The Power of the Test in Three-Level Designs

by

Spyros Konstantopoulos

Northwestern University

Note: This material is based upon work supported by the National Science Foundation under Grant No. 0129365. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. I thank Larry Hedges for his valuable comments.

Abstract

Field experiments that involve nested structures may assign treatment conditions either to entire groups (such as classrooms or schools), or individuals within groups (such as students). Since field experiments involve clustering, key aspects of their design include knowledge of the intraclass correlation structure and the sample sizes necessary to achieve adequate power to detect the treatment effect. This study provides methods for computing power in three-level designs, where for example, students are nested within classrooms and classrooms are nested within schools. The power computations take into account clustering effects at the classroom and at the school level, sample size effects (e.g., number of students, classrooms, and schools), and covariate effects (e.g., pre-treatment measures). The methods are generalizable to quasi-experimental studies that examine group differences in an outcome, or associations between predictors and outcomes.

Keywords: Nested designs, statistical power, randomized trials, treatment effects, random effects

Many populations of interest in psychology, education, and the social sciences exhibit multilevel structure. For example, students are nested within classrooms and classrooms are nested within schools, employees are nested within departments, which are nested within firms, individuals are nested within neighborhoods, which are nested within cities. Because individuals within aggregate units are often more alike than individuals in different units, this nested structure induces an intraclass correlation structure (often called clustering in the sampling literature, see, e.g., Kish, 1965) that needs to be taken into account. In particular, experiments that involve populations with nested structures *must* take this structure into account in both experimental design and analysis.

Field experiments that involve nested population structures, may assign treatment conditions either to individuals or to entire groups (such as classrooms, or schools). Because pre-existing aggregate groups often exhibit statistical clustering, designs that assign intact groups to treatment are often called cluster randomized or group randomized designs. Treatments are sometimes assigned to groups because the treatment is naturally administered to intact groups (such as a curriculum to a classroom or a management system to a firm). In other cases, the assignment of treatments to groups is a matter of convenience, being much easier to implement than assignment to individuals. Analytic strategies for the analysis of designs involving nesting have long been available (see, e.g., Kirk, 1995 and references therein). Such designs involve nested factors (or blocks), which are often considered to have random effects. Cluster randomized trials have also been used extensively in public health and other areas of prevention science and medicine (see, e.g., Donner and Klar, 2000; and Murray, 1998). The use of cluster randomized experiments has increased recently in educational research, mainly because of the increased interest in experiments to evaluate educational interventions (see, e.g., Mosteller and

Boruch, 2002). An example of cluster randomized trials in education is Project STAR, a large scale randomized experiment, where within each school students were randomly assigned to classrooms of different sizes (see Nye, Hedges, and Konstantopoulos, 2000).

One of the most critical issues in designing experiments is to ensure that the design is sensitive enough to detect the intervention effects that are expected if the researchers' hypotheses were correct. In other words, a critical task in planning experimental studies involves making decisions about sample sizes to ensure sufficient statistical power of the test for the treatment effect. Power is defined as the probability of detecting a treatment effect when it exists. There is an extensive literature on the computation of statistical power, (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Murphy & Myors, 2004). Much of this literature involves the computation of power in studies that use simple random samples and hence ignore clustering effects. Recently, software for computing statistical power in single-level designs has become widely available (Borenstein, Rothstein, & Cohen, 2001).

However, the computation of statistical power in designs that involve clustering entails two significant challenges. First, since nested factors are usually taken to have random effects, power computations usually involve variance components structures (often expressed via intraclass correlations) of those random effects. Second, there is not one sample size, but sample sizes at each level of nesting, and these sample sizes at different levels of nesting will affect power differently. For example, in education, the power of the test used to detect a treatment effect depends not only on the sample size of the students within a classroom or a school, but on the sample size of classrooms or schools as well. This design involves nesting at the school or at the classroom level, and is called a two-level design. The last decade the development of advanced statistical methods that account for clustering has equipped social science researchers

with appropriate tools for handling data with nested structures (Goldstein, 2003; Raudenbush & Bryk, 2002). In addition, statistical theory for computing power in two-level designs has been well documented and statistical software for two-level designs is currently available (e.g., Raudenbush, 1997; Raudenbush & Bryk, 2002; Raudenbush & Liu, 2000, 2001; Snijders & Bosker, 1993, 1999).

The work on power analysis for nested designs with random effects has focused thus far solely on designs with two levels of clustering, for example students within schools, or repeated measurements over time within individuals (see Raudenbush & Liu, 2000, 2001). However, designs and data have often more complicated structures that may involve three levels of nesting. For example, in education, students are nested within classrooms, and classrooms are nested within schools. In medicine, patients are nested within wards, and wards are nested within hospitals. These examples are three-level designs where nesting occurs naturally at two levels (classrooms and schools) regardless of whether both sources of clustering are taken into account in the design and the analysis part of the study. For example, an educational researcher may choose to ignore one level of clustering and conduct an experimental study in education following a two-level design (students nested within classrooms, or students nested within schools), and analyze the data obtained using two-level models. However, a more appropriate design and analysis would involve three levels, since clustering effects exist naturally at the classroom and at the school level. Ignoring the three-level structure will result in an inaccurate estimate of power of the test for the treatment effect in the design stage, and an incorrect estimate of the standard error of the treatment effect in the analysis. In three-level designs the computation of power is even more complex than in two-level designs, since there are clustering effects at the second or middle level (e.g., classrooms) and clustering effects at the third of top

level (e.g., schools). The power in these designs is, among other things, a function of three different sample sizes: the number of students within a classroom, the number of classrooms within a school, and the overall number of schools that participate in the experiment. The power is also a function of two different intraclass correlations (at the school and at the classroom level) and the effect size (e.g., the magnitude of the treatment effect). Thus far, methods for computing power and sample size requirements in experiments that involve three-level designs are not available.

In addition, the analyses of nested data do not always involve two levels. For example, it is not uncommon in education to conduct analyses assuming three-level models (e.g., Bryk & Raudenbush, 1988; Nye, Konstantopoulos, & Hedges, 2004). It is crucial that the design and analyses of experimental studies are in congruence, since ignoring or including clustering effects will impact power differently. Analyses that ignore clustering effects provide incorrect estimates of the standard errors of the regression coefficients. Findings from previous work that discuss power in two-level designs (e.g., Hedges & Hedberg, 2006; Raudenbush, & Liu, 2000) indicate that the proportion of the variance in the outcome between schools or clusters is inversely related to the power of the test for the treatment effect. In other words, the larger the clustering effect, the lower the power of the test for the treatment effect. Now consider the case where one more level is added to the hierarchy (e.g., classrooms within schools). If we assume that the clustering effect at the classroom level is different than zero, and the clustering effect at the school level is virtually unchanged, then the total clustering effects are larger than in two-level designs and hence the power has to be smaller.

This paper provides methods that facilitate the computation of statistical power of tests for treatment effects in three-level designs. We provide examples drawn from the field of

education to illustrate the power computations. The remaining of the paper is structured as follows. First, we review the effects of clustering on power. Second, we outline the three-level designs that will be discussed, and we explain the notation we use throughout the paper. Then, we present methods and examples for computing power separately for each design. Finally, we summarize the usefulness of the methods and draw conclusions.

Clustering in Three-Level Designs

Previous methods for power analysis in two-level designs involved the computation of the non-centrality parameter of the non-central *F*-distribution (Raudenbush & Liu, 2000). The power is a function of the non-centrality parameter and higher values of this parameter correspond to higher values of statistical power. The non-centrality parameter in turn is a function of the clustering effect, which is typically expressed as an intraclass correlation. For example, suppose that in a two-level design the total variance of the outcome in a population with nested structure (students nested within schools) is σ_T^2. The total variance is decomposed into a between-school variance ω^2 and a within-school variance σ_e^2, so that $\sigma_T^2 = \sigma_e^2 + \omega^2$. Then $\rho = \omega^2 / \sigma_T^2$ is the intraclass correlation and indicates the proportion of the variance in the outcome between schools.

The same logic holds for three-level designs. The power is a function of the non-centrality parameter, which is a function of the clustering effects. The only difference is that clustering in three-level designs occurs in more than one level. Consider a three-level structure where students are nested within classrooms, and classrooms are nested within schools. The total variance in the outcome is now decomposed into three variances: the between-student-within-classroom and school variance, σ_e^2, the between-classroom-within-school variance, τ^2, and the

between-school variance, ω^2. Then, the total variance in the outcome is defined as $\sigma_T^2 = \sigma_e^2 + \tau^2 + \omega^2$. Hence, in three-level designs we define two intraclass correlations:

$$\rho_c = \frac{\tau^2}{\sigma_T^2} \tag{1}$$

at the classroom level and

$$\rho_s = \frac{\omega^2}{\sigma_T^2} \tag{2}$$

at the school level. These intraclass correlations indicate the effects of clustering at the classroom and at the school level respectively. The subscripts c and s indicate the levels of the hierarchy (e.g., classroom or school).

The computation of statistical power in cluster randomized trials requires knowledge of the intraclass correlations. One way to obtain information about reasonable values of intraclass correlations is to compute these values from cluster randomized trials that have been already 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, Varnell, and Blitstein (2004) give references to 14 very recent studies that provide data on intraclass correlations for health related outcomes. Another strategy is to analyze sample surveys that have used a cluster sampling design. Gulliford, Ukoumunne, and Chinn (1999) and Verma and Lee (1996) presented values of intraclass correlations based on surveys of health outcomes.

A recent study provided plausible values of clustering for educational outcomes using recent large-scale studies that surveyed national probability samples of elementary and secondary students in America (Hedges & Hedberg, 2006). The present study uses intraclass correlation values that are reported in Hedges and Hedberg (2006).

Three-Level Designs

In this study we use examples from education to illustrate the nested structure of the designs. Hence, we use students, classrooms, and schools to define the three levels of the hierarchy. We make use of the terminology of education because it makes the differentiation between smaller and larger units, and the hierarchical structure more obvious. We also assume that the outcome of interest is academic achievement. Nonetheless, the designs and methods discussed here can be used for any hierarchical design with three levels of nesting and it is *not* restricted to educational designs or data.

In two-level designs either students within schools or entire schools are randomly assigned to one of two or more treatment conditions. This indicates that random assignment can occur either at the individual or at the school level. In three-level designs however, the nested structure is more complex, since either students within classrooms, classrooms within schools, or entire schools can be randomly assigned to treatment conditions. This study discusses three-level designs where randomization can occur at any level: student, classroom, or school. In the first design the random assignment is at the school level (the largest unit), that is, schools are assigned to a treatment and a control group. In the second design the random assignment is that classroom level (the second largest unit), that is, within each school classrooms are randomly assigned to a treatment and a control group. In the third design the random assignment is at the student level (the smallest unit), that is, within each school and classroom students are random assigned to a

treatment and a control group. In each design we illustrate the general case, which includes covariates at all levels of the hierarchy, and the restricted case where covariates are not included in the model.

