Educational Evaluation and Policy Analysis
March 2007, Vol. 29, No. 1, pp. 60–87
DOI: 10.3102/0162373707299706

Intraclass Correlation Values for Planning Group-Randomized Trials in Education

Larry V. Hedges
Northwestern University

E. C. Hedberg
University of Chicago

Experiments that assign intact groups to treatment conditions are increasingly common in social research. In educational research, the groups assigned are often schools. The design of group-randomized experiments requires knowledge of the intraclass correlation structure to compute statistical power and sample sizes required to achieve adequate power. This article provides a compilation of intraclass correlation values of academic achievement and related covariate effects that could be used for planning group-randomized experiments in education. It also provides variance component information that is useful in planning experiments involving covariates. The use of these values to compute the statistical power of group-randomized experiments is illustrated.

Keywords: *intraclass correlation, cluster randomized trials, experiments, statistical power*

Many social interventions operate at a group level by altering the physical or social conditions. In such cases, it may be difficult or impossible to assign individuals to receive different intervention conditions. In other cases, it may be possible to assign treatments to individuals, but for practical or political reasons, the assignment of individuals to treatments is not feasible. In either situation, field experiments may assign entire intact groups (such as sites, classrooms, or schools) to the same treatment, with different intact groups being assigned to different treatments. Because these intact groups correspond to what statisticians call clusters in sampling theory, this design is often called a group-randomized or *cluster*-randomized design. Cluster-randomized trials have been used extensively in public health and other areas of prevention science (see, e.g., Donner & Klar, 2000; Murray, 1998). Cluster-randomized trials have become more important in educational research more recently, following increased interest in experiments to evaluate educational interventions (see, e.g., Mosteller & Boruch, 2002). Methods for the design and analysis of group-randomized trials have been discussed extensively by Donner and Klar (2000) and Murray (1998).

The sampling of subjects into experiments via statistical clusters introduces special considerations that need to be addressed in the analysis. For example, a sample obtained from m clusters (such as classrooms or schools) of size n randomized into a treatment group is not a simple random sample of nm individuals, even if it is based on a simple random sample of clusters. Instead, it is a two-stage sample (with one stage of clustering). Consequently, the sampling distribution of statistics on the basis of such

This material is based upon work supported in part by the National Science Foundation under Grant No. 0129365 and the Spencer Foundation Grant Number 200100308.

clustered samples is not the same as that based on simple random samples of the same size. For example, suppose that the (total) variance of a population with clustered structure (such as a population of students within schools) is σ_T^2 and that this total variance is decomposable into a between-cluster variance, σ_B^2, and a within-cluster variance, σ_W^2, so that $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$. Then the variance of the mean of a simple random sample of size mn from that population would be σ_T^2/mn. However, the variance of the mean of a sample of m clusters, each of size n from that population (with the same total sample size mn) would be $[1 + (n - 1)\rho]\sigma_T^2/mn$, where $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ is the intraclass correlation. Thus, the variance of the mean computed from a clustered sample is larger by a factor of $[1 + (n - 1)\rho]$, which is often called the design effect (Kish, 1965) or variance inflation factor (Donner, Birkett, & Buck, 1981).

Several analytical strategies for cluster-randomized trials are possible, but the simplest is to treat the clusters as units of analysis, that is, to compute mean scores on the outcome (and all other variables that may be involved in the analysis) and carry out the statistical analysis as if the site (cluster) means were the data. If all cluster sample sizes are equal, this approach provides exact tests for the treatment effect, but the tests may have lower statistical power than would be obtained by other approaches (see, e.g., Blair & Higgins, 1986). More flexible and informative analyses are also available, including analyses of variance using clusters as a nested factor (see, e.g., Hopkins, 1982) and analyses involving hierarchical linear models (see, e.g., Raudenbush & Bryk, 2002). For general discussions of the design and analyses of cluster-randomized experiments, see Murray (1998); Bloom, Bos, and Lee (1999); Donner and Klar (2000); Klar and Donner (2001); Raudenbush and Bryk (2002); Murray, Varnell, and Blitstein (2004); or Bloom (2005).

Wise experimental design involves the planning of sample sizes so that the test for treatment effects has adequate statistical power to detect the smallest treatment effects that are of scientific or practical interest. There is an extensive literature on the computation of statistical power, (e.g., Cohen, 1977; Kraemer & Thiemann, 1987; Lipsey, 1990). Much of this literature involves the computation of power in studies that use simple random samples. However, methods for the computation of statistical power of tests for treatment effects using the cluster mean as the unit of analysis (Blair & Higgins, 1986), analysis of variance using clusters as a nested factor (Raudenbush, 1997), and hierarchical linear model analyses (Snijders & Bosker, 1993) are available. For all of these analyses, the noncentrality parameter required to compute statistical power involves the intraclass correlation ρ (which was defined above but will be defined formally in Equation 1). More complex analyses involving covariates require corresponding information (covariate effects or the conditional intraclass correlations after adjustment for covariates). Thus, the computation of statistical power in cluster-randomized trials requires knowledge of the intraclass correlation ρ.

Because plausible values of ρ are essential for power and sample-size computations in planning cluster-randomized experiments, there have been systematic efforts to obtain information about reasonable values of ρ in realistic situations. One strategy for obtaining information about reasonable values of ρ is to obtain these values from cluster-randomized trials that have been conducted. Murray and Blitstein (2003) reported a summary of intraclass correlations obtained from 17 articles reporting cluster-randomized trials in psychology and public health, and Murray et al. (2004) gave references to 14 very recent studies that provide data on intraclass correlations for health-related outcomes. Another strategy for obtaining information on reasonable values of ρ is to analyze sample surveys that have used a cluster-sampling design involving the clusters of interest. Gulliford, Ukoumunne, and Chinn (1999) and Verma and Lee (1996) presented values of intraclass correlations on the basis of surveys of health outcomes.

There is much less information about intraclass correlations appropriate for studies of academic achievement as an outcome. Such information is badly needed to inform the design of experiments that measure the effects of interventions on academic achievement by randomizing schools (Schochet, 2005). One compendium of intraclass correlation values on the basis of five large urban school districts in which randomized trials have been conducted has recently become available (see Bloom, Richburg-Hayes, & Black, 2007 [this issue]). The purpose of this article is to

provide a comprehensive collection of intraclass correlations of academic achievement on the basis of national representative samples. We hope that this compilation will be useful in choosing reference values for planning cluster-randomized experiments.

Key Findings

We find that across Grades K–12, the average (unadjusted) intraclass correlation is about .22 for all schools, about .19 for low–socioeconomic status (SES) schools, and about .09 for low-achievement schools. These average intraclass correlations are very similar in reading and mathematics. Note that except in low-achievement schools, these intraclass correlation values are somewhat higher than the guidelines of .05–.15 that are often used. Pretests can explain a substantial amount of the between- and within-school variance when used as covariates. Covariates can substantially increase statistical power by explaining between- and within-school variance. Pretest scores typically explain over three quarters of the between-school variance and over one half of the within-school variance in all schools and in low-SES schools, but they explain somewhat less variance in low-achievement schools. Demographic characteristics are less effective covariates, but they can explain up to one half of the between-school variance in all and low-SES schools. In general, demographic characteristics, when used in addition to pretest scores, explain little additional variance. The remainder of this article gives the methods and data sources that were used, presents the results in detail, and illustrates how to use these results to compute statistical power.

Dimensions of Designs Considered

Our analyses focused on intraclass correlations for designs involving the assignment of schools to treatments. Unfortunately, there is a wide variety of designs that might be used to study education interventions, and each of these designs may have its own intraclass correlation (or conditional intraclass correlation) structure. To attempt to provide a reasonable coverage of the designs most likely to be of interest to researchers planning educational experiments, we considered four dimensions of intervention designs. The first dimension of the design is the grade level. The second dimension of the design is what achievement domain (e.g., reading or mathematics) is the dependent variable. The third dimension of the design is the set of covariates that were used in the analysis, if any. Finally, the fourth dimension is the SES or achievement status of schools sampled in the overall population of schools. These four dimensions of designs can vary independently. We examined all possible combinations of them.

Grade Level of Students and Achievement Domain

We examined each grade level from kindergarten through Grade 12 and both mathematics and reading achievement at each grade level, with one exception. The exception was reading achievement in Grade 11, for which data on a national representative sample were not available to us.

Covariates Used in the Design

We consider four data analysis models involving different covariate sets that we believe are likely to be of considerable interest to educational researchers. The first, the unconditional model, involves the testing of treatment effects with no covariates. This is the minimal design but one that is likely to be of interest in many settings in which researchers have little opportunity to collect prior information about the individuals participating in the experiment.

The second model, which we call the demographic covariates model, involves the testing of treatment effects conditional on covariates that are ascriptive characteristics of students frequently invoked in models of educational achievement, namely, gender, race or ethnicity, and SES. This design may be appropriate when researchers can obtain prior, contemporaneous, or retrospective data from administrative records (appropriate because these covariates are unlikely to change).

The third model, which we call the pretest covariates model, involves the testing of treatment effects using pretest scores on the same achievement domain (mathematics or reading) as a covariate. This design is likely to be considerably

more powerful than the previous designs but involves the additional cost of collecting another wave of test data and the additional organizational burden of making that data collection in a timely manner.

The fourth model, which we call the pretest and demographic covariates model, involves the testing of treatment effects using the ascriptive characteristics of students (gender, race or ethnicity, and SES) and pretest scores on the same achievement domain as the covariates. This design combines both of the sets of covariates in the previous design.

SES or Achievement Status of Schools Within Their Settings

Some experimenters undoubtedly wish to use a representative sample of schools within whatever setting they choose to study. Consequently, one population of schools we considered was the entire collection of schools within a setting.

Researchers sometimes make decisions to carry out their studies in schools that lie within the middle range of outcomes, omitting schools that have had (or are reputed to have had) the very poorest and the very best outcomes, on the rationale that neither the very poorest schools nor the very best schools give a fair test of an intervention. We operationalized this notion by ordering, on average achievement, the entire sample of schools in a setting and selecting the middle 80% of the schools in each setting, omitting the top and bottom 10% of the schools.

Some interventions are designed to be compensatory. Experimenters investigating such interventions might choose only schools within a particular context that have low mean achievement or large numbers of low-SES students to evaluate the intervention. We operationalized low achievement by ordering, on average achievement, the entire sample of schools in a setting and selecting the lower 50% of the schools, omitting the upper 50% of the schools. We operationalized low SES by ordering, on the proportion of students eligible for free or reduced-price lunch, the entire sample of schools in a setting and selecting the upper 50% of the schools, omitting the bottom 50% of the schools. One might argue for a more extreme definition of low-SES or low-achievement schools (e.g., the lower 30% of schools). We chose the lower 50% of schools to achieve a balance between the construct definition (low achievement or low SES) and sufficient sample size to obtain sufficiently precise estimates of the parameters of interest. The choice we made yields some standard errors that are on the order of .02, corresponding to a 2-*SE* band on either side of the estimate (a very crude 95% confidence interval) of width .08. Because even this range is large enough to have important substantive consequences, we judged that restricting the proportion of schools in the definition of the low-SES or low-achievement sample (which would decrease sample sizes of those groups) would lead to unacceptable impreciseness.

Data Sets Used

The object of this article is to estimate intraclass correlations and associated variance components for academic achievement in reading and mathematics for the United States and various subpopulations. Consequently, we relied on data from longitudinal surveys with national probability samples, all of which are described in detail elsewhere. We chose longitudinal surveys because we wished to use achievement data collected in earlier years as pretest data for evaluating conditional intraclass correlations relevant for planning studies that would use a pretest as a covariate. In some cases, more than one survey could have provided data on a given grade level. In such cases, we generally report here results on the basis of the survey with the largest sample size, although we made an exception to this principle when the larger sample was for the base year of a longitudinal study that would have provided no pretest data. Some general information about the surveys used in our main analyses is reported in Table 1.

