Institute for Policy Research
Northwestern University
Working Paper Series

WP-07-04

Computing Power of Tests for the Variability of Treatment Effects in Designs with Two Levels of Nesting

Spyros Konstantopoulos
Faculty Associate, Institute for Policy Research
Assistant Professor, Human Development, Social Policy, and Learning Sciences
Northwestern University

DRAFT
Please do not quote or distribute without permission.

2040 Sheridan Rd. ◆ Evanston, IL 60208-4100 ◆ Tel: 847-491-3395 Fax: 847-491-9916
www.northwestern.edu/ipr ◆ ipr@northwestern.edu

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). Even though typically the interest in field experiments is in determining the significance of the overall treatment effect, it is equally important to examine the inconsistency of the treatment effect in different contexts, and understand how they vary. This study provides methods for computing power of tests for the variability of treatment effects in different clusters 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, and sample size effects (e.g., number of students, classrooms). The methods can also be applied to quasi-experimental studies that examine the significance of the variation of group differences in an outcome, or associations between predictors and outcomes across different clusters.

Often times populations of interest in the social sciences, education, and psychology have multilevel structures. 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. Experiments that involve populations with nested structures may assign treatments to groups (or clusters) because the treatment is naturally administered to intact groups (such as a curriculum to a classroom or a management system to a firm), or because the assignment of treatments to groups is much easier to implement than assignment to individuals. Other times however, treatments are assigned to individuals (such as different forms of computerized instruction to students within classrooms or schools). In education there has recently been an increased interest in large-scale field experiments to evaluate educational interventions (see, e.g., Mosteller & Boruch, 2002). An example of a field experiment in education is Project STAR, a large-scale randomized experiment, where within each school students and teachers were randomly assigned to classrooms of different sizes (see Nye, Hedges, & Konstantopoulos, 2000).

Methods for power computations of tests for treatment effects in multi-level designs have already been developed (Donner, 1984; Hsieh, 1988; Konstantopoulos, 2006; Murray, 1998; Murray, Van Horn, Hawkins, & Arthur, 2006; Raudenbush, 1997; Raudenbush & Liu, 2000, 2001). For example, a recent study by Murray et al. (2006) provided ways for analyzing data with complicated nested structures within the ANCOVA framework and discussed power computations of tests for treatment effects. Although much of the interest in experiments is in determining the significance of an overall treatment effect, it is also important to determine the inconsistency of treatment effects in different contexts, and understand how they vary. That is, to justify the effects of an intervention, one needs to evaluate them in different contexts to

understand how context shapes both the nature of the intervention, as implemented, and the effects expected. This is an important part of the design and the implementation of interventions, since program developers try to ensure that intervention effects are relatively consistent across different contexts. Ideally, a researcher would aspire to an overall positive average treatment effect and a small variation of the treatment effect across clusters. Large variation in the treatment effect across clusters would indicate large differences in treatment effectiveness. A researcher would then be able to identify the clusters where the treatment was more successful and determine the cluster-specific factors that contributed to the success. This could eventually help reconsidering the nature of the intervention and its implementation (see Raudenbush & Wilms, 1991; Turpin & Sinacore, 1991).

This suggests evaluating educational interventions in a larger scale to determine whether treatment effects, that may be evident in smaller scale studies, are generalizable to larger scale studies. The notion of generalizability of treatment effects is closely related to the concept of external validity, which is concerned with the degree to which causal relationships hold across different clusters (see Shadish, Cook, & Campbell, 2002). Even though external validity and generalization have typically been expressed in qualitative terms, as Shadish et al. (2002) argue, there is a conceptual similarity between generalizability of treatment effects and interactions between treatments and clusters. Evidence of an interaction between clusters and treatments indicates low external validity. One way to evaluate the generalizability of an intervention is to examine the inconsistency of the treatment effect across various clusters (e.g., classrooms or schools). Specifically, when the treatment is assigned within clusters it is likely that context may interact with the treatment to produce differential treatment effects across clusters. In educational research for example, these interactions can occur between an intervention and classrooms or

schools. Suppose that an intervention randomly assigns students to treatment conditions within schools. Then, one can compute a treatment effect for each school and since schools may differ in leadership, organization, climate, and commitment to the intervention it is plausible that the effectiveness of the treatment will vary across schools. That is, in some schools the treatment may be more beneficial to students than in other schools. This example illustrates possible variation of the treatment effect across clusters in two-level designs where level-1 units (students) are nested within level-2 units (schools) and the treatment is assigned at level 1. In three-level designs however, the treatment can be assigned either at the first level or at the second level. When the treatment is assigned at the first level the treatment effect can vary across level 2 and level 3 units, and when the treatment is assigned at the second level the treatment effect can vary across level 3 units.

Conducting studies to determine the inconsistency or variability of treatment effects across clusters is a very timely issue, since some research programs are dedicated to understand how treatment effects vary across contexts. For example, the Interagency Educational Research Initiative (IERI), funded jointly by the National Science Foundation, the National Institute of Child Health and Human Development, and the Institute of Educational Sciences is a major program of research devoted to the problem of determining which educational interventions produce consistent effects across classrooms and schools. Most IERI research projects involve implementation of interventions in several clusters to generate evidence about the variation of both implementation and outcomes across clusters.

The Sensitivity of the Design

In order to determine the statistical significance of the between-cluster variation of the treatment effect one needs to ensure that the design is sensitive enough to detect this variation, if

it exists. Hence, a critical task in planning experimental studies to evaluate the inconsistency of treatment effects involves making decisions about sample sizes and clustering effects to ensure sufficient statistical power of the test for the inconsistency or variability of the treatment effect across clusters. The power in this case is the probability of detecting a significant inconsistency of the treatment effect across clusters when the inconsistency is true.

To test the variability of treatments effects, designs in which treatments are crossed with context factors (e.g., classrooms, schools) are necessary. In these multi-level designs the inconsistency of treatment effects across clusters are evaluated as interactions between treatments and factors representing context. These context factors and interactions can be treated as random effects with variance components structures. These variance components, which indicate clustering via intraclass correlations, will influence the power of the inconsistency of the treatment effect. In addition, the sample sizes at different levels of nesting will affect power differently. For example, the power of the test used to detect the variability of the treatment across level 2 units in two-level designs depends on the sample size of the level 1 units, and the clustering effect at the second level (see Raudenbush & Liu, 2000). In addition, the number of level 2 units impacts power via the *df* of the *F*-test. Statistical theory for computing the power of the test for the variability of the treatment effect in two-level designs where the treatment effect varies across level 2 units has already been provided (see e.g., Raudenbush & Liu, 2000). Raudenbush and Liu found that the proportion of the variance between clusters (or clustering effect) is directly and positively related to the power of the test for the variability of the treatment effect. In addition, their work indicated that as the number of level 1 (e.g., students) within clusters (e.g., schools) gets infinitely large, the power tends to one.

