Intraclass Correlations for Planning Group Randomized Experiments in Rural Education

Larry V. Hedges
Northwestern University

E. C. Hedberg
University of Chicago

Experiments that assign intact groups (usually schools) to treatment conditions are increasingly common in educational research. The design of group randomized experiments requires knowledge of the intraclass correlation structure to compute statistical power and to determine the sample sizes required to achieve adequate power. The intraclass correlation structure of academic achievement is shown to be somewhat different in rural schools than in all schools in the nation. 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 rural schools. The use of these values to compute statistical power of group randomized experiments involving rural schools is illustrated.

Randomized experiments are being used with increasing frequency in educational research, in part because of the emphasis on experimental evaluations at the U.S. Institute of Education Sciences (IES). In many situations of practical interest in education, it is difficult or impossible to assign individuals to receive different intervention conditions. In such cases, experiments often assign entire intact groups (such as sites, classrooms, or schools) to the same treatment, with different intact groups being assigned to different treatments. However, individuals in these intact groups are often more alike than individuals in different groups, irrespective of treatment. Thus the intact groups correspond to what statisticians call clusters in sampling theory and designs that assign such groups to treatment conditions are often called group randomized or *cluster* randomized designs. Cluster randomized experiments have been used in education for some time (see, e.g., Coladarci & Gage, 1984; Good & Grouws, 1979) and have also been used extensively in public health and other areas of prevention science (see, e.g., Donner & Klar, 2000; Murray, 1998). Methods for the design and analysis of group randomized trials have been discussed extensively in Donner and Klar (2000) and Murray (1998).

Cluster randomized experiments sample subjects via cluster samples, instead of simple random samples. This has an impact on the sampling distribution of statistics computed in the experiment, and the analysis of the experiments needs to take this sampling design into account. For example, a sample obtained from m clusters (such as classrooms or schools) of size n is not a simple random sample of mn individuals, even if it is based on a simple random sample of both *clusters* and *individuals within clusters*. If the (total) variance of a population (such as a population of students clustered within schools) is σ_T^2, and this total variance is decomposable into a between cluster variance σ_B^2 and a within cluster variance σ_W^2, so that $\sigma_T^2 = \sigma_B^2 + \sigma_W^2$. 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 (also a sample 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 strategies can be used to obtain valid analyses of cluster randomized experiments. The simplest is to treat the clusters as units of analysis by computing the mean scores on the outcome (and all other variables that may be involved in the analysis) for each cluster (e.g., classroom or school) and carrying out the statistical analysis as if the

cluster means were the data. If all cluster sample sizes are equal, this approach provides exact tests for the treatment effect (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) or analyses involving hierarchical linear models (see e.g., Raudenbush & Bryk, 2002). The latter approach is particularly attractive if the final cluster sample sizes in the experiment are not identical (as is almost always the case in field experiments). For general discussions of the design and analyses of cluster randomized experiments see Bloom (2005), Bloom, Bos, and Lee (1999), Donner and Klar (2000), Klar and Donner (2001), Murray (1998), Murray, Varnell, and Blitstein (2004), or Raudenbush and Bryk (2002).

Wise practice (and funding agency requirements) dictate that experiments be designed using sample sizes chosen so that the statistical test for treatment effects has adequate power to detect the smallest treatment effects that are of scientific or practical interest. The literature on the computation of statistical power typically addresses the power of studies that use simple random samples (e.g., Cohen, 1977; Kraemer & Thiemann, 1987; Lipsey, 1990). However methods for the computation of statistical power of tests for treatment effects are available for any of the analyses indicated above (see, e.g., Blair & Higgins, 1986; Raudenbush, 1997; Snijders & Bosker, 1993). In each case computation of statistical power involves the intraclass correlation ρ. Thus the computation of statistical power in cluster randomized experiments requires knowledge of the intraclass correlation ρ in addition to the information on sample size and effect size that is required in experiments employing simpler sampling designs. Moreover, power computations typically depend strongly on ρ, meaning that it is essential to have a clear idea of what the value of ρ might be in order to obtain realistic values of statistical power. Therefore researchers who are designing experiments that randomize schools to assess the effects of interventions on academic achievement have a crucial need for accurate information about intraclass correlations.

Unfortunately, few education researchers have realistic ideas about what value of ρ to expect. For example, there is conventional wisdom suggesting that values of ρ are typically between 0.05 and 0.15. Yet two recent attempts to establish reference values for ρ that could be useful in planning experiments suggested that values between 0.15 and 0.25 were more realistic in the nation as a whole (Hedges & Hedberg, 2007) and in large urban school systems (Bloom, Richburg-Hayes, & Black, 2005).

Neither the national values reported by Hedges and Hedberg (2007), nor the values from large urban school systems reported by Bloom, Richburg-Hayes, and Black (2005) are likely to be applicable to studies of rural schools. However the strategy used by Hedges and Hedberg (2006), which involved the analysis of sample surveys that used nationally representative probability samples with cluster sampling designs using schools as clusters, *can* be used to obtain reasonable values of ρ for rural settings.

The purpose of this article is to provide a comprehensive collection of intraclass correlations of academic achievement based on representative samples of rural schools. This compilation should be useful in choosing reference values for planning cluster randomized experiments in research on rural education. To anticipate, we find that the reference values of ρ obtained for the nation as a whole (and for urban schools) are generally larger than those for rural schools and would therefore lead to larger and more expensive experiments than are necessary in rural settings.