The results reported for kindergarten, Grade 1, and Grade 3 were obtained from three waves of the Early Childhood Longitudinal Survey (ECLS). The ECLS is a longitudinal study that obtained a national probability sample of kindergarten children in 1,591 schools in 1998 and followed them through the fifth grade (see Tourangeau et al., 2005). Achievement test data were collected in both fall and spring of kindergarten and first grade and in spring only in

TABLE 1
Characteristics of Data Sets Used in This Analysis

Grade	Data set domain	Achievement	Achievement test	Number of schools	Number of students
K	ECLS	Reading	ECLS Direct Cognitive Assessment	1,591	20,649
	ECLS	Math	ECLS Direct Cognitive Assessment	1,591	20,649
1	ECLS	Reading	ECLS Direct Cognitive Assessment	689	5,286
	ECLS	Math	ECLS Direct Cognitive Assessment	689	5,286
2	Prospects1	Reading	Comprehensive Test of Basic Skills (4th ed.)	445	8,643
	Prospects1	Math	Comprehensive Test of Basic Skills (4th ed.)	445	8,643
3	ECLS	Reading	ECLS Direct Cognitive Assessment	2,767	14,051
	ECLS	Math	ECLS Direct Cognitive Assessment	2,767	14,051
4	Prospects3	Reading	Comprehensive Test of Basic Skills (4th ed.)	320	10,866
	Prospects3	Math	Comprehensive Test of Basic Skills (4th ed.)	320	10,866
5	Prospects3	Reading	Comprehensive Test of Basic Skills (4th ed.)	605	9,928
	Prospects3	Math	Comprehensive Test of Basic Skills (4th ed.)	605	9,928
6	Prospects3	Reading	Comprehensive Test of Basic Skills (4th ed.)	669	8,335
	Prospects3	Math	Comprehensive Test of Basic Skills (4th ed.)	669	8,335
7	Prospects7	Reading	Comprehensive Test of Basic Skills (4th ed.)	127	7,319
	LSAY7	Math	LSAY Math test based on NAEP items	52	3,116
8	NELS	Reading	NELS Reading Grade 8	1,050	24,562
	LSAY7	Math	LSAY Math test based on NAEP items	264	2,958
9	Prospects7	Reading	Comprehensive Test of Basic Skills (4th ed.)	312	4,704
	LSAY7	Math	LSAY Math test based on NAEP items	442	2,786
10	NELS	Reading	NELS Reading Grade 10	1,288	17,624
	NELS	Math	NELS Math Grade 10	1,288	17,624
11	—	Reading	—		
	LSAY10	Math	LSAY Math test based on NAEP items	163	2,579
12	NELS	Reading	NELS Reading Grade 12	1,138	14,913
	NELS	Math	NELS Math Grade 12	1,138	14,913

Note. ECLS = Early Childhood Longitudinal Survey; LSAY = Longitudinal Study of American Youth; NAEP = National Assessment of Educational Progress; NELS = National Educational Longitudinal Study. Because Prospects and LSAY involve more than one cohort followed longitudinally, each cohort of Prospects and LSAY is identified by the grade level of the base year for that cohort. Thus, Prospects1 is the cohort of Prospects that began in Grade 1, LSAY7 is the cohort of LSAY that began in Grade 7, and so on.

third and fifth grades. There was no data collection in second and fourth grades. Thus, fall achievement test data collected in the same year could serve as a pretest in kindergarten and first grades, while data collected in the spring of the first grade served as pretest data for the third grade.

The results reported for Grade 2 were obtained from the first follow-up to the first grade (base year) sample, and those reported for Grades 4–6 were obtained from the three follow-ups of the third grade (base year) sample in the Prospects study. The results in reading in Grades 7 and 9 were obtained from the base year and the second follow-up of the seventh grade sample in the Prospects study. Prospects was actually a set of three longitudinal studies, starting with (base year) national probability samples of children in 235, 240, and 137 schools, in Grades 1, 3, and 7, respectively, conducted in 1991 (for a complete description of the study design, see Puma, Karweit, Price, Riccuti, & Vaden-Kiernan, 1997). Achievement test data were collected for 3–4 years thereafter for each sample. Thus, the three Prospects studies collected data in Grades 1 (both fall and spring), 2, and 3; Grades 3, 4, 5, and 6; and Grades 7, 8, and 9. There were pretest data in the base year for Grade 1, but no pretest data for the base years in Grades 3 and 7. For all years except the base year, the previous year's achievement test data were used as a pretest, and in Grade 1, the test data collected in fall served as a pretest.

The results reported on reading in Grades 8, 10, and 12 and mathematics in Grades 10 and 12 were obtained from the National Educational Longitudinal Study of the Eighth Grade Class of 1988, a longitudinal study that began in 1988 with a national probability sample of eighth graders in 1,050 schools and collected reading and mathematics achievement test data when the students were in Grades 8, 10, and 12 (Curtin et al., 2002). Thus, no pretest data were available for Grade 8, but for Grade 10, the Grade 8 data were used as a pretest, and for Grade 12, the Grade 10 data were used as a pretest.

Finally, the results on mathematics in Grades 7, 8, 9, and 11 were obtained from the base year and follow-ups of the Longitudinal Study of American Youth (LSAY; see J. D. Miller, Hoffer, Suchner, Brown, & Nelson, 1992). The LSAY is a longitudinal study that began in 1987 with two national probability samples, one of 7th graders and one of 10th graders in 104 schools. Data were collected on mathematics and science achievement each year for 4 years, leading to samples from Grades 7 to 12. There were no pretest data in Grade 7, but the previous year's data served as the pretest for each subsequent year.

Analysis Procedures

The data analysis was carried out using Stata 9.1's XTMIXED routine for mixed linear model analysis. For each sample and achievement domain, analyses were carried out on the basis of four different models, which we call the unconditional model, the pretest covariate model, the demographic covariates model, and the pretest and demographic covariates model. We describe these explicitly below in hierarchical linear model notation.

The Unconditional Model

The unconditional model involves no covariates at either the individual or school (cluster) level. The Level 1 model for the *k*th observation in the *j*th school can be written as

$$Y_{jk} = \beta_{0j} + \varepsilon_{jk},$$

and the Level 2 model for the intercept is

$$\beta_{0j} = \pi_{00} + \zeta_j,$$

where ε_{jk} is an individual-level residual for the *k*th person in the *j*th school, and ζ_j is a random effect (a Level 2 residual) associated with the *j*th school. In this analysis, the between-person, within-school variance component is σ_W^2 (the variance of ε_{jk}), and the between-school variance component is σ_B^2 (the variance of ζ_j).

The Pretest Covariate Model

If pretest scores on achievement are available, they can be a powerful covariate and considerably increase power in experimental designs. The pretest covariate model involves using as a covariate the cluster-centered pretest score at the individual level and the school mean pretest score at the school level. We used group (school) mean centering because it leads to more stable estimates of variance components when, as in the present analyses, the covariate values vary substantially across schools (see Raudenbush & Bryk, 2002, p. 143). Thus, the Level 1 model for the *k*th observation in the *j*th school can be written as

$$Y_{jk} = \beta_{0j} + \beta_{1j}(X_{jk} - \bar{X}_{j\bullet}) + \varepsilon_{jk},$$

and the Level 2 model for the intercept is

$$\beta_{0j} = \pi_{00} + \pi_{01}\bar{X}_{j\bullet} + \zeta_j,$$

where X_{jk} is the achievement pretest score for the *j*th observation in the *k*th school, $\bar{X}_{j\bullet}$ is the pretest mean for the *j*th school, ε_{jk} is an individual-level residual, and ζ_j is a random effect of the *j*th school (a Level 2 residual); the covariate slope β_{1j} was treated as equal in all clusters (schools). The variance components associated with this analysis are σ_{AW}^2 (the variance of ε_{jk}) and σ_{AB}^2 (the variance of ζ_j). In this analysis, the covariate-adjusted between-person, within-school variance component is σ_{AW}^2 (the variance of ε_{jk}), and the covariate-adjusted between-school variance component is σ_{AB}^2 (the variance of ζ_j).

The Demographic Covariates Model

Sometimes pretest scores are not available, but other background information about individuals is available to serve as covariates. The demographic covariates model includes four covariates at each of the individual and group (cluster) levels. At the individual level, the covariates are

dummy variables for male gender and for Black or Hispanic status and an index of mother's and father's levels of education as a proxy for SES. As recommended by Raudenbush and Bryk (2002), each of these individual-level covariates was group centered, that is, transformed by subtracting the group mean as shown in the equation for the Level 1 model below. The school-level covariates were the means of the individual-level variables for each school (cluster). Therefore, the Level 1 model for the *k*th observation in the *j*th school can be written as

$$Y_{jk} = \beta_{0j} + \beta_{1j}(G_{jk} - \bar{G}_{j\bullet}) + \beta_{2j}(B_{jk} - \bar{B}_{j\bullet}) + \beta_{3j}(H_{jk} - \bar{H}_{j\bullet}) + \beta_{4j}(E_{jk} - \bar{E}_{j\bullet}) + \varepsilon_{jk},$$

where G_{jk}, B_{jk}, and H_{jk} are dummy variables for male gender, Black status, and Hispanic status, respectively; E is an index of mother's and father's levels of education (which is a proxy for family SES); and $\bar{G}_{j\bullet}$, $\bar{B}_{j\bullet}$, $\bar{H}_{j\bullet}$, and $\bar{E}_{j\bullet}$ are the means of G, B, H, and E in the *j*th school (cluster). The Level 2 model for the intercept is

$$\beta_{0j} = \pi_{00} + \pi_{10}\,\bar{G}_{j\bullet} + \pi_{20}\,\bar{B}_{j\bullet} + \pi_{30}\,\bar{H}_{j\bullet} + \pi_{40}\,\bar{E}_{j\bullet} + \zeta_j,$$

and the covariate slopes β_{1j}, β_{2j}, β_{3j}, and β_{4j} were treated as equal in all clusters (schools). In this analysis the covariate-adjusted between-person, within-school variance component is σ_{AW}^2 (the variance of ε_{jk}), and the covariate-adjusted between-school variance component is σ_{AB}^2 (the variance of ζ_j).

The Pretest and Demographic Covariates Model

The pretest and demographic covariates model combines the use of an achievement pretest and the individual characteristics of gender, minority group status, and parent's education as individual- and school-level covariates. Therefore, the Level 1 model for the *k*th observation in the *j*th school can be written as

$$Y_{jk} = \beta_{0j} + \beta_{1j}(X_{jk} - \bar{X}_{j\bullet}) + \beta_{2j}(G_{jk} - \bar{G}_{j\bullet}) + \beta_{3j}(B_{jk} - \bar{B}_{j\bullet}) + \beta_{4j}(H_{jk} - \bar{H}_{j\bullet}) + \beta_{5j}(E_{jk} - \bar{E}_{j\bullet}) + \varepsilon_{jk},$$

where all of the symbols are defined as in the models above. The Level 2 model for the intercept is

$$\beta_{0j} = \pi_{00} + \pi_{10}\,\bar{X} + \pi_{20}\,\bar{G}_{j\bullet} + \pi_{30}\,\bar{B}_{j\bullet} + \pi_{40}\,\bar{H}_{j\bullet} + \pi_{50}\,\bar{E}_{j\bullet} + \zeta_j,$$

and the covariate slopes β_{1j}, β_{2j}, β_{3j}, β_{4j}, and β_{5j} were treated as equal in all clusters (schools). In this analysis, the covariate-adjusted between-person, within-school variance component is σ_{AW}^2 (the variance of ε_{jk}), and the covariate-adjusted between-school variance component is σ_{AB}^2 (the variance of ζ_j).

The Intraclass Correlation Data

The (unconditional) intraclass correlation associated with the unconditional model described above is

$$\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2) = \sigma_B^2/\sigma_T^2, \quad (1)$$

where $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$ is the (unconditional) total variance. Note that the residuals ε_{jk} and ζ_j correspond to the within- and between-cluster cluster random effects in an experiment that assigned schools to treatments. Consequently, the variance components associated with these random effects and the intraclass correlation correspond to those in a cluster-randomized experiment that assigned schools to treatments and analyzed the data with no covariates.

In the three models involving covariate adjustment, the (covariate-adjusted) intraclass correlation is

$$\rho_A = \sigma_{AB}^2/(\sigma_{AB}^2 + \sigma_{AW}^2) = \sigma_{AB}^2/\sigma_{AT}^2, \quad (2)$$

where $\sigma_{AT}^2 = \sigma_{AB}^2 + \sigma_{AW}^2$ is the (covariate-adjusted) total variance. Note that the residuals ε_{jk} and ζ_j correspond to the within- and between-cluster cluster random effects in an experiment that assigned schools to treatments and used the same covariates as were used in the models with covariates. Consequently, the variance components associated with these random effects and the conditional intraclass correlation ρ_A correspond to those in a cluster-randomized experiment that assigned schools to treatments

and analyzed the data with these (individual and school mean) characteristics as covariates.

For each combination of design dimensions (i.e., for each grade level, achievement domain, covariate set, setting, and choice of SES or achievement status within setting), we estimated the intraclass correlation (or conditional intraclass correlation) via restricted maximum likelihood using Stata and computed the standard error of that intraclass correlation estimate using the result given in Donner and Koval (1982). This resulted in 13 (grade levels) × 2 (achievement domains) × 4 (covariate sets) × 4 (SES or achievement statuses within settings) = 416 intraclass correlation estimates (each with a corresponding standard error).

For designs that use covariates, we also provide values of

$$\eta_B^2 = \sigma_{AB}^2/\sigma_B^2, \quad (3)$$

the proportion of between-school variance remaining, and

$$\eta_W^2 = \sigma_{AW}^2/\sigma_W^2, \quad (4)$$

the proportion of within-school variance remaining, respectively, after covariate adjustment. For designs involving covariates, these two auxiliary quantities (η_B^2 and η_W^2) are useful in computing statistical power. Their use is illustrated in a subsequent section of this article.

Two alternative parameters that contain the same information as η_B^2 and η_W^2 are $R_B^2 = 1 - \eta_B^2$ and $R_W^2 = 1 - \eta_W^2$, the proportion of between- and within-school variance explained by the covariate. We chose to tabulate the η^2 values instead of the R^2 values because the relation of the η^2 values to the noncentrality parameters used in power analysis is simpler.

Note that each of the four analyses involved slightly different variables, and there were missing values on some of these variables in our survey data. We decided to compute each analysis on the largest set of cases that had all of the necessary variables for the analysis in question. This means that each of the four analyses of a given data set is computed on a slightly different set of cases. Because the quantities η_W^2 and η_B^2 involve a comparison of two different analyses (one with and one without a particular set of covariates), we believed that it was important to make this comparison using estimates derived from exactly the same set of cases. Consequently, for each of the analyses that involved covariates, we recomputed the estimates of the unadjusted variance components, σ_W^2 and σ_B^2, using only the cases that were used to compute the adjusted variance components σ_{AW}^2 and σ_{AB}^2 and used these particular estimates to compute the η_W^2 and η_B^2 values given here.