For simplicity, in each design we assume that there is one treatment and one control group. We also assume that the designs are balanced. In the first design, entire schools are randomly assigned to two groups (treatment and control). We represent the number of schools within each group (treatment or control) by *m*, the number of classrooms within a school by *p*, and the number of students within each classroom by *n*. Then, the sample size for the treatment and the control groups is $N_t = N_c = mpn$ and the total sample size is $N = N_t + N_c = 2mpn$. In the second design, classrooms within schools are assigned to treatment and control conditions. In this case we represent the total number of schools across groups by *m*, the number of classrooms within each group (treatment or control) within each school by *p*, and the number of students within each classroom by *n*. The sample size for the treatment and the control groups is again $N_t = N_c = mpn$ and the total sample size is $N = N_t + N_c = 2mpn$. In the third design, students within classrooms and schools are assigned to treatment and control conditions. In this case we represent the total number of schools across groups by *m*, the number of classrooms within each school by *p*, and the number of students within each group (treatment or control) within each classroom by *n*. The sample size for the treatment and the control groups is again $N_t = N_c = mpn$ and the total sample size is $N = N_t + N_c = 2mpn$. In addition, in each design, we consider the case where *q* school-level covariates, *w* classroom-level covariates, and *r* student-level covariates are included in the analysis. This indicates that in design one for example *q* ($0 \leq q < 2(m - 1)$) school-level covariates, *w* ($0 \leq w < 2mp - q - 2$) classroom-level covariates, and *r* ($0 \leq r < 2mpn - q - w - 2$) student-level covariates can be included in the analysis. We assume

that the covariates at the student and at the classroom level are centered around their classroom and school means respectively (group-mean centering). This ensures that predictors explain variation in the outcome *only* at the level at which they are introduced. Since including covariates in the model will adjust the intraclass correlations and the treatment effect we use the subscript A to indicate adjustment due to covariates. We also use the subscript R to indicate residual variation.

Defining the Effect Size Parameter

In power analysis it is common to define an effect size parameter (or standardized mean difference) that is independent of sample size or the scale of the outcome to examine whether the null hypothesis is true or not. The effect size typically used is Cohen's d (δ for a population parameter), and is expressed in standard deviation units, namely $\delta = (\alpha_1 - \alpha_2)/\sigma$ where α_1 and α_2 are the treatment effect parameters from the usual analysis of variance (ANOVA) model (defined precisely below) and σ is the population standard deviation. In cluster randomized experiments, there are several possible choices for the definition of σ since there are multiple standard deviations (e.g., σ_T, ω, τ, and σ_e) which could be used for the standardization. In academic achievement data, typically, most of the variation in achievement is typically between students within classrooms, and hence the total variation in achievement, σ_T^2, is closer to the error variance in achievement at the student level, σ_e^2. The between classroom variation, τ^2, and the between school variation, ω^2, in achievement are typically much smaller than σ_T^2. For example, let's assume that 70 percent of the variation is between students within classrooms, 15 percent between classrooms within schools, and 15 percent between schools (roughly these are estimates from project STAR). Then, the student-level error term variance is $\sigma_e = 0.84\,\sigma_T$, and the

variances of the school and classroom random effects are $\omega = \tau = 0.39\,\sigma_T$. In this study, we assume that achievement is the potential outcome and use the total standard deviation in achievement to standardize the mean difference in achievement between the two groups. The standard deviation σ_T has the advantage that it is the total standard deviation in the population. Hence, we define the effect size parameter as

$$\delta = \frac{\alpha_1 - \alpha_2}{\sigma_T}. \qquad (3)$$

We follow Cohen's (1988) rules of thump about what constitutes a small and a medium effect size. For example, Cohen considered effect sizes of 0.20 (the treatment effect is 1/5 of a standard deviation) and 0.50 (the treatment effect is ½ of a standard deviation) as small and medium respectively. In this study we compute power for small and medium effect sizes. We focus on the power of treatment contrasts, not omnibus (multiple-degree of freedom) treatment effects, because in our experience, multilevel designs are chosen to ensure the power of particular treatment contrasts. Even when several treatments are being compared, there is typically one contrast that is most important and the design is chosen to ensure adequate sensitivity for that contrast.

Design I: Treatment is Assigned at the School Level

In this design, schools are nested within treatment, and classrooms are nested within schools and treatment (see Kirk, 1995, p. 488). In the discussion that follows, we assume that both schools and classrooms are random effects.

A structural model for a student outcome Y_{ijkl}, the l^{th} student in the k^{th} classroom in the j^{th} school in the i^{th} treatment can be described in ANCOVA notation as

$$Y_{ijkl} = \mu + \alpha_{\text{Ai}} + \boldsymbol{\theta}_I^T \mathbf{X}_{ijkl} + \boldsymbol{\theta}_C^T \mathbf{Z}_{ijk} + \boldsymbol{\theta}_S^T \boldsymbol{\Psi}_{ij} + \beta_{\text{A(i)j}} + \gamma_{\text{A(ij)k}} + \varepsilon_{A(\text{ijk})\text{l}}, \quad (4)$$

where μ is the grand mean, α_{Ai} is the (fixed) effect of the i^{th} treatment (i = 1,2), $\boldsymbol{\theta}_I^T = (\theta_{I1}, \ldots, \theta_{Ir})$ is a row vector of r individual-level covariate effects, $\boldsymbol{\theta}_C^T = (\theta_{C1}, \ldots, \theta_{Cw})$ is a row vector of w classroom-level covariate effects, $\boldsymbol{\theta}_S^T = (\theta_{S1}, \ldots, \theta_{Sq})$ is a row vector of q school-level covariate effects, $\mathbf{X}_{ijkl}$ is a column vector of r classroom mean-centered individual-level covariates (e.g., student characteristics) in the k^{th} classroom in the j^{th} *school* in the i^{th} treatment, $\mathbf{Z}_{ijk}$ is a column vector of w school mean-centered classroom-level covariates (e.g., classroom or teacher characteristics) in the j^{th} school in the i^{th} treatment, $\boldsymbol{\Psi}_{ij}$ is a column vector of q school level covariates (e.g., school characteristics) in the i^{th} treatment, and the last three terms are school, classroom, and student random effects respectively. Specifically, $\beta_{A(i)j}$ is the random effect of school j (j = 1,…,m) within treatment i, $\gamma_{A(ij)k}$ is the random effect of classroom k (k = 1,…,p) within school j within treatment i, and $\varepsilon_{A(ijk)l}$ is the error term of student l (l = 1,…,n) within classroom k, within school j, within treatment i. The subscript A indicates adjustment due to covariate effects. The treatment effect may be adjusted by the effects of the covariates. In principle however, assuming randomization works, the treatment effect is orthogonal to the covariates and the error term, and the adjustment should be zero. The classroom and school random effects are adjusted by classroom and school-level covariates respectively and the student error term is adjusted by student-level covariates. We assume that the adjusted student

error term as well as the adjusted classroom and school random effects are normally distributed with a mean of zero and residual variances σ^2_{Re}, τ^2_R, and ω^2_R respectively.

In a multi-level framework the above ANCOVA model can be expressed as

$$Y_{jkl} = u_{0jk} + \mathbf{u}^T_{rjk}\mathbf{X}_{rjkl} + e_{Ajkl},$$

the level two model for the intercept is

$$u_{0jk} = \pi_{00j} + \boldsymbol{\pi}^T_{0wj}\mathbf{Z}_{wjk} + \zeta_{A0jk},$$

and the level three model for the intercept is

$$\pi_{00j} = \delta_{000} + \delta_{A001}TREATMENT + \boldsymbol{\delta}^T_{00q}\boldsymbol{\Psi}_{qj} + \xi_{A00j}$$

where $TREATMENT_i$ is a dummy variable (treatment is 1, otherwise zero), ζ_{A0jk} is the classroom random effect adjusted by the effects of the classroom predictors $\mathbf{Z}$, ξ_{A00k} is the school random effect adjusted by the effects of the school predictors $\boldsymbol{\Psi}$, and e_{jkl} is a student error term adjusted by the effects of the student predictors $\mathbf{X}$. The student level covariates are treated as fixed, namely $\mathbf{u}_{rjk} = \boldsymbol{\pi}_{rwj}, \boldsymbol{\pi}_{rwj} = \boldsymbol{\delta}_{rwq}$. Similarly the classroom level covariates are treated as fixed, namely $\boldsymbol{\pi}_{0wj} = \boldsymbol{\delta}_{0wq}$. The student residual as well as the adjusted classroom, and school random effects are normally distributed with a mean of zero and residual variances σ^2_{Re}, τ^2_R, and ω^2_R respectively. With the appropriate constraints on the ANCOVA model (i.e., setting $\alpha_i = 0$ for the

control group), 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, $\mu = \delta_{000}$, $\alpha = \delta_{A001}$, $\boldsymbol{\theta}_I = \boldsymbol{\delta}_{rwq}$, $\boldsymbol{\theta}_C = \boldsymbol{\delta}_{0wq}$, $\boldsymbol{\theta}_S = \boldsymbol{\delta}_{00q}$, $\beta_A = \xi_A$, $\gamma_A = \zeta_A$, and $\varepsilon_A = e_A$.

Indexes of Clustering: The Adjusted Intraclass Correlation

When covariates are included in the model the four variances of the outcome are defined as $\sigma^2_{Re}, \tau^2_R, \omega^2_R$, and σ^2_{RT}, where $\sigma^2_{RT} = \sigma^2_{Re} + \tau^2_R + \omega^2_R$. The subscript R indicates residual variances (smaller in magnitude because of the adjustment of the covariates). There are two parameters that summarize the associations between these four variances and indicate the clustering effects at the classroom and at the school level: the adjusted intraclass correlation at the classroom level

$$\rho_{Ac} = \frac{\tau^2_R}{\sigma^2_{RT}} \tag{5}$$

and the adjusted intraclass correlation at the school level

$$\rho_{As} = \frac{\omega^2_R}{\sigma^2_{RT}}, \tag{6}$$

where the subscript A indicates that the intraclass correlations are adjusted for the effects of the covariates and are computed using the residual variances (subscript R). The unadjusted intraclass correlations indicate the effects of clustering, while the adjusted intraclass correlations indicate the remaining effects of clustering, net of the effects of covariates.

Estimation

In balanced designs (that is equal sample sizes for the experimental and the control group) the estimate of the treatment effect is the difference in the means between the treatment and the control group namely $\bar{Y}_{A1...} - \bar{Y}_{A2...}$. The standard error of the estimate of the treatment effect is defined as

$$SE\left(\bar{Y}_{A1...} - \bar{Y}_{A2...}\right) = \sqrt{\frac{2}{mpn}}\sqrt{pn\omega_R^2 + n\tau_R^2 + \sigma_{Re}^2} = \sqrt{\frac{2}{mpn}}\sigma_{RT}\sqrt{1 + \left(pn-1\right)\rho_{As} + \left(n-1\right)\rho_{Ac}},$$

where *m* is the number of schools within each condition, *p* is the number of classrooms within each school, *n* is the number of students within each classroom, and all other terms have been defined previously.

Hypothesis Testing

The objective is to examine the statistical significance of the treatment effect net of the possible effects of covariates, which means to test the hypothesis

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

or equivalently

H_0: $\delta_{A001} = 0$.

Suppose that the researcher wishes to test the hypothesis and carries out the usual *t*-test. This involves computing the test statistic

$$t_A = \frac{\sqrt{\frac{mpn}{2}}\left(\bar{Y}_{A1...} - \bar{Y}_{A2...}\right)}{S_A},$$

where S_A is defined by

$$S_A = \sqrt{\frac{\left[\sum_{i=1}^{m}\left(\bar{Y}_{A1i\bullet\bullet} - \bar{Y}_{A1\bullet\bullet\bullet}\right)^2 + \sum_{i=1}^{m}\left(\bar{Y}_{A2i\bullet\bullet} - \bar{Y}_{A2\bullet\bullet\bullet}\right)^2\right] pn}{2m - q - 2}},$$

and $\bar{Y}_{Aij\bullet\bullet}$ is the adjusted mean of the j^{th} school in the i^{th} treatment group, and $\bar{Y}_{Ai\bullet\bullet\bullet}$ is the adjusted mean of the i^{th} treatment group. When the null hypothesis is true, the test statistic t has a Student's t-distribution with $2m$-q-2 degrees of freedom.