Designs and Populations Considered

The intraclass correlation structure will be different in different populations (for example, rural schools versus urban schools) and at different grade levels. It may, in principle, be different in different achievement domains (e.g., reading versus mathematics). The intraclass correlation structure is also different when different research designs are used (e.g., with or without covariates).

We examined the population of rural schools, at 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 at Grade 11 for which data on a national representative sample was not available to us.

The analyses reported here focused on intraclass correlations for two research designs involving assignment of schools to treatments, which have different intraclass correlation structures. The first design involves no covariates. Formally this is a hierarchical design in which schools are nested within treatments. The second design involves the use of a pretest in the same subject matter as a covariate at both the student and school level. Formally this is a hierarchical design in which schools are nested within treatments with two covariates. One covariate is the school-centered pretest score, and the second covariate is the school mean pretest score. Data collected under these designs could be analyzed via an analysis of variance with school as a factor nested within treatments (possibly including two covariates) or via a hierarchical linear model analysis using students as the first level and schools as the second level with treatment as a predictor at the second level (possibly including a covariate at each level).

Datasets Used

This article provides estimates of intraclass correlations and associated variance components for academic achievement in reading and mathematics for rural schools in the United States. Data from longitudinal surveys with national

probability samples were used because we wished to use achievement data collected in earlier years as pretest data for evaluating conditional intraclass correlations relevant to planning studies that would use a pretest as a covariate. When more than one survey could have provided data on a given grade level, we analyzed all of them but report results based on the survey with the largest sample size or that we believed would provide the most reliable estimates. Generally, we found that the results agreed within sampling error (see Hedges & Hedberg, 2006).

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 1591 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 third and fifth grades. There was no data collection in second and fourth grade. 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 Grades 2, 4 to 7, and 9 were based on data from 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, in 1991 (for a complete description of the study design, see Puma, Karweit, Price, Riccuti, & Vaden-Kiernan, 1997). Achievement test data was collected for three to four 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 was 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 was used as a pretest and in Grade 1 the test data collected in Fall served as a pretest. Results for Grade 2 were obtained from the first followup to the first grade (base year) sample and those reported for Grades 4 to 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 followup of the seventh grade sample in the Prospects study.

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 (NELS: 88). NELS: 88 is a longitudinal study that began in 1988 with a national probability sample of eighth graders in 1050 schools and collected reading and mathematics achievement test data when the students were in Grades 8, 10, and 12 (Curtin, Ingels, Wu, & Heuer, 2002).

Thus no pretest data was available for Grade 8, but for the Grade 10 the Grade 8 data was used as a pretest and for Grade 12 the Grade 10 data was 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 Miller, Hoffer, Suchner, Brown, & Nelson, 1992). The LSAY is a longitudinal study that began in 1987 with two national probability samples, one of seventh graders and one of tenth graders in 104 schools. Data were collected on mathematics and science achievement each year for four years leading to samples from Grades 7 to 12. There was 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 version 9.1's "XTMIXED" routine for mixed linear model analysis. For each sample and achievement domain, analyses were carried out based on two different models, which we call the unconditional model and the pretest covariate 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) levels. The level-one model for the kth observation in the jth school can be written as

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

and the level two model for the intercept is

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

where ε_{jk} is an individual-level residual and ζ_j is a random effect of the jth cluster (a level-two residual). The variance components associated with this analysis are σ_W^2 (the variance of the ε_{jk}) and σ_B^2 (the variance of the ζ_j).

The pretest covariate model. If pretest scores on achievement are available, their use as a covariate can considerably increase power in experimental designs. The pretest covariate model involves using as covariates the cluster-centered pretest score at the individual level and the school mean pretest score at the school level. Thus the level-one model for the kth observation in the jth school can be written as

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

and the level two model for the intercept is

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

where X_{jk} is the achievement pretest score for the jth observation in the kth school, $\overline{X}_{j\bullet}$ is the pretest mean for the jth school, ε_{jk} is an individual-level residual and ζ_j is a random effect of the jth school (a level-two residual) and 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 the ε_{jk}) and σ_{AB}^2 (the variance of the ζ_j).

The Intraclass Correlation

The intraclass correlation associated with the unconditional model described above is

$$\rho = \sigma_B^2/[\,\sigma_B^2 + \sigma_W^2] = \sigma_B^2/\sigma_T^2 \qquad (1)$$

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 random effects in an experiment that assigned schools to treatments and analyzed the data with no covariates. 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 pretest covariate model, the (covariate adjusted) intraclass correlation is

$$\rho_A = \sigma_{AB}^2/[\sigma_{AB}^2 + \sigma_{AW}^2] = \sigma_{AB}^2/\sigma_{AT}^2 \qquad (2)$$

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 random effects in an experiment that assigned schools to treatments and used the pretest at both individual and school level as 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 the pretest (individual and school mean) as covariates.

For each grade level and achievement domain, with and without covariates, 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) x 2 (achievement domains) x 2 (covariate sets) = 52 intraclass correlation estimates (each with a corresponding standard error).

For the pretest covariate model, we also provide values of

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

the percent reduction in between-school variance, and

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

the percent reduction in within-school variance 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.

Note that the parameters are $R_B^2 = 1 - \eta_B^2$ and $R_W^2 = 1 - \eta_W^2$, the proportion of between- and within-group variance explained by the covariate, are used by some authors (e.g., Bloom et al., 2005). We chose to tabulate the η^2 values instead of the R^2 values because of the simpler relation of the η^2 values to the operational effect size parameters used in power analysis.