Although we provide estimates of the standard errors of the intraclass correlations, they should be used with some caution for two reasons. First, the distribution of estimates of the intraclass correlations is only approximately normal. Second, not all of these values are independent of one another, and it is not immediately clear how to carry out a formal statistical analysis of differences between estimates of intraclass correlations computed from the same sample of individuals. Nevertheless, we feel that these standard errors are useful as descriptions of the uncertainty of the individual estimates of intraclass correlations.

Results

We found that the intraclass correlations obtained in the nationally representative sample and the schools in the middle 80% of the achievement distribution had intraclass correleations that were almost identical. Consequently, we present results here only the intraclass correlation data from the entire national sample of schools, those in the upper half of the free and reduced-price lunch distribution (low-SES schools), and those in the lower half of the school mean achievement distribution (low-achievement schools).

The main results of this study are presented in Tables 2–7 and discussed in the sections that follow. Each table is divided into four vertical panels of three columns each, one panel for each of the four analyses described above. The data for each grade level are given in a different row. In the row for each grade, the columns of each panel provide the estimates of the intraclass correlation (ρ), the standard error of the estimate of ρ, and (for all but the unconditional

TABLE 2
Intraclass Correlations (ICCs) and Variance Components for Mathematics Achievement: All Schools

	Unconditional model		Demographic covariates model				Pretest covariate model				Pretest and demographic covariates model			
Grade	ICC	*SE*	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2
K[a]	.243	.010	.110	.007	.384	.920	.107	.007	.143	.379	.102	.007	.143	.371
1[a]	.228	.010	.101	.014	.386	.921	.125	.013	.177	.376	.119	.015	.186	.373
2[b]	.236	.019	.148	.016	.564	.912	.185	.018	.324	.495	.169	.019	.322	.489
3[a]	.241	.010	.102	.009	.361	.912	.130	.008	.195	.406	.113	.009	.175	.387
4[c]	.232	.020	.133	.015	.565	.934	.170	.017	.321	.515	.140	.017	.296	.502
5[c]	.216	.018	.127	.015	.558	.928	.160	.016	.368	.494	.170	.018	.421	.481
6[c]	.264	.019	.174	.042	.883	.931	.139	.015	.260	.498	.194	.048	.458	.525
7[d]	.191	.033	.088	.019	.362	.904	—	—	—	—	—	—	—	—
8[d]	.185	.032	.122	.025	.567	.916	.106	.022	.178	.347	.106	.023	.179	.340
9[d]	.216	.032	.122	.025	.477	.903	.099	.023	.105	.276	.080	.020	.085	.264
10[e]	.234	.010	.067	.006	.220	.908	.066	.006	.081	.351	.062	.006	.076	.345
11[f]	.138	.028	.045	.014	.261	.879	.092	.022	.165	.270	.075	.020	.131	.261
12[e]	.239	.011	.069	.007	.218	.898	.038	.005	.025	.202	.034	.005	.024	.199
M	.220		.108		.447	.913	.118		.195	.384	.114		.208	.378
a	.242		.136		.540	.927	.161		.276	.482	.156		.296	.475
b	–.004		–.005		–.016	–.002	–.007		–.014	–.017	–.007		–.015	–.016
r	–.443		–.514		–.330	–.635	–.694		–.528	–.659	–.600		–.440	–.633

a. These data are from the Early Childhood Longitudinal Survey.
b. These data are from the Prospects cohort from the 1st grade base year.
c. These data are from the Prospects cohort from the 3rd grade base year.
d. These data are from the Longitudinal Study of American Youth (LSAY) cohort from the 7th grade base year.
e. These data are from the National Educational Longitudinal Study.
f. These data are from the LSAY cohort from the 10th grade base year.

model given in the first panel on the left-hand side) estimates of η_B^2 and η_W^2. For example, consider the data in Table 2 for the pretest covariate model for Grade 1, given in the third panel of the table. On the row associated with Grade 1, the values in the columns of the third panel (columns 8–11 of the table) are .125, .0135, .177, and .376, respectively, which correspond to estimates for ρ_A, the standard error of the estimate of ρ_A, η_B^2, and η_W^2.

To help interpret the tables as a whole, the bottom four rows of each table give summary statistics (across grades) of the estimates of ρ_A, η_B^2, and η_W^2, including the mean, the intercept (*a*) and slope (*b*) of an unweighted regression of the estimates on grade level (with kindergarten equaling Grade 0), and the correlation (*r*) between estimates and grade level. For example in Table 2, the mean intraclass correlation in the unconditional model is .220, the correlation between grade and intraclass correlation is –.443, and the regression equation for predicting the unconditional intraclass correlation from grade is .242 – .004(grade).

Mathematics Achievement in the Full Population

Table 2 is a presentation of results from the entire national sample in mathematics. The average unconditional intraclass correlation estimate across all grades is .220. Although there is a tendency of the intraclass correlations to be larger at lower grades, in general, there are not large changes across adjacent grade levels. Few of these differences exceed 2 standard errors of the difference. A notable exception is the unadjusted intraclass correlation for Grade 11, for which the difference between Grade 11 and either of the adjacent grades is about 3

TABLE 3
Intraclass Correlations (ICCs) and Variance Components for Reading Achievement: All Schools

	Unconditional model		Demographic covariates model				Pretest covariate model				Pretest and demographic covariates model			
Grade	ICC	*SE*	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2
K[a]	.233	.010	.144	.008	.566	.919	.166	.009	.258	.379	.165	.009	.268	.361
1[a]	.239	.010	.118	.015	.392	.916	.167	.016	.210	.360	.145	.016	.201	.349
2[b]	.204	.018	.109	.014	.441	.890	.080	.010	.170	.478	.056	.010	.113	.445
3[a]	.271	.011	.089	.008	.259	.921	.135	.009	.241	.522	.083	.008	.159	.521
4[c]	.242	.020	.088	.012	.296	.900	.123	.014	.188	.460	.101	.014	.158	.451
5[c]	.263	.020	.061	.009	.202	.899	.113	.013	.170	.435	.085	.012	.133	.418
6[c]	.260	.019	.065	.033	.366	.924	.072	.010	.118	.490	.025	.031	.089	.578
7[d]	.174	.020	.036	.009	.185	.903	—	—	—	—	—	—	—	—
8[e]	.197	.009	.051	.004	.207	.915	—	—	—	—	—	—	—	—
9[d]	.250	.026	.186	.025	.576	.889	.314	.029	.651	.541	.322	.033	.575	.525
10[e]	.183	.009	.063	.006	.283	.907	.063	.006	.144	.471	.059	.006	.133	.462
12[e]	.174	.010	.053	.006	.252	.909	.055	.006	.108	.383	.050	.006	.101	.382
M	.224		.089		.335	.908	.107		.226	.452	.109		.193	.449
a	.251		.113		.409	.911	.138		.210	.434	.113		.178	.423
b	–.005		–.004		–.013	–.001	–.006		.003	.004	–.001		.003	.005
r	–.505		–.378		–.367	–.191	–.241		.077	.229	–.033		.080	.270

a. These data are from the Early Childhood Longitudinal Survey.
b. These data are from the Prospects cohort from the first grade base year.
c. These data are from the Prospects cohort from the third grade base year.
d. These data are from the Prospects cohort from the seventh grade base year.
e. These data are from the National Educational Longitudinal Study.

standard errors of the difference. None of the differences between adjusted intraclass correlations in adjacent grades is a large as 3 standard errors of the difference, but the values for Grade 2 are somewhat higher (by over 2 standard errors of the difference) and those for Grade 3 somewhat lower than those of adjacent grades.

The linear regression coefficients (the intercept *a* and slope *b*) of each of the tabled quantities on grade given at the bottom of each column of the table permits the computation of smoothed estimates of each quantity $a + b(\text{grade})$. For example, the values of *a* and *b* for the unadjusted intraclass correlation are $a = .242$ and $b = -.004$, so that the smoothed (interpolated) value of the unadjusted intraclass correlation for Grade 11 would be $.242 + (-0.004)11 = .198$, somewhat higher than the tabled value of .138.

The patterns of reduction of between- and within-cluster (school) variances are generally quite different in models involving different covariates. Specifically, the demographic covariate analyses typically reduced the between-cluster variance to one half to one quarter of its value in the unconditional model (e.g., produced η_B^2 from .5 to .25), but typically reduced within-cluster variance by 10% or less (e.g., produced η_W^2 values greater than 0.9). Thus the use of ascriptive characteristics as covariates (as in the demographic covariates model) may lead to increased statistical power. The residualized analyses using pretest score as a covariate typically resulted in larger reductions in between-cluster variance (e.g., produced η_B^2 values from .3 to .1) and typically also reduced within-cluster variance by a much larger amount than the demographic covariates model (e.g., produced η_W^2 values from .25 to .5). In general, demographic characteristics explain little additional variance (at either the student or the school level) beyond what is explained by the pretest, and thus their inclusion in analysis models does not appear to be useful if pretest scores are available.

There is one apparent anomaly in the results reported in Table 2. The η^2 values for the pretest and demographic covariates model are often larger than those for the pretest covariate model. This is equivalent to saying that the estimated

TABLE 4
Intraclass Correlations (ICCs) and Variance Components for Mathematics Achievement: Low–Socioeconomic Status Schools

	Unconditional model		Demographic covariates model				Pretest covariate model				Pretest and demographic covariates model			
Grade	ICC	*SE*	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2
K[a]	.218	.011	.108	.008	.420	.912	.114	.008	.176	.378	.108	.009	.171	.368
1[a]	.223	.011	.088	.015	.352	.924	.116	.015	.179	.382	.108	.017	.181	.380
2[b]	.200	.020	.151	.018	.686	.912	.184	.020	.364	.481	.172	.021	.360	.473
3[a]	.208	.012	.107	.011	.450	.910	.127	.010	.220	.393	.115	.011	.206	.371
4[c]	.217	.021	.144	.018	.702	.934	.184	.020	.388	.522	.159	.021	.386	.505
5[c]	.182	.018	.125	.016	.677	.933	.170	.018	.458	.492	.179	.021	.527	.484
6[c]	.249	.021	.176	.044	1.000	.940	.134	.016	.270	.493	.239	.051	.612	.502
7[d]	.195	.034	.087	.019	.350	.906	—	—	—	—	—	—	—	—
8[d]	.185	.032	.120	.025	.558	.919	.116	.024	.193	.341	.116	.025	.194	.333
9[d]	.177	.034	.039	.016	.198	.921	.082	.024	.102	.274	.048	.018	.064	.265
10[e]	.174	.011	.067	.008	.316	.908	.063	.007	.113	.355	.060	.007	.108	.349
11[f]	.134	.035	.058	.022	.331	.869	.126	.034	.239	.266	.111	.032	.179	.248
12[e]	.172	.012	.065	.009	.324	.896	.037	.007	.038	.200	.041	.008	.045	.195
M	.195		.103		.489	.914	.121		.228	.381	.121		.253	.373
a	.227		.137		.606	.928	.161		.310	.479	.160		.346	.470
b	–.005		–.006		–.019	–.002	–.007		–.014	–.016	–.007		–.016	–.016
r	–.707		–.548		–.339	–.478	–.601		–.452	–.654	–.454		–.356	–.643

a. These data are from the Early Childhood Longitudinal Survey.
b. These data are from the Prospects cohort from the 1st grade base year.
c. These data are from the Prospects cohort from the 3rd grade base year.
d. These data are from the Longitudinal Study of American Youth (LSAY) cohort from the 7th grade base year.
e. These data are from the National Educational Longitudinal Study.
f. These data are from the LSAY cohort from the 10th grade base year.

variance accounted for decreases when ascriptive characteristics are added as covariates to the model that already has pretest as a covariate. It is theoretically possible for this to occur in multilevel models when the actual differences are negligible, as they appear to be here. The difference is particularly large in the sixth grade data, however, and appears to be a consequence of differences between the samples used to estimate the two models. For unknown reasons, there is a considerable amount of missing data on the demographic covariates used to create the demographic covariates and pretest and demographic covariates models in the survey providing the sixth grade data (the third follow-up of the Prospects cohort that began in third grade). The same pattern is evident, but to a lesser extent, in the fifth grade η_B^2 data (based on the second follow-up of the Prospects cohort that began in third grade). We suggest using these values only with great caution. It might be wise to use the smoothed values for the pretest and demographic covariates model in Grade 6 (which would give $\eta_B^2 = .207$ and $\eta_W^2 = .377$) and possibly in Grade 5 (which would give $\eta_B^2 = .222$ and $\eta_W^2 = .393$).

