsen, R. (2004). *Impact evaluation of charter school strategies. Design documents*. Princeton, NJ: Mathematica Policy Research.

Hedberg, E., Santana, R., & Hedges, L. (2004). *The variance structure of academic achievement in America*. Working paper, University of Chicago.

Hedges, L. (2004). *Correcting significance tests for clustering*. Working paper, University of Chicago.

Kane, T. (2004). *The impact of after-school programs: Interpreting the results of four recent evaluations*. Working paper, University of California, Los Angeles.

Kane, T., & Staiger, D. (2002). The promise and pitfalls of using imprecise school accountability measures. *Journal of Economic Perspectives*, *16*(4), 91–114.

Kish, L. (1965). *Survey sampling*. New York: John Wiley.

Krueger, A. (2000). *Economic considerations and class size* (Working Paper No. 447). Princeton, NJ: Princeton University Industrial Relations Section.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment. *American Psychologist*, *48*(12), 1181–1209.

Murnane, R., Willet, J., & Levy, F. (1995). The growing importance of cognitive skills in wage determination. *Review of Economics and Statistics*, *77*, 251–266.

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

Murray, D., Varnell, S., & Blitstein, J. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, *94*(3), 423–432.

Myers, D., & Schirm, A. (1999). *The impacts of upward bound: Final report for Phase I of the National Evaluation*. Princeton, NJ: Mathematica Policy Research.

National Center for Education Statistics. (2004). *The nation's report card: Highlights 2003*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Assessment of Educational Progress.

Neal, D., & Johnson, W. (1996). The role of premarket factors in Black-White wage differentials. *Journal of Political Economy*, *104*, 869–895.

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

Raudenbush, S., Spybrook, J., Liu, X., & Congdon, R. (2004). *Optimal design for longitudinal and multilevel research: Documentation for the Optimal Design software*. Ann Arbor: University of Michigan.

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

Schochet, P., Burghardt, J., & Glazerman, S. (2001). *National Job Corps Study: The impacts of Job Corps on participants' employment and related outcomes*. Princeton, NJ: Mathematica Policy Research.

Schochet, P., James-Burdumy, S., & Kisker, E. (2004). *Social and Character Development (SACD) Research Program: Analysis plan for the multisite impact analysis*. Princeton, NJ: Mathematica Policy Research.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley.

Ukoumunne, O., Gulliford, M., Chinn, S., Sterne, J., & Burney, P. (1999). Methods for evaluating area-wide and organisation-based interventions in health and health care: A systematic review. *Health Technology Assessment*, *3*(5), 1–98.

Varnell, S., Murray, D., Janega, J., & Blitstein, J. (2004). Design and analysis of group-randomized trials: A review of recent practices. *American Journal of Public Health*, *94*(3), 393–399.

Walsh, J. (1947). Concerning the effects of the intra-class correlation on certain significance tests. *Annals of Mathematical Statistics*, *18*, 88–96.

Author

PETER Z. SCHOCHET, PhD, is a senior fellow at Mathematica Policy Research, Inc., P.O. Box 2393, Princeton, NJ 08543-2393; pschochet@mathematica-mpr.com. His areas of interest include conducting random assignment impact evaluations of education, employment, and welfare programs; statistical power; and the use of propensity scoring and regression adjustment in experimental designs.

Manuscript received July 8, 2005
Accepted November 5, 2006