Computing Power

When the null hypothesis is false, the test statistic t_A has the non-central t-distribution with $2m$-q-2 degrees of freedom and non-centrality parameter λ_A. The non-centrality parameter is defined as the expected value of the estimate of the treatment effect divided by the square root of the variance of the estimate of the treatment effect namely

$$\lambda_A = \frac{\alpha_{A1} - \alpha_{A2}}{\sigma_{RT}} \sqrt{\frac{mpn}{2}} \sqrt{\frac{1}{1 + (pn - 1)\rho_{As} + (n - 1)\rho_{Ac}}}.$$

Notice that the new adjusted effect size parameter may not be known when computing the power. Typically, the effect size estimate that is expressed in units of the unadjusted standard deviation, that is $\alpha_1 - \alpha_2 / \sigma_T$, is more likely to be known than $\alpha_{A1} - \alpha_{A2} / \sigma_{RT}$. In principle, in randomized trials including covariates should not adjust the treatment effect estimate since the treatment is orthogonal to all covariates. This means that $\alpha_{A1} - \alpha_{A2} = \alpha_1 - \alpha_2$. However that the

standard deviation (the denominator) changes when covariates are included, and is now a residual standard deviation. This means that the effect size is not the same as the unadjusted effect size in equation 3. The covariates will adjust the variances at each level and ultimately the total variance. Hence, we express the non-centrality parameter as a function of the unadjusted effect size and the unadjusted intraclass correlations, namely

$$\lambda_A = \sqrt{\frac{mpn}{2}}\frac{\alpha_1-\alpha_2}{\sigma_T}\frac{\sigma_T}{\sigma_{RT}}\sqrt{\frac{1}{1+(pn-1)\rho_{As}+(n-1)\rho_{Ac}}} = \\ = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{\eta_e+(pn\eta_s-\eta_e)\rho_s+(n\eta_c-\eta_e)\rho_c}}. \quad (7)$$

where

$$\eta_s = \omega_R^2/\omega^2,\ \eta_c = \tau_R^2/\tau^2,\ \eta_e = \sigma_{Re}^2/\sigma_e^2\ . \quad (8)$$

The ηs indicate the proportion of the variances at each level of the hierarchy that is still unexplained (percentage of residual variation). For example when η_e = 0.25, this indicates that the variance at the student level decreased by 75 percent due to the inclusion of covariates such as pre-treatment measures. In the case where no covariates are included the non-centrality parameter reduces to

$$\lambda = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{1+(pn-1)\rho_s+(n-1)\rho_c}}, \quad (9)$$

since $\eta_e = \eta_c = \eta_s = 1$.

When the clustering at the classroom level, ρ_c, is zero, then equations 7 and 9 reduce to the non-centrality parameter used to compute power in two-level designs. For example, when the clustering at the classroom level is zero, $\rho_c = 0$, the two-level design involves students nested within schools where the treatment is assigned at the school level.

The equations 7 and 9 indicate that the non-centrality parameter of the *t*-test is a function of the number of students, classrooms, schools, and the intraclass correlations at the classroom and at the school level. Power is inversely related to the clustering effects, that is the larger the intraclass correlations the smaller the non-centrality parameter, and the power. The factor $1+\left(pn-1\right)\rho_s+\left(n-1\right)\rho_c$ in equation 9 can be thought of as a design effect since it reflects the factor by which the variance of the mean difference from a clustered sample exceeds the variance computed from a simple random sample of the same total sample size. Notice that when the intraclass correlations are zero (no clustering) equation 9 reduces to $\lambda = \sqrt{mpn/2\delta}$ and this holds for all three designs considered in this study.

The power of the one-tailed *t* test at level α is $p_1 = 1 - \mathrm{H}[c(\alpha, 2m\text{-}q\text{-}2), (2m\text{-}q\text{-}2), \lambda_\mathrm{A}]$, where $c(\alpha, \nu)$ is the level α one-tailed critical value of the *t*-distribution with ν degrees of freedom [e.g., $c(0.05,20) = 1.72$], and $\mathrm{H}(x, \nu, \lambda)$ is the cumulative distribution function of the non-central *t*-distribution with ν degrees of freedom and non-centrality parameter λ_A.

The power of the two-tailed test at level α is $p_2 = 1 - \mathrm{H}[c(\alpha/2, 2m\text{-}q\text{-}2), (2m\text{-}q\text{-}2), \lambda_\mathrm{A}] + \mathrm{H}[-c(\alpha/2, 2m\text{-}q\text{-}2), (2m\text{-}q\text{-}2), \lambda_\mathrm{A}]$. In the case of no covariates (that is $q, w, r = 0$) the degrees of freedom for the *t*-test are slightly changed, and so is the non-centrality parameter $\lambda_\mathrm{A} = \lambda$. Alternatively and equivalently, the test for the treatment effect and statistical power can be computed using the *F*-statistic. In this case the *F*-statistic has a non-central *F*-distribution with 1

degree of freedom in the numerator and $2m - q - 2$ degrees of freedom in the denominator and non-centrality parameter λ_A^2 or λ^2.

How Does Power Depend on Student, Classroom, and School Units?

In three-level designs, the number of units at different levels of the hierarchy will have different effects on power. In two-level designs we know that the number of schools have a larger impact on power than the number of students within schools (see Raudenbush & Liu, 2000). Similarly, in three-level designs schools, classrooms, and students will impact power differently. One way to examine this impact is to compute the non-centrality parameter when the number of units at each level of the hierarchy gets infinitely large. The power is a direct function of the non-centrality parameter, and hence, other things being equal, when the non-centrality parameter converges to a real number, the power is smaller than one, and when the non-centrality parameter gets infinitely large the power approaches one. We illustrate the effect of the number of students, classrooms, and schools on statistical power in figures 1 to 3.

Figure 1 illustrates that, as the number of students in each classroom becomes larger, power increases, but then levels off at a value that is smaller than one. This suggests that for a fixed number of schools and classrooms, increasing student sample size beyond a certain point has only a small effect on statistical power. In fact, as the number of students in each class becomes indefinitely large, the value of the non-centrality parameter tends to

$$\lambda_{(max)_{st}} = \sqrt{\frac{mp}{2}}\delta\sqrt{\frac{1}{(p\eta_s\rho_s + \eta_c\rho_c)}}$$

and in the case of no covariates to

$$\lambda_{(max)_{st}} = \sqrt{\frac{mp}{2}}\delta\sqrt{\frac{1}{(p\rho_s + \rho_c)}}.$$

Therefore the maximum power for a design with *m* schools per condition and *p* classrooms per school is the power associated with $\lambda_{(max)_{st}}$, which is less than one. Figure 1 illustrates power computations for one tailed *t*-tests at the 0.05 level for small and medium effect sizes (0.2 and 0.5 SD) with and without covariates as a function of the number of students per classroom holding constant the number of classrooms per school and the number of schools per condition ($p = 2$, $m = 8$).

Figure 2 illustrates that, as the number of classrooms in each school becomes larger, power increases, but then levels off at a value that is smaller than one. This suggests that for a fixed number of schools and students in each classroom, increasing classroom sample size beyond a certain point has a small effect on statistical power. In fact, as the number of classrooms in each school becomes indefinitely large, the value of the non-centrality parameter tends to

$$\lambda_{(max)_c} = \sqrt{\frac{mn}{2}}\delta\sqrt{\frac{1}{n\eta_s\rho_s}}$$

and in the case of no covariates to

$$\lambda_{(max)_c} = \sqrt{\frac{mn}{2}}\delta\sqrt{\frac{1}{n\rho_s}}.$$

Therefore the maximum power for a design with *m* schools per condition and *n* students in each classroom is the power associated with $\lambda_{(max)_c}$, which is less than one. Figure 2 illustrates power computations for one tailed *t*-tests at the 0.05 level for small and medium effect sizes (0.2 and 0.5 SD) with and without covariates as a function of the number of classrooms per school holding constant the number of students per classroom and the number of schools per condition ($n = 10$, $m = 8$).

Figure 3 illustrates that, as the number of schools becomes larger, power increases dramatically, and approaches one. This suggests that for a fixed number of classrooms in each school and students in each classroom, increasing school sample size has a substantial effect on statistical power. In fact, as the number of schools becomes indefinitely large, the value of the non-centrality tends to infinity $\lambda_{(max)_s} \to \infty$. Therefore the maximum power for a design with *p* classrooms per school and *n* students in each classroom is the power associated with $\lambda_{(max)_s}$, which tends to one. This indicates that the number of schools impacts power much more than the number of classrooms and the number of students. Notice that the number of schools influences power via the degrees of freedom of the test statistic as well, and this holds for all three designs. Corresponding results hold for two-level designs (e.g., students are nested within schools, and schools are assigned to treatments). The power again depends much more on the number of schools than the number of students in each school. Raudenbush and Liu (2000) have demonstrated empirically that for a fixed number of schools, the power did not tend to one when the number of students becomes large. In contrast, for a fixed number of students in each school, the power tends to one when the number of schools becomes large. Figure 3 illustrates power computations for one tailed *t*-tests at the 0.05 level for small and medium effect sizes (0.2 and

0.5 SD) with and without covariates as a function of the number of schools holding constant the number of students per classroom and the number of classrooms per school ($n = 10$, $p = 2$).

To select plausible values of clustering at the school level, we follow the findings of a recent study that computed a large amount of intraclass correlations in two-level designs (where students are nested within schools) using recent large-scale studies that surveyed national probability samples of elementary and secondary students in America (Hedges & Hedberg, 2006). The results indicated that most of the school-level intraclass correlations ranged between 0.1 and 0.2. Evidence from two-level analysis (where students are nested within schools) of the National Assessment of Educational Progress (NAEP) trend data, and Project STAR data also points to school-level intraclass correlations between 0.1 and 0.2. Hence, we compute power using two values for the intraclass correlation at the school level: 0.1 and 0.2. In addition, evidence from NAEP main assessment and Project STAR using three-level models (where students are nested within classrooms and classrooms are nested within schools) indicate that the clustering at the classroom level is nearly 2/3 as large as the clustering at the school level. Hence, we compute power using two values for the intraclass correlation at the classroom level: 0.067 and 0.134.

Including covariates also influences power. Specifically, the larger the proportion of the variation explained at the student, classroom, and school level, the larger the value of the non-centrality parameter and in turn the higher the power. Again, following Hedges and Hedberg (2006) we assume that five covariates are included at each level of the hierarchy (e.g, $q = w = r = 5$), and that the covariates explain 50 percent of the variation in achievement at each level. Hence we compute power for $\eta_e = \eta_c = \eta_s = 1$, and for $\eta_e = \eta_c = \eta_s = 0.5$. The power computations illustrated are for one-tailed tests for all three designs discussed. It is plausible to

hypothesize that the treatment effect is positive and that the mean outcome of the treatment group will be larger than the mean outcome of the control group (one-directional research hypothesis). The power computations of two-tailed tests yield smaller values of power.