Reading Achievement in the Full Population

Table 3 is a presentation of results from the entire national sample in reading, organized in the same way as Table 2 which reports results for mathematics. The intraclass correlation and adjusted intraclass correlation values in reading are generally quite similar to those in mathematics. The mean (across grade levels) unconditional intraclass correlation in reading was .224. As in mathematics, there is a tendency of the intraclass correlations in reading to become smaller at higher grades, but the changes across adjacent grade levels are often larger. The results for Grade 9 are particularly inconsistent with (having larger

TABLE 5
Intraclass Correlations (ICCs) and Variance Components for Reading Achievement: Low–Socioeconomic Status Schools

	Unconditional model		Demographic covariates model				Pretest covariate model				Pretest and demographic covariates model			
Grade	ICC	*SE*	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2
K[a]	.215	.011	.144	.010	.617	.910	.168	.010	.307	.397	.166	.011	.314	.377
1[a]	.227	.011	.118	.018	.383	.919	.152	.017	.196	.366	.145	.019	.199	.357
2[b]	.181	.018	.119	.016	.533	.891	.066	.010	.155	.484	.050	.010	.108	.449
3[a]	.223	.012	.098	.011	.355	.908	.123	.010	.267	.495	.085	.010	.197	.493
4[c]	.214	.021	.096	.014	.385	.896	.138	.017	.253	.471	.113	.017	.217	.467
5[c]	.230	.021	.061	.011	.246	.905	.123	.015	.222	.440	.089	.014	.165	.420
6[c]	.221	.020	.059	.033	.500	.920	.070	.011	.137	.494	.023	.027	.125	.576
7[d]	.173	.023	.052	.014	.230	.908	—	—	—	—	—	—	—	—
8[e]	.137	.010	.057	.006	.361	.905	—	—	—	—	—	—	—	—
9[d]	.236	.032	.131	.027	.410	.897	.213	.033	.412	.538	.231	.038	.363	.524
10[e]	.131	.010	.056	.007	.381	.905	.047	.007	.163	.470	.047	.007	.166	.463
12[e]	.131	.011	.044	.008	.297	.906	.050	.007	.134	.367	.041	.008	.118	.365
M	.193		.086		.391	.906	.115		.225	.452	.099		.197	.449
a	.231		.122		.477	.908	.142		.242	.441	.119		.214	.430
b	–.007		–.006		–.015	.000	–.005		–.003	.002	–.004		–.003	.004
r	–.620		–.687		–.508	–.126	–.373		–.150	.143	–.242		–.154	.205

a. These data are from Early Childhood Longitudinal Survey.
b. These data are from the Prospects cohort from the first grade base year.
c. These data are from the Prospects cohort from the third grade base year.
d. These data are from the Prospects cohort from the seventh grade base year.
e. These data are from the National Educational Longitudinal Study.

values of the intraclass correlations than) the results from either Grade 8 or Grade 10. The results from Grade 2 are also somewhat different (having smaller values of the intraclass correlations) than the results from either Grade 1 or Grade 3. Several of these differences exceed 3 standard errors of the difference. Few of the other differences exceed 2 standard errors of the difference.

There is less consistency in reading than in mathematics among the adjusted intraclass correlations for the three models involving covariates. However, the general pattern of reduction in between- versus within-cluster variance was similar in reading and in mathematics. That is, there was somewhat greater reduction in between-cluster variance and much greater reduction in within-cluster variance in the pretest covariate model than in the demographic covariates model. As in the case of mathematics achievement in the full population, the pretest and demographic covariates model leads to little additional variance explained at either the school or the individual level compared with the model using only pretest as a covariate.

Mathematics Achievement in Low-SES Schools

Table 4 is a presentation of results in mathematics computed for the schools in the bottom half of the school SES distribution (operationalized by the proportion of students eligible for free or reduced-price lunch) and is organized in the same way as Tables 2 and 3. The mean (across grade levels) unconditional intraclass correlation in mathematics was .195. There is a tendency for the intraclass correlation values in this sample to be a bit smaller than those reported in Table 2 for the entire national population, a tendency that does not hold for the conditional (adjusted) intraclass correlations.

There is one substantial anomaly in the results reported in Table 4 that is similar to that in Table 2: The η^2 values for the pretest and demographic covariates model are sometimes larger than those for the pretest covariate model, a difference that is particularly large at Grade 6. This anomaly (like that in Table 2) appears to be a consequence of differences between the samples used

TABLE 6
Intraclass Correlations (ICCs) and Variance Components for Mathematics Achievement: Low-Achievement Schools

	Unconditional model		Demographic covariates model				Pretest covariate model				Pretest and demographic covariates model			
Grade	ICC	*SE*	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2
K[a]	.113	.009	.044	.008	.347	.959	.073	.008	.382	.612	.064	.009	.329	.625
1[a]	.089	.009	.053	.017	.556	.969	.085	.016	.506	.568	.068	.018	.459	.594
2[b]	.111	.015	.067	.014	.804	.982	.092	.014	.480	.641	.088	.018	.635	.675
3[a]	.102	.010	.050	.011	.503	.976	.077	.010	.411	.554	.069	.012	.411	.553
4[c]	.134	.016	.081	.015	.864	.989	.127	.017	.709	.796	.101	.019	.815	.826
5[c]	.059	.010	.041	.011	.811	.981	.080	.013	.838	.767	.075	.016	.888	.784
6[c]	.082	.013	.078	.042	1.000	.924	.098	.015	1.000	.771	.147	.054	1.000	.660
7[d]	.045	.015	.037	.014	.794	.982	—	—	—	—	—	—	—	—
8[d]	.085	.023	.073	.022	.876	.958	.067	.020	.552	.685	.056	.019	.486	.666
9[d]	.081	.024	.066	.023	.790	.953	.056	.021	.429	.558	.054	.021	.418	.550
10[e]	.076	.008	.050	.008	.641	.972	.065	.009	.622	.752	.065	.009	.641	.736
11[f]	.081	.024	.042	.018	.531	.930	.085	.025	.525	.502	.072	.024	.466	.484
12[e]	.080	.010	.051	.010	.626	.962	.042	.008	.234	.443	.050	.010	.288	.448
M	.087		.056		.703	.964	.079		.557	.638	.076		.570	.633
a	.106		.058		.645	.976	.095		.582	.678	.085		.607	.694
b	–.003		.000		.010	–.002	–.003		–.004	–.007	–.002		–.006	–.010
r	–.511		–.079		.205	–.370	–.494		–.082	–.239	–.250		–.113	–.361

a. These data are from the Early Childhood Longitudinal Survey.
b. These data are from the Prospects cohort from the 1st grade base year.
c. These data are from the Prospects cohort from the 3rd grade base year.
d. These data are from the Longitudinal Study of American Youth (LSAY) cohort from the 7th grade base year.
e. These data are from the National Educational Longitudinal Study.
f. These data are from the LSAY cohort from the 10th grade base year.

to estimate the two models. As in Table 2, the same pattern is also evident, but to a lesser extent, in the fifth grade η_B^2 data. We suggest using these values only with great caution. It might be wise to use the smoothed values for the pretest and demographic covariates model in Grade 6 (which would give $\eta_B^2 = .195$ and $\eta_W^2 = .453$) and possibly in Grade 5 (which would give $\eta_B^2 = .192$ and $\eta_W^2 = .448$).

Reading Achievement in Low-SES Schools

Table 5 is a presentation of results in reading computed for the schools in the bottom half of the school SES distribution (operationalized by the proportion of students eligible for free or reduced-price lunch) and is organized in the same way as Tables 2–4. The mean (across grade levels) unconditional intraclass correlation is .193. As in the case of mathematics, there is a tendency for the intraclass correlation values in this sample to be a bit smaller than those reported in Table 3 for the entire national population, a tendency that does not hold for the conditional (adjusted) intraclass correlations.

Mathematics Achievement in Low-Achievement Schools

Table 6 is a presentation of results in mathematics computed for the schools in the bottom half of the distribution of school mean mathematics achievement and is organized in the same way as Tables 2–5. The mean (across grade levels) unconditional intraclass correlation in mathematics was .087. The intraclass correlation values in this sample are considerably smaller than those reported in Table 2 for the entire national population, a tendency that also holds for the conditional (adjusted) intraclass correlations. There is some variation of intraclass correlations across grade levels, but

TABLE 7
Intraclass Correlations (ICCs) and Variance Components for Reading Achievement: Low-Achievement Schools

	Unconditional model		Demographic covariates model				Pretest covariate model				Pretest and demographic covariates model			
Grade	ICC	*SE*	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2	ICC	*SE*	η_B^2	η_W^2
K[a]	.104	.008	.079	.009	.817	.948	.118	.010	.807	.712	.111	.011	.843	.707
1[a]	.142	.010	.066	.018	.472	.967	.158	.020	.592	.529	.129	.022	.572	.539
2[b]	.109	.014	.092	.016	.816	.967	.038	.009	.278	.783	.032	.012	.219	.780
3[a]	.139	.011	.080	.012	.494	.972	.075	.010	.381	.649	.057	.012	.301	.670
4[c]	.103	.013	.066	.013	.694	.978	.090	.014	.557	.717	.094	.018	.629	.742
5[c]	.071	.011	.027	.009	.477	.978	.085	.013	.764	.727	.057	.014	.707	.734
6[c]	.058	.011	.066	.039	1.000	.966	.056	.011	.734	.794	.025	.030	.395	.855
7[d]	.063	.012	.076	.020	.954	.968	—	—	—	—	—	—	—	—
8[e]	.070	.007	.044	.006	.636	.978	—	—	—	—	—	—	—	—
9[d]	.154	.023	.221	.031	.987	.964	.216	.028	1.000	.853	.292	.036	1.000	.873
10[e]	.050	.007	.044	.008	.882	.961	.050	.008	.895	.848	.056	.008	.949	.831
12[e]	.047	.008	.036	.009	.774	.956	.046	.009	.663	.684	.050	.010	.792	.685
M	.093		.075		.750	.967	.093		.667	.730	.091		.641	.742
a	.123		.074		.634	.967	.103		.526	.668	.081		.462	.681
b	–.006		.000		.021	.000	–.002		.027	.012	.002		.034	.012
r	–.545		.013		.401	–.037	–.134		.491	.490	.093		.515	.468

a. These data are from Early Childhood Longitudinal Survey.
b. These data are from the Prospects cohort from the first grade base year.
c. These data are from the Prospects cohort from the third grade base year.
d. These data are from the Prospects cohort from the seventh grade base year.
e. These data are from the National Educational Longitudinal Study.

only the difference between Grades 4 and 5 is larger than 2 standard errors of the difference. In general, the intraclass correlations at kindergarten through Grade 4 range from about .09 to .13, in Grades 5–7 from about .05 to .08, and in Grades 8–12 from .075 to .085.

The use of covariates resulted in a much smaller reduction in both between- and within-school variances in this sample than in the unrestricted sample. Specifically, the demographic covariates analyses typically reduced the between-school variance to no less than one half of its value in the unconditional model (e.g., produced η_B^2 from .5 to .8) but typically reduced within-cluster variance by 5% or less (e.g., produced η_W^2 values greater than .95). The pretest covariate analyses using pretest score as a covariate typically (but not always) resulted in modestly larger reductions in between-cluster variance (e.g., produced η_B^2 values from .3 to .8) but typically reduced within-cluster variance by a larger amount than the demographic covariates model (e.g., produced η_W^2 values from .5 to .8). As in the case of mathematics achievement in the full population, the pretest and demographic covariates model leads to little additional variance explained at either school or individual level compared with the model using only pretest as a covariate. Overall, we find that the intraclass correlation is smaller in this sample than in the full sample, but the explanatory power of pretest and other covariates is also smaller. These two tendencies have opposite effects on statistical power. The smaller intraclass correlation generally leads to larger statistical power, but the smaller explanatory power of covariates generally leads to less statistical power, one partially offsetting the effects of the other.

There is one substantial anomaly in the results reported in Table 6 that is similar to those in Tables 2 and 4: The Grade 2 η^2 values for the pretest and demographic covariates model are larger than those for the pretest covariate model. This anomaly (like that in Tables 2 and 4) appears to be a consequence of differences between the samples used to estimate the two models. We suggest using the values for the pretest and demographic covariates model with

some caution. It might be wise to use the smoothed values for the pretest and demographic covariates model in Grade 6 (which would give $\eta_B^2 = .195$ and $\eta_W^2 = .453$) and possibly in Grade 5 (which would give $\eta_B^2 = .192$ and $\eta_W^2 = .448$).

Reading Achievement in Low-Achievement Schools

Table 7 is a presentation of results in reading computed for the schools in the bottom half of the distribution of school mean reading achievement and is organized in the same way as Tables 2–6. The mean (across grade levels) unconditional intraclass correlation in mathematics was .093, and as in the case of reading, the intraclass correlation values in this sample are considerably smaller than those reported in Table 3 for the entire national population, a tendency that also holds for the conditional (adjusted) intraclass correlations.

There is some variation of intraclass correlations across grade levels. The intraclass correlation in Grade 9 is larger (by over 3 standard errors of the difference) than that in either of the adjacent grades. Similarly, the intraclass correlation in Grade 1 is more than 2 standard errors greater than that in kindergarten but less than 2 standard errors of the difference from that in Grade 2. None of the other differences between grades is this large in comparison with their uncertainty. In general, the intraclass correlations at Grades K–4 range from about .10 to .14 and in Grades 5–8 from about .06 to .07, and in Grades 10–12, they are about .05.