Nonetheless, 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. In these examples 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, a researcher may choose to ignore one level of clustering, conduct an experimental study following a two-level design 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 both level 2 and level 3. Ignoring the three-level structure will result in an inaccurate estimate of power of the test for the variability of the treatment effect in the design stage as we demonstrate in the results section.

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 and at the third. The power in these designs is, among other things, a function of two different sample sizes: the number of level 1 and level 2 units. The power is also a function of two different intraclass correlations (at levels 2 and 3). This paper provides methods that facilitate the computation of statistical power of tests for the inconsistency of the treatment effect 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. 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 of the variance of the treatment effect in two-level designs involved the central *F*-distribution (see Raudenbush & Liu, 2000). The power is a function of the clustering effect, typically expressed as an intraclass correlation, and the number of observations within clusters. Suppose that in a two-level design the total variance of the outcome in a population with nested structure (e.g., students nested within schools) is σ_T^2. The total variance is decomposed into a between-cluster variance ω^2 and a within-cluster 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 clusters.

The same logic holds for three-level designs. The only difference is that clustering in three-level designs occurs in more than one level. In this case the total variance in the outcome is decomposed into three components: the within level 2 and between level 1 units variance, σ_e^2, the between level 2 and within level 3 units variance, τ^2, and the between level 3 units 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:

$$\rho_2 = \frac{\tau^2}{\sigma_T^2} \tag{1}$$

at level 2

$$\rho_3 = \frac{\omega^2}{\sigma_T^2} \tag{2}$$

at level 3 (and the subscripts 2 and 3 indicate the level of the hierarchy).

When covariates are included in the model the variances are defined as $\sigma^2_{Re}, \tau^2_R, \omega^2_R$, and σ^2_{RT}, where $\sigma^2_{RT} = \sigma^2_{Re} + \tau^2_R + \omega^2_R$ (and R indicates residual variances 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 levels 2 and 3: the adjusted intraclass correlation at the level 2

$$\rho_{A2} = \frac{\tau^2_R}{\sigma^2_{RT}} \tag{3}$$

and the adjusted intraclass correlation at level 3

$$\rho_{A3} = \frac{\omega^2_R}{\sigma^2_{RT}}, \tag{4}$$

(where subscript A indicates adjustment).

The computation of statistical power for the variability of the treatment effect in field experiments 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 two-level designs either level 1 or level 2 units are randomly assigned to one of two or more treatment conditions. When treatment is assigned at level 1 the treatment effect can vary across level 2 units. In three-level designs the nested structure is more complex, since the treatment can be assigned either at level 1, at level 2, or at level 3. When the assignment is either at the first or the second level the treatment effect can vary across level 2 and level 3 units. This study discusses three-level designs where randomization can occur at two levels: the first and the second level. In the first design, the random assignment occurs at level 1 and hence, the treatment effect can vary across level 2 and level 3 units. In the second design, the random assignment occurs at level 2 and hence the treatment effect can vary across level 3 units. In each design we illustrate the general case, which includes covariates at all levels of the hierarchy and then the simplest case, where covariates are not included at any level.

For simplicity, in each design we assume that there is one treatment and one control group and that the designs are balanced. In the first design, level 1 units are randomly assigned to treatment and control conditions. In this case we represent the total number of level 3 units by *m*, the number of level 2 units within each level 3 unit by *p*, and the number of level 1 units within each group (treatment or control) within each level 2 unit by *n*. 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, level 2 units are randomly assigned to treatment and control groups. In this case we also represent the total number of level 3 units by *m*, the number of level 2 units within each group (treatment or control) within each level 3 unit by *p*, and the number of level 1 unit

within each level 2 unit 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 level 3 covariates, w level 2 covariates, and r level 1 covariates are included in the analysis. We assume that the covariates at the first and second levels are centered around their level specific means respectively (that is we assume group-mean centering). This ensures that predictors explain variation in the outcome *only* at the level at which they are introduced. In addition, we assume that the covariates at the first and second levels are fixed.

Design I: Treatment is Assigned at Level 1

In this design level 2 units are nested within level 3 units, and the treatment is crossed with level 2 and level 3 units (see Kirk, 1995, p. 489). The level 1 units are randomly assigned to treatment and control conditions within each level 2 and level 3 units. In this case, n is the number of level 1 units within each condition within each level 2 unit. In the discussion that follows, we assume that the level 2 and level 3 units as well as the treatment by level 2 and the treatment by level 3 units interactions are random effects.

The structural model in ANCOVA notation is

$$Y_{ijkl} = \mu + \alpha_{Ai} + \boldsymbol{\theta}_I^T \mathbf{X}_{ijkl} + \boldsymbol{\theta}_C^T \mathbf{Z}_{ijk} + \boldsymbol{\theta}_S^T \boldsymbol{\Psi}_j + \beta_{Aj} + \alpha\beta_{Aij} + \gamma_{A(j)k} + \alpha\gamma_{Ai(j)k} + \varepsilon_{A(ijk)l}, \qquad (5)$$

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 level 1 covariate effects, $\boldsymbol{\theta}_C^T = (\theta_{C1}, \ldots, \theta_{Cw})$ is a row vector of w level 2 covariate effects, $\boldsymbol{\theta}_S^T = (\theta_{S1}, \ldots, \theta_{Sq})$ is a row vector of q level 3 covariate effects, $\mathbf{X}_{ijkl}$ is a column vector of r level 1 covariates, $\mathbf{Z}_{ijk}$ is a column vector of w level 2 covariates, $\boldsymbol{\Psi}_{ij}$ is a column vector of q level 3 covariates, the last five terms represent level 3, treatment by level 3 units, level 2, treatment by level 2 units, and level 1 random effects respectively. Specifically,