The following patterns are consistent in figures 1 to 3. First, the power increases as the effect size increases, other things being equal. For example, consider a design with a total sample size of 640, where there are 20 students per classroom, two classrooms per school, and eight schools per treatment condition. When no covariates are included, the clustering effects of schools and classrooms are respectively ρ_s = 0.1 and ρ_c = 0.067, and the effect size is δ = 0.2 SD, the power of a one-tailed test is 0.16, and nearly four times larger, 0.66, when the effect size is δ = 0.5 SD. Second, the power decreases as the intraclass correlations increase, other things being equal. In the previous example, when the effect size is δ = 0.5 SD, the intraclass correlations at the school and classroom level are respectively ρ_s = 0.2 and ρ_c = 0.134, and all else is equal, the power of a one-tailed test is 0.42 (an absolute decrease of 24 percent from 0.66). Third, the number of classrooms has a larger impact on power than the number of students, other things being equal. For example, when the total number of schools is 16, there are eight classrooms per school and five students per classroom, no covariates are included, the intraclass correlations at the school and at the classroom level are respectively ρ_s = 0.2 and ρ_c = 0.134, and the effect size is δ = 0.5 SD, the power of a one-tailed test is 0.49. This indicates a relative increase in power of 17 percent (from 0.42 to 0.49).

Fourth, the number of schools influences power much more than the number of classrooms and the number of students. For example, when the total number of schools is 32, there are two classrooms per school, 10 students per classroom, no covariates are included, the intraclass correlations at the school and at the classroom level are respectively ρ_s = 0.2 and ρ_c =

0.134, and the effect size is δ = 0.5 SD, the power of a one-tailed test is 0.70 (a 67 percent relative increase from 0.42 to 0.70). When the total number of schools is 32, there is one classroom per school, 20 students per classroom, no covariates are included, the intraclass correlations at the school and at the classroom level are respectively ρ_s = 0.2 and ρ_c = 0.134, and the effect size is δ = 0.5 SD, the power of a one-tailed test is 0.62 (nearly a 50 percent relative increase from 0.42 to 0.62). When the total number of schools is 22, there are two classrooms per school, 20 students per classroom, and the clustering effects for classrooms and schools are respectively ρ_s = 0.1 and ρ_c = 0.067, the power of a one-tailed test is just above 0.80, the threshold established by Cohen (1992). Fifth, as one would expect, the power is higher when covariates are included in the model. For example, when the total number of schools is 16, there are two classrooms per school, 20 students per classroom, five covariates are included at each level, 50 percent of the variance is explained by the covariates at each level, the intraclass correlations at the school and at the classroom level are respectively ρ_s = 0.1 and ρ_c = 0.067, and the effect size is δ = 0.5 SD, the power of a one-tailed test is 0.89. This indicates a 35 percent relative increase in power from 0.66 (no covariates) to 0.89. Overall, the larger the proportion of variance explained at each level, the higher the power, other things being equal. Also, the smaller the number of covariates at the school level the larger the power, other things being equal.

Insert Figures 1 to 3 About Here

Design II: Treatment is Assigned at the Classroom Level

In this design, schools and treatments are crossed, and classrooms are nested within treatments and schools (see Kirk, 1995, p. 491). Within each school, classrooms are randomly assigned to a treatment and a control group. Since classrooms within schools are assigned to treatment and control groups, *p* in this design is the number of classrooms in each condition in each school. In the discussion that follows, we assume that schools, classrooms, and the treatment by school interaction are random effects.

The structural model in ANCOVA notation is

$$Y_{ijkl} = \mu + \alpha_{Ai} + \boldsymbol{\theta}_I^T \mathbf{X}_{ijkl} + \boldsymbol{\theta}_C^T \mathbf{Z}_{ijk} + \boldsymbol{\theta}_S^T \boldsymbol{\Psi}_j + \beta_{Aj} + \alpha\beta_{Aij} + \gamma_{A(ij)k} + \varepsilon_{A(ijk)l}, \quad (10)$$

where the last four terms represent school, treatment by school, classroom, and student random effects respectively and all other terms are as defined previously. The new term, $\alpha\beta_{Aij}$, is the treatment by school random effect and follows a normal distribution with a mean of zero and a variance ω_{Rt}^2 (the subscript *t* indicates treatment).

In a multi-level framework the model becomes

$$Y_{jkl} = u_{0jk} + \mathbf{u}_{rjk}^T \mathbf{X}_{rjkl} + e_{Ajkl},$$

the level two model for the intercept is

$$u_{0jk} = \pi_{00j} + \pi_{A01j} Treatment_{jk} + \boldsymbol{\pi}_{0wj}^T \mathbf{Z}_{wjk} + \zeta_{A0jk},$$

and the level three model for the intercept and the treatment effect is

$$\pi_{00j} = \delta_{000} + \boldsymbol{\delta}_{00q}^{T}\boldsymbol{\Psi}_{qj} + \xi_{A00j}$$
$$\pi_{A01j} = \delta_{A010} + \boldsymbol{\delta}_{01q}^{T}\boldsymbol{\Psi}_{qj} + \xi_{A01j}.$$

where ξ_{A01j} is the treatment by school random effect, and all other parameters are as defined in the first design. All covariates are treated as fixed as in the first design.

Estimation

As in the first design, the estimate of the treatment effect is the difference in the means between the treatment and the control group, $\overline{Y}_{A1...} - \overline{Y}_{A2...}$. and the standard error of the estimate of the treatment effect in the second design is defined as

$$SE\left(\overline{Y}_{A1...} - \overline{Y}_{A2...}\right) = \sqrt{\frac{2}{mpn}}\sqrt{pn\omega_{Rt}^2 + n\tau_R^2 + \sigma_{Re}^2} = \sqrt{\frac{2}{mpn}}\sigma_{RT}\sqrt{1 + \left(pn\vartheta_{Rs} - 1\right)\rho_{As} + \left(n-1\right)\rho_{Ac}},$$

where ω_{Rt}^2 is the treatment by school random effect residual variance, and $\vartheta_{Rs} = \omega_{Rt}^2 / \omega_R^2$ is the proportion of the treatment by school random effect residual variance to the total between-school residual variance and $0 \le \vartheta_{Rs} \le 1$, m is the total number of schools, p is the number of classrooms within each condition in each school, n is the number of students within each classroom, and all other terms have been defined previously.

Hypothesis Testing

As previously we test the hypothesis

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

or equivalently

H_0: $\delta_{A010} = 0$.

In this design, an exact test does not exist. However, two modified tests for the treatment effect can be computed (see Kirk, 1995). First, when the null hypothesis is true and the variance of the treatment by school random effect ($\alpha\beta_{A(j)ij}$) is zero, the test statistic *t* that examines the significance of the treatment effect has a Student's *t*-distribution with *2m(p-1)-w* degrees of freedom. Second, when the null hypothesis is true and the variance of the classroom random effect ($\gamma_{A(ij)k}$) is zero, the test statistic *t* that examines the significance of the treatment effect has a Student's *t*-distribution with *m-q-1* degrees of freedom. However, the restrictions in both tests are quite strong. That is, even if the treatment effect is consistent across schools and the variance of the treatment by school interaction is not significant, it does not mean that it is *exactly* zero. Similarly, it is highly unlikely that the variance of the classroom random effect is *exactly* zero. Hence an approximate *t*-test is employed instead within the maximum likelihood estimation framework. This *t*-test involves dividing the estimate of the treatment effect by its standard error. When the null hypothesis is true, this ratio has approximately a *t*-distribution with *m-q-1* degrees of freedom.

Computing Power

When the null hypothesis is false the test statistic that examines the significance of the treatment effect has approximately a non-central *t*-distribution with *m-q-1* degrees of freedom and a non-centrality parameter λ_A. As in design one, we express the non-centrality parameter as a function of the unadjusted intraclass correlations and the unadjusted effect size, namely,

$$\lambda_A = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{\eta_e + (pn\vartheta_{Rs}\eta_s - \eta_e)\rho_s + (n\eta_c - \eta_e)\rho_c}}. \quad (11)$$

The adjusted intraclass correlation at the school level is defined as $\rho_{As} = \omega_{Rs}^2 / \sigma_{RT}^2 = \omega_{Rt}^2 / \vartheta_{Rs}\sigma_{RT}^2$, the adjusted intraclass correlation at the classroom level is as defined in equation 5, and the *ηs* are as defined previously in equation 8. In the case where no covariates are included the non-centrality parameter becomes

$$\lambda = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{1 + (pn\vartheta_s - 1)\rho_s + (n-1)\rho_c}}, \quad (12)$$

where $\vartheta_s = \omega_t^2 / \omega_s^2$ is the proportion of the treatment by school random effect variance to the total between-school variance. The power in design two should typically be larger than the power in design one for two reasons. First, $\vartheta_s, \vartheta_{Rs}$ are typically less than one, that is, the between-school variance of the treatment effect is typically smaller than the total between-school variance. Second, the term *p*n* in design two is always a smaller number than in design one, since the number of classrooms within schools in design one is larger than the number of classrooms per condition in design two and the number of students per classroom remains the same.

The power of the one-tailed *t* test at a significance level *α* is $p_1 = 1 - H[c(\alpha, m\text{-}q\text{-}1), (m\text{-}q\text{-}1), \lambda_A]$, where $c(\alpha, \nu)$ is the level *α* one-tailed critical value of the *t*-distribution with *ν* degrees of freedom [e.g., $c(0.05,20) = 1.72$], and H (x, ν, λ) is the cumulative distribution function of the non-central *t*-distribution with ν degrees of freedom and non-centrality parameter λ_A. The power

of the two-tailed t test at level α is $p_2 = 1 - \mathrm{H}[c(\alpha/2,\ m\text{-}1\text{-}q), (m\text{-}q\text{-}1), \lambda_A] + \mathrm{H}[-c(m\text{-}1\text{-}q), (m\text{-}q\text{-}1), \lambda_A]$. When no covariates are included at any level (that is $q, w, r = 0$) the degrees of freedom for the t-test are slightly changed, and $\lambda_A = \lambda$. Alternatively and equivalently, the test for the treatment effect and statistical power can be computed using the F-statistic. In this case the F-statistic has a non-central F-distribution with 1 degree of freedom in the numerator and $m - q - 1$ degrees of freedom in the denominator and non-centrality parameter λ_A^2 or λ^2.

How Does Power Depend on Student, Classroom, and School Units?

As in design one, we illustrate the effect of the number of students, classrooms, and schools on statistical power in figures 4 to 6. Figure 4 illustrates that, as the number of students in each classroom becomes larger, power increases, but then as in design one it levels off at a value that is less than one. Nonetheless the trajectory is steeper than that in design one. This suggests that for a fixed number of schools and classrooms, increasing student sample size beyond a certain point has a small effect on statistical power. In fact, as the number of students in each class becomes indefinitely large, the value of the non-centrality parameter tends to

$$\lambda_{(max)_{st}} = \sqrt{\frac{mp}{2}}\delta\sqrt{\frac{1}{\left(p\vartheta_{Rs}\eta_s\rho_s + \eta_c\rho_c\right)}}$$

and in the case of no covariates to

$$\lambda_{(max)_{st}} = \sqrt{\frac{mp}{2}}\delta\sqrt{\frac{1}{\left(p\vartheta_{Rs}\rho_s + \rho_c\right)}}\ .$$

Therefore the maximum power for a design with *m* schools and *p* classrooms per condition per school, is the power associated with $\lambda_{(max)_{st}}$, which is smaller than one. Figure 4 illustrates power computations for one tailed *t*-tests at the 0.05 level when the effect size is 0.2 or 0.5 SD as a function of the number of students per classroom holding constant the number of classrooms per condition per school and the number of schools ($p = 1$, $m = 16$).