Note that the two analyses involved slightly different variables, and there were missing values on some of these variables in our survey data. Because the quantities η_W^2 and η_B^2 involve a comparison of two different analyses (one with and one without pretest as a covariate), we believed 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. However, we used all cases that could be used (including those with missing pretest values) for computing the unadjusted intraclass correlation estimate.

We provide estimates of the standard errors of the intraclass correlations to provide some idea of their sampling uncertainty. Note, however, that the distribution of estimates of the intraclass correlations is only approximately normal. Moreover, 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 on different variables computed from the same sample of individuals.

Results

We found that the patterns of intraclass correlations for reading and mathematics in rural schools were quite similar. However, there were some differences, and consequently we present intraclass correlation data for both reading and mathematics achievement. For both mathematics and reading achievement, the intraclass correlation structure is substantially different for rural schools than for all schools in the nation, and we present the unconditional intraclass correlation for all schools to provide a comparison.

Mathematics achievement. Table 1 is a presentation of results for mathematics achievement. The table is divided into two panels of two columns (for the unconditional model in the entire nation and for rural schools) and one panel of

Table 1

Intraclass correlations for rural schools in mathematics achievement

| | All Schools | | Rural Schools Only | | | | | |
| | Unconditional Model | | Unconditional Model | | Pretest Covariate Model | | | |
Grade	ICC	(SE)	ICC	(SE)	ICC	(SE)	η_B^2	η_W^2
K	243	(9.8)	205	(20.0)	103	(14.5)	172	368
1	228	(9.8)	196	(36.0)	69	(22.1)	115	383
2	236	(19.4)	167	(29.6)	145	(28.0)	358	485
3	241	(10.4)	214	(21.8)	156	(18.9)	275	401
4	232	(19.6)	153	(28.5)	96	(22.3)	278	528
5	216	(17.9)	134	(26.1)	134	(26.9)	478	513
6	264	(19.4)	132	(28.9)	93	(23.2)	283	477
7	191	(33.0)	84	(32.7)	n/a	n/a	n/a	n/a
8	185	(31.5)	96	(36.3)	51	(23.9)	201	386
9	216	(32.3)	118	(43.2)	349	(69.7)	50	258
10	234	(10.0)	171	(22.6)	95	(17.1)	188	375
11	138	(28.3)	130	(45.9)	58	(26.6)	105	255
12	239	(10.9)	135	(15.0)	54	(10.0)	72	194
$M =$	220		149		117		215	385
$a =$	242		188		119		294	482
$b =$	-4		-7		0		-13	-16
$r =$	-0.44		-0.63		-0.02		-0.43	-0.62

Note: All values in this table are multiplied by 1,000, thus a value listed as 243 is 0.243.

a. The values for all schools are from Hedberg and Hedges (2006).

four columns (for the pretest covariate model). The data for each grade level is given in a different row. In the row for each grade, the columns of the first panel provide the estimates of the intraclass correlation (ρ) and the standard errors of the estimates of ρ (in parentheses after the estimate of ρ) for all schools in the nation. The columns of the second panel provide the estimates of the intraclass correlation (ρ) and the standard errors of the estimates of ρ (in parentheses after the estimate of ρ) for rural schools. The columns in the third panel provide the estimates of the conditional intraclass correlation (ρ_A), the standard errors of the estimates of ρ_A (in parentheses after the estimate of ρ_A), and estimates of η_B^2 and η_W^2 for rural schools. For example, consider the data for the pretest covariates 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 6 to 9 of the table) are 69, 22.1, 115, and 383, respectively, which correspond to estimates of 0.069, 0.0221, 0.115, and 0.383 for the estimate of ρ_A, the standard error of the estimate of ρ_A, η_B^2, and η_W^2.

The bottom four rows of each table give summary statistics (across grades) for the estimates of ρ, η_B^2, and η_W^2 to help interpret the table as a whole. The summary statistics are 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, the mean intraclass correlation in the unconditional model in rural schools is 0.149, the correlation between grade and intraclass correlation is –0.63, and the regression equation for predicting the unconditional intraclass correlation from grade is 0.188 – 0.007(Grade).

Comparing the unconditional intraclass correlations in all schools with those of rural schools, we see that the intraclass correlations in rural schools are generally smaller and often substantially so. The average intraclass correlation in all schools is 0.220, but in rural schools it is only 0.149. This illustrates the principle that design parameters suitable for experiments intending to represent all schools are generally not appropriate for experiments intended to represent rural schools. In particular, since intraclass correlations in rural schools are generally smaller, experiments involving fewer schools will have the same statistical power as larger experiments intended to represent the nation as a whole.

Although there is a tendency of the intraclass correlations in rural schools to be larger at lower grades, in general there are not large changes across adjacent grade levels. None of these changes exceed two standard errors of the change. The pretest covariate analyses typically reduced between cluster variance to from one tenth to one third of its value in the unconditional model (e.g., produced η_B^2 from 0.1 to 0.3), but typically reduced within-cluster variance to only from one third to one half of its value in the unconditional model (e.g., produced η_W^2 values of 0.3 to 0.5). Generally, smaller between- and within- (but especially between-) schools variance and smaller intraclass correlation lead to higher statistical power.

There is a slight anomaly in the results for the pretest covariate model in Grade 9. The adjusted intraclass correlation is larger than the unadjusted intraclass correlation. This is theoretically possible if the covariate adjustment reduces within-school variation more than between-school variation; however this appears to happen in Grade 9 only. It might be wise to use smoothed values estimated from the regression coefficients given in the bottom rows of the table. These smoothed values of the ninth grade adjusted intraclass correlation (ρ_A), η_B^2, and η_W^2 are 0.116, 0.173, and 0.335, respectively.