β_{Aj} is the random effect of level 3 unit j (j = 1,…,m), $\alpha\beta_{Aij}$, is the treatment by level 3 units random effect, $\gamma_{A(j)k}$ is the random effect of level 2 unit k (k = 1,…,p) within level 3 unit j, $\alpha\gamma_{Ai(j)k}$ is the treatment by level 2 units random effect within level 3 unit j, and $\varepsilon_{A(ijk)l}$ is the error term of level 1 unit l (l = 1,…,n) within treatment i, within level 2 unit k, within level 3 unit j. We assume that the adjusted level 1 error term as well as the adjusted level 2, level 2 units by treatment, level 3, and level 3 units by treatment random effects are normally distributed with a mean of zero and residual variances σ^2_{Re}, τ^2_R, τ^2_{Rt}, ω^2_R, ω^2_{Rt}, respectively (where 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

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

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

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

where $TREATMENT_i$ is a dummy variable centered at its mean (treatment is 0.5, otherwise - 0.5), ζ_{A0jk} is the level 2 random effect adjusted by the effects of the level 2 covariates $\mathbf{Z}$, ζ_{A1jk} is the treatment by level 2 units random effect adjusted by the effects of the level 2 covariates $\mathbf{Z}$, ξ_{A00j} is the level 3 random effect adjusted by the effects of the level 3 covariates $\boldsymbol{\Psi}$, ξ_{A01j} is the treatment by level 3 units random effect adjusted by the effects of the level 3 covariates $\boldsymbol{\Psi}$, and e_{Ajkl} is a level 1 error term adjusted by the effects of the level 1 covariates $\mathbf{X}$. The level 1 and

level 2 covariates are treated as fixed. The level 1 residual as well as the adjusted level 2, level 2 units by treatment, level 3 unit, and level 3 units by treatment random effects are normally distributed with a mean of zero and residual variances σ^2_{Re}, τ^2_R, τ^2_{Rt}, ω^2_R, ω^2_{Rt}, respectively.

In this design two tests are employed to determine the varaibility of the treatment effect. In particular, since the treatment is assigned at level 1, the treatment effect can vary across level 2 units (within level 3 units) and within level 3 units. The first test will determine the significance of the variance of the treatment effect between level 2 units and the second test will determine the significance of the variance of the treatment effect between level 3 units.

Hypothesis Testing for the Variance of the Treatment Effect between Level 2 Units

The significance of the variance τ^2_{Rt} of the treatment effect across level 2 units can be tested with an *F*-test. In particular, we test the hypothesis

H_0: $\tau^2_{Rt} = 0$,

and compute

$$F = \frac{\sum_{j=1}^{m}\sum_{k=1}^{p}\left(\Delta\bar{Y}_{Ajk\bullet\bullet} - \Delta\bar{Y}_{Aj\bullet\bullet\bullet}\right)^2 / m(p-1) - w}{\left[\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{l=1}^{n}\left(Y_{Ajk1l} - \bar{Y}_{Ajk1\bullet}\right)^2 + \sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{l=1}^{n}\left(Y_{Ajk2l} - \bar{Y}_{Ajk2\bullet}\right)^2\right] / 2mp(n-1) - r}, \qquad (6)$$

where $\Delta\bar{Y}_{Ajk\bullet\bullet} = \bar{Y}_{A1jk\bullet} - \bar{Y}_{A2jk\bullet}$ is the adjusted mean difference in the outcome between the treatment and the control group for level 2 unit *k* within level 3 unit *j*, $\bar{Y}_{A1jk\bullet}$ is the adjusted mean of the outcome in the treatment group for level 2 unit *k* within level 3 unit *j*, $\bar{Y}_{A2jk\bullet}$ is the adjusted mean of the outcome in the control group for level 2 unit *k* within level 3 unit *j*, $\Delta\bar{Y}_{Aj\bullet\bullet\bullet} = \bar{Y}_{A1j\bullet\bullet} - \bar{Y}_{A2j\bullet\bullet}$ is the adjusted mean difference in the outcome between the treatment

and the control group for level 3 unit j, $\bar{Y}_{A1j\bullet\bullet}$ is the adjusted mean of the outcome in the treatment group for level 3 unit j, $\bar{Y}_{A2j\bullet\bullet}$ is the adjusted mean of the outcome in the control group for level 3 unit j, Y_{Ajk1l} is the adjusted outcome for level 1 unit l in condition i within level 2 unit k within level 3 unit j, $\bar{Y}_{Ajki\bullet}$ is the adjusted mean in the outcome of condition i within level 2 unit k within level 3 unit j, m is the total number of level 3 units, p is the number of level 2 units within each level 3 unit, n is the number of level 1 units in condition i within level 2 unit k within level 3 unit j, w is the number of level 2 covariates, and r is the number of level 1 covariates. Under the null hypothesis the test statistic in equation 6 follows an F distribution with $df = (m(p-1)-w, 2mp(n-1)-r)$.

The alternative hypothesis is

H_a: $\tau^2_{Rt} > 0$.

Following Kirk (1995) and Raudenbush and Liu (2000) the ratio of the expectation of the numerator to the expectation of the denominator in equation 6 is

$$\upsilon_A = \frac{n\tau^2_{Rt} + \sigma^2_{Re}}{\sigma^2_{Re}} = 1 + \frac{n\tau^2_{Rt}}{\sigma^2_{Re}}, \tag{7}$$

which indicates that under the null hypothesis $\upsilon_A = 1$. The parameter σ^2_{Re} can be expressed as a function of the adjusted intraclass correlations at level 2 and level 3 namely

$$\upsilon_A = 1 + \frac{n\tau^2_{Rt}}{\sigma^2_{RT}(1-\rho_{A2}-\rho_{A3})}. \tag{8}$$

Suppose that the residual variance of the treatment effect across level 2 units, τ^2_{Rt}, is a proportion of the total residual variance between level 2 units, τ^2_{R2}, namely $\tau^2_{Rt} = \vartheta_{R2}\tau^2_{R2}$ and $0 \le \vartheta_{R2} \le 1$. Then equation 8 becomes

$$\upsilon_A = 1 + \frac{n\vartheta_{R2}\tau_{R2}^2}{\sigma_{RT}^2(1-\rho_{A2}-\rho_{A3})} \tag{9}$$

and in turn equation 9 can be expressed as a function of the unadjusted intraclass correlations

$$\upsilon_A = 1 + \frac{n\vartheta_{R2}\eta_2\rho_2}{\eta_1(1-\rho_2-\rho_3)}, \tag{10}$$

where

$$\eta_2 = \tau_R^2/\tau^2,\ \eta_1 = \sigma_{Re}^2/\sigma_e^2. \tag{11}$$

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_1 = 0.20, this indicates that the variance at level 1 decreased by 80 percent due to the inclusion of covariates.