As in the case of mathematics, the use of covariates resulted in a much smaller reduction in both between- and within-school variances in this sample than in the entire national sample. Specifically, the demographic covariates analyses typically reduced the between-school variance to no less than one half of its value in the unconditional model (e.g., produced η_B^2 values from .5 to .8) but typically reduced within-cluster variance by 5% or less (e.g., produced η_W^2 values greater than .95). The analyses using pretest score as a covariate typically (but not always) resulted in modestly larger reductions in between-cluster variance (e.g., produced η_B^2 values from .3 to .8) and typically reduced within-cluster variance by a larger amount than the demographic covariates model (e.g., produced η_W^2 values from .5 to .8). As in the case of mathematics achievement in the full population, the use of both pretest and demographic covariates leads to little additional variance explained at either the school or the individual level compared with the model using only pretest as a covariate. Thus, we find, as in the case of mathematics, that the intraclass correlation is smaller in this sample, but the explanatory power of pretest and other covariates is also smaller, one of these differences partially offsetting the effects of the other on statistical power.

There are several small anomalies in the results reported in Table 7 that are similar to those in Table 6, in which the η_B^2 values for the pretest and demographic covariates model are slightly larger than those for the pretest covariate model. These anomalies (like those in Table 6) appear to be a consequence of instability in variance component estimates in the sample of low-achievement schools.

Comparison With Published Experiments

Although the estimates presented in this article are derived from national probability samples, few experiments actually use national probability samples. Thus, one might question if intraclass correlations obtained from national samples resemble those of experiments actually conducted in education. To obtain some empirical evidence on this question, we searched the two most prestigious education journals that publish experimental studies, the *American Educational Research Journal* and *Educational Evaluation and Policy Analysis*, from 1995 to 2005 to find the cluster-randomized experiments with academic achievement as an outcome variable. We found eight reports of experiments that had randomized schools. We were able to obtain at least one unconditional intraclass correlation estimate from seven of these experiments (which required contacting authors in several cases). The eighth study did not treat schools as a random effect in the analyses and therefore could not provide an intraclass correlation value. This yielded a total of 41 intraclass correlation estimates, 14 in mathematics outcomes and 27 in reading outcomes. They ranged from .07 to .31 in mathematics achievement (with

a mean of .17) and .05 to .74 in reading achievement (with a mean of .19). Eliminating the largest estimate in reading reduced the average value, but only to .17. Some of this variation is surely due to sampling error of estimation. None of the studies provided a standard error for the intraclass correlation estimates, but the form of the standard error is proportional to the square root of the number of schools (see, e.g., Donner & Koval, 1982). Therefore, these standard errors of the experimental estimates must be considerably larger than the largest of those we report on the basis of survey data (i.e., considerably bigger than .03), because the experiments involved considerably fewer schools than our surveys.

The average (unconditional) intraclass correlation in Tables 2 and 3 for the full national sample is about .22, the average value in Tables 4 and 5 for low-SES schools is about .19, and the average value in Tables 6 and 7 for low-achieving schools is about .09. Therefore, the average value of the intraclass correlation estimates from the published experiments is roughly consistent with the national values for low-SES schools but somewhat larger than the national values for low-achieving schools. This is consistent with the fact that most of the published experiments explicitly targeted, or realized, substantial samples of low-SES or disadvantaged students. It would not be appropriate to draw strong conclusions from such a small sample of empirical evidence, but this evidence does not suggest that the intraclass correlations obtained in published experiments are substantially different than those obtained from corresponding national (e.g., low-SES) samples.

Agreement Among Intraclass Correlation Estimates From Different Data Sets

When it was possible to estimate intraclass correlations for the same grade and achievement domain from more than one survey, we computed estimates from all surveys from which it was possible. Table 8 is a presentation of these estimates for the unconditional and demographic covariates models, along with the difference between each pair of intraclass correlation estimates that should estimate the same value and the standard error of the difference. Too few estimates from the other models could be computed for meaningful comparisons. Because the estimated intraclass correlations are approximately normally distributed in large samples, the difference divided by its standard error should have approximately a standard normal distribution if the two estimates are estimating the same population quantity, and thus a difference larger than 2 standard errors for any particular comparison should happen only about 5% of the time by chance.

Although some of the differences are large enough to have practical implications, they are subject to considerable sampling uncertainty. We found that most of the results agreed within sampling error. Overall, 14 of the 18 differences of unadjusted intraclass correlation estimates (across both reading and mathematics) were less than 2 standard errors of the difference. Three of the 13 differences in mathematics exceeded 2 standard errors (ECLS – Prospects1 at Grade 3 and LSAY10 – NELS in Grades 10 and 12). One of the five differences in reading (ECLS – Prospects1 at Grade 3) exceeded 3 standard errors.

However, it is crucial to recognize that the conceptual hypothesis of agreement among data sets that we are testing is that all of the pairs of intraclass correlations are equal. Although the criterion that "differences exceeding 2 standard errors are statistically significant at the 5% level" is (approximately) valid for any single comparison, it is not appropriate for evaluating several comparisons at the same time. To evaluate whether at least one of the comparisons implies a reliable difference, a multiple comparison procedure is needed (see, e.g., R. Miller, 1977). A Bonferroni adjustment for 13 comparisons would require a difference of 2.89 standard errors to be significant at the 5% level, and none of the difference in mathematics is that large. The difference in reading between the estimates from ECLS and Prospects1 at Grade 3 is large enough to be statistically significant, even taking multiple comparisons into account. However, we interpret these comparisons as suggesting that there is a reasonable degree of agreement among the intraclass correlations in these surveys, even though they were conducted as much as a decade apart, by different research organizations, and using different achievement measures.

TABLE 8
Comparisons of Intraclass Correlations (ICCs) Estimated From Different Surveys

Grade	Survey	Unconditional model		Demographic covariates model	
		ICC	*SE*	ICC	*SE*
Mathematics					
1	ECLS[a]	.228	.010	.101	.014
	Prospects1	.193	.018	.147	.016
	ECLS – Prospects1	**.035**	**.021**	**–.046**	**.021[b]**
3	ECLS[a]	.241	.010	.102	.009
	Prospects1	.189	.016	.121	.014
	Prospects3	.222	.019	.155	.037
	ECLS – Prospects1	**.052[b]**	**.019**	**–.019**	**.016**
	ECLS – Prospects3	**.019**	**.021**	**–.053**	**.038**
	Prospects1 – Prospects 2	**–.033**	**.024**	**–.034**	**.040**
7	LSAY7[a]	.191	.033	.088	.019
	Prospects7	.161	.019	.081	.015
	LSAY7 – Prospects7	**.030**	**.038**	**.006**	**.024**
8	LSAY7[a]	.185	.032	.122	.025
	NELS	.246	.010	.072	.005
	Prospects7	.212	.025	.156	.022
	LSAY7 – NELS	**–.061**	**.033**	**.050**	**.025**
	LSAY7 – Prospects7	**–.027**	**.040**	**–.034**	**.033**
	NELS – Prospects7	**.034**	**.027**	**–.084**	**.023[b]**
9	LSAY7[a]	.216	.032	.122	.025
	Prospects7	.253	.026	.189	.025
	LSAY7 – Prospects7	**–.037**	**.041**	**–.067**	**.035**
10	LSAY7	.205	.032	.120	.026
	LSAY10	.162	.030	.076	.018
	NELS[a]	.234	.010	.067	.006
	LSAY7 – LSAY10	**.043**	**.044**	**.045**	**.032**
	LSAY7 – NELS	**–.029**	**.034**	**.053**	**.026**
	LSAY10 – NELS	**–.072[b]**	**.032**	**.008**	**.019**
12	LSAY10	.153	.032	.049	.017
	NELS[a]	.239	.011	.069	.007
	LSAY10 – NELS	**–.086[b]**	**.034**	**–.019**	**.018**
Reading					
1	ECLS[a]	.239	.010	.118	.010
	Prospects1	.207	.019	.143	.019
	ECLS – Prospects1	**.032**	**.021**	**–.025**	**.021**
3	ECLS[a]	.271	.011	.089	.008
	Prospects1	.209	.017	.077	.010
	Prospects3	.243	.019	.078	.031
	ECLS – Prospects1	**.062[b]**	**.020**	**.012**	**.013**
	ECLS – Prospects3	**.028**	**.022**	**.011**	**.032**
	Prospects1 – Prospects3	**–.034**	**.025**	**–.001**	**.033**
8	NELS[a]	.197	.009	.051	.004
	Prospects7	.181	.022	.064	.012
	NELS – Prospects7	**.016**	**.024**	**–.013**	**.013**

Note. The use of bold in the table indicates differences. ECLS = Early Childhood Longitudinal Survey; LSAY = Longitudinal Study of American Youth; NELS = National Educational Longitudinal Study. Because Prospects and LSAY involve more than one cohort followed longitudinally, each cohort of Prospects and LSAY is identified by the grade level of the base year for that cohort. Thus, Prospects1 is the cohort of Prospects that began in Grade 1, LSAY7 is the cohort of LSAY that began in Grade 7, and so on.
a. Data from this survey are included in Tables 2 to 7.
b. This difference exceeds 2 standard errors of the difference.

TABLE 9
Minimum Detectable Effect Sizes With Power 0.80 and n *= 60 as a Function of* m*: All Schools*

Grade	Covariate	Mathematics achievement					Reading achievement				
		m = 10	*m* = 15	*m* = 20	*m* = 25	*m* = 30	*m* = 10	*m* = 15	*m* = 20	*m* = 25	*m* = 30
K	None	0.67	0.54	0.46	0.41	0.38	0.66	0.53	0.46	0.41	0.37
	Pretest	0.27	0.22	0.19	0.17	0.15	0.34	0.28	0.24	0.21	0.19
1	None	0.66	0.53	0.45	0.40	0.37	0.67	0.54	0.46	0.41	0.37
	Pretest	0.29	0.23	0.20	0.18	0.16	0.32	0.25	0.22	0.19	0.18
2	None	0.67	0.53	0.46	0.41	0.37	0.62	0.50	0.43	0.38	0.35
	Pretest	0.39	0.31	0.27	0.24	0.22	0.27	0.22	0.19	0.17	0.15
3	None	0.67	0.54	0.46	0.41	0.38	0.71	0.57	0.49	0.44	0.40
	Pretest	0.31	0.25	0.21	0.19	0.17	0.36	0.29	0.25	0.22	0.20
4	None	0.66	0.53	0.45	0.41	0.37	0.67	0.54	0.46	0.41	0.38
	Pretest	0.38	0.31	0.26	0.24	0.21	0.31	0.25	0.21	0.19	0.17
5	None	0.64	0.51	0.44	0.39	0.36	0.70	0.56	0.48	0.43	0.39
	Pretest	0.39	0.32	0.27	0.24	0.22	0.30	0.24	0.21	0.19	0.17
6	None	0.70	0.56	0.48	0.43	0.39	0.70	0.56	0.48	0.43	0.39
	Pretest	0.37	0.30	0.25	0.23	0.21	0.26	0.21	0.18	0.16	0.15
7	None	0.60	0.48	0.42	0.37	0.34	0.58	0.46	0.40	0.36	0.32
	Pretest	—	—	—	—	—	—	—	—	—	—
8	None	0.60	0.48	0.41	0.37	0.33	0.61	0.49	0.42	0.38	0.34
	Pretest	0.26	0.21	0.18	0.16	0.15	—	—	—	—	—
9	None	0.64	0.51	0.44	0.39	0.36	0.68	0.55	0.47	0.42	0.38
	Pretest	0.22	0.18	0.15	0.14	0.12	0.55	0.44	0.38	0.34	0.31
10	None	0.66	0.53	0.46	0.41	0.37	0.59	0.47	0.41	0.36	0.33
	Pretest	0.21	0.17	0.14	0.13	0.12	0.25	0.20	0.17	0.15	0.14
11	None	0.52	0.42	0.36	0.32	0.29	—	—	—	—	—
	Pretest	0.22	0.18	0.15	0.14	0.13	—	—	—	—	—
12	None	0.67	0.54	0.46	0.41	0.37	0.58	0.46	0.40	0.36	0.32
	Pretest	0.13	0.10	0.09	0.08	0.07	0.21	0.17	0.15	0.13	0.12

Minimum Detectable Effect Sizes

One way to summarize the implications of these results for statistical power is to use them to compute the smallest effect size for which a target design would have adequate statistical power. This effect size is often called the minimum detectable effect size (MDES; see Bloom, 1995, 2005). In computing the MDES values reported in this article, we used the value 0.8 with a two-sided test at a significance level of .05 as the definition of adequate power. We considered designs with no covariates and with pretest as a covariate at both the individual and group levels. We considered both reading and mathematics achievement as potential outcomes. Finally, we considered a balanced design with a sample of size of $n = 60$ per school and $m = 10$, 15, 20, 25, or 30 schools randomized to each treatment group.

Table 9 gives the MDESs on the basis of parameters given in Tables 2 and 3 that were estimated from the full national sample. Perhaps the most obvious finding is that the corresponding MDES values for mathematics and reading are quite similar. With no covariates, the MDES values typically exceed 0.60 for $m = 10$ and typically exceed 0.35 even for $m = 30$. However, the use of pretest as a covariate reduces the MDES values to less than 0.40 for $m = 10$ and 0.20 or less for $m = 30$. Although Cohen (1977) proposed the values 0.20 to define small-sized effects and 0.50 to define medium-sized effects, these labels can be misleading in educational policy contexts, in which effect sizes of 0.20 or smaller are often of policy interest, and consequently, experiments may well be designed to detect effects in this range. Effect sizes used in power analyses should be informed by the magnitude of effects that would be policy relevant and by prior empirical evidence about the likely effect of an intervention being evaluated.