Reading achievement. Table 2 is a presentation of results for reading achievement, organized in the same way as Table 1, which reported results for mathematics. The intraclass correlation and adjusted intraclass correlation values in reading are generally quite similar to those in mathematics. As in mathematics, there is a tendency of the intraclass correlations in reading to become smaller at higher grades, but relative to mathematics the changes across adjacent grade levels are often larger. The results for Grade 2 are particularly inconsistent (having smaller values of the intraclass correlations) with the results from either Grade 1 or Grade 3. These differences exceed two standard errors of the difference. Because the values at Grade 2 do not fit the pattern of variation across the other grades, it might be wise to use "smoothed" values estimated from the regression coefficients given in the bottom rows of the table. These smoothed values of the second grade unadjusted intraclass correlation, the adjusted intraclass correlation (ρ_A), η_B^2, and η_W^2 are 0.204, 0.101, 0.182, and 0.441, respectively. None of the other differences exceed two standard errors of the difference.

There appears to be somewhat less consistency among the adjusted intraclass correlations in reading than in mathematics. There also appears to be somewhat smaller reduction in between-cluster variance in reading and than in mathematics. The pretest covariate analyses typically reduced the between cluster variance to from one tenth to one third of its value in the unconditional model (e.g., produced η_B^2 from 0.1 to 0.3), but typically reduced within-cluster variance by onlt slightly over one half (e.g., produced η_W^2 values less than 0.5).

There is a slight anomaly in the results for the pretest covariate model in Grade 9. The adjusted intraclass correlation is larger than the unadjusted intraclass correlation. This is theoretically possible if the covariate adjustment reduces within-school variation more than between-school variation; however this appears to happen in Grade 9 only. It might be wise to use smoothed values estimated from the regression coefficients given in the bottom rows of the table. These smoothed values of the ninth grade adjusted

Table 2

Intraclass correlations for rural schools in reading achievement

	All Schools		Rural Schools Only					
	Unconditional Model		Unconditional Model		Pretest Covariate Model			
Grade	ICC	(SE)	ICC	(SE)	ICC	(SE)	η_B^2	η_W^2
K	233	(9.7)	240	(21.7)	192	(20.4)	306	412
1	239	(10.0)	255	(40.3)	111	(27.5)	133	367
2	204	(17.9)	144	(26.9)	79	(18.6)	236	439
3	271	(10.8)	229	(22.5)	178	(20.2)	367	507
4	242	(19.9)	188	(32.6)	40	(12.5)	71	459
5	263	(19.5)	181	(31.6)	86	(20.0)	171	413
6	260	(19.2)	148	(31.0)	35	(12.5)	85	439
7	174	(20.0)	76	(19.7)	n/a	n/a	n/a	n/a
8	197	(8.5)	111	(11.6)	n/a	n/a	n/a	n/a
9	250	(25.5)	103	(26.6)	341	(56.1)	50	494
10	183	(8.9)	119	(18.6)	53	(13.1)	198	456
11	-- --		n/a	n/a	n/a	n/a	n/a	n/a
12	174	(9.5)	99	(12.9)	69	(11.0)	245	358
M =	224		158		118		186	434
a =	258		229		95		186	443
b =	-6		-13		3		-2	-1
r =	-0.53		-0.79		0.10		-0.09	-0.07

Note: All values in this table are multiplied by 1,000, thus a value listed as 233 is 0.233.

a. The values for all schools are from Hedberg and Hedges (2006).

intraclass correlation (ρ_A), $\eta_B{}^2$, and $\eta_W{}^2$ are 0.119, 0.166, and 0.434, respectively.

Minimum Detectable Effect Sizes

A simple way to describe the implications of these results for statistical power is to use them to derive the smallest effect size for which a design that randomized schools to treatments would have adequate statistical power. This effect size is often called the minimum detectable effect size (MDES) (Bloom, 1996, 2005). We used 0.8 (with a two-sided test at significance level 0.05) as the criterion for adequate statistical power in computing the MDES values reported in this article. A power value of 0.8 has been widely used in research design and was recommended, for example, by Cohen (1977). We considered designs with no covariates and with pretest as a covariate at both the individual and group level and 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 with $m = 10, 15, 20, 25,$ or 30 schools randomized to each treatment group.

Table 3 gives the minimum detectable effect sizes for studies of mathematics achievement based on the parameters given in Table 1. With no covariates, the MDES values are in the range of 0.45 to 0.65 for $m = 10$ schools assigned to each treatment and typically exceed 0.25 even for $m = 30$. The use of a pretest as a covariate typically reduces the MDES values to less than 0.40 for $m = 10$ and 0.30 or less for $m = 15$. MDES values tend to be smaller in the higher grades, so that the MDES values are less than 0.25 for $m = 15$ in Grades 6 to 12. Table 4 gives the minimum detectable effect sizes for studies of reading achievement based on the parameters given in Table 2. These values are somewhat larger than, but quite comparable to, those for mathematics achievement given in Table 3.

Although there is no universally adequate standard for evaluating the importance of effect sizes, applying Cohen's (1977) widely used labels of 0.20 as small and 0.50 as medium would imply that an experiment randomizing $m = 10$ schools to each treatment should easily be adequate to detect effects of "medium" size in a design that used pretest as a covariate and that an experiment randomizing $m = 15$ schools to each treatment should be adequate to detect effects of a size slightly larger than "small."