Computing Power of the Test for the Variance of the Treatment Effect between Level 2 Units

When the null hypothesis is false, the ratio of the *F* test statistic in equation 6 to υ_A in equation 8 follows an *F* distribution with *df = m(p – 1) – w, 2mp(n – 1) – r.* The power of the *F*-test is computed as

$$p_1 = prob\{F > F_a\} = prob\{F/\upsilon_A > F_a/\upsilon_A\}$$

$$= 1 - H\{F_a/\upsilon_A,\ m(p-1)-w,\ 2mp(n-1)-r\},$$

where F_a is the critical value of the *F* distribution for a certain level of significance a, and H is the cumulative distribution function of the *F* distribution. When no covariates are included at any level (that is *q, w, r* = 0) the degrees of freedom are slightly changed and the ratio of the expectations becomes

$$\upsilon = 1 + \frac{n\vartheta_2\rho_2}{1-\rho_2-\rho_3},$$

since $\eta_1 = \eta_2 = 1$, and $\vartheta_2 = \tau_t^2 / \tau^2$ is the proportion of the between-level 2 units variance of the treatment effect to the overall between-level 2 units variance. Notice that when the clustering at level 3 is ignored the power of the test refers to a two-level design where level 1 units are nested within level 2 units.

How Does Power Depend on Level 1, Level 2, and Level 3 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 for example we know that the number of level 1 units in each level 2 unit have a larger impact on power than the number of level 2 units (see Raudenbush & Liu, 2000). Similarly, in three-level designs level 1, level 2, and level 3 units will impact power differently. One way to examine this impact is to compute υ_A (in 10) when the number of units at different levels of the hierarchy gets infinitely large. The power is a direct function of υ_A and hence, other things being equal, when υ_A converges to a real number, the power is smaller than one, and when υ_A gets infinitely large the power approaches one. We illustrate the effect of the number of level 1, level 2, and level 3 units on statistical power in figures 1 to 2. In our computations we assume achievement data and clustering effects at the level 2 and level 3.

Notice that in this test the power is influenced by the term

$$\frac{n\vartheta_{R2}\eta_2\rho_2}{\eta_1(1-\rho_2-\rho_3)}. \qquad (12)$$

Figure 1 illustrates that, as the number of level 1 units in each level 2 unit becomes larger, power increases dramatically and tends to one. In fact, as the number of level 1 units becomes infinitely large, υ_A (or υ) tends to infinity and hence power tends to one. Figure 1 illustrates power computations for the *F*-test without covariates as a function of the number of

level 1 units holding constant the number of level 2 and level 3 units ($p = 4$, $m = 20$). Figure 2 illustrates that, as the number of level 3 units becomes larger power increases considerably and when the number of level 3 units gets infinitely large the power tends to one. In fact, as the number of level 3 units becomes infinitely large power tends to one. Figure 2 illustrates power computations for the *F*-test without covariates as a function of the number of level 3 units holding constant the number of level 1 and level 2 units ($n = 10$, $p = 4$). The number of level 2 units has a similar effect on power as the number of level 3 units. Notice that the number of level 2 and level 3 units influence power only via the *df* of the *F*-test.

To select plausible values of clustering at the third level (school), we use the findings of a recent study that computed a large amount of intraclass correlations in two-level designs (Hedges & Hedberg, 2006). The findings indicated that with educational data most of the level 3 intraclass correlations ranged between 0.1 and 0.2. Evidence from two-level analysis of the National Assessment of Educational Progress (NAEP) trend data, and Project STAR data also points to level 2 (school) intraclass correlations between 0.1 and 0.2. Hence, we compute power using two values for the intraclass correlation at level 3: 0.1 and 0.2. In addition, evidence from NAEP main assessment and Project STAR using three-level models indicate that the clustering at level 2 is nearly 2/3 as large as the clustering at level 3. Hence, we compute power using two values for the intraclass correlation at level 2: 0.07 and 0.14. In addition, evidence from project STAR indicates that the variance of the treatment effect between level 3 units is nearly 15 percent of the overall variance between level 3 units (and this estimate is used in our power computations). We also assume that the variance of the treatment effect between level 2 units is nearly 15 percent of the total variance between level 2 units.

The following patterns are consistent in Figures 1 to 2. First, the number of level 1 units has the largest impact on power and increases power at a faster rate than the number of level 2 and level 3 units. The level 1 units impact power via the *df* of the denominator also, and when the *df* become infinitely large the critical value of *F* tends to a χ^2 / ν_1, where $\nu_1 = m(p - 1) - w$, and in turn power tends to one. The number of level 2 and level 3 units influences power *only* via the *df* of the *F*-statistic. As the *df* of the numerator and the denominator become infinitely large, the critical value of *F* tends to one, and in turn power tends to one. Second, the clustering effects (e.g., intraclass correlations) also impact power with larger clustering effects producing larger power. For example, consider a design with a total sample size of 1600, where there are 20 level 1 units per level 2 unit, four level 2 units per level 3 unit, and 20 level 3 units. When no covariates are included, and the clustering effects at level 3 and level 2 are respectively $\rho_3 = 0.1$ and $\rho_2 = 0.07$, the power is 0.16, but nearly three times larger, 0.44, when the clustering effects are respectively $\rho_3 = 0.2$ and $\rho_2 = 0.14$. Third, the larger the variance of the treatment effect between level 2 units the larger the power. In addition, the proportion of the variance in the outcome explained at level 1 and level 2 also impacts power, but differently. In particular, covariates that explain variation at level 1 influence power positively, and covariates that explain variation of the treatment effect at level 2 affect power inversely.

Insert Figures 1 and 2 About Here

Hypothesis Testing for the Variance of the Treatment Effect between Level 3 Units

The variance ω_{Rt}^2 of the treatment effect across level 3 units can be tested with and *F*-test. In particular, we test the hypothesis

H_0: $\omega_{Rt}^2 = 0$,

and compute

$$F = \frac{\sum_{j=1}^{m}\left(\Delta\bar{Y}_{Aj\cdots} - \Delta\bar{Y}_{A\cdots\cdot}\right)^2 / m - q - 1}{\sum_{j=1}^{m}\sum_{k=1}^{p}\left(\Delta\bar{Y}_{Ajk\cdot\cdot} - \Delta\bar{Y}_{Aj\cdots}\right)^2 / m(p-1) - w}, \tag{13}$$

where $\Delta\bar{Y}_{A\cdots\cdot}$ is the adjusted mean difference in the outcome between the treatment and the control group across all level 3 units, q is the number of covariates at level 3, and all other terms have been defined in equation 6. Under the null hypothesis, the test statistic in equation 13 follows an F distribution with $df = (m - q - 1, m(p - 1) - w)$.

