Contract No.: ED-01-CO-0038-0009
MPR Reference No.: 6046-310

MATHEMATICA
Policy Research, Inc.

Statistical Power for Random Assignment Evaluations of Education Programs

June 22, 2005

Peter Z. Schochet

Submitted to:

Institute of Education Sciences
U.S. Department of Education
80 F Street, NW
Washington, DC 20208

Project Officer:
Elizabeth Warner

Submitted by:

Mathematica Policy Research, Inc.
P.O. Box 2393
Princeton, NJ 08543-2393
Telephone: (609) 799-3535
Facsimile: (609) 799-0005

Project Director:
Peter Z. Schochet

CONTENTS

TABLES

This paper examines issues related to the statistical power of impact estimates for experimental evaluations of education programs. We focus on "group-based" experimental designs, because many studies of education programs involve random assignment at the group level (for example, at the school or classroom level) rather than at the student level. The clustering of students within groups (units) 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 (that is, are not independent of each other). 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). Consequently, there is currently much concern among education policymakers 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. Cochrane (1963) and Kish (1964) 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 (see, for example, the books by Donner and Klar 2000 and Murray 1998, and Raudenbush 1997). Much of this literature has focused on cluster randomized trails of medical and public health interventions (such as community intervention trials (Koepsell et al. 1992), interventions against infectious diseases (Hayes et al. 2000), and family practice research (Campbell 2000)). Despite this literature, however, Varnell et al. (2004) found that only about 15 percent of the published impact studies that they reviewed in the public health field used appropriate methods to account for clustering; Ukoumunne et al. (1999) came to similarly pessimistic conclusions based on their review of publications in seven health science journals.

Less attention has focused specifically on statistical power analyses in the education field. Bryk and Raudenbush (1992), Bloom et al. 1999, and Raudenbush et al. 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.

In this paper, we apply the analytic methods found in the literature to examine appropriate school sample sizes in random assignment evaluations of education interventions. We provide a

unified theoretical framework for examining statistical power under various types of commonly-used experimental designs that are conducted in a school setting, and discuss appropriate precision standards. In our discussion, we provide examples from recent large-scale experimental evaluations of education programs. We provide also empirical estimates of key parameters (such as intraclass correlations and regression R^2 values) that are required to estimate power levels. Using conservative values of these estimates, we conduct a power analysis for each of the considered designs.

Our empirical analysis focuses on achievement test scores of elementary school and preschool school in low-performing school districts due to the accountability provisions of the No Child Left Behind Act of 2001. The Act mandates the annual testing of all students in grades 3 to 8 and the development of initiatives to improve the literacy of preschool and K-3 children. Thus, there has been an ensuing federal emphasis on testing interventions to improve reading and mathematics scores of young students. Furthermore, more information exists to determine appropriate precision standards for student test scores than other student outcomes. Our analysis focuses also on designs with a single treatment and control group per site, which is the most common design used in education evaluations. Our methods, however, can be easily generalized to experimental designs with multiple treatment groups.

This paper is in five sections. First, we discuss general issues for a statistical power analysis, including procedures for assessing appropriate precision levels. Second, we discuss reasons that a clustered design reduces the statistical power of impact estimates and provide a simple mathematical formulation of the problem. Third, we discuss procedures that can be used to reduce design effects. Fourth, we present power calculations for impact estimates under various design options and parameter assumptions. Finally, we present our conclusions.

A. 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. Precision levels for most evaluations of education interventions are a particularly important issue, because it is often the case that schools or classrooms are randomly assigned to a research condition rather than students, which generates design effects from the clustering of students within groups.

In order 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. Researchers often scale nominal impact estimates into standard deviation units to facilitate the comparison of findings across outcomes that are measured on different scales. Hereafter, we denote minimum detectable impacts in effect size units as "MDEs."

This paper focuses on the calculation of MDEs. Next, we discuss the structure of MDEs and appropriate precision standards for standardized effect sizes.

1. Structure of MDEs

MDEs represent the smallest program effects that can be detected with a high degree of confidence. MDEs are a function of the standard errors of the impact estimates, the assumed significance level (Type I error), the assumed power level (Type II error), and the number of degrees of freedom for conducting tests gauging the statistical significance of the program impacts. Mathematically, the MDE formula can be expressed as follows:

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

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)*.[1] *Factor(.)* becomes larger as the significance level is decreased and as the power level is increased. Appendix Table A.1 displays values for *Factor(.),* by the number of degrees of freedom, for one-tailed and two-tailed tests, at 80 and 85 percent power and a 5 percent significance level (which are typical assumptions that are used in MDE calculations).

We note that equation (1) ignores the estimation error in the standard deviation (that is, it assumes that σ is known). Hedges (2004) uses a more sophisticated ratio estimator that accounts for the estimation error in the standard deviation. His resulting variance formulas are very similar to the case where σ is assumed to be known, except that it includes an additive correction factor that reflects the estimation error in σ. This correction factor, however, is very small in most practical applications and also depends on the true (but unknown) effect size. Thus, for simplicity, we do not account for it in our presentation.

Before discussing issues pertaining to *Var(impact)*, we first discuss several issues pertaining to *Factor(.)* that affect our power calculations, including the use of one-tailed or two-tailed tests, accounting for multiple comparisons, and the number of degrees of freedom.

a. Using a One-Tailed or Two-Tailed Test

For a given significance level and power level, the use of one-tailed tests produces smaller MDEs than the use of two-tailed tests (see Appendix Table A.1).[2] This is because under a one-tailed test, the rejection region for the null hypothesis of no program impact is concentrated in only one tail of the distribution of the outcome measure, whereas the rejection region under a two-tailed test is concentrated in both the lower and upper tails of the distribution.

[1] Specifically, *Factor(.)* can be expressed as $[T^{-1}(\alpha) + T^{-1}(\beta)]$ for a one-tailed test and $[T^{-1}(\alpha/2) + T^{-1}(\beta)]$ 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).

[2] The value of *Factor(.)* is the same for a two-tailed test at an α significance level and for a one-tailed test at an $\alpha/2$ significance level.

For several reasons, however, our illustrative power calculations presented in this paper focus on two-tailed tests rather than one-tailed tests. First, it is often unclear a priori whether a particular intervention will improve all student outcomes. Second, a two-tailed test provides more conservative estimates to help guard against unexpected events that might reduce the size of the analysis samples. Third, researchers typically employ two-tailed tests when conducting statistical tests in impact analyses (even if one-tailed tests were used in the initial power calculations).

We note that power calculations in program evaluations are sometimes conducted using one-tailed tests. The use of one-tailed tests is often justified on the grounds that an intervention should be supported only if it produces beneficial impacts, so that harmful impacts have the same policy significance as zero impacts.

b. Adjusting Significance Levels for Multiple Comparisons

MDE calculations are typically performed assuming a 5 percent significance level. However, this Type I error can be viewed as being too large when experiments test the relative effectiveness of more than one intervention by randomly assigning multiple treatments to units (such as schools or classrooms). This is because with multiple comparisons, the chance of finding *any* statistically significant impact, even when none actually exists, is much higher than 5 percent. For example, suppose four different interventions and a control condition were randomly assigned to schools. In this example, there are 5(5-1)/2 = 10 pairs of treatment and control group means to compare, each with a 5 percent probability of a Type I error. In this case, if several t-tests are performed, the probability that *at least one* of these tests is significant is much greater than five percent. For example, assuming independent t-tests, the probability that at least one of these 10 tests is significant is 40 percent $[(1 - (1-.05)^{10}]$. Although this estimate is an upper bound (because it assumes independent tests), it demonstrates that there is a good chance that the evaluation will conclude that a particular intervention is superior, when in fact, all interventions are indistinguishable from each other and from the control condition. This erroneous finding could have important policy ramifications.

To correct for this multiple comparisons problem, the α level could be set lower than 5 percent when calculating MDEs. A lower α level, however, increases *Factor(.)*, and hence, increases MDEs and the required sample sizes for the evaluation. One widely-used method is to use the Bonferroni inequality and to set the α level at 5 percent divided by the number of tests that are conducted. This approach is conservative because it assumes independent tests, but ensures that the probability of erroneously finding *any* significant impacts across the multiple tests will be less than 5 percent. Less conservative methods have been developed to adjust for correlations among the tests (see, for example, Ramsey 2002).

Similar correction procedures could be used also in education evaluations that examine impacts on multiple outcome measures. The corrections could be made when examining outcomes within a similar domain or for priority outcomes. An alternative procedure is to use factor or cluster analytic techniques to construct a small number of composite outcome measures to help reduce the multiple comparisons problem.

Finally, a related issue concerns the estimation of impacts for subgroups defined by baseline student characteristics (such as gender, race/ethnicity, family income, baseline test scores, etc.), that are often calculated in experimental evaluations of education programs. Whether to adjust probability levels for multiple comparisons for these subgroup analyses depends on the research question. If the research question is, "Does the intervention work for a subgroup in isolation," then α level corrections are not needed. On the other hand, if the research question is, "Does the intervention work better for one subgroup than another," and if the program intends to use the subgroup results to target services to selected students only, then it is appropriate to make the multiple comparison corrections.

c. Number of Degrees of Freedom

As shown in Appendix Table A.1, *Factor(.)* is essentially constant if the number of degrees of freedom is relatively large (for a given α and β). However, *Factor(.)* becomes larger if the number of degrees of freedom is small. For instance, for a two-tailed test at 80 percent power and a 5 percent significance level, *Factor(.)* is about 3.1 for 10 degrees of freedom, 2.9 for 20 degrees of freedom, and 2.8 for 100 degrees of freedom, but is 3.7 for 4 degrees of freedom.[3] *Factor(.)* is about 7 percent larger for tests at 85 percent power than 80 percent power.

In a *nonclustered* experimental design, where students within a given population are randomly assigned directly to a research group, the number of degrees of freedom, df_{NC}, can be expressed as follows:

$$(2)\quad df_{NC} = Total\ Number\ of\ Students - Number\ of\ Strata - 1.$$

Thus, under this design, *Factor(.)* is effectively constant if the sample contains at least 25 or 30 sample members, which is usually the case.

Under a group-based design with a *single* treatment and control group, the number of degrees of freedom, df_C, is typically expressed as (Murray 1998):

$$(3)\quad df_C = Total\ Number\ of\ Groups - Number\ of\ Strata - 1.$$

Thus, under a clustered design, *Factor(.)* does not vary if the number of groups is relatively large and if the number of strata is relatively small. The situation, however, is different if only a small number of groups (schools or classrooms) are randomly assigned to a research condition. In this case, *Factor(.)* becomes larger (that is, precision levels are reduced).

[3] The corresponding figures for a one-tailed test are somewhat smaller: 2.7 for 10 degrees of freedom, 2.6 for 20 degrees of freedom, 2.5 for 100 degrees of freedom, and 3.1 for 4 degrees of freedom.

For example, in the Social and Character Development (SACD) Research Program (Schochet et al. 2004), about 10 elementary schools per site were randomly assigned to either a treatment group (who will offer a promising SACD intervention designed to improve positive social and character development) or to a control group (who will offer the current curriculum), with equal numbers of schools assigned to each research group. Furthermore, pairwise matching was used to select the treatment and control group schools (that is, five stratum of school pairs were formed, and one school within each pair was randomly assigned to the treatment group and the other to the control group). Thus, for the SACD evaluation, the number of degrees of freedom at the site level is 4 (10 schools minus 5 strata minus 1). Consequently, *Factor(.)* is about 3.7 rather than the typical 2.8 value, which has important power implications.