Figure 5 illustrates that, as the number of classrooms in each school becomes larger, power increases, but then levels off at a value that is less than one. This suggests that for a fixed number of schools and students in each classroom, increasing classroom sample size beyond a certain point has a small effect on statistical power. Nonetheless the trajectory is steeper than that in design one. In fact, as the number of classrooms in each school becomes indefinitely large, the value of the non-centrality parameter tends to

$$\lambda_{(max)_c} = \sqrt{\frac{mn}{2}}\delta\sqrt{\frac{1}{n\vartheta_{Rs}\eta_s\rho_s}}$$

and in the case of no covariates to

$$\lambda_{(max)_c} = \sqrt{\frac{mn}{2}}\delta\sqrt{\frac{1}{n\vartheta_{Rs}\rho_s}}\,.$$

Therefore the maximum power for a design with *m* schools and *n* students in each classroom is the power associated with $\lambda_{(max)_c}$, which is less than one. Figure 5 illustrates power computations for one tailed *t*-tests at the 0.05 level when the effect size is 0.2 or 0.5 SD as a function of the

number of classrooms per condition per school holding constant the number of students per classroom and the number of schools (n = 10, m = 16).

Figure 6 illustrates that, as the number of schools becomes larger, power increases dramatically, and approaches one. This suggests that for a fixed number of classrooms per condition per school and students in each classroom, increasing school sample size has a substantial effect on statistical power. In fact, as the number of schools becomes indefinitely large, the value of the non-centrality tends to infinity $\lambda_{(max)_s} \to \infty$. Therefore the maximum power for a design with p classrooms per condition per school and n students in each classroom is the power associated with $\lambda_{(max)_s}$, which tends to one. This indicates that as in design one the number of schools impacts power more than the number of classrooms and the number of students. Nonetheless in design two the impact of students and classrooms is more evident than in design one. Figure 6 illustrates power computations for one tailed t-tests at the 0.05 level when the effect size is 0.2 or 0.5 SD as a function of the number of schools holding constant the number of students and classrooms per condition (n = 10, p = 1).

The patterns observed in figures 4 to 6 are generally similar to those in figures 1 to 3. Overall, other things being equal, the power in the second design is higher than the power in the first design since the between-school variance of the treatment effect is typically smaller than the overall between-school variance, that is, ϑ_s or ϑ_{Rs} is typically smaller than one. Evidence from project STAR indicates that the between-school variance of the treatment effect is nearly 1/7 of the overall between-school variance (and this is the estimate used in our computations). It appears that in the second design the number of classrooms impacts power much more than the number of students (compared to the first design). For example, assume a design with a total sample size of 600, where there are overall 10 schools (m = 10), one classroom within each

condition in each school (p = 1) and 30 students per classroom (n = 30). When there are no covariates, the intraclass correlations at the school and classroom level are respectively ρ_s = 0.2, and ρ_c = 0.134, and the effect size is δ = 0.5 SD, the power of a one-tailed test is 0.64. However, when the number of classrooms per condition increases to three, the number of student per classroom reduces to 10 (that is the total sample size remains 600) and all other values are the same the power of a one-tailed test is 0.90 (a 40% relative increase from 0.65 to 0.90). Also, as in the first design the number of schools has the largest impact on power.

Insert Figures 4 to 6 About Here

Design III: Treatment is Assigned at the Student Level

In this design classrooms are nested within schools, and the treatment is crossed with classrooms and schools (see Kirk, 1995, p. 489). Specifically, students are randomly assigned to treatment and control conditions within classrooms and schools. In this case, *n* is the number of students within each condition within each classroom. In the discussion that follows, we assume that schools, classrooms, the treatment by school and the treatment by classroom interactions are random effects.

The structural model in ANCOVA notation is

$$Y_{ijkl} = \mu + \alpha_{Ai} + \boldsymbol{\theta}_I^T \mathbf{X}_{ijkl} + \boldsymbol{\theta}_C^T \mathbf{Z}_{ijk} + \boldsymbol{\theta}_S^T \boldsymbol{\Psi}_j + \beta_{Aj} + \alpha\beta_{Aij} + \gamma_{A(j)k} + \alpha\gamma_{Ai(j)k} + \varepsilon_{A(ijk)l}, \qquad (13)$$

where the last five terms represent school, treatment by school, classroom, treatment by classroom, and student random effects respectively and all other terms are defined previously.

The new term $\alpha\gamma_{Ai(j)k}$ is the treatment by classroom random effect which follows a normal distribution with a mean of zero and variance τ^2_{Rt} (the subscript *t* indicates treatment).

In a multi-level framework the model becomes

$$Y_{jkl} = u_{0jk} + u_{A1jk}TREATMENT_{jkl} + \mathbf{u}^T_{rjk}\mathbf{X}_{rjkl} + e_{Ajkl},$$

the level two model for the intercept and the treatment effect is

$$\begin{aligned} u_{0jk} &= \pi_{00j} + \boldsymbol{\pi}^T_{0wj}\mathbf{Z}_{wjk} + \zeta_{A0jk} \\ u_{A1jk} &= \pi_{A10j} + \boldsymbol{\pi}^T_{1wj}\mathbf{Z}_{wjk} + \zeta_{A1jk} \end{aligned},$$

and the level three model for the intercept and the treatment effect is

$$\begin{aligned} \pi_{00j} &= \delta_{000} + \boldsymbol{\delta}^T_{00q}\boldsymbol{\Psi}_{qj} + \xi_{A00j} \\ \pi_{A10j} &= \delta_{A010} + \boldsymbol{\delta}^T_{01q}\boldsymbol{\Psi}_{qj} + \xi_{A01j}. \end{aligned}$$

where ζ_{A1jk} is the treatment by classroom random effect, and all other terms have been defined in designs one and two. As in designs one and two the covariates are treated as fixed.

Estimation

As in the first two designs the estimate of the treatment effect is the difference in the means between the treatment and the control group, $\overline{Y}_{A1...} - \overline{Y}_{A2...}$. The standard error of the estimate of the treatment effect in the third design is defined as

$$SE\left(\bar{Y}_{A1\ldots}-\bar{Y}_{A2\ldots}\right)=\sqrt{\frac{2}{mpn}}\sqrt{pn\omega_{Rt}^2+n\tau_{Rt}^2+\sigma_{Re}^2}=\sqrt{\frac{2}{mpn}}\sigma_{RT}\sqrt{1+\left(pn\theta_{Rs}-1\right)\rho_{As}+\left(n\theta_{Rc}-1\right)\rho_{Ac}}\,,$$

where τ_{Rt}^2 is the treatment by classroom random effect residual variance, $\vartheta_{Rc}=\tau_{Rt}^2/\tau_R^2$ is the proportion of the treatment by classroom random effect residual variance to the total between-classroom residual variance and $0\le\vartheta_{Rc}\le 1$, m is the total number of schools, p is the number of classrooms within each school, n is the number of students within each condition in each classroom, and all other terms have been defined previously.

Hypothesis Testing

The goal is to examine the statistical significance of the treatment effect net of the possible effects of covariates, that is to test the hypothesis

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

or equivalently

H_0: $\delta_{A010}=0$.

The *t*-test involves computing the test statistic

$$t_A=\frac{\sqrt{\frac{mpn}{2}}\left(\bar{Y}_{A1\ldots}-\bar{Y}_{A2\ldots}\right)}{S_A},$$

where S_A is defined by

$$S_A = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\bar{Y}_{Aij1k} - \bar{Y}_{Aij1\bullet}\right)^2 + \sum_{i=1}^{m}\sum_{j=1}^{p}\sum_{k=1}^{n}\left(\bar{Y}_{Aij2k} - \bar{Y}_{Aij2\bullet}\right)^2}{m-q-1}},$$

and $\bar{Y}_{Ajkil}$ is the adjusted y value of the l^{th} student in the i^{th} treatment in the k^{th} classroom in the j^{th} school, and $\bar{Y}_{Ajki\bullet}$ is the adjusted mean of the i^{th} treatment in the k^{th} classroom in the j^{th} school. If the null hypothesis is true, the test statistic t has Student's t-distribution with $m–q-1$ degrees of freedom.

Computing Power

When the null hypothesis is false, the test statistic t that examines the significance of the treatment effect has a non-central t-distribution with $m-q-1$ degrees of freedom and non-centrality parameter λ_A. As in designs one and two, the non-centrality parameter can be expressed as a function of the unadjusted intraclass correlations and the unadjusted effect size, namely,

$$\lambda_A = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{\eta_e + \left(pn\vartheta_{Rs}\eta_s - \eta_e\right)\rho_s + \left(n\vartheta_{Rc}\eta_c - \eta_e\right)\rho_c}}. \qquad (14)$$

The adjusted intraclass correlation at the classroom level is $\rho_{Ac} = \tau_R^2 / \sigma_{RT}^2 = \tau_{Rt}^2 / \vartheta_{Rc}\sigma_{RT}^2$, the adjusted intraclass correlation at the school level is as defined in design two, and the ηs are as defined previously in equation 8. In the case where no covariates are included the non-centrality parameter reduces to

$$\lambda = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{1+\left(pn\vartheta_s-1\right)\rho_s+\left(n\vartheta_c-1\right)\rho_c}}, \qquad (15)$$

where $\vartheta_c = \tau_t^2/\tau^2$ is the proportion of the treatment by classroom random effect variance to the total between-classroom variance, and all other terms are defined previously. Notice that the power in design three should always be larger than the power in designs one and two, so long as ϑ_s, ϑ_c, ϑ_{Rs}, and ϑ_{Rc} are smaller than one. When the clustering at the classroom level, ρ_c, is zero, equations 14 and 15 reduce to the non-centrality parameter used to compute power in two-level designs where students are nested within schools and the treatment is assigned at the student level.

The power of the one-tailed *t* test at level α is $p_1 = 1 - \mathrm{H}[c(\alpha, m\text{-}q\text{-}1), (m\text{-}q\text{-}1), \lambda_\mathrm{A}]$, where $c(\alpha, \nu)$ is the level α one-tailed critical value of the *t*-distribution with ν degrees of freedom [e.g., $c(0.05,20) = 1.72$], and $\mathrm{H}(x, \nu, \lambda)$ is the cumulative distribution function of the non-central *t*-distribution with ν degrees of freedom and non-centrality parameter λ_A. The power of the two-tailed test at level α is $p_2 = 1 - \mathrm{H}[c(\alpha/2, m\text{-}q\text{-}1), (m\text{-}q\text{-}1), \lambda_\mathrm{A}] + \mathrm{H}[-c(\alpha/2, m\text{-}q\text{-}1), (m\text{-}q\text{-}1), \lambda_\mathrm{A}]$. When no covariates are included at any level (that is $q, w, r = 0$) the degrees of freedom for the *t*-test are slightly changed, and $\lambda_\mathrm{A} = \lambda$. Alternatively and equivalently, the test for the treatment effect and statistical power can be computed using the *F*-statistic. In this case the *F*-statistic has a non-central *F*-distribution with 1 degree of freedom in the numerator and $m - q - 1$ degrees of freedom in the denominator and non-centrality parameter λ_A^2 or λ^2.