The alternative hypothesis is

H_a: $\omega_{Rt}^2 > 0$.

Following Kirk (1995) and Raudenbush and Liu (2000) the ratio of the expectation of the numerator to the expectation of the denominator in equation 13 is

$$\upsilon_A = \frac{pn\omega_{Rt}^2 + n\tau_{Rt}^2 + \sigma_{Re}^2}{n\tau_{Rt}^2 + \sigma_{Re}^2} = 1 + \frac{pn\omega_{Rt}^2}{n\tau_{Rt}^2 + \sigma_{Re}^2}, \tag{14}$$

which indicates that under the null hypothesis $\upsilon_A = 1$. Equation 14 is eventually expressed as a function of the unadjusted intraclass correlations at the levels 2 and 3 namely

$$\upsilon_A = 1 + \frac{pn\vartheta_{R3}\eta_3\rho_3}{\eta_1 + (n\vartheta_{R2}\eta_2 - \eta_1)\rho_2 - \eta_1\rho_3}, \tag{15}$$

where $\eta_3 = \omega_R^2 / \omega^2$ and $\vartheta_{R3} = \omega_{Rt}^2 / \omega_R^2$ is the proportion of the residual variance of the treatment effect across level 3 units to the total residual variance between level 3 units, and $0 \le \vartheta_{R3} \le 1$.

Computing Power of the Test for the Variance of the Treatment Effect between Level 3 Units

When the null hypothesis is false, the ratio of the *F* test statistic in 13 to υ_A in equation 14 follows an *F* distribution with *df* = (*m - q – 1, m(p – 1) – w).* The power of the *F*-test is computed as

$$p_1 = prob\{F > F_a\} = prob\{ F/\upsilon_A > F_a/\upsilon_A \}$$

$$= 1 - H\{ F_a/\upsilon_A, m - q - 1, m(p-1) - w\},$$

where F_a is the critical value of the *F* distribution for a certain level of significance a, and H is the cumulative distribution function of the *F* distribution. When no covariates are included at any level (that is *q, w, r* = 0) the degrees of freedom are slightly changed and the ratio of the expectations becomes

$$\upsilon = 1 + \frac{pn\vartheta_3\rho_3}{1 + (n\vartheta_2 - 1)\rho_2 - \rho_3},$$

where $\vartheta_3 = \omega_t^2 / \omega^2$ is the proportion of the variance of the treatment effect between level 3 units to the total variance between level 3 units.

How Does Power Depend on Level 1, Level 2, and Level 3 Units?

Notice that in this test the power is influenced by the term

$$\frac{pn\vartheta_{R3}\eta_3\rho_3}{\eta_1 + (n\vartheta_{R2}\eta_2 - \eta_1)\rho_2 - \eta_1\rho_3}. \qquad (16)$$

Figure 3 illustrates that, as the number of level 1 units becomes larger, power increases dramatically and tends to one. In fact, as the number of level 1 units becomes infinitely large, υ_A (or υ) tends to infinity and hence power tends to one. Figure 3 illustrates power computations for the *F*-test without covariates as a function of the number of level 1 units holding constant the number of level 2 and level 3 units (*p* = 4, *m* = 20). Figure 4 illustrates that, as the number of level 2 units becomes larger, power increases dramatically and tends to one. As the number of

level 2 units becomes infinitely large power tends to one. Figure 4 illustrates power computations for the *F*-test without covariates as a function of the number of level 2 units holding constant the number of level1 and level 3 units ($n = 10$, $m = 20$).

The following patterns are consistent in figures 3 to 4. First, the number of level 1 and level 2 units have the largest impact on power. The number of level 2 units impacts power via the *df* of the denominator also and when the *df* become infinitely large the critical value of *F* tends to a χ^2 / ν_1, where $\nu_1 = m - q - 1$, and in turn power tends to one. The number of level 3 units influences power *only* via the *df* of the *F*-statistic and, as the *df* of the numerator and the denominator become infinitely large, the critical value of *F* tends to one, and in turn power tends to one. Second, the clustering effects (e.g., intraclass correlations) also affect power, with larger clustering effects producing larger power. Third, the larger the variance of the treatment effect between level 3 units the larger the power. In addition, the proportion of the variance in the outcome explained at level 1, level 2, and level 3 also impacts power, but differently. In particular, covariates that explain variation at level 1 and level 2 influence power positively, whereas covariates that explain part of the variation of the treatment effect between level 3 units, affect power inversely.

Insert Figures 3 and 4 About Here

Design II: Treatment is Assigned at Level 2

In this design, level 3 units and treatments are crossed, and level 2 units are nested within treatments and level 3 units (see Kirk, 1995, p. 491). Within each level 3 unit, level 2 units are randomly assigned to a treatment and a control group. In this design *p* is the number of level 2

units in each condition. In the discussion that follows, we assume that level 2, level 3 units, and the treatment by level 3 units interaction are random effects.

The structural model in ANCOVA notation is

$$Y_{ijkl} = \mu + \alpha_{Ai} + \boldsymbol{\theta}_I^T \mathbf{X}_{ijkl} + \boldsymbol{\theta}_C^T \mathbf{Z}_{ijk} + \boldsymbol{\theta}_S^T \boldsymbol{\Psi}_j + \beta_{Aj} + \alpha\beta_{Aij} + \gamma_{A(ij)k} + \varepsilon_{A(ijk)l}, \quad (17)$$

where the last four terms represent level 3, treatment by level 3 units, level 2, and level 1 random effects respectively (and all other terms have been defined previously).

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 all parameters are as defined in the first design. All covariates at level 1 and 2 are treated as fixed as in the first design.

Hypothesis Testing for the Variance of the Treatment Effect between Level 3 Units

The significance of the variance ω_{Rt}^2 of the treatment effect across level 3 units can be tested with an *F*-test. In particular, we test the hypothesis

H_0: $\omega_{Rt}^2 = 0$,

and compute

$$F = \frac{\sum_{j=1}^{m}\left(\Delta\bar{Y}_{Aj\ldots} - \Delta\bar{Y}_{A\ldots\ldots}\right)^2 / m - q - 1}{\left[\sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{l=1}^{n}\left(Y_{Ajk1l} - \bar{Y}_{Ajk1\cdot}\right)^2 + \sum_{j=1}^{m}\sum_{k=1}^{p}\sum_{l=1}^{n}\left(Y_{Ajk2l} - \bar{Y}_{Ajk2\cdot}\right)^2\right] / 2mp\left(n-1\right) - r}, \qquad (18)$$

where m is the total number of level 3 units, p is the number of level 2 units within condition i within level 3 unit j, n is the number of level 1 units, and all other terms are defined in equations 6 and 13. Under the null hypothesis the test statistic in equation 18 follows an F distribution with $df = (m - q - 1,\ 2mp(n - 1) - r)$.