2. Precision Standards

A key issue for any evaluation is the precision standard to adopt for the impact estimates. There are two key factors that need to be considered in selecting a precision standard for a particular study. First, it depends on what impact is deemed meaningful in terms of future, longer-term student outcomes (such as high school graduation, college attendance, earnings, welfare receipt, criminal behavior etc.). Second, the precision standard should depend on what intervention effects are realistically attainable.

These two factors will depend on the key study outcome measures and the study context. For example, in a medical trial where death is the key outcome, small impacts are clearly meaningful, whereas larger standardized effect sizes might be appropriate in education trials. Similarly, in terms of attainability, some student outcomes are harder to influence than others. For instance, it might be more difficult for an intervention to improve test scores than student attitudes, so smaller effect size targets are more appropriate for studies focusing on test scores.

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.

a. Examine Impact Results from Previous Evaluations

One approach for adopting a precision standard is to use impact results found in previous evaluations similar to the one under investigation. For instance, to evaluate the impacts of a reading intervention on elementary school children, one could adopt a precision standard based on impact results from previous evaluations of similar reading interventions that were tested on a similar student population. This approach is appropriate if the previous impact studies produced credible results based on rigorous evaluation designs, and if the studies found beneficial and meaningful program impact estimates.

Another 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) suggested that effect sizes of .20 are small, effect sizes of .50 are moderate, and effect sizes of .80 are large. In an important study, Lipsey and Wilson (1993) examined the

distribution of effect size estimates reported in 9,400 studies (with more than 1 million individual subjects) testing the efficacy of various psychological, educational, and behavioral interventions. They found that one-third of the effect sizes were smaller than .32, one-third were between .33 and .55, and one-third were between .56 and 1.20.

Based on these studies, many evaluations of education programs adopt standardized effect sizes of .20, .25, or .33 as the precision standard. While this meta-analysis approach can be used to determine what impacts could be attainable for a particular intervention, it does not necessarily address what impacts are meaningful. As discussed next, we believe that these precision standards are somewhat high for testing the efficacy of education interventions on student test scores.

b. Adopt a Benefit-Cost Framework

One approach for assessing meaningful standardized effect sizes, and which suggests smaller benchmark precision standards are appropriate, is to select samples large enough to detect impacts such that program benefits would offset program costs. This approach could be used in studies where a dollar value can be assigned to key program benefits. 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 percent higher earnings when the students join the labor market (Currie and Thomas (1999); Murnane, Willet, and Levy (1995); Neal and 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.[4,5] Consequently, the present value of lifetime earnings would be $12,375 if the intervention improved test scores by .33 of a standard deviation, $7,500 for an impact of .20 standard deviations, and $3,750 for an impact of .10 standard deviations. Stated differently, if an intervention improved test scores by .20 standard deviations, then Krueger's estimates suggest that program benefits would exceed program costs if the intervention cost less than $7,500 per pupil—which was roughly the nationwide *total* expenditures per pupil in 1997-98. 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. Stated differently, the evaluation could miss an effect worth finding if the precision standard was .20.

These results, however, must be interpreted cautiously, because there is only a small literature on the long-term economic returns to test score increases for elementary school children. Furthermore, many of the studies cited above pertain to older students only, and it is likely that the test score-earnings relationship is stronger for older students than younger ones. For instance, Murnane, Willet, and Levy (1995) used data from the High School and Beyond survey to estimate the economic returns to test score increases using male high school *seniors*.

[4] This figure was calculated (1) using the age-earnings profile in the March 1999 Current Population Survey, (2) a 4 percent discount rate, (3) assuming workers begin wage at age 18 and retire at age 65, and (4) a productivity (wage) growth rate of 1 percent per year.

[5] Kane and Staiger (2002) find similar results.

Similarly, Neal and Johnson (1996) used the National Longitudinal Survey of Youth to estimate the effect of students' AFQT scores at age *15 to 18* on the students' earnings at age 26 to 29. Furthermore, although Currie and Thomas (1999) examined the relationship between test scores at age 7 and earnings at age 33, they used data from the British National Child Development Study. Thus, their results may not pertain to students in the United States.

Consequently, although these studies suggest that relatively small test score gains for elementary school children are associated with relatively large lifetime earnings gains, these results must be deemed tenuous. Much more research is needed to examine the test score-earnings relationship using data collected on samples of pre-school and elementary school children as they enter adulthood and beyond.

c. Examine the Natural Progression of Students

Another approach is to adopt a precision standard based on the natural growth of student outcomes over time to get a sense of intervention effects that can be realistically attained and that are meaningful. This approach, however, can only be used for those outcome measures that can be compared over time and that naturally change over time.

Several studies suggest that the test performance of elementary school students in math and reading grows by about .70 standard deviations per grade. Kane (2004) compared Stanford 9 achievement reading and math test scores across elementary school grades (where the scores were "scaled" to allow comparisons of scores across grades). He found that test performance grew by approximately .70 standard deviations in math and .80 standard deviations in reading *per grade level*. However, the rise in test scores was smaller after third grade; between fifth and sixth grades, performance grew by only .30 standard deviations in both math and reading. We found similar results using scaled SAT-9 test score data from the Longitudinal Evaluation of School Change and Performance (LESCP) in Title I schools; the average reading and math test score gain between the third and fourth grades was about .70 standard deviations.[6]

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 to 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. For instance, an effect size of .10 corresponds to about 1 to 1.5 months of instruction.

d. Examine the Distribution of Outcomes Across Schools

Another metric for assessing an appropriate precision standard is to assess what an MDE implies about movements in mean student outcomes in a typical school relative to the

[6] Kane's results are based on separate cross-sections of students (which could be affected by cohort effects), whereas the LESCP results are based on the same students over time.

distribution of outcomes across a broader set of schools. This approach again suggests that effect sizes of .20 to .33 are large.

For instance, we 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 percent 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.[7] Similarly, a 20 percent effect size would move that school to the 33rd percentile. A more attainable 10 percent effect size would move the school from the 25th to 29th percentile. LESCP data for SAT-9 scores of third graders in Title I schools yield similar findings.

A related method is to assess the magnitude of MDEs by translating them into nominal impacts for *binary* outcomes (such as the percentage of students with test scores below a certain threshold level). For example, according to the National Assessment of Educational Progress (NAEP), nearly 70 percent of fourth graders nationally performed below the *Proficient* level in reading and math in 2003 (NCES 2004). For this binary outcome, effect sizes of .33, .25, and .20 translate into impacts of about 15.0, 11.5, and 9.2 percentage points, respectively. Stated differently, an effect size of .33 implies that the intervention must reduce the percentage of students scoring below the *Proficient* level from 70 to 55 percent, which is a large reduction. A smaller 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.

Finally, our discussion has focused on precision standards for comparing one treatment group to one control group. It is more difficult to develop rules for adopting a precision standard for comparing *across* treatment groups in experiments with multiple treatments, because this will depend on the nature of the interventions being tested. However, we can expect impacts to be smaller when treatments are compared to each other than when a treatment is compared to the control condition. Thus, MDEs should be set lower for power analyses that focus on between-treatment contrasts.

B. VARIANCE CALCULATIONS FOR GROUP-BASED EXPERIMENTAL DESIGNS

As discussed, MDEs are proportional to the standard errors of the impact estimates. Under a group-based, clustered design, standard errors are typically larger than those under a nonclustered design of the same size, and thus, clustering usually increases MDEs. The

7 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.

clustering of students within schools increases standard errors, because students who live in the same communities and face the same school environments tend be similar. The clustering of students within classrooms also increases standard errors, because of teacher effects and the possibility that schools group similar types of students in the same classrooms. Thus, the precision of estimates under a clustered design is reduced, because the variance expressions must account not only for the variance of outcomes across students, but *also* for the variance of average student outcomes across schools and across classrooms within schools.

In this section, we present a simple, unified mathematical formulation to demonstrate the sources of variance under various types of designs that are typically used in impact evaluations of education programs. The designs are, in general, ordered from least to most clustered. To make the presentation concrete, we consider experimental designs where the following units are randomly assigned to a research status:

- Students within sites (schools or districts)
- Classrooms within schools
- Schools within districts

Clustering in these designs comes from two potential sources: (1) the random *assignment* of units to the treatment and control groups, and (2) the random *sampling* of units from a broader universe of units before or after random assignment takes place. As part of our presentation, we discuss the important issue of when it is appropriate in the variance calculations to treat group effects as random or fixed, which has important implications for the statistical power of the designs.

For ease of exposition, we first demonstrate the variance formulas assuming that the evaluation sample is selected from a single participating site—such as a school or school district. We then indicate how to generalize the variance formulas when aggregating the sample across multiple sites to obtain pooled estimates. As discussed, we assume designs where a *single* treatment is tested against the control condition within each site.

Table 1 summarizes the various designs that we consider, and displays equation numbers in the text for the variance formulas for each design. These designs can all be estimated using standard statistical packages (see, for example, Murray (1998) and Singer (1998)).

1. 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. This design was used in the evaluation of the 21st Century Community Learning Centers Program (Dynarski et al. 2004), where interested students within each of the study schools were randomly assigned to a either a treatment group (who could attend an after-school program) or a control group (who could not). Another example of this design is the Impact

TABLE 1

SUMMARY OF ALTERNATIVE DESIGNS

Design Designation and Unit of Random Assignment	Fixed or Random Site, School, and Classroom Effects	Sources of Clustering	Equation Numbers for Variance Formulas
Design I: Students Within Sites (Schools or Districts)	Fixed Site Effects	None	Equation 6
II: Students Within Sites	Random Site Effects	Sites	8, 10
III: Students Within Sites	Random Site and Subunit Effects	Sites; Subunits[b]	13
IV: Classrooms Within Schools[a]	Fixed or No School Effects[b]	Classrooms	14
V: Classrooms Within Schools (At Least 2 Classrooms per School)	Random School Effects	Schools, Classrooms	15
VI: Classrooms Within Schools (Only 2 Classrooms per School)	Random or Fixed School Effects	Schools	15 with $\rho_2 = 0$
VII: Schools Within Districts	Fixed Classroom Effects	Schools	16
VIII: Schools Within Districts	Random Classroom Effects	Schools, Classrooms	17

Note: All designs assume fixed school district effects, except for Designs I and II where sites are school districts.

[a]This design is pertinent if (1) there are at least two treatment and control classrooms per school and school fixed effects are included in the analysis, or (2) if there is only 1 classroom per condition per school, but school fixed effects are not included in the analysis.

[b]Subunits are classrooms when sites are schools, and schools when sites are school districts.

Evaluation of Charter Schools Strategies (Gleason et al. 2004) where, within each charter school area, students interested in attending a charter school will be randomly assigned to either a treatment group (who will be allowed to enroll in a charter school) or a control group (who will not).

Clearly, 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 (that is, there is no "diffusion of treatments," as denoted by Cook and Campbell 1979).

Under these types of designs, an important issue is whether the variance calculations should account for school- or classroom-level clustering. There are two views on this issue. First, if the impact findings are to be generalized *only* to the specific classrooms and schools included in the study (the fixed effects case), then clustering is not present, even though sample members are grouped in the same classrooms and schools. This is because students in the treatment and control groups are expected to be spread across all classrooms and schools in the sample. Thus, bypassing the selection of classrooms or schools removes the link between students and classrooms/schools, and thus, direct inferences can be made about intervention effects that pertain only to students in the study samples.