One reason for this article is that the intraclass correlation structure found in rural schools is somewhat different than that of the nation as a whole. Thus the design requirements for experiments intending to represent rural schools are somewhat different than for experiments intended to represent the nation as a whole. Comparing the minimum detectable effect sizes in Tables 3 and 4 to the MDES values for the entire nation given in Table 7 of Hedges and Hedberg (2006), we see that the MDES values in Tables 3 and 4 are generally smaller. For example, consider an experiment in which the outcome is mathematics achievement at Grade 6 and which uses pretest as a covariate. The MDES for an experiment randomizing $m = 10$ schools to each treatment and intending to generalize to all schools is $d = 0.37$, but Table 3 of this article reveals that the MDES for the same experiment conducted in rural schools is $d = 0.28$. This means that, to obtain the same statistical power, experiments intending to represent rural schools require smaller sample sizes than experiments intending to represent the nation as a whole. To illustrate the difference, it would be necessary to have $m = 17$ schools per treatment in the national study to obtain the same power to detect an effect of $d = 0.28$ as is possible with an experiment involving $m = 10$ rural schools. This is a very substantial (70%) difference in sample size requirements.

Using the Results of this Article to Compute Statistical Power of Cluster Randomized Experiments

In this section, we illustrate the use of the results in this article to compute the statistical power of cluster randomized experiments comparing two treatments. We first show how to use conventional power tables and power calculation software (which assume simple random samples), then show how to carry out the computations using the noncentral t-distribution function that is available in many software packages such as SPSS, SAS, and STATA. In each case we show how to compute power with and without covariates.

Using Power Tables and Power Calculation Software

Tables giving statistical power (e.g., Cohen, 1977) and computer software for computing statistical power (e.g., Borenstein, Rothstein, & Cohen, 2001) are available for many designs using simple random samples, but those designed for obtaining power from the independent-groups t-test are the most widely available. Following Cohen's framework, such tables typically provide power values based on treatment and control group 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 these tables are based translate the sample sizes and effect size into the degrees of freedom and noncentrality parameter of the noncentral t-distribution on which the statistical power depends.

Because the power of tests for treatment effects in cluster randomized experiments also depends on the noncentral t-distribution, tables like 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

Table 3

Minimum detectable mathematics achievement effect sizes in rural settings with power 0.80 and n = 60 as a function of m

Grade	Covariates	Number of schools per treatment group (m)				
		10	15	20	25	30
K	none	0.62	0.50	0.43	0.38	0.35
	pretest	0.27	0.22	0.19	0.17	0.15
1	none	0.61	0.49	0.42	0.38	0.34
	pretest	0.23	0.18	0.16	0.14	0.13
2	none	0.57	0.46	0.39	0.35	0.32
	pretest	0.35	0.28	0.24	0.21	0.19
3	none	0.64	0.51	0.44	0.39	0.36
	pretest	0.34	0.27	0.24	0.21	0.19
4	none	0.55	0.44	0.38	0.34	0.31
	pretest	0.30	0.24	0.21	0.19	0.17
5	none	0.52	0.41	0.36	0.32	0.29
	pretest	0.36	0.29	0.25	0.22	0.20
6	none	0.51	0.41	0.35	0.31	0.29
	pretest	0.28	0.23	0.20	0.18	0.16
7	none	0.42	0.34	0.29	0.26	0.24
	pretest	--	--	--	--	--
8	none	0.45	0.36	0.31	0.27	0.25
	pretest	0.22	0.17	0.15	0.13	0.12
9	none	0.49	0.39	0.34	0.30	0.27
	pretest	0.14	0.11	0.09	0.08	0.08
10	none	0.57	0.46	0.40	0.35	0.32
	pretest	0.26	0.21	0.18	0.16	0.15
11	none	0.51	0.41	0.35	0.31	0.28
	pretest	0.18	0.14	0.12	0.11	0.10
12	none	0.52	0.41	0.36	0.32	0.29
	pretest	0.15	0.12	0.11	0.10	0.09

Table 4
Minimum detectable reading achievement effect sizes in rural settings with power 0.80 and n = 60 as a function of m

Grade	Covariates	Number of schools per treatment group (*m*)				
		10	15	20	25	30
K	none	0.67	0.54	0.46	0.41	0.37
	pretest	0.38	0.30	0.26	0.23	0.21
1	none	0.69	0.55	0.48	0.42	0.39
	pretest	0.27	0.21	0.18	0.16	0.15
2	none	0.53	0.43	0.37	0.33	0.30
	pretest	0.27	0.22	0.19	0.17	0.15
3	none	0.66	0.53	0.45	0.40	0.37
	pretest	0.41	0.32	0.28	0.25	0.23
4	none	0.60	0.48	0.41	0.37	0.34
	pretest	0.19	0.15	0.13	0.12	0.11
5	none	0.59	0.47	0.41	0.36	0.33
	pretest	0.26	0.21	0.18	0.16	0.15
6	none	0.54	0.43	0.37	0.33	0.30
	pretest	0.19	0.15	0.13	0.12	0.11
7	none	0.41	0.33	0.28	0.25	0.23
	pretest	--	--	--	--	--
8	none	0.47	0.38	0.33	0.29	0.27
	pretest	--	--	--	--	--
9	none	0.46	0.37	0.32	0.28	0.26
	pretest	0.15	0.12	0.11	0.10	0.09
10	none	0.49	0.39	0.34	0.30	0.27
	pretest	0.24	0.19	0.16	0.15	0.13
12	none	0.45	0.36	0.31	0.28	0.25
	pretest	0.23	0.19	0.16	0.14	0.13

(here called the *operational effect size*) that will yield the appropriate degrees of freedom and noncentrality parameter. We describe this process first for analyses without covariates and then for analyses with the pretest as a covariate.