Table 10 gives the MDESs on the basis of parameters given in Tables 4 and 5 that were

TABLE 10
Minimum Detectable Effect Sizes With Power 0.80 and n *= 60 as a Function of* m*: Low–Socioeconomic Status Schools*

Grade	Covariate	Mathematics achievement					Reading achievement				
		$m = 10$	$m = 15$	$m = 20$	$m = 25$	$m = 30$	$m = 10$	$m = 15$	$m = 20$	$m = 25$	$m = 30$
K	None	0.64	0.51	0.44	0.39	0.36	0.64	0.51	0.44	0.39	0.36
	Pretest	0.28	0.23	0.19	0.17	0.16	0.36	0.29	0.25	0.22	0.20
1	None	0.65	0.52	0.45	0.40	0.36	0.65	0.52	0.45	0.40	0.37
	Pretest	0.29	0.23	0.20	0.18	0.16	0.30	0.24	0.21	0.18	0.17
2	None	0.62	0.49	0.42	0.38	0.34	0.59	0.47	0.41	0.36	0.33
	Pretest	0.38	0.30	0.26	0.23	0.21	0.25	0.20	0.17	0.16	0.14
3	None	0.63	0.50	0.43	0.39	0.35	0.65	0.52	0.45	0.40	0.36
	Pretest	0.31	0.24	0.21	0.19	0.17	0.35	0.28	0.24	0.21	0.19
4	None	0.64	0.51	0.44	0.39	0.36	0.64	0.51	0.44	0.39	0.36
	Pretest	0.41	0.33	0.28	0.25	0.23	0.33	0.27	0.23	0.20	0.19
5	None	0.59	0.47	0.41	0.36	0.33	0.66	0.53	0.45	0.40	0.37
	Pretest	0.40	0.32	0.28	0.25	0.23	0.32	0.26	0.22	0.20	0.18
6	None	0.68	0.55	0.47	0.42	0.38	0.65	0.52	0.44	0.40	0.36
	Pretest	0.37	0.29	0.25	0.22	0.20	0.26	0.21	0.18	0.16	0.15
7	None	0.61	0.49	0.42	0.37	0.34	0.58	0.46	0.40	0.35	0.32
	Pretest	—	—	—	—	—	—	—	—	—	—
8	None	0.60	0.48	0.41	0.37	0.33	0.52	0.42	0.36	0.32	0.29
	Pretest	0.27	0.22	0.19	0.17	0.15	—	—	—	—	—
9	None	0.58	0.47	0.40	0.36	0.33	0.67	0.53	0.46	0.41	0.37
	Pretest	0.20	0.16	0.14	0.12	0.11	0.43	0.35	0.30	0.27	0.24
10	None	0.58	0.46	0.40	0.36	0.32	0.51	0.41	0.35	0.31	0.29
	Pretest	0.21	0.17	0.15	0.13	0.12	0.23	0.18	0.16	0.14	0.13
11	None	0.52	0.41	0.36	0.32	0.29	—	—	—	—	—
	Pretest	0.26	0.21	0.18	0.16	0.14	—	—	—	—	—
12	None	0.58	0.46	0.40	0.35	0.32	0.51	0.41	0.35	0.31	0.29
	Pretest	0.13	0.11	0.09	0.08	0.08	0.21	0.17	0.14	0.13	0.12

estimated from the national sample of low-SES schools. These results are remarkably similar to those in Table 9.

Table 11 gives the MDESs on the basis of parameters given in Tables 6 and 7 that were estimated from the national sample of schools in the lower half of the achievement distribution. Because the unconditional intraclass correlations are lower, the MDES values for designs with no covariates are smaller. However, because the covariates are less effective in reducing between- and within-school variance in this sample, the MDES values with pretest as a covariate are not always smaller than in the national sample of all schools. With no covariates, the MDES values typically less than 0.50 for $m = 10$ and less than 0.30 for $m = 30$. However, the use of pretest as a covariate typically reduces the MDES values to about 0.30 for $m = 10$ and 0.20 or less for $m = 30$.

Using the Results of This Study to Compute the Statistical Power of Cluster-Randomized Experiments

Specialized software for computing statistical power in group-randomized designs can use the intraclass correlation values and R_B^2 and R_W^2 values (where $R^2 = 1 - \eta^2$) presented in this article to compute statistical power. Such programs include Optimal Design (Raudenbush & Liu, 2000) and PinT (Snijders & Bosker, 1993). However, such software is not necessary to compute power for studies that randomize schools. In this section, we illustrate the use of the results in this article to compute the statistical power of cluster-randomized experiments. Consider the two-treatment-group design with q ($0 \le q < M - 2$) group-level (cluster-level) covariates and p ($0 \le p < N - q - 2$) individual-level covariates in the analysis. Note that we

TABLE 11
Minimum Detectable Effect Sizes With Power 0.80 and n *= 60 as a Function of* m*: Low-Achievement Schools*

		Mathematics achievement					Reading achievement				
Grade	Covariate	$m = 10$	$m = 15$	$m = 20$	$m = 25$	$m = 30$	$m = 10$	$m = 15$	$m = 20$	$m = 25$	$m = 30$
K	None	0.48	0.38	0.33	0.29	0.27	0.46	0.37	0.32	0.28	0.26
	Pretest	0.31	0.25	0.21	0.19	0.17	0.41	0.33	0.28	0.25	0.23
1	None	0.43	0.35	0.30	0.27	0.24	0.53	0.42	0.36	0.32	0.30
	Pretest	0.31	0.25	0.22	0.19	0.18	0.41	0.33	0.28	0.25	0.23
2	None	0.47	0.38	0.33	0.29	0.27	0.47	0.38	0.32	0.29	0.26
	Pretest	0.34	0.27	0.23	0.21	0.19	0.28	0.22	0.19	0.17	0.16
3	None	0.46	0.37	0.32	0.28	0.26	0.52	0.42	0.36	0.32	0.29
	Pretest	0.30	0.24	0.21	0.19	0.17	0.34	0.27	0.23	0.21	0.19
4	None	0.52	0.41	0.36	0.32	0.29	0.46	0.37	0.32	0.28	0.26
	Pretest	0.44	0.35	0.30	0.27	0.25	0.35	0.28	0.24	0.22	0.20
5	None	0.37	0.29	0.25	0.23	0.21	0.39	0.32	0.27	0.24	0.22
	Pretest	0.33	0.27	0.23	0.21	0.19	0.35	0.28	0.24	0.21	0.19
6	None	0.42	0.34	0.29	0.26	0.23	0.36	0.29	0.25	0.22	0.20
	Pretest	0.41	0.33	0.28	0.25	0.23	0.32	0.25	0.22	0.19	0.18
7	None	0.33	0.27	0.23	0.20	0.19	0.38	0.3	0.26	0.23	0.21
	Pretest	—	—	—	—	—	—	—	—	—	—
8	None	0.42	0.34	0.29	0.26	0.24	0.39	0.31	0.27	0.24	0.22
	Pretest	0.32	0.26	0.22	0.20	0.18	—	—	—	—	—
9	None	0.42	0.33	0.29	0.26	0.23	0.55	0.44	0.38	0.34	0.31
	Pretest	0.28	0.23	0.19	0.17	0.16	0.55	0.44	0.38	0.33	0.30
10	None	0.41	0.33	0.28	0.25	0.23	0.34	0.28	0.24	0.21	0.19
	Pretest	0.33	0.26	0.23	0.20	0.18	0.33	0.26	0.22	0.20	0.18
11	None	0.42	0.33	0.29	0.26	0.23	—	—	—	—	—
	Pretest	0.30	0.24	0.21	0.19	0.17	—	—	—	—	—
12	None	0.41	0.33	0.29	0.25	0.23	0.34	0.27	0.23	0.21	0.19
	Pretest	0.22	0.17	0.15	0.13	0.12	0.28	0.22	0.19	0.17	0.16

specifically include the possibility that there are zero (no) covariates at a given level. For example, a design with $p = 1$ and $q = 1$ might arise, for example, if there was a pretest that was used as an individual-level covariate and cluster means on the covariate were used as a group-level covariate. We assume also that the individual-level covariate has been centered about cluster means. The structural model for Y_{ijk}, the kth observation in the jth cluster in the ith treatment might be described in analysis of covariance (ANCOVA) notation as

$$Y_{ijk} = \mu + \alpha_{A_i} + \boldsymbol{\theta}'_{1}\mathbf{x}_{ijk} + \boldsymbol{\theta}'_{G}\mathbf{z}_{ij} + \gamma_{A(i)j} + \varepsilon_{Aijk},$$

where μ is the grand mean, α_{A_i} is the covariate-adjusted effect of the ith treatment, $\boldsymbol{\theta}_1 = (\theta_{11}, \ldots, \theta_{1p})'$ is a vector of p individual-level covariate effects, $\boldsymbol{\theta}_G = (\theta_{G1}, \ldots, \theta_{Gq})'$ is a vector of q group-level covariate effects, $\mathbf{x}_{ijk}$ is a vector of p group (cluster) centered individual-level covariate values for the jth cluster in the ith treatment, $\mathbf{z}_{ij}$ is a vector of q group-level (cluster-level) covariate values for the jth cluster in the ith treatment, $\gamma_{(i)j}$ is the random effect of cluster j within treatment i, and ε_{Aijk} is the covariate-adjusted within-cell residual. Here, we assume that both of the random effects (clusters and the residual) are normally distributed.

The analysis might be carried out either as an ANCOVA with clusters as a nested factor or by viewing the model as a hierarchical linear model and using software for multilevel models such as HLM. In multilevel model notation, it would be conventional to specify a Level 1 (individual-level) model as

$$Y_{ijk} = \beta_{0j} + \boldsymbol{\beta}'_{j}\mathbf{x}_{ijk} + \varepsilon_{Aijk}$$

and a Level 2 (cluster-level) model for the intercept as

$$\beta_{0j} = \pi_{00} + \pi_{A01}\text{TREATMENT}_i + \boldsymbol{\pi}'_{02}\mathbf{z}_{ij} + \zeta_{Aj},$$

where TREATMENT_i is a dummy variable for the treatment group, while the covariate slopes in $\boldsymbol{\beta}_j$ would be treated as fixed effects ($\boldsymbol{\beta}_j = \boldsymbol{\theta}_1$), and ζ_{Aj} is the random effect of the jth cluster (a Level 2 residual). With the appropriate constraints on the ANCOVA model (i.e., setting $\alpha_{Ai} = 0$ for the control group and constraining the mean of the $\gamma_{A(i)j}$ values to be 0), these two models are identical, and there is a one-to-one correspondence between the parameters and the random effects in the two models. That is, $\mu = \pi_{00}$, $\alpha_{Ai} = \pi_{A01}$, $\boldsymbol{\theta}_G = \boldsymbol{\pi}_{02}$, $\boldsymbol{\theta}_1 = \boldsymbol{\beta}_j$ (for all j), $\gamma_{A(i)j} = \zeta_{Aj}$ (with a suitable redefinition of the index j), and ε_{Aijk} is identical in both models. The variance components associated with this analysis are σ_{AW}^2 (the variance of ε_{Aijk}) and σ_{AB}^2 (the variance of ζ_j), where the A in the subscript denotes that these variance components are adjusted for the covariate.

The Intraclass Correlations

Note that if in the experiment, schools were sampled at random, students were sampled at random within schools, and $q = p = 0$, then $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ is exactly the intraclass correlation that would obtain in a survey that sampled first schools and then students at random. Similarly, if there are covariates in the experiment, schools were sampled at random, students were sampled at random within schools, and $q \neq 0$ or $p \neq 0$, then $\rho_A = \sigma_{AB}^2/(\sigma_{AB}^2 + \sigma_{AW}^2)$ is exactly the adjusted intraclass correlation that would obtain in the analysis of the survey (with appropriate covariates) that sampled first schools and then students at random.

Hypothesis Testing

The object of the statistical analysis is to test the statistical significance of the intervention effect, that is, to test the following hypothesis:

Hypothesis H_0: $\alpha_{A1} - \alpha_{A2} = 0$.

Or, equivalently,

Hypothesis H_0: $\pi_{A01} = 0$.

The ANCOVA t-test statistic is

$$t_A = \frac{\sqrt{\tilde{m}}(\bar{Y}_{A1\bullet\bullet} - \bar{Y}_{A2\bullet\bullet})}{S_A}, \tag{5}$$

where $\tilde{m}$ is defined in terms of the number of clusters assigned to the treatment and control groups (m_1 and m_2, respectively) as

$$\tilde{m} = m_1 m_2/(m_1 + m_2),$$

$\bar{Y}_{A1\bullet\bullet}$ and $\bar{Y}_{A2\bullet\bullet}$ are the adjusted means, S_A is the pooled within-treatment-groups adjusted standard deviation of cluster means, and the subscript A is used to denote that the means and standard deviation are adjusted for the covariates. The F-test statistic from a one-way ANCOVA using cluster means is of course

$$F_A = \frac{MS_{AB}}{MS_{AC}} = t_A^2. \tag{6}$$

In this case, $MS_{AB} = n\tilde{m}(\bar{Y}_{A1\bullet\bullet} - \bar{Y}_{A2\bullet\bullet})^2$ and $MS_{AC} = nS_A^2$, where S_A is the pooled within-treatment-groups standard deviation of the covariate-adjusted cluster means (the standard deviation of the Level 2 residuals). If the null hypothesis is true, the test statistic t_A has Student's t distribution with $M - q - 2$ degrees of freedom. Equivalently, the test statistic F_A has the central F distribution with 1 degree of freedom in the numerator and $M - q - 2$ degrees of freedom in the denominator when the null hypothesis is true.