The alternative hypothesis is

$\mathrm{H_a}$: $\omega_{Rt}^2 > 0$.

Following Kirk (1995) and Raudenbush and Liu (2000) the ratio of the expectation of the numerator to the expectation of the denominator in equation 18 is

$$\upsilon_A = \frac{pn\omega_{Rt}^2 + \sigma_{Re}^2}{\sigma_{Re}^2} = 1 + \frac{pn\omega_{Rt}^2}{\sigma_{Re}^2}, \qquad (19)$$

which indicates that under the null hypothesis $\upsilon_A = 1$. Equation 19 is eventually expressed as a function of the unadjusted intraclass correlations at levels 2 and 3 namely

$$\upsilon_A = 1 + \frac{pn\vartheta_{R3}\eta_3\rho_3}{\eta_1(1 - \rho_2 - \rho_3)}. \qquad (20)$$

Computing Power of the Test for the Variance of the Treatment Effect between Level 3 Units

When the null hypothesis is false, the ratio of the F test statistic in 18 to υ_A in equation 19 follows an F distribution with $df = (m - q - 1,\ 2mp(n - 1) - r)$. The power of the F-test is computed as

$$p_1 = prob\{F > F_{\mathrm{a}}\} = prob\{F / \upsilon_A > F_{\mathrm{a}} / \upsilon_A\}$$

$$= 1 - \mathrm{H}\{ F_a / \upsilon_A, m - q - 1, 2mp(n - 1) - r \},$$

where F_a is the critical value of the F distribution for a certain level of significance a, and H is the cumulative distribution function of the F distribution. When no covariates are included at any level (that is $q, w, r = 0$) the degrees of freedom are slightly changed and the ratio of the expectations becomes

$$\upsilon = 1 + \frac{pn\vartheta_3\rho_3}{1 - \rho_2 - \rho_3}.$$

Notice that when the clustering at level 2 is ignored the power of the test refers to a two-level design where level 1 units are nested within level 3 units.

How Does Power Depend on Level 1, Level 2, and Level 3 Units?

Notice that in this case the term that impacts power is

$$\frac{pn\vartheta_{R3}\eta_3\rho_3}{\eta_1(1 - \rho_2 - \rho_3)}. \tag{21}$$

Figure 5 illustrates that, as the number of level 1 units becomes larger, power increases dramatically and tends to one. Figure 5 illustrates power computations for the F-test without covariates as a function of the number of level 1 units holding constant the number of level 2 and level 3 units ($p = 4$, $m = 20$). Figure 6 illustrates that, as the number of level 2 units becomes larger, power increases dramatically and tends to one. Figure 6 illustrates power computations for the F-tests without covariates as a function of the number of level 2 units holding constant the number of level 1 and level 3 units ($n = 10$, $m = 20$).

The following patterns are consistent in figures 5 and 6. First, the number of level 1 and level 2 units have the largest impact on power. The number of level 1 and level 2 units impacts power via the *df* of the denominator also and when the *df* become infinitely large the critical value of F tends to a χ^2 / ν_1 where $\nu_1 = m - q - 1$, and in turn power tends to one. The number of

level 3 units influences power *only* via the *df* of the *F*-statistic and as the number of level 3 units (and *df*) become infinitely large power tends to one. Second, the clustering effects (e.g., intraclass correlations) also affect power with larger clustering effects producing larger power. Third, the larger the variance of the treatment effect between level 3 units the larger the power. In addition, the proportion of the variance in the outcome explained at level 1 and at level 3 also impacts power, but differently. In particular, covariates that explain variation at level 1 influence power positively, whereas covariates that explain part of the variation of the treatment effect between level 3 units, affect power inversely.

Insert Figures 5 and 6 About Here

The Importance of Conducting Three-Level Power Analysis

In three-level designs three tests can be constructed for the variability of the treatment effect. When the treatment is assigned at level 1, the first test examines the significance of the variation of the treatment effect across level 2 units and the second test examines the significance of the variation of the treatment effect across level 3 units. When the treatment is assigned at level 2, the third test examines the significance of the variation of the treatment effect across level 3 units. Power analysis for all tests take into account both clustering effects at levels 2 and 3. Power computations that ignore a level of clustering will always produce power estimates that are different than those in three-level designs. If the effects of clustering on statistical power in three-level designs are mainly due to one level of clustering and the other level of clustering has little additional effect, then, power computations that ignore a level of clustering will produce estimates similar to 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 or underestimation of statistical power that can arise when one of the levels of clustering in the design is ignored.

The first test determines the significance of the variance of the treatment effect between level 2 units. In this case as equation 12 indicates, ignoring the clustering at level 3 (that is the third level is omitted) indicates that power becomes smaller and hence, a two-level design under-predicts power. Suppose that we have a three-level design where randomization occurs at level 1 and involves 20 level 3 units, four level 2 units per level 3 unit, and 30 level 1 units per level 2 unit (total sample size is 2400). Suppose that no covariates are included at any level, and that the intraclass correlations at levels 2 and 3 are $\rho_2 = 0.2$ and $\rho_3 = 0.2$ respectively. If we were to ignore the third level and consider a two-level design, we would still have 80 level 2 units with 30 (15 in each condition) level 1 unit each (total sample size of 2400). There would still be no covariates, and let's assume that the level 2 intraclass correlation would still be $\rho_2 = 0.2$. Power computations assuming two levels yield a power of 0.76. Power computations assuming three levels however yield much larger power of 0.90. Thus there is a 14 percent absolute difference in power, and an 18 percent relative increase in power from 0.76 to 0.90. This is a considerable difference in power.

The second test determines the significance of the variance of the treatment effect between level 3 units (when treatment is assigned at level 1). In this case as equation 16 indicates, ignoring the clustering at level 2 (that is level 2 is omitted), suggests that power becomes larger, so long as $n\vartheta_{R2}\eta_2 > \eta_1$ (which is the most likely case) and hence the two-level design over-predicts power. In contrast, if $n\vartheta_{R2}\eta_2 < \eta_1$ the two-level model under-predicts power.