No covariates. If the actual numbers of clusters assigned are m_1 and m_2, then entering the power table with sample sizes $N_1^T = m_1$ and $N_2^T = m_2$ yields the correct degrees of freedom for the test. The relevant operational effect size using our choice of degrees of freedom is

$$\Delta^T = \delta \sqrt{\frac{n}{1 + (n - 1)\rho}} \qquad (5)$$

where δ is the effect size (standardized mean difference) and ρ is the *unadjusted* intraclass correlation.

Pretest as a covariate. In this design, we assume that the pretest has been used as a covariate at both the cluster level (as the school mean of the covariate) and at the individual level (as an individual value centered within each cluster). If the actual numbers of clusters assigned are m_1 and m_2, then entering the power table with sample sizes $N_1^T = m_1$ and $N_2^T = m_2 - 1$ yields the correct degrees of freedom for the test, since one degree of freedom is lost for the introduction of the cluster-level covariate. The relevant operational effect size is

$$\Delta^T = \delta \sqrt{\frac{nm_2 (m_1 + m_2 - 1)}{(m_1 + m_2)(m_2 - 1)}} \sqrt{\frac{1}{\eta^2_W + (n\eta_B^2 - \eta_W^2)\rho}} \qquad (6)$$

where δ is the (unadjusted) effect size, ρ is the *unadjusted* intraclass correlation, and η_B^2 and η_W^2 are the ratios of adjusted to unadjusted between- and within-cluster variances defined in Equations (3) and (4). Reference values for all of these parameters are given in Tables 1 and 2 of this article. Using ρ, 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 based on clustered samples using these tables and computer programs designed for the two group t-test.

Note that, if the covariate had been used at the school level only, the procedure for computing statistical power would be exactly the same as that given above, except that the value of η_W^2 used in Equation (6) would be $\eta_W^2 = 1$. If the covariate had been used at the individual level only (centered within schools), there would be two differences in the procedure for computing statistical power. The first difference is that the value of η_B^2 used in Equation (6) would be $\eta_B^2 = 1$. The second difference would be that N_1^T and N_2^T would be set equal to m_1 and m_2 as in the case with no covariates.

Using the Noncentral t-*Distribution Function to Compute Power*

Although tables of the power of the two sample t-test are common, they do not cover every possible situation. However most statistical packages (including SPSS, SAS, and STATA) include a noncentral t-distribution function, which can be used to compute power directly. When the null hypothesis is false, the t-statistic has a noncentral t-distribution which depends on two parameters: degrees of freedom and a noncentrality parameter. Suppose that there are m_1 clusters in the control group, m_2 clusters in the treatment group, and the effect size (before covariate adjustment) is δ.

No covariates. When there are no covariates, the degrees of freedom are $\nu = m_1 + m_2 - 2$ and the noncentrality parameter is

$$\lambda = \delta \sqrt{\frac{m_1 m_2 n}{m_1 + m_2}} \frac{1}{\sqrt{[1 + (n - 1)\rho]}} \qquad (7)$$

where ρ is the *unadjusted* intraclass correlation.

The power of the one-tailed test at level α is

$$p_1 = 1 - H[c(\alpha, \nu), \nu, \lambda] \qquad (8)$$

where $c(\alpha, \nu)$ is the level α one-tailed critical value of the t-distribution with ν degrees of freedom [e.g., $c(0.05,10) = 1.81$], 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

$$p_2 = 1 - H[c(\alpha/2, \nu), \nu, \lambda] + H[-c(\alpha/2, \nu), \nu, \lambda]. \qquad (9)$$

Pretest as a covariate. If the pretest has been used as a covariate at both the cluster level (as cluster mean on the covariate) and at the individual level (as an individual value centered within each cluster), the degrees of freedom are $\nu = m_1 + m_2 - 3$ and the noncentrality parameter is

$$\lambda_A = \delta \sqrt{\frac{m_1 m_2 n}{m_1 + m_2}} \sqrt{\frac{1}{\eta_W^2 + (n\eta_B^2 - \eta_W^2)\rho}} \qquad (10)$$

where δ is the effect size before covariate adjustment, ρ is the *unadjusted* intraclass correlation, and η_B^2 and η_W^2 are the ratios of adjusted to unadjusted between- and within-cluster variances defined in Equations (3) and (4).

The power of the one-tailed test at level α is

$$p_1 = 1 - H[c(\alpha, \nu), \nu, \lambda_A] \qquad (11)$$

where $c(\alpha, \nu)$ is the level α one-tailed critical value of the t-distribution with ν degrees of freedom [e.g., $c(0.05,10) = 1.81$], 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

$$p_2 = 1 - H[c(\alpha/2, \nu), \nu, \lambda_A] + H[-c(\alpha/2, \nu), \nu, \lambda_A]. \quad (12)$$

Note that, if the covariate had been used at the school level only, the procedure for computing statistical power would be exactly the same as that given above, except that the value of η_W^2 used in Equation (8) would be $\eta_W^2 = 1$. If the covariate had been used at the individual level only (centered within schools), the value of η_B^2 used in Equation (8) would be $\eta_B^2 = 1$, and the degrees of freedom would become $\nu = m_1 + m_2 - 2$ as in the case with no covariates.