When the null hypothesis is false, the test statistic t_A has for this analysis a noncentral t distribution with $M - q - 2$ degrees of freedom and noncentrality parameter

$$\begin{aligned}\lambda_A &= \frac{\sqrt{\tilde{m}n}(\alpha_{A1} - \alpha_{A2})}{\sigma_{AT}}\sqrt{\frac{1}{1 + (n-1)\rho_A}} \\ &= \frac{\sqrt{\tilde{m}n}\delta_A}{\sqrt{[1 + (n-1)\rho_A]}},\end{aligned} \tag{7}$$

where $\delta_A = (\alpha_{A1} - \alpha_{A2})/\sigma_{AT}$.

Alternatively (and equivalently), the F statistic has the noncentral F distribution with 1 degree of freedom in the numerator and $M - q - 2$ degrees of freedom in the denominator and noncentrality parameter

$$\omega_A = \frac{\tilde{m}n(\alpha_{A1} - \alpha_{A2})^2}{\sigma_{AT}^2[1 + (n-1)\rho_A]}.$$

For the purposes of power computation, expression 7 is not convenient, because the minimum effect size of interest is likely to be known in units of the unadjusted standard deviation rather than the adjusted standard deviation; that is, we are more likely to know $\delta = (\alpha_1 - \alpha_2)/\sigma_T$ rather than $\delta_A = (\alpha_{A1} - \alpha_{A2})/\sigma_{AT}$. In a randomized experiment, covariate adjustment should not affect the treatment effect parameter, so that $\alpha_{A1} - \alpha_{A2} = \alpha_1 - \alpha_2$, but the covariate adjustment necessarily affects the standard deviation. This is true even if the covariates operate at only one level of the design. Because $\sigma_{AT}^2 = \sigma_{AB}^2 + \sigma_{AW}^2$, a covariate adjustment at the individual level will affect σ_{AT}^2 via σ_{AW}^2, and a covariate adjustment at the cluster level will affect σ_{AT}^2 through σ_{AB}^2.

To express λ_A in terms of δ, we need only express σ_{AT} in terms of σ_T. A direct derivation shows that

$$\lambda_A = \frac{\sqrt{\tilde{m}n}(\alpha_{A1} - \alpha_{A2})}{\sigma_T}\left(\frac{\sigma_T}{\sigma_{AT}}\right)\sqrt{\frac{1}{1 + (n-1)\rho_A}}$$

$$= \sqrt{\tilde{m}n}\delta\sqrt{\frac{\eta_B^2 + (\eta_W^2 - \eta_B^2)\rho_A}{\eta_B^2\eta_W^2[1 + (n-1)\rho_A]}}. \quad (8)$$

An alternative, but equivalent, expression of λ_A that is considerably more revealing involves η_B^2, η_W^2, and the unadjusted intraclass correlation ρ. This expression is

$$\lambda_A = \delta\sqrt{\tilde{m}n}\sqrt{\frac{1}{\eta_W^2 + (n\eta_B^2 - \eta_W^2)\rho}}. \quad (9)$$

Note that the quantity $[\eta_W^2 + (n\eta_B^2 - \eta_W^2)\rho]$ is analogous to $[1 + (n-1)\rho]$, Kish's (1965) design effect. We see that $[\eta_W^2 + (n\eta_B^2 - \eta_W^2)\rho]$ reduces to $[1 + (n-1)\rho]$ in the analysis without covariates (because $\eta_W^2 = \eta_B^2 = 1$), and Equation 9 reduces to the expression given (e.g., in Blair & Higgins, 1986) for the *t* test conducted using cluster means as the unit of analysis.

We illustrate the use of the *t* statistic. The power of the one-tailed test at level α is

$$p_1 = 1 - H[c(\alpha, M - q - 2), (M - q - 2), \lambda_A], \quad (10)$$

where $c(\alpha, \nu)$ is the level α one-tailed critical value of the *t* distribution with ν degrees of freedom (e.g., $c[.05, 10] = 1.81$), and $H(x, \nu, \lambda)$ is the cumulative distribution function of the noncentral *t* distribution with ν degrees of freedom and noncentrality parameter λ. The power of the two-tailed test at level α is

$$p_2 = 1 - H[c(\alpha/2, M - q - 2), (M - q - 2), \lambda_A] + H[-c(\alpha/2, M - q - 2), (M - q - 2), \lambda_A]. \quad (11)$$

Using Power Tables and Power Calculation Software

Many tabulations (e.g., Cohen, 1977) and programs (e.g., Borenstein, Rothstein, & Cohen, 2001) are available for computing statistical power from designs involving simple random samples, but tables for computing power from the independent-groups *t* test are the most widely available. Following Cohen's (1977) framework, such tables typically provide power values on the basis of sample sizes N_1^T and N_2^T (often assumed to be equal for simplicity) and effect size Δ^T, where the superscript T indicates that these quantities are what is used in the power tables. The calculations on which they are based translate the sample sizes and effect size into degrees of freedom ν^T and noncentrality parameter λ^T to compute statistical power. In the case of the two-sample *t* test, they do so via

$$\nu^T = N_1^T + N_2^T - 2$$

and

$$\lambda^T = \sqrt{\tilde{N}^T}\Delta^T,$$

where

$$\tilde{N}^T = \frac{N_1^T N_2^T}{N_1^T + N_2^T}.$$

Tables such as Cohen's (or the corresponding software) can be used to compute the power of the test used in the case of clustered sampling by judicious choice of sample sizes and effect size. We have to enter the table with a configuration of sample sizes and a synthetic effect size (here called the *operational effect size*) that will

yield the appropriate degrees of freedom and noncentrality parameter.

If the actual numbers of clusters assigned are m_1 and m_2, then entering the power table with sample sizes $N_1^T = m_1 - q$ and $N_2^T = m_2$ yields $\nu^T = (m_1^T + m_2^T - 2) = M - q - 2$, the correct degrees of freedom for the test. Of course, many other combinations of sample sizes will also yield the correct degrees of freedom as well and will yield equivalent results as long as the operational effect size is modified in a corresponding manner. The relevant operational effect size using our choice of degrees of freedom is

$$\Delta^T = \delta\sqrt{\frac{\tilde{m}n}{\tilde{N}^T}}\sqrt{\frac{\eta_B^2 + (\eta_W^2 - \eta_B^2)\rho_A}{\eta_B^2\eta_W^2[1 + (n-1)\rho_A]}}$$
$$= \delta\sqrt{\frac{\tilde{m}n}{\tilde{N}^T}}\sqrt{\frac{1}{\eta_W^2 + (n\eta_B^2 - \eta_W^2)\rho}}, \quad (12)$$

where δ is the unadjusted effect size, ρ is the unadjusted intraclass correlation, and η_B^2 and η_W^2 are defined in Equations 5 and 6. If the analysis makes a covariate adjustment at the cluster level, η_B^2 is the appropriate value given in the tables of this article, but if the analysis makes no covariate adjustment at the cluster level (i.e., $q = 0$), then $\eta_B^2 \equiv 1$. Similarly, if the analysis makes a covariate adjustment at the individual (within-cluster) level, η_W^2 is the appropriate value given in the tables of this article, but if the analysis makes no covariate adjustment at the individual level (that is if $p = 0$), then $\eta_W^2 \equiv 1$. Note that the value of Δ^T given in Equation 12 is appropriate, because when this is multiplied by $\sqrt{\tilde{N}^T}$, it yields the noncentrality parameter λ_A given in Equation 9. Using ρ or ρ_A, the cluster sample size n, and the variance ratios η_B^2 and η_W^2 to compute operational effect size makes it possible to compute statistical power and sample size requirements for analyses on the basis of clustered samples using these tables and computer programs designed for the two-group t test.

Example With No Covariates at Either Level

Consider an experiment that will randomize $m_1 = m_2 = 10$ schools to receive an intervention to improve mathematics achievement so that $n = 20$ students in each school would be part of the experiment. There are no covariates at either individual or group level, so that $p = q = 0$ and $\eta_W^2 = \eta_B^2 = 1$. The analysis will involve a two-tailed t test with significance level $\alpha = .05$. Suppose that the smallest educationally significant effect size for this intervention is assumed to be $\delta = 0.50$. Suppose further that the schools were chosen to attempt to be represent first graders nationally.

Entering Table 2 on the first row for Grade 1 and the panel for the unconditional model (columns 2–3) gives the intraclass correlation for first graders as $\rho = .228$. Then the variance inflation factor is

$$1 + (20 - 1)(.228) = 5.332,$$

so that the noncentrality parameter from Equation 7 is

$$\lambda = \frac{0.50\sqrt{(10/2)20}}{\sqrt{5.332}} = 2.165.$$

Using Equation 11 and the noncentral t-distribution function (e.g., the function NCDF.T in SPSS), with $M - 2 = 18$ degrees of freedom, $c(.05/2, 18) = 2.101$, and $\lambda = 2.165$, we obtain a two-sided power of $p_2 = 1 - 0.467 + 0.000 = 0.53$.

Alternatively, we could compute the power from tables of the power of the t test such as those given by Cohen (1977). To do so, we first compute the operational effect size given in Equation 12 as

$$\Delta^T = \frac{0.50\sqrt{20}}{\sqrt{5.332}} = 0.968.$$

Cohen's tables give the statistical power in terms of sample size (in each treatment group) and effect size. Examining Cohen's Table 2.3.5, we see that the operational effect size of 0.968 is between tabled effect sizes of 0.8 and 1.0. Entering the table with sample size $N_1^T = N_2^T = 10$, we see that a power of 0.39 is tabulated for the effect size of $\Delta^T = 0.80$, and a power of 0.56 is tabulated for an effect size of $\Delta^T = 1.00$. Interpolating between these two values, we obtain a power of 0.53 for $\Delta^T = 0.97$.

Note that in this case (and many others), the operational effect size for the tests based on clustered samples is larger than the actual effect size (in this case 0.97 vs. 0.50). This does not mean that the power of the test for the design based on the clustered sample is larger than that based on a simple random sample with the same total sample size. The reason is that the test using the clustered sample has many fewer degrees of freedom in the error term. For example, a test based on an effect size of $\Delta^T = 0.50$ and a simple random sample of $nm = (10)(20) = 200$ in each group would have power essentially 1.0.

Example With Pretest as a Covariate at Both Individual and Cluster Levels

Consider an experiment that will randomize $m_1 = m_2 = 10$ schools to receive an intervention to improve first grade reading achievement and that $n = 20$ students in each school would be part of the experiment. An ANCOVA will be used with pretest as a covariate at both individual and school level (so that $p = q = 1$) using a two-tailed test with significance level $\alpha = .05$. Suppose that the smallest educationally significant effect size for this intervention is $\delta = 0.25$. Suppose further that the schools were chosen in an attempt to be representative of first graders nationally.

Entering Table 3 on the first row for Grade 1 and the panel for the unconditional model (columns 3–5) gives the intraclass correlation for first graders as $\rho = .239$. Entering Table 2 on the second row for Grade 1 and the panel for the pretest and demographic covariates model (columns 9–11) gives the between- and within-school variance ratios after covariate adjustment as $\eta_B^2 = .210$ and $\eta_W^2 = .360$. Then the variance inflation factor is

$$0.360 + [(20)(.210) - .360](.239) = 1.2778,$$

so that the noncentrality parameter from Equation 9 is

$$\lambda_A = \frac{0.25\sqrt{(10/2)20}}{\sqrt{1.278}} = 2.211.$$

Using Equation 11 and the noncentral *t*-distribution function (e.g., the function NCDF.T in SPSS), with $M - 2 - 1 = 17$ degrees of freedom, $c(.05/2, 17) = 2.110$, and $\lambda_A = 2.211$, we obtain a two-sided power of $p_2 = 1 - 0.450 + 0.000 = 0.55$.

Alternatively, we could compute the power from tables of the power of the *t* test such as those given by Cohen (1977). Because there is $q = 1$ covariate at the school level, $N_1^T = m_1 - 1 = 10 - 1 = 9$ and $N_2^T = m_2 = 10$. Because Cohen's tables give the statistical power in terms of equal sample sizes (in each treatment group), we will need to interpolate between sample sizes $N_1^T = N_2^T = 9$ and $N_1^T = N_2^T = 10$. Here we compute $\tilde{m} = (10 \times 10)/(10 + 10) = 5$. For $N_1^T = N_2^T = 9$, $\tilde{N}^T = (9 \times 10)/(9 + 10) = 4.737$, and the operational effect size is

$$\Delta^T = 0.25\sqrt{\frac{(4.5)(20)}{4.737}}\sqrt{\frac{1}{1.2778}} = 1.016.$$

Examining Cohen's Table 2.3.5, we see that the effect size $\Delta^T = 1.02$ is between tabled values of effect sizes of 1.0 and 1.2. Entering the table with sample size $N_1^T = N_2^T = 9$, we see that a power of 0.51 is tabulated for the effect size of $\Delta^T = 1.0$, and a power of 0.65 is tabulated for an effect size of $\Delta^T = 1.2$. Interpolating between the two power values (0.51 and 0.65) for $N_1^T = N_2^T = 9$, we obtain a power of 0.524 for $\Delta^T = 1.02$. This value (0.524) corresponds to the power associated with the effect size of $\delta = 0.25$ and a test based on 16 degrees of freedom.