The other view is that the impact findings can be generalized to a broader population (or "superpopulation") of classrooms and schools "similar" to the ones included in the study. In this view, students and teachers change over the short term, and the ones that are observed at a fixed time point are a representative sample from this larger population. In this case, the variance estimates should account for classroom- or school-level clustering.

In this section, we consider designs *without* school- or classroom-level clustering—which we label *nonclustered, stratified* designs. First, we discuss the appropriate variance calculations for impact estimates within sites (strata), and then for impact estimates pooled across sites.

a. Variance of Impact Estimates Within Sites

Under a nonclustered, stratified design, the variance of an impact estimate within a site—that is, the variance of the difference between a mean outcome across the treatment and control groups—must account for between-student variance only, and can be expressed as follows:

$$(4)\quad Var(\textit{impact in site } p) = \frac{2\sigma_p^2}{m_p},$$

where m_p is the size of the treatment (control) group in site p and σ_p^2 is the variance of the outcome measure.[8,9]

b. Variance of Pooled Impact Estimates

Pooled impact estimates are often obtained in impact evaluations conducted in multiple sites in order to examine the extent to which, taken together, the tested interventions change student outcomes relative to what they would have been otherwise. In many instances, estimating pooled impacts is appropriate, 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, a charter or magnet school, a teacher preparation model, or a social and character development initiative), 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, and site-specific impacts must be examined separately to assess whether the pooled impacts are driven by a small number of sites.

A central issue in the variance calculations for the pooled estimates is whether site effects should be treated as fixed or random. For most evaluations of education programs, sites (such as schools or school districts) 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). In these instances, the variance calculations hinge critically on whether the pooled estimates are viewed as generalizing to the study sites only (the fixed effects case) or to a broader population of sites similar to the study sites (the random effects case). In the fixed effects case, between-site variance terms do not enter the variance calculations (because in repeated "sampling," the same, fixed, set of sites would always be "selected"), unlike the random effects case where the study sites constitute a random sample, or a least a representative sample, from some larger population.

Although this issue needs to be addressed for each study, we believe that the fixed effects case is usually more realistic in evaluations of education interventions. Most evaluations are efficacy trials where a relatively small number of purposively-selected sites are 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 policymakers on potentially promising interventions. Instead, we 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

[8] We assume equal numbers of treatment and control group students for illustrative simplicity, and because for a given total research sample size, a 50:50 split between the treatment and control groups yields the most precise estimates. We follow this approach for the remainder of this section, although we discuss unbalanced designs later in this paper. The formulas for unbalanced designs are very similar to the ones presented in this paper.

[9] We discuss the use of the finite sample correction (equal to 1 minus the proportion of the population being selected) in the variance calculations later in this paper.

impact estimates across sites (for example, by calculating the percentage of sites with beneficial impacts). 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 *are* effective.

Pooled impact estimates in the fixed-effects case are calculated as a weighted average of the impact estimates in each site. The associated variances are obtained by aggregating the site-specific variances in equation (4) as follows:

$$(5)\quad Var(\textit{pooled impact}) = \sum_{p=1}^{s} w_p^2 \frac{2\sigma_p^2}{m_p},$$

where w_p is the weight associated with site p and 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 (Fleiss 1986).

To reduce notation, and to facilitate comparisons with the other designs discussed below, we will refer to the following simplified version of equation (5):

$$(6)\quad Var(\textit{pooled impact}) = \frac{2\sigma^2}{sm},$$

where σ^2 is the average variance across the s sites, and m is the average number of treatment or control group members per site.

To further demonstrate the appropriate calculations in the fixed-effects case, the following regression (ANOVA) model can be used for estimating pooled impacts across sites (schools or districts):

$$(7)\quad Y_{ip} = \lambda_0 + \sum_{p=2}^{s} \lambda_{0p} D_{ip} + \sum_{p=1}^{s} \lambda_{1p} (D_{ip} * T_{ip}) + e_{ip},$$

where Y_{ip} is a continuous, posttreatment outcome measure for student i nested in site p (nested in the treatment or control condition), D_{ip} is an indicator variable equal to 1 for those in site p, T_{ip} is an indicator variable equal to 1 for treatment group members, and e_{ip} are assumed to be *iid* $N(0,\sigma^2)$ student-level random error terms. In this model, site-specific impacts are treated as *fixed* (not random) and are represented by the λ_{1p} parameters. Pooled impact estimates are then calculated as a simple (or weighted) average of the site-specific impact estimates, and similarly

for the estimated variances. Thus, in this model, the variance estimates are *not* inflated to account for between-site effects.

We note that the fixed versus random effects issue is more complex in instances where a large number of purposively-selected sites are included in the study. For example, in the Impact Evaluation of Charter School Strategies (Gleason et al. 2004), a large number of geographically-dispersed school districts will be included in the evaluation. If available data indicate that the characteristics of these school districts are similar to the larger population of school districts with charter schools, then it might be reasonable to include between-district effects in the variance calculations. However, even in these instances, we believe that this approach should be supplementary to the primary approach discussed above, and should be used only to check the robustness of study findings.

Finally, for several reasons, it might be appropriate to account for between-site variance terms when conducting the power analysis during the design phase of an evaluation. First, this approach provides a guide to the number of sites that should be selected for the study. This is important, because for a given sample size, the fixed-effects approach generates the *same* precision levels for a design with many sites and only a few students per site, and a design with a smaller number of sites but with more students per site. Second, incorporating between-site variance terms is conservative, because it will generate larger sample sizes to help guard against unexpected events that could reduce the size of the analysis sample during the follow-up period.

2. Random Assignment of Students Within Sites: Random-Effects Case

In this section, we consider designs where students are randomly selected to a research condition within sites, and where site effects are treated as random. This random-effects case can occur in two ways. First, as discussed, purposively-selected sites could be considered representative of a broader population of similar sites. Second, in some evaluations, sites are *randomly sampled* from a larger pool of sites. This type of design is typically employed in large-scale studies of a well-established program or intervention that require externally-valid impact estimates (and where the burden of evidence of program effectiveness is set high). For example, for the national evaluation of Upward Bound (Myers et al. 1999), a nationally representative sample of eligible program applicants was selected in two stages. In the first stage, a random sample of Upward Bound sites (projects) was selected from all sites nationwide, and in the second stage, students within each of the selected sites were randomly assigned to either a treatment or control group. For this evaluation, the impact results are generalizable to all Upward Bound projects nationwide. Similarly, for the National Job Corps Study (Schochet et al. 2001), all eligible Job Corps applicants nationwide in 1995 were randomly assigned to a research condition.

In these random-effects designs, study results can be generalized more broadly than in the fixed-effects designs. However, this generalization involves a cost in terms of precision levels: the variance formulas must be inflated to account for between-site effects. Stated differently, site effects must be treated in the variance formulas as *random*, not fixed. Intuitively, in repeated sampling, a different set of sites would be selected for the evaluation, which could influence the

impact findings. Hence, the variance expressions must account for the extent to which mean student outcomes vary across sites.

To illustrate the variance calculations under a random-effects design, we first consider the scenario where (1) site (school or district) effects are random, (2) students within sites are randomly selected to a research condition, and (3) there is no clustering of students within subunits (classrooms or schools). In this case, the variance formula for a pooled impact estimate can be expressed as follows:

$$(8)\quad Var(pooled\ impact) = \frac{\sigma_\tau^2}{s} + \frac{2\sigma_e^2}{sm},$$

where s is the number of sites in the sample, m is the average number of treatment or control group members in each site, σ^2_e is the variance of the outcome measure for students within sites, and σ^2_τ is the variance of the *impacts* (treatment effects) across sites.

The within-site variance term in equation (8) is the conventional variance expression for an impact estimate under a nonclustered design (see equation (6)). Design effects in a clustered design arise because of the first variance term (that is, the between-site term), and can be large because the divisor in this term is the number of sites rather than the number of sample members. Thus, precision levels can usually be improved by selecting more sites (for example, schools) and fewer students per site (to the extent that project resources allow). The optimal allocation of sites and students can be obtained by minimizing equation (8) subject to a budget constraint that includes unit costs of including an additional site and an additional student (Raudenbush 1997).

To make equation (8) more operational for our power calculations (and for purposes of comparing variance formulas across other designs), we use the following expression for σ^2_τ:

$$(9)\quad \sigma_\tau^2 = 2\sigma_u^2(1-c_1),$$

where σ^2_u is the variance of the mean outcome measure (not impacts) across sites (which is assumed to be equal for the treatment and control groups), and c_1 is the *correlation* between the treatment and control group means within a site.[10] This correlation is likely to be positive because students in the same site (for example, school) are likely to have similar characteristics, have similar teachers, and face similar environments.

[10] This expression can be derived by noting that the variance of the impact across sites is the sum of the (1) the variance of the mean outcome for the treatment group across sites; (2) the variance of the mean outcome for the control group across sites (which is assumed to be roughly the same as that for the treatment group in step (1)); and (3) –2 times the covariance of the treatment and control group means within a site.

If we insert equation (9) into equation (8), and define ρ_1 as the between-site variance in the outcome measure (σ^2_u) as a proportion of the total variance of the outcome measure (σ^2), then the variance formula can be expressed as follows:

$$(10)\quad Var(\textit{pooled impact}) = \frac{2\sigma^2\rho_1(1-c_1)}{s} + \frac{2\sigma^2(1-\rho_1)}{sm},$$

where $\sigma^2 = \sigma^2_u + \sigma^2_e$. The term, ρ_1, is the *intraclass correlation (ICC)*, which tends to be large if mean student outcomes vary considerably across sites, and tends to be small if site means are similar.

In this formulation, the design effect from clustering is small (that is, near 1) if either the mean of the outcome measure does not vary across sites (that is, if ρ_1 is small), or if the correlation between the treatment and control group means within a site is large and positive (that is, if c_1 is near 1). A large correlation implies that impacts do not vary across sites.

The variance expressions in equations (8) and (10) can be derived using the following two-level hierarchical linear (HLM) model (Bryk and Raudenbush 1992):

$$(11)\quad \begin{aligned} \textit{Level}\,1:\ & Y_{ip} = \lambda_{0p} + \lambda_{1p}T_{ip} + e_{ip} \\ \textit{Level}\,2:\ & \lambda_{0p} = \lambda_0 + u_p \\ & \lambda_{1p} = \lambda_1 + \tau_p, \end{aligned}$$

where Level 1 corresponds to students and Level 2 corresponds to sites (schools or districts). The term, Y_{ip}, is the continuous outcome measure for student i in site p; T_{ip} is an indicator variable equal to 1 for treatment group members and 0 for controls; u_p are assumed to be *iid* $N(0,\sigma^2_p)$ site-specific random error terms; τ_p are *iid* $N(0,\sigma^2_\tau)$ error terms which represent the extent to which *treatment effects vary across sites*; e_{ip} are *iid* $N(0,\sigma^2_e)$ within-site error terms that are distributed independently of u_p and τ_p; and the λ terms are parameters.