How Does Power Depend on Student, Classroom, and School Units?

As in designs one and two, as the number of students in each classroom becomes larger, power increases, but then levels off at a value that is less than one. Nonetheless the trajectory is

steeper than those in designs one and two. In fact, as the number of students in each class becomes indefinitely large, the value of the non-centrality parameter tends to

$$\lambda_{(max)_{st}} = \sqrt{\frac{mp}{2}}\delta\sqrt{\frac{1}{\left(p\vartheta_{Rs}\eta_s\rho_s + \vartheta_{Rc}\eta_c\rho_c\right)}}$$

and in the case of no covariates to

$$\lambda_{(max)_{st}} = \sqrt{\frac{mp}{2}}\delta\sqrt{\frac{1}{\left(p\vartheta_{Rs}\rho_s + \vartheta_{Rc}\rho_c\right)}}\,.$$

Therefore the maximum power for a design with *m* schools and *p* classrooms per school is the power associated with $\lambda_{(max)_{st}}$, which is less than one.

Similarly, as the number of classrooms in each school becomes larger, power increases, but then levels off at a value that is less than one. Nonetheless the trajectory is steeper than those in designs one and two. In fact, as the number of classrooms in each school becomes indefinitely large, the value of the non-centrality parameter tends to

$$\lambda_{(max)_{c}} = \sqrt{\frac{mn}{2}}\delta\sqrt{\frac{1}{n\vartheta_{Rs}\eta_s\rho_s}}$$

and in the case of no covariates to

$$\lambda_{(max)_c} = \sqrt{\frac{mn}{2}}\delta\sqrt{\frac{1}{n\vartheta_{Rs}\rho_s}}\ .$$

Therefore the maximum power for a design with *m* schools and *n* students in each condition in each classroom is the power associated with $\lambda_{(max)_c}$, which is less than one. Notice that $\lambda_{(max)_c}$ is identical to that in the second design.

Finally, as the number of schools becomes larger, power increases dramatically, and approaches one. In fact, as the number of schools becomes indefinitely large, the value of the non-centrality tends to infinity $\lambda_{(max)_s} \to \infty$. Therefore the maximum power for a design with *p* classrooms and *n* students in each condition in each classroom is the power associated with $\lambda_{(max)_s}$, which tends to one. Corresponding results are true in two-level designs (e.g., when students are nested within schools, and students within schools are assigned to treatments). The power in this case depends much more on the number of schools than the number of students in each school. Nonetheless, overall, in this design the impact of students and classrooms on power is more evident than in designs one and two.

The power computations in the third design follow overall very similar patterns with the first and the second design. As in designs one and two the number of classrooms within schools has a larger impact on power than the number of students per classroom. As in designs one and two the number of schools has the largest impact on power computations. In addition, the smaller ϑ_s or ϑ_{Rs} and ϑ_c or ϑ_{Rc} are, the higher the power of the test. The power in the third design is always larger that the power in the first design if ϑ_s or ϑ_{Rs}, and ϑ_c or ϑ_{Rc} are smaller than one. In addition, the terms *p*n* and *n* are always larger in design one since the number of students per classroom is always larger than the number of students within a condition. Similarly, if ϑ_s or

ϑ_{Rs} and ϑ_c or ϑ_{Rc} are smaller than one, the power in design three is always higher than the power in design two. In addition, the term *n* is always larger in design two since the number of students per classroom is always larger than the number of students within a condition. The same logic holds for two-level models. Specifically, the power is typically higher in two-level models where the treatment is assigned at the lowest level or unit (e.g., student), other things being equal.

The Importance of Conducting Three-Level Power Analysis

Power computations that ignore a level of clustering will always overestimate the power of tests in multilevel designs (unless the intraclass correlation of the omitted level is *exactly* zero). The most dramatic example is the well-known and large overestimation in statistical power that takes place when one ignores the effects of clustering in two-level designs. Of course, in three-level designs it could be that the effects of clustering on statistical power are mainly due to the first or one level of clustering (e.g., either classrooms or schools), and that other levels of clustering have little additional effect. If so, power computations in three-level designs that ignore a level of clustering might produce only slight overestimates of the actual power. One way to address this question is to compare estimates from power computations in three-level designs to power computations in two-level designs that ignore one level of clustering. The examples below illustrate the degree of overestimation of statistical power that can arise when one of the levels of clustering in the design is ignored.

In this section we examine power computations in the first design, where schools are randomly assigned to treatment conditions. First, we compute power estimates when the school level is omitted (e.g., students are nested within classrooms), and then when the classroom level is omitted (e.g., students are nested within schools). Suppose that we have a three-level design

where randomization occurs at the school level and involves 30 schools, two classrooms per school, and 20 students per classroom (the total sample size is 1200 students). Suppose that no covariates are included at any level, that the effect size is δ = 0.5 SD, and that the intraclass correlations at the classroom and at the school level are ρ_c = 0.2 and ρ_s = 0.2 respectively.

First consider the consequences for statistical power if we ignore clustering at the classroom level. If we were to ignore the classroom level in this design and consider it as if it were a two-level design, we would still have 30 schools randomly assigned to treatment conditions with 40 students per school (for a total sample size of 1200 students). There would still be no covariates, the effect size would be δ = 0.5, and let's assume that the school-level intraclass correlation would still be ρ_S = 0.2. If we computed the power of the one-tailed test in this design assuming only two levels (by ignoring clustering at the classroom level), we would obtain a power of 0.80. However, if we computed the power of the test by correctly recognizing the three levels in the design, the power would be only 0.65. Thus there is a 15 percent absolute difference in power, and a 23 percent relative increase in power from 0.65 to 0.80. This is *not* a trivial difference in power.

Now consider the consequences for statistical power if we ignore clustering at the school level. If we were to ignore the school level in this design and consider it as if it were a two-level design, there would still be 60 classrooms in 30 schools (30 classrooms per condition), and 20 students per classroom (the total sample size is again 1200 students). There would still be no covariates, the effect size would be δ = 0.5 SD, and let's assume that the classroom-level intraclass correlation would still be ρ_s = 0.2. If we computed the power of a one-tailed test in this design assuming only two levels (by ignoring clustering at the school level), we would obtain a statistical power of 0.97. Thus, there is a 32 percent absolute difference in power

between this design and the three-level design described above, and nearly a 50 percent relative increase in power from 0.65 to 0.97. This is a *very substantial* difference in power.

These power computations are intuitive. Specifically, in a three-level design where schools are assigned to treatments and there are no covariates, when the null hypothesis is false the *t*-test has a non-centrality parameter $\lambda_{3L} = \sqrt{mpn/2}\delta\sqrt{1/1+(pn-1)\rho_s+(n-1)\rho_c}$ and *2m – 2* degrees of freedom (where *m* is the number of schools within each condition, each school with *p* classrooms, and each classroom with *n* students). When the school-level clustering is ignored and no covariates are included, the *t*-test has a non-centrality parameter $\lambda_{2LC} = \sqrt{mpn/2}\delta\sqrt{1/1+(n-1)\rho_c}$ and *2mp – 2* degrees of freedom (where *p* is the number of classrooms within each condition in each school and *m* is the total number of schools). When the classroom-level clustering is ignored and no covariates are included, the *t*-test has a non-centrality parameter $\lambda_{2LS} = \sqrt{mpn/2}\delta\sqrt{1/1+(pn-1)\rho_s}$ and *2m – 2* degrees of freedom (where *m* is the number of schools within each condition). When comparing the non-centrality parameters of the two-level designs to the non-centrality parameter of the three-level design, it is evident that $\lambda_{3L} < \lambda_{2LS}$ and $\lambda_{3L} < \lambda_{2LC}$, and hence the power computed when one level of clustering is ignored will always be larger than that of a design including both levels of clustering. The above comparisons assume that the intraclass correlations in two and three-level designs remain unchanged and are larger than zero. In addition, since $\lambda_{2LS} < \lambda_{2LC}$ when $\rho_c \leq \rho_s$ and $p > 1$, and *2mp-2 > 2m-2* since $p > 1$ the power computed when clustering at the school level is ignored will always be larger than that computed when clustering at the classroom level is ignored.

These results indicate that, in designs that involve naturally three levels (e.g., students nested within classrooms, and classrooms nested within schools), ignoring a level of clustering can lead to substantial overestimates of statistical power. The amount of overestimation of power depends on the intraclass correlation structure (and the degrees of freedom). If a researcher chooses to ignore a level of clustering (even though it is present in the design) our computations indicate that, with intraclass correlations that are plausible in educational achievement data, the power computations are less overestimated when the classroom (lowest) level of clustering is ignored. However, in other areas and with different kinds of outcome data, it is not obvious which omitted level of clustering will produce more optimistic power estimates.

Conclusion

Three-level designs are increasingly common in educational research. Experiments that involve multiple schools with multiple classrooms in each school are inherently three-level designs, whether or not the investigators choose to treat them as such. The appropriate power computations of three-level data structures need to include clustering effects at all levels. Similarly, the appropriate analyses of three-level data need to take into account this multi-level structure, because otherwise the standard errors of estimates and statistical tests of such analyses are incorrect. Specifically, the standard errors of treatment effect estimates ignoring clustering are typically smaller, which translates to higher values of *t*-tests and higher probabilities of finding a significant effect. The correct analyses would take into account clustering at all levels of the hierarchy (e.g., the classroom and school level). The present study provided methods for computing power of tests for treatment effects in various three-level designs where clustering occurs at two levels (e.g., classrooms and schools).

Several interesting findings emerged from this study. First, other things being equal, the number of classrooms impacts power more than the number of students, and the number of schools influences power more than the number of classrooms and the number of students and this holds for all three designs. For example, in a design with a total sample size of 640, where schools are randomly assigned to treatments (and students and classrooms are nested within schools) assume that there are no covariates, that the school and classroom level intraclass correlations are respectively ρ_s = 0.1 and ρ_c = 0.067, and that the effect size is δ = 0.5 SD. If the total sample size of 640 is obtained by allocating n = 20 students per classroom, p = 2 classrooms per school, and m = 8 schools per treatment group, the power of a one-tailed test is 0.66. If the total sample size of 640 is obtained by doubling the number of classrooms and halving the number of students per classroom so that p = 4 and n = 10 and everything else is unchanged the power is 0.71 (an eight percent relative increase). However, if the total sample size of 640 is obtained by doubling the number of schools and halving the number of students per class so that m = 16, p = 2, and n = 10 and everything else is unchanged the power of a one-tailed test is 0.91 (an additional relative increase of 28 percent). Thus, if we keep the total sample size constant, doubling the number of schools results in a 38 percent relative increase in power (from 0.66 to 0.88), whereas doubling the number of classrooms only resulted in an eight percent relative increase (from 0.66 to 0.71). This indicates clearly that even though the number of classrooms matters in computing power, the number of schools overwhelms the computation of power in three-level designs. In addition, the number of schools impacts power via the degrees of freedom of the *t*-test. In designs that assign treatments to classrooms or individual students within classrooms (that is designs two and three) however, the difference in the impact on power between the number of schools and the number of classrooms per school is smaller.

Second, power is typically higher in three-level designs that assign treatments at lower levels or units (e.g., classrooms or students), but the gain in power will depend on the magnitude of the variance of the treatment effect across classrooms and schools. For example, in a design that assigns classrooms to treatments the smaller the variance of the treatment effect across schools the higher the power. Similarly, in a design that assigns treatments to students within classrooms the smaller the variances of the treatment effects across classrooms and schools the higher the power. In addition, useful covariates increase power dramatically.