Example with No Covariates

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. The analysis will involve a two-tailed t-test with significance level $\alpha = 0.05$. 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 represent a national sample of rural first graders. Entering Table 1 on the second row for Grade 1 and the panel for the unconditional model (columns 4 and 5) gives the intraclass correlation for first graders as $\rho = 0.196$.

Using power tables. 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 (5) as

$$\Delta^T = \frac{0.50\sqrt{20}}{\sqrt{1 + (20 - 1)(0.196)}} = 1.029.$$

Cohen's tables give the statistical power in terms of sample size (in each treatment group) and effect size. Examining Cohen's (1977) Table 2.3.5, we see that the operational effect size of 1.029 is between tabled effect sizes of 1.0 and 1.2. 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 these two values, we obtain a power of 0.58 for $\Delta^T = 1.03$.

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 1.03 versus 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.

Using the noncentral t-distribution. Alternately, we could compute the power from the noncentral t-distribution function. The noncentrality parameter from Equation (7) is

$$\lambda = \frac{0.50\sqrt{(10/2)20}}{\sqrt{1 + (20 - 1)(0.196)}} = 2.300.$$

Using Equation (9) and the noncentral t-distribution function, (for example the function NCDF.T in SPSS), with $\nu = 10 + 10 - 2 = 18$ degrees of freedom, $c(0.05/2, 18) = 2.101$, and $\lambda = 2.300$, we obtain a two-sided power of $p_2 = 1 - 0.41 + 0.00 = 0.59$. The slight difference between the value computed using the noncentral t-distribution directly and that from the power tables (0.58 versus 0.59) is due to rounding and interpolation from the values in the power tables.

Example with Pretest as a Covariate at Both Individual and Cluster Level

Consider an experiment that will randomize $m_1 = m_2 = 10$ schools to receive an intervention to improve third grade reading achievement and that $n = 20$ students in each school would be part of the experiment. An analysis of covariance will be used with pretest as a covariate at both individual and school level using a two-tailed test with significance level $\alpha = 0.05$. Suppose that the smallest educationally significant effect size for this intervention is $\delta = 0.40$. Suppose further that the schools were chosen to attempt to be representative of rural first graders nationally.

Entering Table 2 on the fourth row for Grade 3 and the panel for the unconditional model (columns 3 and 4) gives the intraclass correlation for third graders as $\rho = 0.229$. Entering Table 2 on the fourth row for Grade 3 and the panel for the pretest covariate model (columns 6 to 9) gives the between- and within-school variance ratios after covariate adjustment as $\eta_B^2 = 0.367$ and $\eta_W^2 = 0.507$.

Using power tables. To compute the power from tables of the power of the t-test such as those given by Cohen (1977), we face a complication in that the tables assume equal integer-valued treatment and control group sample sizes, but we have to use operational sample sizes

of $N_1{}^T = 10$ and $N_2{}^T = 9$. Because Cohen's tables give the statistical power in terms of equal sample sizes (in each treatment group), we will need to interpolate between the tabled sample sizes $N_1{}^T = N_2{}^T = 9$ and $N_1{}^T = N_2{}^T = 10$. The operational effect size depends on $N_1{}^T$ and $N_2{}^T$, so we have to compute a different value of Δ^T for each of the sample sizes between which we will interpolate. For $N_1{}^T = N_2{}^T = 9$, the operational effect size computed using Equation (6) is

$$\Delta^T = 0.4 \sqrt{\frac{(20)(9 + 9 - 1)}{(9 + 9)(9 - 1)}} \cdot$$

$$\sqrt{\frac{1}{0.507 + [(20)(0.367) - 0.507](0.229)}} = 1.281.$$

For $N_1{}^T = N_2{}^T = 10$, the operational effect size computed using Equation (6) is

$$\Delta^T = 0.4 \sqrt{\frac{(20)(10 + 10 - 1)}{(10 + 10)(10 - 1)}} \cdot$$

$$\sqrt{\frac{1}{0.507 + [(20)(0.367) - 0.507](0.229)}} = 1.277.$$

Note that, although the operational effect sizes depend slightly on $N_1{}^T$ and $N_2{}^T$, both values round to 1.28, illustrating that the operational effect sizes are essentially identical for consecutive sample sizes found in power tables.

Examining Cohen's (1977) Table 2.3.5, we see that the effect size $\Delta^T = 1.28$ is between tabled values of effect size of 1.2 and 1.4. Entering the Table with sample size $N_1{}^T = N_2{}^T = 9$, we see that a power of 0.65 is tabulated for the effect size of $\Delta^T = 1.2$ and a power of 0.79 is tabulated for an effect size of $\Delta^T = 1.4$. Interpolating between the two power values (0.65 and 0.79) for $N_1{}^T = N_2{}^T = 9$, we obtain a power of 0.71 for $\Delta^T = 1.28$. This value (0.71) corresponds to the power associated with the effect size of $\delta = 0.40$ and a test based on 16 degrees of freedom.

Examining Cohen's (1977) Table 2.3.5 again with sample size $N_1{}^T = N_2{}^T = 10$, we see that a power of 0.71 is tabulated for the effect size of $\Delta^T = 1.2$ and a power of 0.84 is tabulated for an effect size of $\Delta^T = 1.4$. Interpolating between the two power values (0.71 and 0.84) for $N_1{}^T = N_2{}^T = 10$, we obtain a power of 0.76 for $\Delta^T = 1.28$. This value (0.76) corresponds to the power associated with the effect size of $\delta = 0.40$ and a test based on 18 degrees of freedom.