Assume the three-level design described above with a total sample size of 2400 and clustering effects at levels 2 and 3 $\rho_2 = \rho_3 = 0.15$. If we were to ignore level 2 and consider a two-level design, we would still have 20 level 3 units with 120 (60 in each condition) level 1 units each (total sample size of 2400). There would still be no covariates, and let's assume that the level 3 intraclass correlation would still be $\rho_3 = 0.15$. Power computations assuming two levels yield power of 0.80. Power computations assuming three levels however yield smaller power of 0.72. Thus there is an eight percent absolute difference in power, and a 10 percent relative increase in power from 0.72 to 0.80. In this case the two-level design over-predicts power, and this difference in power is *not* trivial.

The third test determines the significance of the variance of the treatment effect between level 3 units (when treatment is assigned at level 2). As equation 21 indicates, ignoring the clustering at level 2 (level 2 is omitted) suggests that the power becomes smaller and hence the two-level design under-predicts power. Assume again the three-level design described above with a total sample size of 2400 and clustering effects at levels 2 and 3 $\rho_2 = \rho_3 = 0.15$. If we were to ignore the level 2 and consider a two-level design, we would still have 20 level 3 units with 120 (60 in each condition) level 1 units each (total sample size of 2400). There would still be no covariates, and let's assume that the level 3 intraclass correlation would still be $\rho_3 = 0.15$. Power computations assuming two levels yield power of 0.88. Power computations assuming three levels however, yield somewhat larger power of 0.93. Thus there is a five percent absolute difference in power, and a six percent relative increase in power from 0.88 to 0.93. This is a smaller difference in power than in the previous cases. Notice that when the level 2 clustering effect is ignored, this test for the treatment effect variance between level 3 units has higher power (0.88) than that in the previous test (the second test produced a power of 0.80) since the *df*

in the denominator are larger in this test and in turn the critical F value is smaller (which means that the power is larger).

These results indicate that, in designs that involve naturally three levels ignoring a level of clustering results in inaccurate estimates of statistical power and the amount of overestimation or underestimation of power depends on the intraclass correlation structure (and the degrees of freedom). If a researcher chooses to ignore a level of clustering our computations indicate that, with intraclass correlations that are plausible in educational achievement data, the power computations ignoring one level of clustering are closer to the power estimates that take into account both levels of clustering when level 2 is ignored. The difference in power estimates is larger when level 3 is ignored.

Conclusion

In three-level designs the treatment can be assigned at level 1 or level 2 and hence the treatment effect can vary among level 2 and level 3 units. The appropriate power computations of tests for the inconsistency of the treatment effect in three-level designs need to include clustering effects at both the second and the third level. The present study provided methods for computing power of tests for the inconsistency or variability of the treatment effect in three-level designs where clustering occurs at the second and the third level.

Several interesting findings emerged from the power analyses: First, the number of level 1 units has an overwhelming impact on power and this holds for both designs and all *F*-tests discussed. Similarly, the number of level 2 units has an overwhelming impact on power for both designs and for the *F*-tests that determine the significance of the variance of the treatment effect between level 3 units. The number of level 3 units also impacts power in all designs and tests, but to a smaller degree. In general, the larger the number of level 1, level 2, and level 3 units, the

larger the power. The clustering effects have a positive impact on power as well. In particular, the larger the clustering effects at level 2 and level 3, the larger the power. Similarly, the larger the variation of the treatment effect at level 2 and level 3, the higher the power.

In addition, covariates at level 1, which explain level 1 variation, increase power in all designs and tests. For the *F*-tests that determine the significance of the variance of the treatment effect between-level 3 units, including covariates at level 3 decreases power since the covariates explain part of the variation of the treatment effect between level 3 units. Including covariates at level 2, when the treatment is assigned at level 1 and the *F*-test determines the significance of the variance of the treatment effect between level 3 units increases power. Finally, including covariates at level 2 when the treatment is assigned at level 1 and the F-test determines the significance of the variance of the treatment effect between level 2 units decreases power, since these covariates explain part of the variation of the treatment effect between level 2 units.

Power computations that ignore one level of clustering in the design will either overestimate or underestimate 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 or underestimation of statistical power is not trivial. When clustering occurs naturally at two levels, three-level power computations are the most appropriate and accurate.

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

References

Donner, A. (1984). Approaches to sample size estimation in the design of clinical trials-a review. *Statistics in Medicine, 3,* 199-214.

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.

Hsieh, F. Y. (1988). Sample size formulae for intervention studies with the cluster as unit of randomization. *Statistics in Medicine, 8,* 1195-1201.

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

Konstantopoulos, S. (2006). *The power of tests for treatment effects in three-level designs*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

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.

Murray, D. M., Van Horn, M. L., Hawkins, J D., & Arthur, M. W. (2006). Analysis strategies for a community trial to reduce adolescent ATOD use: A comparison of random coefficient and ANOVA/ANCOVA models. *Contemporary Clinical Trials, 27,* 188-206.

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.

Raudenbush, S. W., & Wilms, J. D. (1991). *Schools, classrooms, and pupils*. San Diego, CA: Academic Press.

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.

Shadish, W. R., Cook, T. D. & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.

Turpin, R. S., & Sinacore, J. M. (1991). *Multisite evaluations*. San Francisco, CA: Jossey-Bass.Verma, V., & Lee, T. (1996). An analysis of sampling errors for demographic and Health surveys. *International Statistical Review*, *64*, 265-294.