Inserting the Level 2 equations into the Level 1 equation yields the following unified regression model:

$$(12)\quad Y_{ip} = \lambda_0 + \lambda_1 T_{ip} + [u_p + T_{ip}\tau_p + e_{ip}].$$

In this formulation, λ_1 represents the pooled impact estimate (that is, $[Y_{..T} - Y_{..C}]$ where $Y_{..T}$ and $Y_{..C}$ represent mean outcomes for the treatment and control groups, respectively), and its associated variance is $(\sigma^2_\tau/s + \sigma^2_e/ms)$ which is identical to equation (8). In this model, the random school and treatment effects—u_p and τ_p—are a component of the error structure and

account for the clustering of students within sites. This is very different from the fixed-effects specification (see equation (7)) where site effects are *not* included in the error structure, but are treated as fixed parameters in the regression model.

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 this design could be represented as a three-level HLM model, where Level 1 corresponds to students (the level of random assignment), Level 2 to classrooms, and Level 3 to schools. This design effectively treats students as if they were randomly assigned to the treatment and control groups within classrooms. This framework yields the following variance formula:

$$(13)\quad Var(\textit{pooled impact}) = \frac{2\sigma^2\rho_1(1-c_1)}{s} + \frac{2\sigma^2\rho_2(1-c_2)}{sk} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{sk(.5n)},$$

where ρ_2 is the between-classroom effect, c_2 is the correlation between the outcomes of treatment and control group students within classrooms, k is the average number of classrooms per school, n is the average number of students per classroom (split evenly between the treatment and control groups), and where other parameters are defined as above. This expression accounts not only for the extent to which treatment effects vary across schools, but also the extent to which treatment effects vary across classrooms within schools. In this random-effects framework, additional variance terms could be included to account for potential "treatment-induced" correlations between the outcomes of treatment group members if the intervention is administered in small groups, thereby creating potential correlations between the outcomes of treatment group members within each small group (Murray et al. 2004; Raudenbush 1997). These correlations could be modeled as another level in the HLM framework. Finally, additional variance terms at the level of the school district could also be included in the variance formulas if district effects were treated as random.

Importantly, simulation studies (Murray et al. 1996) suggest that that Type I errors for statistical tests of intervention effects are similar if the variance expressions account for clustering only at the highest level of clustering, and if they account also for clustering of intermediate nested subunits. These findings suggest that empirical results based on equations (10) and (13) could be similar (as long as covariates at the classroom level are not included in the regression models).

Finally, the above analysis suggests that in some evaluations, researchers face a conundrum about whether or not to randomly select study sites. For example, suppose an evaluation is being conducted in a small purposively-selected city. Furthermore, suppose that *all* schools in that city agree to participate in the study. In this case, one can argue that the study schools should be selected randomly, so that the impact results can be generalized to all schools in that city. On the other hand, for a fixed sample size, selecting schools randomly rather than purposively will inflate the standard errors of the impact estimates, which will reduce the chance that the study will find statistically significant impact estimates. Furthermore, one might argue that in an efficacy study, it might not make much difference from a policy standpoint whether the results

can be generalized to all schools in the small city or to only those schools that are selected for the study. Clearly, the choice of whether to select sites randomly or not will depend on the scope and objectives of the study. However, in making this important design decision, it is important to consider the tradeoff between statistical power and the generalizability of study findings.

3. 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, in the Evaluation of the Effectiveness of Educational Technology Interventions (Dynarski et al. 2004), teachers in participating schools will be assigned at random to use a technology intervention or not. This type of design is appropriate for interventions that are administered at the classroom level and where potential spillover effects are deemed to be small.

One way to interpret this design is that a "mini-experiment" is being conducted within each school. Under a design with purposively-selected schools and where school effects are treated as fixed, pooled impact estimates across schools are calculated as a simple or weighted average of the impact estimates from each mini-experiment. Accordingly, the variance formula for these pooled impact estimates can be expressed as follows:

$$(14)\quad Var(pooled\ impact) = \frac{2\sigma^2\rho_2}{s(.5k)} + \frac{2\sigma^2(1-\rho_2)}{s(.5k)n},$$

where s is the total number of schools in the sample, k is the number of classrooms per school (split evenly between the treatment and control groups), n is the average number of students per classroom, and ρ_2 is the between-classroom variance as a proportion of the total variance.[11] Design effects arise because of the between-classroom variance term.

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 (because there is no sampling of students within classrooms). 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.

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, for the Teach For America (TFA) Evaluation (Decker et al. 2004), students within each of the study schools and

[11] The variance of an impact estimate within a single school can be obtained by setting s equal to 1 in equation (14).

grades were randomly assigned to classrooms taught by TFA teachers or to classrooms taught by other teachers. This design ensures that the average baseline characteristics of students in the treatment and control group classrooms are similar. While this design reduces classroom effects, it does *not* remove them. This is because classroom effects arise from two sources: (1) differences in the quality of teachers within schools, and (2) 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 both school-level and classroom-level clustering are present. Using results from the previous section, the variance expression can be expressed as follows:

$$(15)\quad Var(pooled\ impact) = \frac{2\sigma^2\rho_1(1-c_3)}{s} + \frac{2\sigma^2\rho_2}{s(.5k)} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{s(.5k)n},$$

where c_3 is the correlation between the treatment and control group *classroom* means within a school, and where other parameters are defined as above. For two reasons, this variance is larger than the corresponding variance in equation (13) under the design where students are the unit of random assignment. First, twice as many treatment and control group classrooms are in the sample when students are the unit of random assignment. Second, unlike equation (15), the classroom-level term in equation (13) is deflated by the correlation between the outcomes of treatment and control group students within the same classrooms.

Finally, due to limitations in the number of available classrooms, it is often the case that only one treatment and control classroom can be selected per school. In this case, there are not enough degrees of freedom to estimate between-classroom effects within schools, which are confounded with between-school effects (Murray 1998). One approach is to set ρ_2 to zero in equation (15) and to use the resulting variance formula for *either* the random or fixed effects specifications. Another possibility for the fixed effects specification is to use equation (14) and ignore the stratification by school (that is, by not including fixed school effects in the regression models). In this case, the between-classroom effect is estimated by combining classrooms across schools, which could increase design effects due to clustering. To mitigate these precision losses, another approach is to combine similar schools into larger strata, thereby making it possible to estimate between-classroom effects within stratum.

4. Random Assignment of Schools

In some designs, *schools* within districts are randomly selected to the treatment and control groups. These designs are necessary for testing interventions that are school-based. For instance, the Social and Character Development (SACD) Research Program is testing, in seven sites, promising interventions designed to promote positive social and character development and prevent negative behaviors among elementary school students. In each site, 10 to 18 schools were randomly assigned to either a treatment group (who are offering the SACD intervention) or to a control group (who are offering the current curriculum), with equal numbers of schools

assigned to each research group. This design is necessary to avoid contamination of the control group, because the SACD interventions include components aimed at changing schoolwide outcomes. Another example is the Evaluation of the Impact of Teacher Induction Programs (Johnson et al. 2005) where two models of high-intensity teacher induction are being tested in 20 high-poverty, large school districts across the country. Within each district, 10 elementary schools will be randomly selected to implement the high-intensity program, and 10 will be randomly selected to continue to receive whatever induction program their respective districts normally provide.

Next, we discuss the variance formulas for impact estimates under school-based experimental designs with and without classroom effects. We focus on designs where school districts volunteer for the study (that is, are selected purposively) and where district effects are treated as fixed, not random.

a. Clustering at the School Level Only

For a school-based experimental evaluation, one design option is not to sample classrooms within the treatment and control group 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. For example, under the SACD design, all consenting students in third-grade classrooms were included in the research sample.

In these designs, if the impact findings are to be generalized only to the study schools and classrooms at the time of sampling, there is clustering at the school level, but not at the classroom level. Intuitively, if sampling were repeated, a different random allocation of schools would be selected to the treatment and control groups, but not a different set of classrooms within schools. Consequently, the variance of an impact estimate within a district can be expressed as follows:

$$(16)\quad Var(\textit{impact in a district}) = \frac{2\sigma^2\rho_1}{.5s} + \frac{2\sigma^2(1-\rho_1)}{(.5s)kn},$$

where all parameters are defined as above.

The variance estimates under this school-level design are larger than those previously considered for two main reasons. First, there are now *half* as many treatment (control) schools (because random assignment occurs between schools rather than within them). Second, the between-school variance term is no longer deflated by the correlation between the treatment and control group means within schools.

Pooled impact estimates across districts can be calculated as a simple or weighted average of the district-specific impact estimates, and similarly for the associated variance estimates. The treatment of district effects as random would introduce additional design effects, because the variance formulas would need to contain district-level variance terms.

b. Clustering at the School and Classroom Level

For a school-based experimental evaluation, 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 those schools.

In the presence of both school- and classroom-level clustering, the variance formula can be now expressed as follows:

$$(17)\quad Var(\textit{impact in a district}) = \frac{2\sigma^2\rho_1}{.5s} + \frac{2\sigma^2\rho_2}{(.5s)k^*} + \frac{2\sigma^2(1-\rho_1-\rho_2)}{(.5s)k^*n},$$

where k^* is the number of sampled classrooms, and all other parameters are defined as above. Design effects arise in this design from the first and second variance terms, and hence, are larger than in the previous design with clustering at the school level only. It is noteworthy that neither the school- or classroom-level terms are deflated by correlations between the outcomes of the treatment and control groups. Additional variance terms are required if school district effects are treated as random. We note again, however, that there is some empirical evidence that in multi-stage clustered designs, variance estimates are similar if the variance formulas account for clustering at the highest level only or if they account also for clustering at lower levels (Murray et al. 1996).

5. Estimating Correlations

A critical issue for the MDE calculations is what estimates to use for the following correlations that enter the variance formulas:

- ρ_1 = The extent to which mean outcomes differ across schools (that is, the ICC at the school level)[12]
- ρ_2 = The extent to which mean outcomes differ across classrooms within schools (that is, the ICC at the classroom level)
- c_1 = The correlation between the mean outcomes of treatment and control group students within schools
- c_2 = The correlation between the mean outcomes of treatment and control group students within classrooms

[12] We have also considered designs that require ICCs at the district level (see Design II in Table 1). However, for this paper, we focus on ICCs at the school level, because this type of design is more common, and there is more empirical evidence on ICCs across schools than across school districts.

- c_3 = The correlation between the mean outcomes of treatment and control group classrooms within schools

As discussed, for policy reasons and the current research emphasis at ED, we focus our presentation on obtaining plausible correlation values for standardized math and reading test scores of elementary school and preschool students in low-performing schools. For context, we also discuss plausible correlation values for behavioral outcomes.

a. Intraclass Correlations

To obtain plausible values for ρ_1 for student achievement measures, we examined results found in the literature, and performed new tabulations using reading and math test score data from several recent evaluations conducted by Mathematica Policy Research Inc. (see Table 2).

We find that ICCs for standardized test scores vary somewhat by data source, and differ somewhat by grade level. The ICCs, however, typically become smaller when adjusted for district fixed effects, because these figures pertain to ICCs *within* districts rather than across all districts. Hedberg et al. (2004) show also that ICCs vary by region of the country and urban/rural status, although the pattern of the estimates across subgroups is not always clear. Consequently, the ICCs that are applicable for a specific power analysis will depend on the study context, and, in particular, on the homogeneity of the schools in the sample.

Nonetheless, the examined data sources suggest that values for ρ_1 often range from .10 to .20 for standardized test scores. Thus, in our illustrative power calculations below, we use the midpoint, .15, as a reasonable approximation for ρ_1. Because of the uncertainty in this parameter, however, we also present selected calculations assuming a more optimistic value of .10 and a less optimistic value of .20.