To obtain the power associated with an effect size of $\delta = 0.4$ and a test based on 17 degrees of freedom, we must interpolate once again between these two values. We obtain a power value for $N_1{}^T = 9$ and $N_2{}^T = 10$ of $p_2 = 0.74$.

Using the noncentral t-distribution. Alternatively, we could use the noncentral t-distribution to compute power. Then the noncentrality parameter from Equation (10) is

$$\lambda_A = 0.4 \sqrt{\frac{(10)(10)(20)}{10 + 10}} \cdot$$

$$\sqrt{\frac{1}{0.507 + ((20)(0.367) - 0.507)(0.229)}} = 2.779.$$

Using Equation (12) and the noncentral t-distribution function, (for example the function NCDF.T in SPSS), with $10 + 10 - 2 - 1 = 17$ degrees of freedom, $c(0.05/2, 17) = 2.110$, and $\lambda_A = 2.779$, we obtain a two-sided power of $p_2 = 1 - 0.25 + 0.00 = 0.75$. The slight difference between the value computed using the noncentral t-distribution directly and that from the power tables (0.74 versus 0.75) is due to rounding and interpolation from the values in the power tables.

Conclusions

The values of intraclass correlations presented in this article suggest that values for rural schools are smaller than those in the nation as a whole, typically ranging from 50 to 85% as large in mathematics and from 55 to 100% as large in reading. For Kindergarten through Grade 4 in math and through Grade 5 in reading, somewhat larger values of the interclass correlation (roughly 0.15 to 0.25) may be appropriate in rural education than the 0.05 to 0.15 guidelines that have sometimes been used. The guideline of 0.05 to 0.15 is more consistent with the values of intraclass correlations we found in Grades 5 to 12 in math and Grades 6 to 12 in reading.

The differences between rural schools and all schools in intraclass correlations and covariate effectiveness in reducing between- and within-school variation are large enough to have important consequences for the design of experiments. Because statistical power depends so much on these design parameters, these differences translate into differences of 50% or more in required sample sizes, with smaller sample sizes typically being required to achieve the same statistical power in studies representing rural schools.

The principal application of the results given in this article will probably be for planning randomized experiments in rural education that assign schools (rather than individuals) to treatments. However, it is important to recognize that the between-district and between-state components of variance are not estimated here. Consequently, these two components of variance are implicitly included here as part of the between-school variance. This is desirable if the values are to be used to plan experiments that involve schools from several districts or states. However, the estimates reported

here may overestimate the relevant intraclass correlations to some degree if the design involves schools from only a single district or state. Similarly, whether multiple districts or states are involved will have some impact on the effectiveness of the covariates in explaining between- and within-school variation. It is unclear just how much of an impact this may have, but the consistency of Hedges and Hedberg's (2007) results based on national data with those of Bloom et al. (2005) based on single school districts suggest that district and state effects on intraclass correlations may not be large.

There are also other potential applications of the results reported in this article. One involves the use of information external to an experiment to adjust the degrees of freedom of significance tests in designs involving group randomization (see Murray, Hannan, & Baker, 1996). Murray, et al. caution that users should have good reasons to assume that any external estimates used actually estimate the same intraclass correlation as that in the experiment. If the data from this compilation meets that assumption in any given situation involving rural schools, these intraclass correlations should substantially increase the degrees of freedom used in the test for treatment effects because they tend to have relatively small standard errors.

A second potential application of these intraclass correlations is for the adjustment of test statistics based on analyses that incorrectly ignored clustering. Ignoring clustering (when it is present), generally makes results look more significant than they actually are. Adjustments to the *t*-statistic for the effects of sample clustering are available, but these corrections require knowledge of ρ (Hedges, 2007). For experiments that have been conducted in rural schools, the values in this compilation provide some guidelines on values of ρ that might be plausible in attempting to determine if a conclusion about the statistical significance of a treatment effect might have held if clustering had been taken into account.

A third potential application of the reference values given in this article is to the computation of standardized effect size estimates and their standard errors in group randomized trials. Depending on the statistics reported, the computation of effect size estimates and their standard errors in multilevel designs may require knowledge of ρ (see Hedges, in press). If the report of the experiment itself does not include information that can be used to compute an estimate of ρ, a reviewer conducting a meta-analysis will need to impute values of ρ, and this compilation may provide some idea of values that are plausible in rural schools.

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 morefrom social experiments: Evolving analytic approaches.* New York: Russell Sage Foundation.

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. (2005). *Using covariates to improve precision: Empirical guidelines for studies that randomize schools to measure the impacts of educational interventions.* New York, NY: MDRC.

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

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

Coladarci, T., & Gage, N. L. (1984). Effects of a minimal intervention on teacher behavior and student achievement. *American Educational Research Journal, 21*, 539-556.

Curtin, T. R., Ingels, S. J., Wu, S., & Heuer, R. (2002). *User's manual: Nels:88 base-year to fourth followup*. Washington, DC: US 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., and J. J. Koval. (1982). Design considerations in the estimation of intraclass correlation. *Annals of Human Genetics, 46*, 271-77.

Good, T. L. & Grouws, D. A. (1979). The Missouri mathematics effectiveness project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology, 71*, 355-362.

Hedges, L.V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32,* 151-179.

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

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

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: John 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.

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

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

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

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

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, volume 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 Publications.

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: US National Center for Education Statistics.