Third, power computations that ignore one level of clustering in the design will always overestimate the power of a three-level design (unless the intraclass correlation of the omitted level is *exactly* zero). Moreover our computations indicated that the degree of overestimation of statistical power may be substantial. When clustering occurs naturally at two levels (e.g., classrooms and schools), three-level power computations are the most appropriate and accurate. Avoiding bias due to omitting a level of clustering in computations of statistical power is particularly important, since there are often other reasons to think that power computations in field experiments are likely to be optimistic (Boruch & Gomez, 1977).

The methods provided here apply to both experimental designs and any non-experimental studies that involve nesting and estimate either the association between a predictor and an outcome or group differences in the outcome. The logic of power computations remains the same and one can compute the power of a test that examines an association or a group difference of interest using the results presented in this study.

Although the present study provides methods for computing power in experimental designs with three levels of nesting, additional methodological work concerning power in these designs is needed. First, even though this study provided methods for computing power in three

level designs, it did not account for the costs involved in designing nested studies. It would be useful for future work to provide estimates of power within an optimal design framework and extend previous work (see e.g., Hedrick & Zumbo, 2005; Raudenbush, 1997; Raudenbush & Liu, 2000). Second, the computation of power for tests that examine the consistency of the treatment effect across classrooms and schools within a three-level design would also be useful. Finally, extensions of the methods presented here to longitudinal studies or growth curve modeling would be valuable.

References

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

Boruch, R. E, & Gomez, H. (1977). Sensitivity, bias, and theory in impact evaluations. *Professional Psychology, 8*, 411-434.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects. A three-level hierarchical linear model. *American Journal of Education*, *97*, 65-108.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159.

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

Guilliford, 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., & Hedberg, E. (2006). *Intraclass correlation values for planning group randomized trials in Education*. Manuscript submitted for publication.

Hedrick, T. C., & Zumbo, B. D. (2005). On optimizing multi-level designs: Power under budget constraints. *Australian and New Zealand Journal of Statistics, 47*, 219-229.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London: Arnold.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.

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

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.

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

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum.

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., 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.

Nye, B, Hedges, V. E., & Konstantopoulos, S. (2000). The effects of small classes on academic achievement: The results of the Tennessee class size experiment. *American Educational Research Journal*, *37*, 123-151.

Nye, B, Konstantopoulos, S., & Hedges, V. E. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*, 237-257.

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

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

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

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

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

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis*. London: Sage.

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

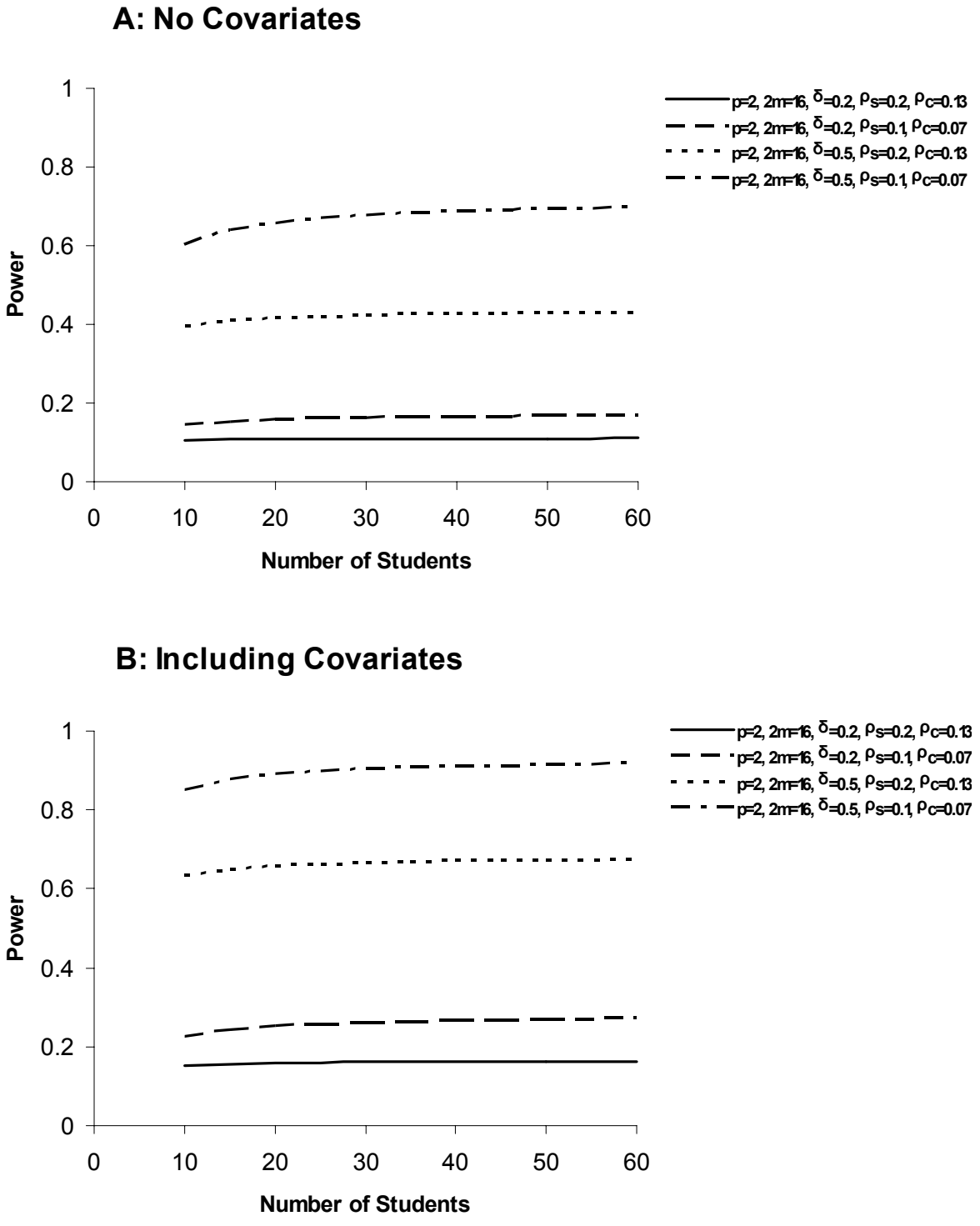

Figure 1. The power of the treatment effect in design one as a function of the number of students within classrooms holding constant the number of classrooms per school and the number of schools per condition. A: No covariates at any level. B: Including five covariates at each level (50% reduction in the variances at each level).

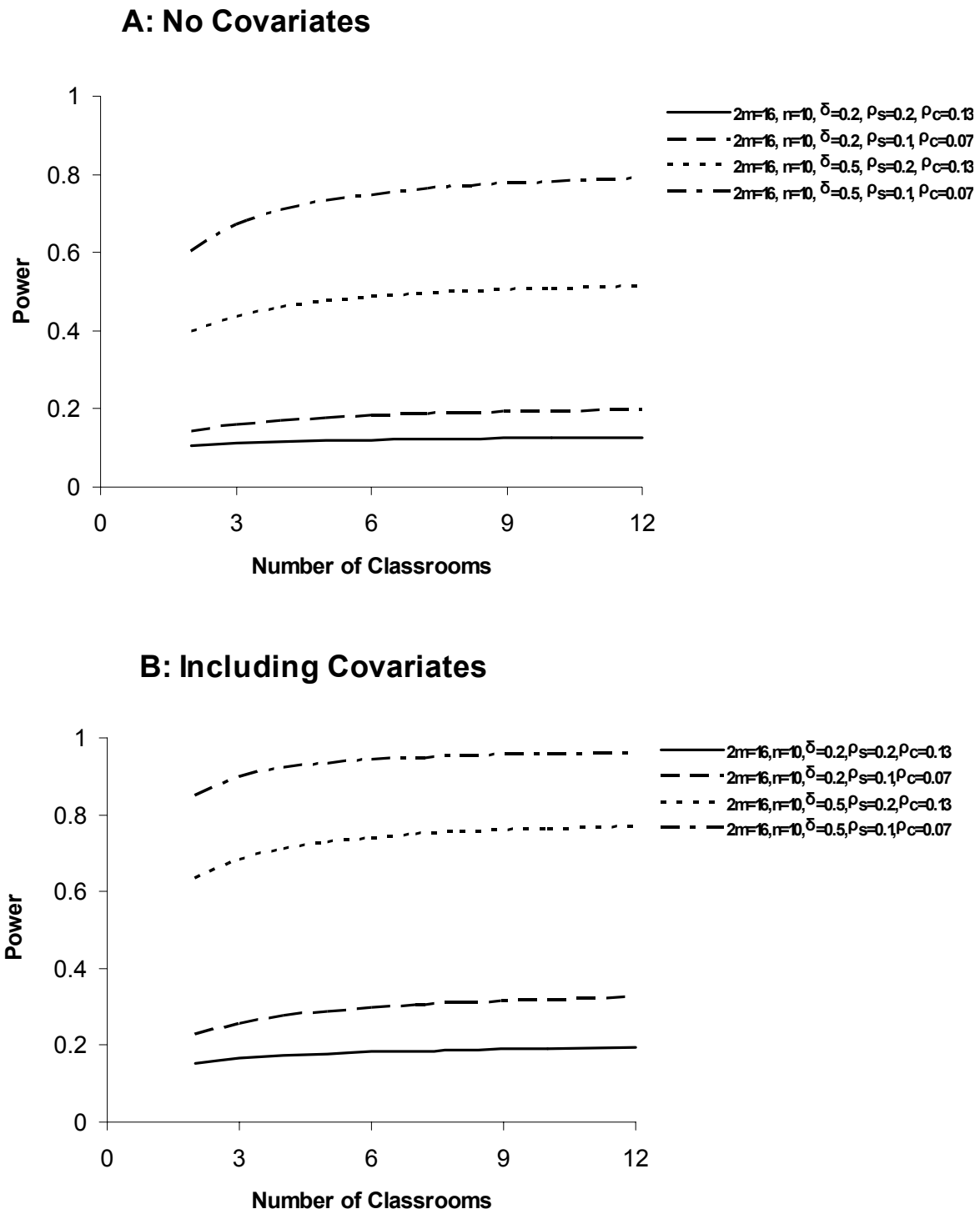

Figure 2. The power of the treatment effect in design one as a function of the number of classrooms within schools holding constant the number of students per classroom, and the number of schools per condition. A: No covariates at any level. B: Including five covariates at each level (50% reduction in the variances at each level).