There is less evidence on plausible ρ_2 values because there are fewer data sources that have student-level data on multiple classrooms within schools (within a treatment condition). LESCP and TFA data suggest values of about .16 for ρ_2. Thus, ρ_2 values appear to be similar to ρ_1 values. Stated differently, mean student test scores tend to differ as much across classrooms within schools as they do across schools. This could be due to the fact that the examined data sources contain relatively homogenous schools in low-income districts and with low aggregate test scores. Thus, differences in teacher quality within these schools might have a large effect on student academic achievement. Our estimates for ρ_2, however, are based on only a small number of data sources, and an important future research topic is to estimate ICCs at the classroom level using additional data sources. In our power calculations, we assume the same values for ρ_2 and ρ_1.

Finally, for context, we examined the much larger literature on ICCs based on behavioral outcomes (see Murray et al. 2004 for a review). Siddiqui et al. (1996) present ICCs from a study of smoking prevention programs based on 6,695 seventh-graders in 287 classrooms from 47 schools. Outcomes examined include students' knowledge of health and tobacco, student's

TABLE 2

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 et al. (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
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
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	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

TABLE 2 *(continued)*

Data Source	Description of Data	Standardized Test Measure	Grade and Year	ICC Estimate
Data from Rochester: Figures Calculated from MDEs Reported in Bloom et al. (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 and Cook (2005)	22 schools	KCCT developed for Kentucky students	Grade not reported: 2003	Reading: .11
Preschools				
Early Reading First Evaluation	162 preschools in 68 sites	Expressive One Word Picture Vocabulary (EOW) Test; 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	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 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	ECLS-K	Kinder-garteners in spring 1999	**Adjusted**[a] Math: .17 Reading: .23

Note: Tabulations were conducted using SAS PROC MIXED.

[a] Adjusted for SES, race, and gender.

knowledge of social influences and resistance skills, and the prevalence of student smoking. Their analysis suggests a wide range of intraclass correlations, with ρ_1 ranging from .01 to .09 and ρ_2 ranging from .04 to .14 on the three outcomes listed above. Similarly, Aber et al. (1999) found in the evaluation of the Resolving Conflict Creatively (RCCP) Program that intraclass correlations for reports of students' aggressiveness, pro-social behavior, and hostile attribution bias ranged from about .02 to .06. Murray et al. (2003) found similar estimates based on 1,881 ICCs from 17 studies across a variety of outcome measures (tobacco, drug, and alcohol use; diet and nutrition, general health and personal factors). Finally, Ukoumunne et al. (1999) found using data from the Health Survey of England that ICCs for lifestyle risk factors were generally below .01 at the district health authority level. Thus, ICCs for behavioral outcomes appear to be somewhat smaller than those for academic outcomes.

b. Correlations Between Treatment and Control Group Means Within Schools

The parameters, c_1 to c_3 measure the correlation between treatment and control group outcomes under designs where school effects are treated as random. It is more difficult to estimate values for these correlations than for ρ_1 and ρ_2, because they depend on the relative effectiveness of the tested interventions across sites. However, these correlations tend to be positive and large, because intervention effects do not typically vary substantially across sites. For instance, in the evaluation of the 21st Century Community Learning Centers Program, the value of c_1 for math and reading test scores was about .85 across elementary schools and .70 across middle schools. Similarly, in the evaluation of the School Dropout Demonstration Assistance Program, the value of c_1 for student grades was .80. Finally, in the Early Head Start evaluation, c_1 was about .80 for Bayley scores and .70 for the MacArthur CDI. However, because of the uncertainty of this correlation, we assume a conservative value of .50 for c_1 in our power calculations below, and assume the same value for c_2.

We have less information on plausible values for c_3. ITBS test score data from the TFA evaluation suggests a value of about .50 for c_3. However, we assume a more conservative value of .30 in our power calculations to reflect the uncertainty in this parameter.

C. WAYS TO IMPROVE PRECISION UNDER A CLUSTERED DESIGN

As discussed, clustering at the school and classroom levels substantially reduces the precision of estimates. There are, however, several design and estimation strategies that can be used in clustered designs to reduce design effects. In this section, we discuss these strategies.

1. Using a Balanced Sample Size Allocation

For a given total sample size of schools, classrooms, and students, a 50-50 split of the treatment and control groups yields more precise estimates than other splits. Bloom (2004) demonstrates, however, that precision levels do not erode substantially unless the proportion of the total sample that is allocated to the treatment or control groups exceeds 80 percent or is less than 20 percent. This is an important finding, because selecting a larger control group could

reduce study costs associated with implementing the tested interventions. Conversely, a larger treatment group sample might be preferred, because district and school staff might be more willing to participate in a random assignment study if the size of the control group is as small as possible. Furthermore, larger treatment groups increase the precision of impact estimates for subgroups defined by program experiences and program features. Nonetheless, a balanced allocation produces the most precise estimates, and thus, many evaluations adopt this design.

Another reason to adopt a balanced sample allocation is that statistical tests under this design are robust to deviations from the usual assumption that the variances of the outcome measures are the same for the treatment and control groups. Traditional t-tests are strictly valid only under the homoescadicity assumption that treatment and control group variances are the same. However, if the variances differ (because of intervention effects on the distribution of the outcome variables), the literature suggests that t-tests are approximately valid under balanced sample allocations, but are not valid under unbalanced sample allocations (Snedecor 1956; Gail et al. 1996).

2. Using Stratified Sampling Methods

The use of stratified sampling methods to select treatment and control groups can reduce design effects. This is because under a stratified design, the ICCs pertain to clustering effects *within* strata (assuming that fixed stratum effects are included in the regression or ANOVA models). Thus, to the extent that strata are formed using group-level measures that are correlated with the outcome measures, stratified sampling will diminish clustering effects.

Many of the designs that we have considered in this paper are stratified designs where random assignment occurs within fixed strata defined by purposively-selected school districts or schools. Additional stratification can further reduce design effects. For instance, under a design where classrooms within schools are randomly assigned to a research condition, classroom strata could first be formed based on available teacher characteristics, and the treatment and control groups would then be selected within each strata. As another example, under a design where schools are randomly assigned within districts, schools could first be grouped on the basis of their average test scores and locations.

Stratified sampling, however, reduces the number of degrees of freedom for statistical tests if stratum effects are included in the regression models (see equation (2) above) which could offset some of the precision gains from stratification. This precision loss, however, is meaningful only if small numbers of groups are randomly assigned to a research condition.

An extreme form of stratification occurs when, prior to random assignment, only *two* units are assigned to each strata. This pairwise matching approach is sometimes used when only small numbers of units are randomly assigned to a research condition to avoid the possibility of obtaining a "bad draw." For example, the SACD evaluation used this pairwise matching design to allocate the 10 to 18 schools within each site to the treatment and control groups. Schools with similar characteristics were first paired, and one school in a pair was then randomly assigned to the treatment group and the other in the pair was randomly assigned to the control group. As

discussed, this sampling approach is also used, by necessity, in designs where only two classrooms within a school are available for random assignment.

Under designs with only one treatment and control group per stratum, there are no degrees of freedom available for estimating within-stratum group effects (Murray 1998). As discussed above for the case with only one classroom per condition per school, there are several approaches for dealing with this problem. One approach is to ignore the stratification in the analysis (which could increase the ICC estimates), while another approach is to use a random effects framework where stratum effects are treated as random (with the associated loss in degrees of freedom). For the second approach, the leading term in the variance formula for an impact estimate represents the extent to which impacts vary across strata (pairs). Diehr et al. (1995), Martin et al. (1993), Klar and Donner (1997) discuss the benefits of the various approaches when the number of pairs is small (and hence, where statistical power losses from pairwise matching could be severe).

3. Using Regression Models

For a given sample design, the most effective strategy for improving precision levels for group-based random assignment designs is to use regression models to estimate program impacts. The inclusion of relevant *baseline* student-, classroom-, and school-level explanatory variables in the regression models 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 the power improvements from using regression (ANCOVA) models, we consider the design where the school is the unit of random assignment, and generalize equation (17) as follows:

$$(18)\quad Var(impact) = \frac{2\sigma^2\rho_1(1-R^2_{BS})}{.5s} + \frac{2\sigma^2\rho_2(1-R^2_{BC})}{(.5s)k^*} + \frac{2\sigma^2(1-\rho_1-\rho_2)(1-R^2_W)}{(.5s)k^*n}.$$ [13]

In this expression, R^2_{BS} is the proportion of the between-school variance that is explained by the regression model, R^2_{BC} is the proportion of the between-classroom variance within schools that is explained by the regression model, and R^2_W is the proportion of the within-classroom variance that is explained by the regression model. Thus, the inclusion of explanatory variables that have significant predictive power in the regression models can substantially improve the precision levels of the impact estimates. The most effective explanatory variables are likely to be pre-intervention measures of the outcome variables, measured at the student, classroom, and aggregate school levels.

[13] 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.

It is important to note that it is possible, although unlikely, that R^2_{BS} or R^2_{BC} (but not R^2_W) are *negative* if the distribution of the covariates across groups exacerbates differences across the groups (Murray 1998). Thus, regression adjustment methods do not necessarily reduce ICCs.

The groups of covariates that can be included in the regression models will depend on the design. For instance, for the fixed-effects design where students or classrooms are randomly assigned to a research condition within volunteer schools and districts, the covariates cannot include school-level (or district-level) measures. This is because these measures will be perfectly collinear with the school indicator variables (see equation (7)). However, school-level covariates should be included if school effects are treated as random.

The inclusion of covariates decreases the degrees of freedom available for statistical tests, but in ways that depend on the level at which the covariates are measured. For estimating variances at the group level, one degree of freedom is lost for each group-level covariate included in the model. Individual-level covariates, however, reduce the degrees of freedom for estimating the individual-level variance terms, but not the group-level terms. Thus, if available, individual-level covariates are preferred to group-level ones. Furthermore, because the degrees of freedom at the group level are critical for power, use of group-level covariates should be limited to those that have significant explanatory power in the regression models, and that adjust for residual measurable differences between the treatment and control groups.

To obtain benchmark regression R^2 values, we examined the fit of models using baseline and follow-up test score data on elementary school students from various data sources: (1) the LESCP, (2) the national evaluation of the 21st Century Community Learning Centers program, and (3) the TFA evaluation. Our 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, KY and Rochester, NY, respectively. However, in the absence of these pre-intervention measures of the outcome variables, R^2 values were closer to .20. Because the amount and quality of baseline data vary across evaluations, we conduct our power calculations assuming conservative R^2 values of 0, .20 and .50.

4. Including Finite Population Corrections

When samples of students, classrooms, and schools are considered to be sampled from a *finite* population, the use of a finite population correction (fpc) reduces the variance of a sample mean by a factor equal to 1 minus the proportion of the population being selected. The gains from using the fpc can be substantial if a significant proportion of all population units are selected to the sample (because, under repeated sampling, there would be considerable overlap in the analysis samples). This is relevant for many group-based evaluations of education programs, because it is often the case that a *large* percentage of all relevant units are randomly assigned to a research status. For instance, for the SACD evaluation, half the population of schools per site were randomly assigned to the treatment group and half were randomly assigned to the control group. Similarly, for the Evaluation of the Effectiveness of Educational Technology Interventions (Dynarski et al. 2004), most teachers in participating schools and grades will be assigned at random to use a technology intervention or not. Thus, it is worth considering whether

the use of a finite population correction increases precision levels for impact estimates under group-based experimental designs.