Entering the table with sample size $N_1^T = N_2^T = 10$, we see that a power of 0.56 is tabulated for the effect size of $\Delta^T = 1.00$, and a power of 0.71 is tabulated for an effect size of $\Delta^T = 1.20$. Interpolating between the two power values (0.56 and 0.71) for $N_1^T = N_2^T = 10$, we obtain a power of 0.575 for $\Delta^T = 1.02$. This value (0.575) corresponds to the power associated with the effect size of $\delta = 0.25$ and a test based on 18 degrees of freedom.

To obtain the power associated with an effect size of $\delta = 0.25$ and a test based on 17 degrees of freedom, we must interpolate once again between these two values (0.524 and 0.575), and we obtain a power value for $N_1^T = 9$ and $N_2^T = 10$ of $p_2 = 0.55$.

It is worth noting that if no covariates had been used at either level of this analysis (i.e., if $p = q = 0$ and therefore $\eta_B^2 = \eta_W^2 = 1$), the power would have been 0.17. If the pretest as a covariate had been used only at the individual level (i.e., if $p = 1$, $q = 0$, $\eta_B^2 = 1$, but $\eta_W^2 = .360$), the power would

have increased to 0.18. But if the pretest had been used as a covariate only at the school level (i.e., if $p = 0$, $q = 1$, $\eta_W^2 = 1$, but $\eta_B^2 = .210$), the power would have increased to 0.43. This illustrates the fact that covariates at the (group) cluster level can have far more impact on the power than covariates at the individual level.

Conclusions

The values of intraclass correlations and variance components presented in this article provide some guidance for the selection of intraclass correlations for planning cluster-randomized experiments. These values suggest that for experiments that have samples as diverse as the nation as a whole and for those using low-SES schools, somewhat larger values of the intraclass correlation (roughly .15–.25) may be appropriate than the .05–.15 guidelines that have sometimes been used. The guideline of .05–.15 is more consistent with the values of unadjusted intraclass correlations among low-achieving schools and those of covariate-adjusted intraclass correlations we found.

In using these values, it is important to keep in mind that these analyses do not separately estimate the between-district and between-state components of variance. Therefore, these two components of variance are included here as part of the between-school variance. This is desirable if the values are to be used in connection with designs that involve schools from several districts or states. However, if the design involves schools from only a single district or state, the estimates reported here may overestimate the relevant intraclass correlations to some degree. Unfortunately, it is unclear just how much of an impact this may have. We suspect that these influences are not large, because a general rule of thumb in both sample surveys and cluster-randomized experiments is that variance components (and therefore contributions to intraclass correlations) of larger units tend to be smaller in magnitude, even though their impact on design effects may be large (because effects on variance inflation factors are proportional to the unit sample size multiplied by the intraclass correlation). Our attempts to explore this question by calculating intraclass correlations with the inclusion of state dummy variables in some of the surveys yielded only negligible effects. Note that the inclusion of multiple districts and states in national samples is also likely to have some impact on the effectiveness of the covariates in explaining between- and within-school variation. It is likely that the somewhat greater between-school variation in national samples leads to a larger intraclass correlation but also to larger covariate effects, so that these impacts partially cancel one another in their effects on statistical power.

A more detailed compilation is available from the authors providing values for regions of the country, settings with different levels of urbanicity, and regions crossed with levels of urbanicity. However it is important to recognize that there is a trade-off between bias (estimating exactly the right value of the intraclass correlation in a particular context) and variance (the sampling uncertainty of that estimate). The variance of the intraclass correlation estimate is driven primarily by the number of clusters (in this case, schools). Although the intraclass correlations we computed in a particular region and setting are more specific and therefore likely to have less bias as estimates of the intraclass correlation in an experiment that is to be conducted within a particular region and context, the sample size used to estimate the intraclass correlations is smaller, and thus the estimate is subject to greater sampling uncertainties. Our analyses suggest that although there is often statistically significant variation in intraclass correlations between regions and settings, the magnitude of this variation is typically small. Thus, it is not completely clear whether more specific estimates are always better (i.e., more accurate) for planning purposes.

It is important to note that the power computations illustrated in this article apply to two-level experiments in which students are nested within schools. If the sampling design used is actually a three-level design (e.g., if students are sampled by classrooms within schools) then the power computations given here (or given by specialized software for computing power in two-level designs) would not be correct. Consider a sample (e.g., for a treatment group) obtained by selecting m schools, then p classrooms within each school, and then n students within each classroom. This is not a simple random sample of mpn individuals, nor is it a (two-stage) clustered sample obtained by randomly

selecting *pn* students within each cluster (school). Instead, it is a three-stage cluster sample of *m* clusters (schools) and *p* subclusters (classrooms), with *n* students randomly selected within each subcluster (classroom). The sampling distribution of statistics based on such three-stage clustered samples is not the same as those based on two-stage clustered samples of the same size. For example, suppose that the (total) variance of a population with clustered structure (such as a population of students within classrooms within schools) is σ_T^2, and that this total variance is decomposable into a between-school variance σ_S^2, a between-classroom variance σ_C^2, and a within-classroom variance σ_W^2, so that $\sigma_T^2 = \sigma_S^2 + \sigma_C^2 + \sigma_W^2$. Then the variance of the mean of a simple random sample of size *mpn* from this population would be σ_T^2/mpn, and the variance of the mean of a two-stage cluster sample of *m* clusters, each of size *pn* from that population (with the same sample size *pn* per school and the same total sample size *mpn*) would be $[1 + (pn - 1)\rho_S]\sigma_T^2/mpn$, where $\rho_S = \sigma_S^2/\sigma_T^2$ is the cluster-level (school-level) intraclass correlation. The variance of the mean computed from a three-stage clustered sample of *m* schools, *p* classrooms within each school, and *n* students within each classroom would be $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]\sigma_T^2/mpn$, where $\rho_C = \sigma_C^2/\sigma_T^2$ is the subcluster-level (classroom-level) intraclass correlation. Note that the design effect in the three-stage cluster sample $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]$ is larger than that in the two-stage cluster sample of the same size $[1 + (pn - 1)\rho_S]$, which implies that the estimated treatment effect (which is just a difference between means) estimated from the three-stage cluster sample, is less precise.

This difference in precision of treatment effect estimates leads to a difference in the noncentrality parameters that determine statistical power. In a two-level experiment, the treatment effects are estimated from two-stage cluster samples, leading to the noncentrality parameter (with no covariates) of

$$\lambda = \frac{\sqrt{\tilde{m}pn}(\alpha_1 - \alpha_2)}{\sigma_T}\sqrt{\frac{1}{1 + (pn - 1)\rho_S}} = \delta\sqrt{\frac{\tilde{m}pn}{2}}\sqrt{\frac{1}{1 + (pn - 1)\rho_S}}, \quad (13)$$

where δ is the effect size (mean difference standardized by σ_T). In a three-level experiment, the treatment effects are estimated from three-stage cluster samples, leading to the noncentrality parameter (with no covariates) of

$$\lambda = \frac{\sqrt{\tilde{m}pn}(\alpha_1 - \alpha_2)}{\sigma_T}\sqrt{\frac{1}{1 + (pn - 1)\rho_S + (n - 1)\rho_C}} = \delta\sqrt{\frac{\tilde{m}pn}{2}}\sqrt{\frac{1}{1 + (pn - 1)\rho_S + (n - 1)\rho_C}}, \quad (14)$$

which is generally smaller than that computed from Equation 13. Therefore, the statistical power of three-level experiments that assign schools to treatments is generally smaller than that of the analogous experiments with two-level designs having the same number of schools and students (see Konstantopoulos, 2006). Note, however, that the issue here is not in which analysis is used (two- vs. three-level) but which sampling design is used (one vs. two stages of clustering within a two- vs. three-stage sampling design).

Although we anticipate that the principal use of the results given in this article will be for planning randomized experiments in education that assign schools (rather than individuals) to treatments, there are other potential applications. One involves the use of information external to an experiment to adjust the degrees of freedom of significance tests in designs involving group randomization, called the df* method by its originators (see Murray, Hannan, & Baker, 1996). Although the originators of this method caution that it is important that users should have good reasons to assume that any external estimates used should estimate the same intraclass correlation as that in the experiment, there may be situations in which data from this compilation meet that assumption. Because they are based on relatively large samples, the intraclass correlation estimates reported in this article tend to have small standard errors. Consequently, if they are thought to be appropriate for use in a particular df* computation, they should substantially increase the degrees of freedom used in the test for treatment effects.

A second potential application is to evaluate whether the conclusions of statistical analyses

that incorrectly ignored clustering might have changed if those significance tests had taken clustering into account. Hedges (in press-a) has shown how to compute the actual significance level of the usual *t* statistic when it has been computed from clustered samples (by incorrectly ignoring clustering). The computation of this actual significance level depends on ρ. The values in this compilation provide some guidelines on values of ρ that might be used for sensitivity analyses to see if a conclusion about the statistical significance of a treatment effect might not have held if clustering had been taken into account.

A third potential application involves the computation of standardized effect size estimates and their standard errors in group-randomized trials. There are several approaches to the computation of effect size estimates in multilevel designs, but in some cases, the computation of estimates and the computation of standard errors requires knowledge of ρ (see Hedges, in press-b). In cases in which the report of the experiment itself does not include information that can be used to compute an estimate of ρ, this compilation may provide some idea of a range of plausible values to incorporate into sensitivity analyses used in connection with effect sizes from experiments that assign schools to treatment.

References

Blair, R. C., & Higgins, J. J. (1986). Comment on "Statistical power with group mean as the unit of analysis." *Journal of Educational Statistics*, *11*, 161–169.

Bloom, H. S. (1995). Minimum detectable effects: A simple way to report statistical power of experimental designs. *Evaluation Review*, *19*, 547–556.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches*. New York: Russell Sage.

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

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis, 29*(1), 30–59.

Borenstein, M., Rothstein, H., & Cohen, J. (2001). *Power and precision*. Teaneck, NJ: Biostat.

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

Curtin, T. R., Ingels, S. J., Wu, S., & Heuer, R. (2002). *User's manual: NELS:88 base-year to fourth followup*. Washington, DC: National Center for Education Statistics.

Donner, A., Birkett, N., & Buck, C. (1981). Randomization by cluster. *American Journal of Epidemiology*, *114*, 906–914.

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

Donner, A., & Koval, J. J. (1982). Design considerations in the estimation of intraclass correlation. *Annals of Human Genetics*, *46*, 271–277.

Gulliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*, 876–883.

Hedges, L. V. (in press-a). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*.

Hedges, L. V. (in press-b). Effect sizes in cluster randomized designs. *Journal of Educational and Behavioral Statistics*.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal*, *19*, 5–18.

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

Klar, N., & Donner, A. (2001). Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine*, *20*, 3729–3740.

Konstantopoulos, S. (2006). *Statistical power in three-level designs* (Working paper). Evanston, IL: Northwestern University, Institute for Policy Research.

Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

Lipsey, M. W. (1990). *Design sensitivity: Statistical power analysis for experimental research*. Newbury Park, CA: Sage.

Miller, J. D., Hoffer, T., Suchner, R. W., Brown, K. G., & Nelson, C. (1992). *LSAY codebook*. DeKalb: Northern Illinois University.

Miller, R. (1977). *Simultaneous statistical inference*. New York: Springer-Verlag.

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

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

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, *27*, 79–103.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials. *Evaluation Review*, *20*, 313–337.

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

Puma, M. J., Karweit, N., Price, C., Riccuti, A., & Vaden-Kiernan, M. (1997). *Prospects: Final report on student outcomes, Vol. II: Technical report*. Cambridge, MA: Abt Associates.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage.

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

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

Snijders, T., & Bosker, J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237–259.

Tourangeau, K., Brick, M., Le, T., Nord, C., West, J., & Hausken, E. G. (2005). *Early childhood longitudinal study, kindergarten class of 1998–99*. Washington, DC: National Center for Education Statistics.

Verma, V., & Lee, T. (1996). An analysis of sampling errors for demographic and health surveys. *International Statistical Review*, *64*, 265–294.

Authors

LARRY V. HEDGES is currently Board of Trustees Professor of Statistics, Professor of Education and Social Policy, and faculty fellow at the Institute for Policy Research at Northwestern University, 2040 North Sheridan Road, Evanston, IL 60610; l-hedges@northwestern.edu. His interests include methods for educational and social policy research.

E. C. HEDBERG is currently an advanced graduate student in the Department of Sociology at the University of Chicago, NORC Research Centers, 1155 East 60th Street, Chicago, IL 60637; ech@uchicago .edu. He is part of many projects that span a wide variety of interests that include the sociology of family and the life course, education and methods. His dissertation research focuses on using context-effect models and dyadic analysis to understand familial social exchange between kin.

Manuscript received March 16, 2006
Final revision received December 13, 2006
Accepted January 2, 2007