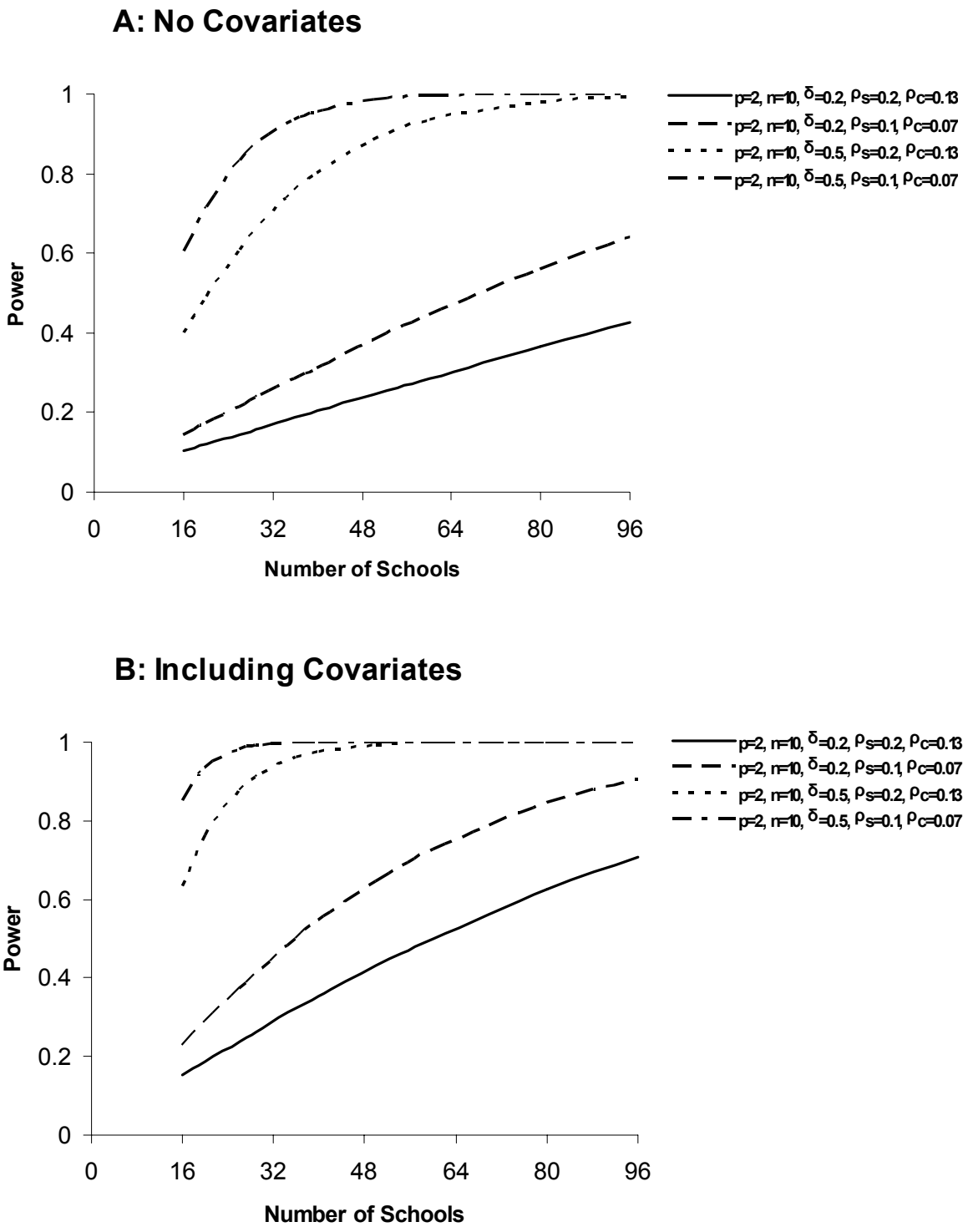

Figure 3. The power of the treatment effect in design one as a function of the number of schools holding constant the number of students per classroom and the number of classrooms per school. A: No covariates at any level. B: Including five covariates at each level (50% reduction in the variances at each level).

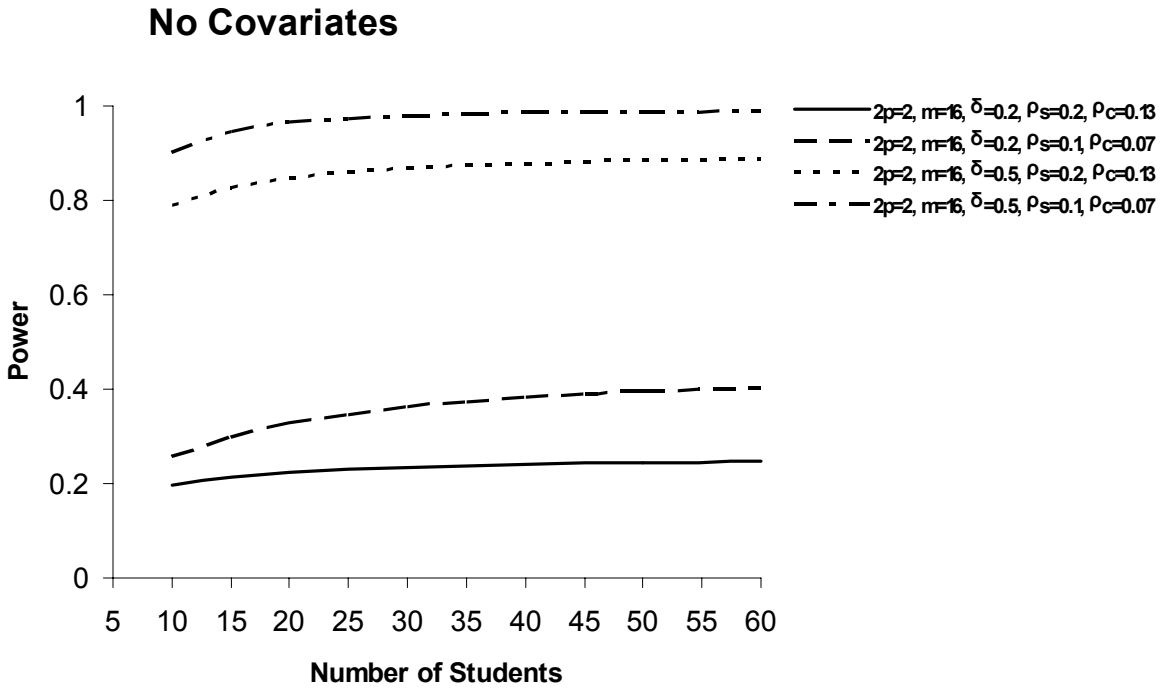

Figure 4. The power of the treatment effect in design two as a function of the number of students per classroom holding constant the number of classrooms per condition per school and the number of schools.

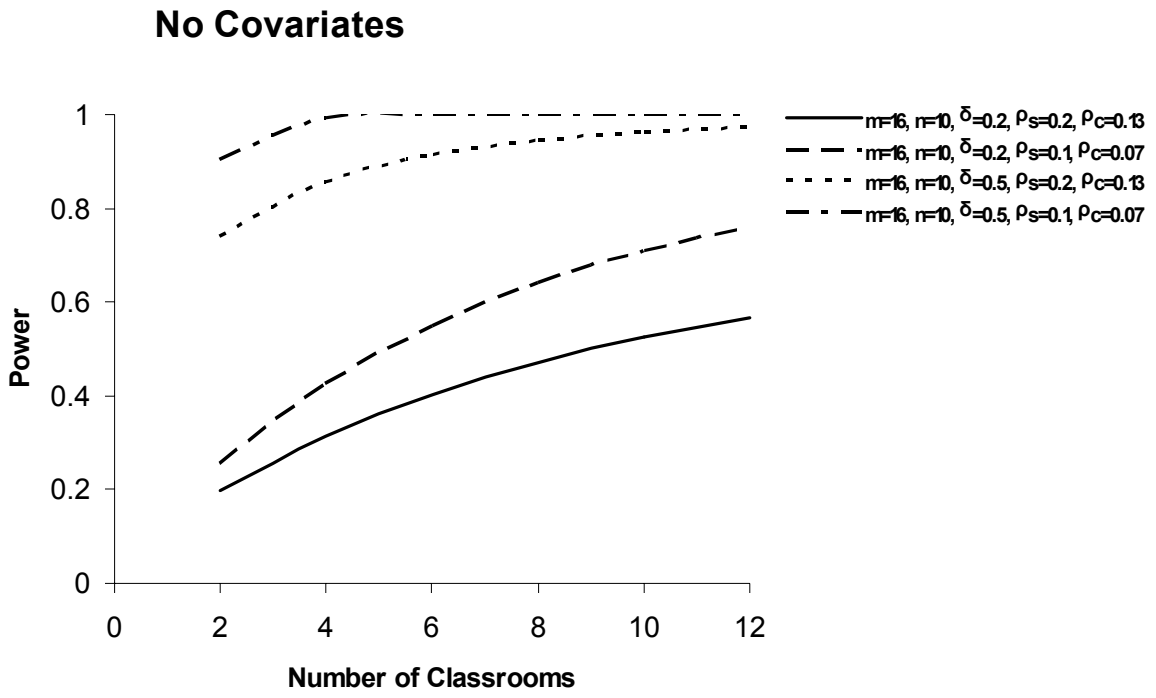

Figure 5. The power of the treatment effect in design two as a function of the number of classrooms within conditions within schools holding constant the number of students per classroom, and the number of schools.

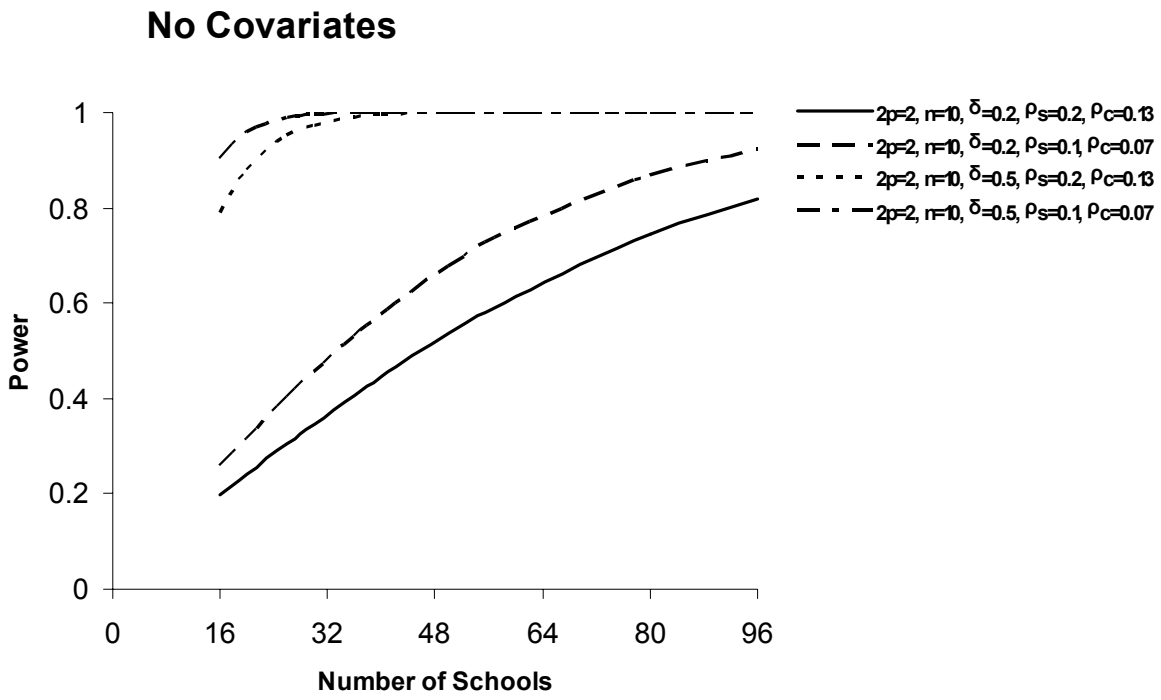

Figure 6. The power of the treatment effect in design two as a function of the number of schools holding constant the number of students per classroom and the number of classrooms per condition per school.