In order to address this issue, we first note that for the multi-stage designs that we have considered, randomization at each stage takes one of two forms: (1) the random *assignment* of units to a research condition, or (2) the random *selection* of units to the sample from a larger universe of units. For example, under some school-based experimental designs, schools are randomly assigned to a research condition (the first type of randomization) and classrooms are then randomly selected within the study schools (the second type of randomization). The variance formulas for such multi-stage designs include terms that account for both sources of randomization.

The fpc *does not apply* to variance terms associated with the random assignment of units (the first type of randomization) when a large percentage of all units are randomly assigned. This is because there is a negative correlation between the treatment and control group means that cancels the gains from using the finite population correlation. To fix ideas, consider a design where 100 students within purposively-selected schools are randomly assigned to a treatment or control group. Then, if, by chance, average test scores for the 50 treatment group students are larger than average test scores of all 100 students, then, by definition, the 50 control group students will have lower-than-average test scores. Because the variance of an impact estimate equals the sum of the variances of the treatment and control group means minus twice the covariance between the two means, a negative correlation between the means increases the variance of the impact estimates, and directly offsets the precision gains from using the fpc.

The fpc, however, *does apply* for variance terms associated with the random selection of units (the second type of randomization). For example, consider a three-stage sample design where schools are randomly assigned to a treatment or control group in the first stage, classrooms are randomly sampled within schools in the second stage, and students are randomly sampled within classrooms in the third stage. Then, the finite population correction applies to the classroom- and student-level variance terms, but not to the school-level term. Using earlier results, the variance formula can be written as follows:

$$(19)\quad Var(impact) = \frac{2\sigma^2\rho_1}{.5s} + \frac{2\sigma^2\rho_2(1-\frac{k^*}{K})}{(.5s)k^*} + \frac{2\sigma^2(1-\rho_1-\rho_2)(1-\frac{n}{N})}{(.5s)k^*n},$$

where K is the total number of classrooms per school and N is the total number of students per classroom (and where we have omitted the regression R^2 terms). Thus, the classroom-level effect is reduced as the sampling fraction of classrooms is increased, and similarly for the student-level effect. However, the finite population correction does not affect the school-level term. Clearly, if the population universe is assumed to be infinite, the finite sample corrections do not enter the variance formulas.

5. Accounting for Longitudinal Observations and Repeated Measures

For many evaluations of education programs, longitudinal data are collected on sample members at baseline and at various follow-up time points to examine changes in impacts over time. In this section, we discuss appropriate variance formulas for impact estimates using longitudinal observations where time is modeled either as a fixed effect or a linear time trend in the HLM framework. We examine also the case where repeated measures are collected on units within a time period. We use results found in Koepsell et al. (1991), Murray et al. (1998), Murray and Blitstein (2003), Klar and Darlington (2004), and Janega et al. (2004). For illustrative simplicity, we focus on the design where schools are the unit of random assignment and where classroom-level clustering is not present, although the results can be easily applied to other designs that we have considered.

a. Modeling Time as a Fixed Effect

Suppose that comparable test score data are available at baseline and at several follow-up points. As discussed, one procedure for incorporating the baseline data into the posttest analysis is to include the baseline test scores as covariates in the regression models. Another procedure is to treat the baseline test scores as a dependent variable along with the follow-up test scores and to include time effects in the regression models.

Consider first a design where, within the study schools, data are collected on *different* cohorts of students during the baseline and follow-up periods (which would occur, for instance, if data were collected on only third grade students in each period). Consider also the following two-level HLM model, where Level 1 pertains to the student and Level 2 pertains to the school (the unit of random assignment):

$$(20)\quad Level\,1:\ Y_{ipq} = \lambda_{0p} + \sum_{q=2}^{l} \lambda_{1pq} F_{ipq} + e_{ipq}$$

$$Level\,2:\ \lambda_{0p} = \gamma_0 + \gamma_1 T_p + u_p$$

$$\lambda_{1pq} = \delta_{0q} + \delta_{1q} T_p + \tau_{pq},$$

where Y_{ipt} are standardized test scores of student i in school p at follow-up point q $(q=1,..,l)$, where period q=1 corresponds to the baseline period; F_{ipq} is an indicator variable equal to 1 for observations at follow-up point q; T_p is a treatment status indicator variable for school p; u_p are *iid* $N(0,\sigma^2_u)$ school-specific random error terms (at baseline); τ_{pq} are *iid* $N(0,\sigma^2_\tau)$ error terms which represent the extent to which *school effects vary over time* during the follow-up period (relative to the baseline period); e_{ip} are *iid* $N(0,\sigma^2_e)$ student-level residual error terms that are distributed independently of u_p and τ_{pq}; and the remaining terms are parameters.

Inserting the Level 2 equations into the Level 1 equation yields:

$$(21)\quad Y_{ipq} = \gamma_0 + \gamma_1 T_p + \sum_{q=2}^{l} \delta_{0q} F_{ipq} + \sum_{q=2}^{l} \delta_{1q} (T_p * F_{ipq}) + [u_p + \sum_{q=2}^{l} \tau_{pq} F_{ipq} + e_{ipq}].$$

In this formulation, δ_{1q} represents the impact in follow-up period q, and is the treatment-control difference between the mean posttest score in period q relative to the mean pretest score in period 1 (that is, $[Y_{..qT} - Y_{..1T}] - [Y_{..qC} - Y_{..1C}]$). Because the u_p terms cancel in this difference-in-difference estimator, the variance of the impact estimate is:

$$(22)\quad Var(impact) = 2[\frac{\sigma_\tau^2}{.5s} + \frac{2\sigma_e^2}{(.5s)kn}].$$

Using earlier results, this variance formula can also be expressed as follows:

$$(23)\quad Var(impact) = 2[\frac{2\sigma_u^2(1-c_4)}{.5s} + \frac{2\sigma_e^2}{(.5s)kn}],$$

where c_4 represents the correlation between mean test scores within a school over time.

It is an empirical issue whether conducting a pretest-posttest analysis yields more efficient estimates than conducting a posttest analysis only with the pretest scores included as covariates in the regression models. This issue largely depends on the extent to which student outcomes within a school vary over time (that is, on c_4) and the predictive power of the pretest scores in the posttest regression models. Janega et al. (2004) provide evidence using data from the TEENS study that the regression-adjusted posttest analysis is the more powerful technique.

Finally, we note that equations (22) and (23) are applicable also to the case where data on the *same* students are collected over time in the study schools, and where there is no repeated testing of students within the same time period (Murray 1998). This is because, although time-by-student random effects can be included in the models, time-by-student and within-student variability are not separable.

b. Linear Trend Analysis

In education research, growth-curve analyses are often conducted to examine intervention effects on the growth trajectories of student outcomes. In these analyses, longitudinal observations are modeled as a function of time (measured, for example, as the number of months

or years from random assignment until data collection). In its simplest form, time can be modeled as a linear trend, in which case equation (21) can be modified as follows:

$$(24)\quad Y_{ipq} = \gamma_0 + \gamma_1 T_p + \delta_0 t_q + \delta_1 (T_p * t_q) + [u_p + \sum_{q=2}^{l} \tau_{pq} F_{ipq} + e_{ipq}],$$

where t_q is the time between random assignment and the collection of observation q (appropriately centered).

In this model, the intervention effect is δ_1, which represents the treatment-control difference in the estimated *slopes* from regressions of test scores on time. Standard regression theory shows that this impact estimate can be expressed as a weighted sum of the (l-1) difference-in-difference estimators discussed above with weights $(t_q - t_.)/\Sigma(t_q - t_.)^2$. The variance of this impact estimate is:

$$(25)\quad Var(impact) = \frac{L^2}{\sum_{q=1}^{l} (t_q - t_.)^2} [\frac{2\sigma_u^2 (1 - c_4)}{.5s} + \frac{2\sigma_e^2}{(.5s)kn}],$$

where L is the length of the follow-up period (see Koepsell et al. 1991 for a similar expression). This variance expression tends to decrease as the number of time periods increases. Murray (1998) provides more general versions of variance formulas for growth curve models that allow for covariates and random time trends.

c. Accounting for Repeated Measures

In some evaluations, repeated measures are collected on subjects at *each* data collection point. For example, in the Evaluation of the Impact of Teacher Induction Programs, researchers plan to observe teacher practices twice per data collection point. The presence of repeated measures increases the effective sample size for the analysis, and hence, increases the precision of the impact estimates. The effective sample size will depend on the correlation of the repeated measures.

To quantify the extent to which repeated measures improve precision levels, we consider a three-stage HLM model, where Level 1 refers to measurement *m*, Level 2 refers to subjects, and Level 3 to schools (the unit of random assignment). In this case, the treatment effect for a single posttest period can be estimated using the following expression:

$$(26)\quad Y_{ipm} = \gamma_0 + \gamma_1 T_p + [u_p + \tau_{ip} + e_{ipm}].$$

The associated variance of the impact estimate is:

$$(27)\quad Var(\textit{impact in a district}) = \frac{2\sigma^2\rho_1}{.5s} + \frac{2\sigma^2\rho_3(1-\rho_1)}{(.5s)kn} + \frac{2\sigma^2(1-\rho_3)(1-\rho_1)}{(.5s)knf},$$

where ρ_3 is the ICC for student outcomes within schools (that is, the proportion of the total student-level variance that is not due to measurement error), and f is the number of repeated measures. A large value for ρ_3 signifies that the repeated measures on subjects are highly correlated , and thus, that precision gains from the repeated measures are negligible. Conversely, a small value for ρ_3 suggests that the repeated measures can effectively be treated as separate observations. Using the last two variance terms in equation (27), the effective number of students can be calculated by dividing the total number of observations [(.5s)knf] by the design effect [1+ρ_3(f-1)].

Importantly, the presence of repeated measures on students influences the student-level variance terms, but not the larger school- or classroom-level variance components. Thus, for a group-based experimental design, the presence of repeated measures usually has only have a modest effect on overall precision levels.

D. ILLUSTRATIVE PRECISION CALCULATIONS

In this section, we collate results from above and calculate illustrative MDE calculations for the most common designs that we have considered. The purpose of this analysis is to provide estimates of appropriate sample sizes that are required for experimental impact evaluations of education programs that aim to improve the standardized test scores of elementary school students in low-income schools.

Next, we discuss the presentation of findings and the assumptions underlying the power calculations. Then, we summarize our main findings.

1. Presentation and Assumptions

Tables 3 to 8 display, under various conservative 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 quality of baseline data will vary across evaluations, each table presents school sample sizes assuming R^2 values of 0, .20, and .50 at each group level. We consider all designs from Table 1 except Design V, because in evaluations where classrooms within schools are the unit of random assignment, it is often the case that there are only *two* applicable classrooms per school; thus, it is usually not possible to estimate separate classroom-level and school-level variance terms.

TABLE 3

REQUIRED SCHOOL SAMPLE SIZES TO DETECT TARGET EFFECT SIZES, BY DESIGN

Assumes a Two-Tailed Test, a Value of .15 for the Intraclass Correlations, a Balanced Allocation of the Research Groups, and No Subsampling of Students Within Units

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:			
Unit of Random Assignment: Sources of Clustering	.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 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, a 2:1 Split of the Research Groups, and No Subsampling of Students Within Units

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:			
Unit of Random Assignment: Sources of Clustering	.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.

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, a Balanced Allocation of the Research Groups, and Subsampling of 33 Percent of Students Within Units

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:			
Unit of Random Assignment: Sources of Clustering	.10	.20	.25	.33
I: Students Within Schools: No Clustering				
R^2 = 0	170	43	27	16
R^2 = .2	136	34	22	13
R^2 = .5	85	21	14	8
II: Students Within Schools: School-Level Clustering				
R^2 = 0	262	68	44	27
R^2 = .2	210	54	36	22
R^2 = .5	131	35	24	14
III: Students Within Schools: School- and Classroom-Level Clustering				
R^2 = 0	276	71	46	28
R^2 = .2	221	57	38	23
R^2 = .5	138	37	25	15
IV: Classrooms Within Schools: Classroom-Level Clustering (Ignoring School Fixed Effects)				
R^2 = 0	302	75	48	29
R^2 = .2	241	60	39	23
R^2 = .5	151	38	25	15
VI: Classrooms Within Schools: School-Level Clustering				
R^2 = 0	310	80	51	31
R^2 = .2	248	64	42	25
R^2 = .5	155	41	27	17
VII: Schools Within Districts: School-Level Clustering				
R^2 = 0	615	154	100	58
R^2 = .2	492	123	81	47
R^2 = .5	308	79	51	31
VIII: Schools Within Districts: School- and Classroom-Level Clustering				
R^2 = 0	747	187	119	71
R^2 = .2	597	149	98	57
R^2 = .5	373	96	62	37

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

TABLE 6

REQUIRED SCHOOL SAMPLE SIZES TO DETECT TARGET EFFECT SIZES, BY DESIGN

Assumes a One-Tailed Test, a Value of .15 for the Intraclass Correlations, a Balanced Allocation of the Research Groups, and No Subsampling of Students Within Units

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:			
Unit of Random Assignment: Sources of Clustering	.10	.20	.25	.33
I: Students Within Schools: No Clustering				
$R^2 = 0$	45	11	7	4
$R^2 = .2$	36	9	6	3
$R^2 = .5$	23	6	4	2
II: Students Within Schools: School-Level Clustering				
$R^2 = 0$	132	34	23	14
$R^2 = .2$	106	28	19	11
$R^2 = .5$	67	18	12	7
III: Students Within Schools: School- and Classroom-Level Clustering				
$R^2 = 0$	157	40	27	16
$R^2 = .2$	125	33	22	13
$R^2 = .5$	79	21	14	8
IV: Classrooms Within Schools: Classroom-Level Clustering (Ignoring School Fixed Effects)				
$R^2 = 0$	163	41	27	15
$R^2 = .2$	131	33	21	12
$R^2 = .5$	82	21	13	8
VI: Classrooms Within Schools: School-Level Clustering				
$R^2 = 0$	170	43	29	17
$R^2 = .2$	136	35	23	14
$R^2 = .5$	86	23	15	9
VII: Schools Within Districts: School-Level Clustering				
$R^2 = 0$	413	103	67	39
$R^2 = .2$	331	84	54	32
$R^2 = .5$	207	53	34	21
VIII: Schools Within Districts: School- and Classroom-Level Clustering				
$R^2 = 0$	532	133	86	50
$R^2 = .2$	425	106	69	40
$R^2 = .5$	266	67	44	26

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

TABLE 7

REQUIRED SCHOOL SAMPLE SIZES TO DETECT TARGET EFFECT SIZES, BY DESIGN

Assumes a Two-Tailed Test, a Value of .10 for the Intraclass Correlations, a Balanced Allocation of the Research Groups, and No Subsampling of Students Within Units

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:			
Unit of Random Assignment: Sources of Clustering	.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	130	35	23	14
R^2 = .2	104	28	19	12
R^2 = .5	67	19	12	7
III: Students Within Schools: School- and Classroom-Level Clustering				
R^2 = 0	150	40	27	16
R^2 = .2	120	32	22	13
R^2 = .5	77	21	14	9
IV: Classrooms Within Schools: Classroom-Level Clustering (Ignoring School Fixed Effects)				
R^2 = 0	156	39	26	15
R^2 = .2	125	32	21	12
R^2 = .5	78	20	13	8
VI: Classrooms Within Schools: School-Level Clustering				
R^2 = 0	161	42	28	17
R^2 = .2	129	35	23	14
R^2 = .5	83	23	15	9
VII: Schools Within Districts: School-Level Clustering				
R^2 = 0	365	94	60	36
R^2 = .2	292	75	49	29
R^2 = .5	182	48	32	19
VIII: Schools Within Districts: School- and Classroom-Level Clustering				
R^2 = 0	464	116	77	45
R^2 = .2	371	95	61	36
R^2 = .5	232	60	39	24

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

TABLE 8

REQUIRED SCHOOL SAMPLE SIZES TO DETECT TARGET EFFECT SIZES, BY DESIGN

Assumes a Two-Tailed Test, a Value of .20 for the Intraclass Correlations, a Balanced Allocation of the Research Groups, and No Subsampling of Students Within Units

	Number of Schools Required to Detect an Impact in Standard Deviation Units of:			
Unit of Random Assignment: Sources of Clustering	.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	202	52	35	21
R^2 = .2	162	43	28	17
R^2 = .5	102	28	19	11
III: Students Within Schools: School- and Classroom-Level Clustering				
R^2 = 0	243	63	41	25
R^2 = .2	195	51	34	20
R^2 = .5	122	33	22	13
IV: Classrooms Within Schools: Classroom-Level Clustering (Ignoring School Fixed Effects)				
R^2 = 0	255	64	41	24
R^2 = .2	204	51	33	19
R^2 = .5	127	33	21	12
VI: Classrooms Within Schools: School-Level Clustering				
R^2 = 0	265	68	44	27
R^2 = .2	212	55	36	22
R^2 = .5	132	35	24	14
VII: Schools Within Districts: School-Level Clustering				
R^2 = 0	673	168	108	64
R^2 = .2	538	135	89	51
R^2 = .5	336	87	56	33
VIII: Schools Within Districts: School- and Classroom-Level Clustering				
R^2 = 0	870	218	139	82
R^2 = .2	696	174	111	66
R^2 = .5	435	109	72	42

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

Our *benchmark* estimates assume the following:

- A two-tailed test
- A balanced allocation across the treatment and control groups
- A design without sampling of students within units
- A value of .15 for the between-school and between-classroom ICCs (ρ_1 and ρ_2, respectively)

However, we also present selected tables assuming a one-tailed test, an unbalanced 2-to-1 split at the point of random assignment, a design where one-third of students are sampled, and values of .10 and .20 for the ICCs.

All calculations were conducted using the MDE formula in equation (1) and the variance formulas presented in the text (see Table 1 for the pertinent equation numbers for each design, and equation (18) for the treatment of regression R^2 values). We assume that one intervention is being tested against the control condition within each site; the sample sizes in the tables, however, can be inflated multiplicatively to account for multiple treatments (in the absence of multiple comparison corrections). Finally, all calculations were conducted under the following (conservative) assumptions:

- An 80 percent power level and a 5 percent significance level
- A value of .50 for the correlation between treatment and control group students within schools (c_1) and classrooms (c_2), and a value of .30 for the correlation between treatment and control group classrooms within schools (c_3).
- No finite sample corrections
- The intervention is being tested in a single grade
- An average of 3 classrooms per school per grade
- An average of 23 students per classroom
- 80 percent of students in the sample will provide follow-up (posttest) data
- No adjustments for longitudinal observations or repeated measures on students

For example, under our benchmark assumptions in Table 2, we calculated the number of schools required to detect an impact of *MDE* standard deviations for Design VII (school-based random assignment) by combining equations (1) and (16) and solving for the number of schools (s) as follows :

$$(28)\quad s = \frac{Factor(.05,.80,s-2)^2}{MDE^2}\left[\frac{2*.15}{.5} + \frac{2*(1-.15)}{.5*3*23*.8}\right](1-R^2),$$

where MDE is .10, .20, .25, or .33; R^2 is 0, .2, or .5 at the school and student levels; *Factor(.)* is obtained using the figures in Table A.1; and .8 is included in the denominator of the student-level term to reflect the assumed 80 percent response rate to the follow-up interview. Similar calculations were performed for the other designs.

2. Results

Our results can be summarized as follows:

- ***Clustering Matters.*** Precision levels decrease substantially as clustering effects increase. For instance, assuming a zero R^2 value and a value of .15 for the ICCs, 14 schools under Design I are required to detect an effect size of .20 standard deviations, compared to 55 schools for Design VI, and 130 schools for Design VII (Table 3). Similarly, required school sample sizes are considerably smaller when the ICCs are .10 rather than .15 (Tables 3 and 7), and considerably larger when the ICCs are .20 (Tables 3 and 8).

- ***Relatively large school sample sizes are required under group-based experimental designs.*** Consequently, because of resource constraints, many evaluations will only have sufficient power to detect precise impacts for relatively large subgroups of sites, and can rigorously address *broad* research questions only.

- ***Achieving effect sizes of .10 may not be attainable in many evaluations.*** As discussed, relatively small standardized 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 and the distribution of test scores across schools. However, our 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 (Table 3). Consequently, 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.*** The most 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 to only 44 schools for an R^2 value of .50 (Table 3). Thus, the availability of detailed baseline data at the aggregate school or individual student level (and in particular, data on pre-intervention 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 3 and 4).

- ***The subsampling of students within schools matters less as clustering effects increase.*** This is because student-level variance terms become a smaller share of the total variance estimates as clustering effects increase. However, even under designs with large clustering effects, the subsampling of students has some effect on precision levels (Tables 3 and 5).

- ***Precision levels are greater for one-tailed tests than two-tailed tests.*** However, the differences are not large (Tables 3 and 6). For example, assuming an R^2 value of .20 and a .15 effect size precision standard, Design VI requires 45 schools under a two-tailed test, compared to 35 schools under a one-tailed test. The differences are not large, because the inflation factor in the MDE formula is about 2.5 under a one-tailed test, compared to about 2.8 under a two-tailed test (Table A.1).

E. SUMMARY AND CONCLUSIONS

In this paper, we 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 student's standardized test scores. Our main conclusion is that clustering effects *cannot* be ignored when groups—such as schools or classrooms—are randomly *assigned* to a research status, or if groups are considered to be randomly *sampled* to the research sample from a larger universe. We 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, which are somewhat larger than comparable ICCs found for key outcomes in the public health literature.

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). Consequently, 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 must be deemed heuristic.

Our discussion has stressed that sources of clustering under group-based experimental designs must be examined and treated carefully. Ignoring clustering effects can lead to serious

overestimates of precision levels. Conversely, introducing spurious sources of clustering can lead to serious underestimates of precision levels. Thus, education researchers who conduct power and impact analyses must carefully specify the sources of clustering under their designs and the assumptions underlying them. In particular, as discussed in this paper, the treatment of group effects as fixed or random is an important issue that has major implications for sample size requirements. We emphasize also that the collection of detailed baseline data is an important way to reduce clustering effects under group-based experimental designs. Thus, for evaluations of education programs, researchers, whenever possible, should obtain detailed baseline data on student, teacher, and aggregate school characteristics.

There are several avenues for future research in this area. More rigorous empirical research is needed on the association between student academic achievement gains in the early grades and medium- and long-term improvements in students' school- and employment-related outcomes. This information (especially for students in low-performing schools) is critical for assessing appropriate precision benchmarks for impact evaluations of education programs that are typically conducted. A more complete compilation of empirical evidence is also needed on plausible parameter values for ICCs at the school and classroom levels for different types of student outcome measures and in different settings (which, as discussed, has been done much more extensively in the public health literature). Furthermore, an important area for future research is to identify baseline measures that are most effective in reducing clustering effects when regression models are used to estimate program impacts, and to compare the relative effectiveness of student measures compared to aggregate school measures.

REFERENCES

Aber, Larry J., L. Brown, and S. M. Jones. "Developmental Trajectories Toward Violence in Middle Childhood: Course, Demographic Differences and Response to School-Based Intervention." *Developmental Psychology*, vol. 39, 2003.

Aber, Larry, et al. "Teaching Conflict Resolution: An Effective School-Based Approach to Violence Prevention." Report submitted to the National Center for Children in Poverty. New York, NY: Columbia University, 1999.

Bloom, H., J. Bos, and S. Lee. "Using Cluster Random Assignment to Measure Program Impacts: Statistical Implications for Evaluation of Education Programs." *Evaluation Review*, vol. 23, no. 4, 1999.

Bloom, Howard. *Randomizing Groups to Evaluate Place-Based Programs*. New York, NY: MDRC, 2004.

Bloom, Howard. Unpublished Tabulations of Intraclass Correlations. Presented at a conference at the U.S. Department of Education, Institute of Education Sciences, Washington, D.C., March 2005.

Byrk, A., and S. Raudenbush. *Hierarchical Linear Models for Social and Behavioral Research. Applications and Data Analysis Methods.* Newbury Park, CA: Sage, 1992.

Campbell, M. "Cluster Randomised Trials in General (Family) Practice Research." *Statistical Methods in Medical Research,* vol. 9, 2000.

Cochran, William. *Sampling Techniques*. New York: John Wiley and Sons, 1977.

Cohen, J. *Statistical Power Analysis for Behavioral Sciences.* Hillside, NJ: Lawrence Erlbaum, 1988.

Cook, T., and D. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago: Rand-McNally, 1979.

Cornfield, J. "Randomization by Group: A Formal Analysis. *American Journal of Epidemiology,* vol. 108, no. 2, 1978.

Currie, Janet, and Duncan Thomas. "Early Test Scores, Socioeconomic Status and Future Outcomes." NBER Working Paper No. 6943, 1999.

Decker, Paul, et al. "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, 2004.

Diehr, P., D. Martin, T. Koepsell, and A. Cheadle. "Breaking the Matches in a Paired t-Test for Community Interventions When the Number of Pairs Is Small. *Statistics in Medicine*, vol. 14, 1994.

Donner, Allan, and Neil Klar. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold, 2000.

Dynarski, Mark, Mary Moore, Linda Rosenberg, Susanne James-Burdumy, John Deke, and Wendy Mansfield. "When Schools Stay Open Late: The National Evaluation of the 21st Century Community Learning Centers Program, New Findings. Final report submitted to the U.S. Department of Education. Princeton, NJ: Mathematica Policy Research, Inc, 2004.

Dynarski, Mark, and R. Agodini. "The Effectiveness of Educational Technology: Issues and Recommendations for the National Study." Report submitted to the U.S. Department of Education, Institute of Education Sciences. Princeton, NJ: Mathematica Policy Research, Inc., 2003.

Dynarksi, Mark, P. Gleason, A. Rangarajan, and R. Wood. "Impacts of Dropout Prevention Programs." Report submitted to the U.S. Department of Education, Planning and Evaluation Service. Princeton, NJ: Mathematica Policy Research., 1998.

Flay, Brian, C. Allred and N. Ordway. "The Positive Action Program for Improving Achievement." *Prevention Science*, vol. 2, 2001.

Fleiss, J. L. *The Design and Analysis of Clinical Experiments*. New York: John Wiley & Sons, 1986.

Gargani, J., and T. Cook. "How Many Schools? Limits of the Conventional Wisdom about Sample Size Requirements for Cluster Randomized Trials." University of California, Berkeley Working Paper, 2005.

Glazerman, Steven, L. Tarullo, S. Sprachman, C. Tuttle, J. Love, C. Rowland, and A. KewalRamani. "Preschool Curriculum Evaluation Research." Design Documents. Princeton, NJ: Mathematica Policy Research, Inc., 2004.

Gleason, Philip, and R. Olsen. "Impact Evaluation of Charter School Strategies." Design Documents. Princeton, NJ: Mathematica Policy Research, Inc., 2004.

Hayes, R., N. Alexander, S. Bennett, et al. "Design and Analysis Issues in Cluster-Randomized Trials of Interventions Against Infectious Diseases." *Statistical Methods in Medical Research*, vol. 9, 2000.

Hedberg, Eric, R. Santana, and L. Hedges. "The Variance Structure of Academic Achievement in America." Working Paper. Chicago, IL: University of Chicago, 2004.

Hedges, Larry. "Effect Sizes in Multi-Site Designs Using Assignment by Cluster." Working Paper. Chicago, IL: University of Chicago, 2004.

Hedges, Larry. "Correcting Significance Tests for Clustering. Working Paper. Chicago, IL: University of Chicago, 2004.

Janega, J., D. Murray, S. Varnell, J. Blitstein, A. Birnbaum, and L. Lytle. "Assessing Intervention Effects in a School-Based Nutrition Intervention Trial: Which Analytic Model Is Most Powerful?" *Health Education and Behavior*, vol. 31, 2004.

Kane, Thomas. "The Impact of After-School Programs: Interpreting the Results of Four Recent Evaluations." Working Paper. Los Angeles, CA: UCLA, 2004.

Kane, Thomas, and Douglas Staiger. "The Promise and Pitfalls of Using Imprecise School Accountability Measures." *Journal of Economic Perspectives*, vol. 16, no. 4, 2002.

Kish, Leslie. *Survey Sampling*. New York: John Wiley and Sons, 1965.

Klar, N., and A. Donner. "The Merits of Matching in Community Intervention Trials: A Cautionary Tale." *Statistics Medicine*, vol. 16, 1997.

Klar, N., and G. Darlington. "Methods for Modelling Change in Cluster Randomization Trials." *Statistics in Medicine*, vol. 23, 2004.

Koepsell, T., D. Martin, P. Diehr, B. Psaty, E. Wagner, E. Perrin, and A. Cheadle. "Data Analysis and Sample Size Issues in Evaluations of Community-Based Health Promotion and Disease Prevention Programs: A Mixed-Model Analysis of Variance Approach. *Journal of Clinical Epidemiology*, vol. 44, no. 7, 1991.

Koepsell, T., E. Wagner, and A. Cheadle. "Selected Methodological Issues in Evaluating Community-Based Health Promotion Programs." *Annual Review of Public Health*, vol. 13, 1992.

Krueger, Alan B. "Economic Considerations and Class Size." Working Paper #447. Princeton, NJ: Princeton University Industrial Relations Section, 2000.

Lipsey, M.W., and D.B. Wilson. "The Efficacy of Psychological, Educational, and Behavioral Treatment." *American Psychologist*, vol. 48, no. 12, 1993.

Love, John, et al. "Building Their Futures: How Early Head Start Programs Are Enhancing the Lives of Infants and Toddlers in Low-Income Families." Report submitted to the Office of Research and Evaluation of the Administration on Children, Youth and Families, U.S. Department of Health and Human Services. Princeton, NJ: Mathematica Policy Research, Inc., 2002.

Martin, D., P. Diehr, E. Perrin, and T. Koepsell. "The Effect of Matching on the Power of Randomized Community Intervention Studies. *Statistics in Medicine*, vol. 12, 1993.

Murnane, Richard, J. Willet, and F. Levy. "The Growing Importance of Cognitive Skills in Wage Determination." *Review of Economics and Statistics*, vol. 77, 1995.

Murray, D., P. Hannan, and W. Baker. "A Monte Carlo Study of Alternative Responses to Intraclass Correlation in Community Trials: Is It Ever Possible to Avoid Cornfield's Penalties?" *Evaluation Review*, vol. 20, no. 3, 1996.

Murray, David. *Design and Analysis of Group-Randomized Trials*. Oxford: Oxford University Press, 1998.

Murray, D., and J. Blitstein. "Methods to Reduce the Impact of Intraclass Correlations in Group-Randomized Trials." *Evaluation Review*, vol. 27, no. 1, 2003.

Murray, D., S. Varnell, and J. Blitstein. "Design and Analysis of Group-Randomized Trials: A Review of Recent Methodological Developments." *American Journal of Public Health*, vol. 94, no. 3, 2004.

Myers, David, and Allen Schirm. "The Impacts of Upward Bound: Final Report for Phase I of the National Evaluation." Princeton, NJ: Mathematica Policy Research, Inc., 1999.

National Center for Education Statistics. "The Nation's Report: Highlights 2003." U.S. Department of Education, Institute of Education Sciences, National Assessment of Educational Progress (NAEP), 2003.

Neal, Derek, and William Johnson. "The Role of Premarket Factors in Black-White Wage Differentials." *Journal of Political Economy*, vol. 104, 1996.

Ramsey, Philip. "Comparison of Closed Testing Procedures for Pairwise Testing of Means. *Psychological Method*, vol. 7, no. 4, 2002.

Raudenbush, Stephen. "Statistical Analysis and Optimal Design for Cluster Randomized Trials." *Psychological Methods*, vol. 2, no. 2, 1997.

Raudenbush, Stephen, J. Spybrook, X. Liu, R. Congdon. "Optimal Design for Longitudinal and Multilevel Research: Documentation for the Optimal Design Software." University of Michigan, 2004.

Schochet, Peter Z., Susanne James-Burdumy, and Ellen Kisker. "Social and Character Development (SACD) Research Program: Analysis Plan for the Multisite Impact Analysis." Princteon, NJ: Mathematica Policy Research, Inc., 2004.

Schochet, Peter, John Burghardt, and Steven Glazerman. "National Job Corps Study: The Impacts of Job Corps on Participants' Employment and Related Outcomes." Princeton, NJ: Mathematica Policy Research, Inc., 2001.

Singer, Judith. "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics*, vol. 24, no. 4, 1998.

Ukoumunne, O., M. Gulliford, S. Chinn, J. Sterne, and P. Burney. "Methods for Evaluating Area-Wide and Organisation-Based Interventions in Health and Health Care: A Systematic Review." *Health Technology Assessment*, vol. 3, no. 5, 1999.

Varnell, S., D. Murray, J. Janega, and J. Blitstein. "Design and Analysis of Group-Randomized Trials: A Review of Recent Practices. *American Journal of Public Health*, vol. 94, no. 3, 2004.

Walsh, J. "Concerning the Effects of the Intra-Class Correlation on Certain Significance Tests." *Annals of Mathematical Statistics*, vol. 18, 1947.

APPENDIX A

VALUES FOR FACTOR(.) IN EQUATION (2)

TABLE A.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 PERCENT POWER

	One-Tailed Test		Two-Tailed Test	
Number of Degrees of Freedom	80 Percent Power	85 Percent Power	80 Percent Power	85 Percent 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 percent significance level.