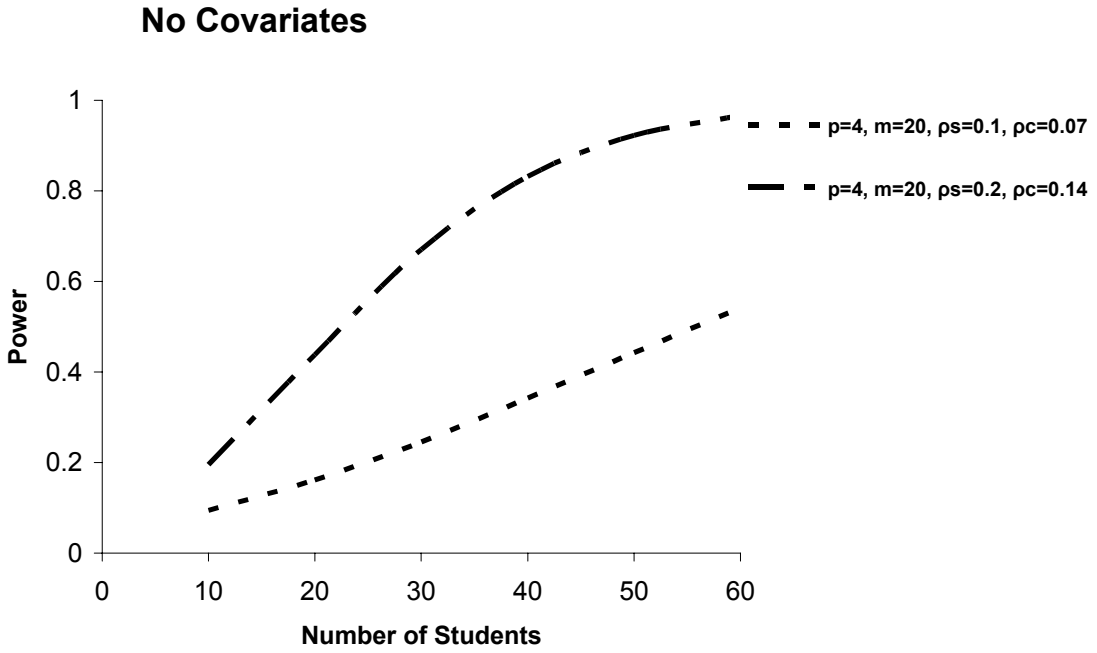

Figure 1. The power of the between-classroom variance of the treatment effect in design one as a function of the number of students within classrooms holding constant the number of classrooms per school and the number of schools.

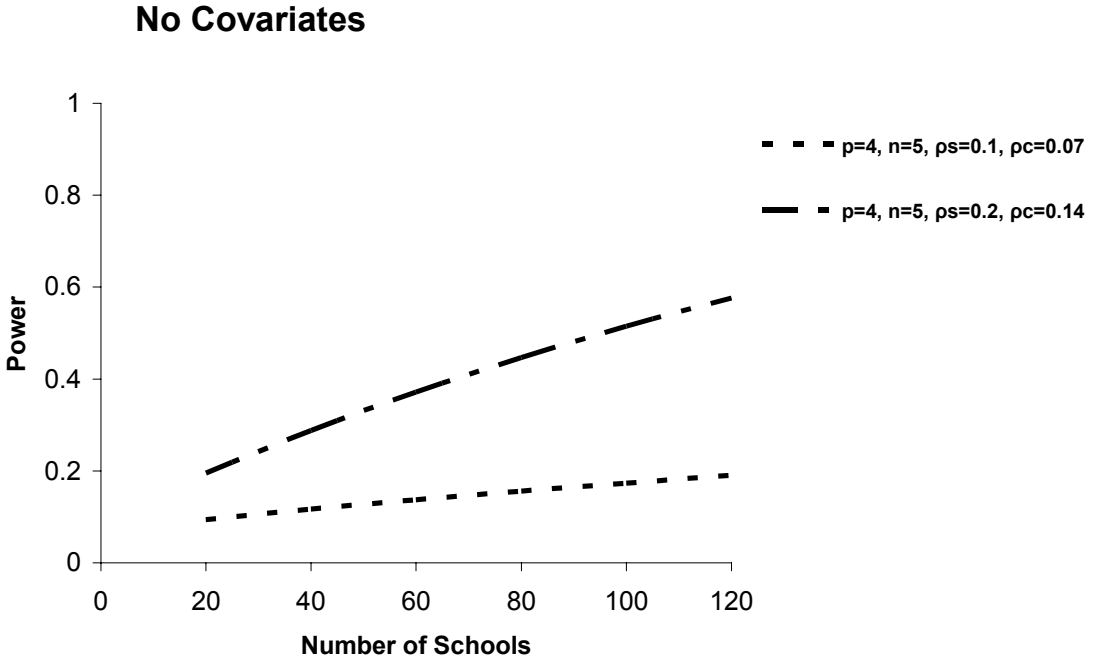

Figure 2. The power of the between-classroom variance of the treatment effect in design one as a function of the number of schools holding constant the number of students within classrooms and the number of classrooms per school.

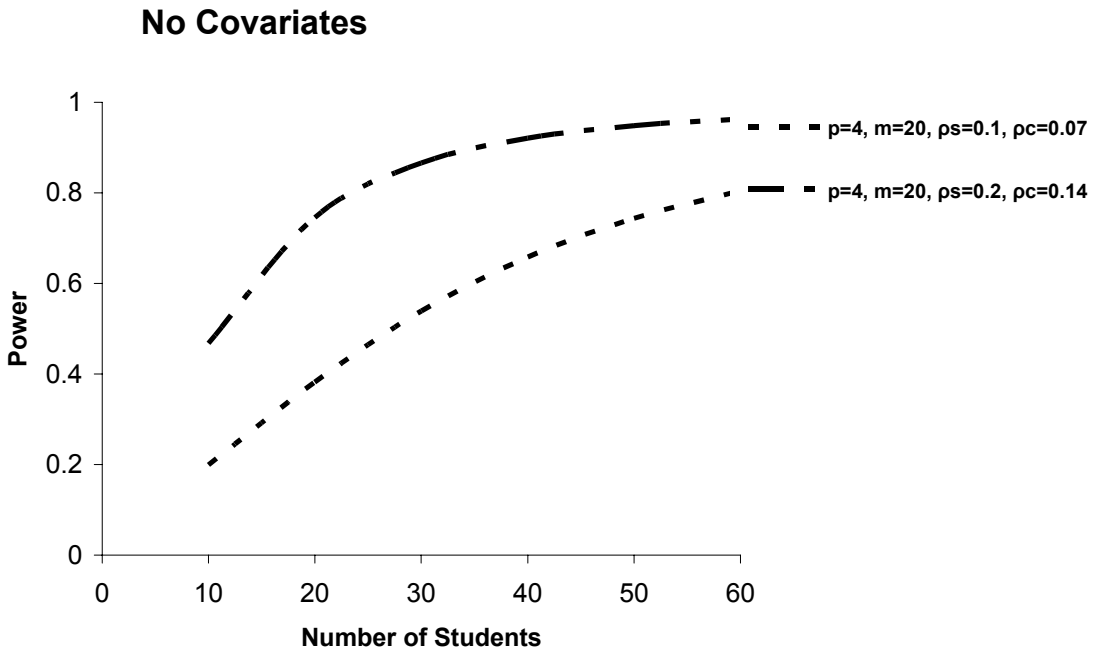

Figure 3. The power of the between-school variance of the treatment effect in design one as a function of the number of students within classrooms holding constant the number of classrooms per school and the number of schools.

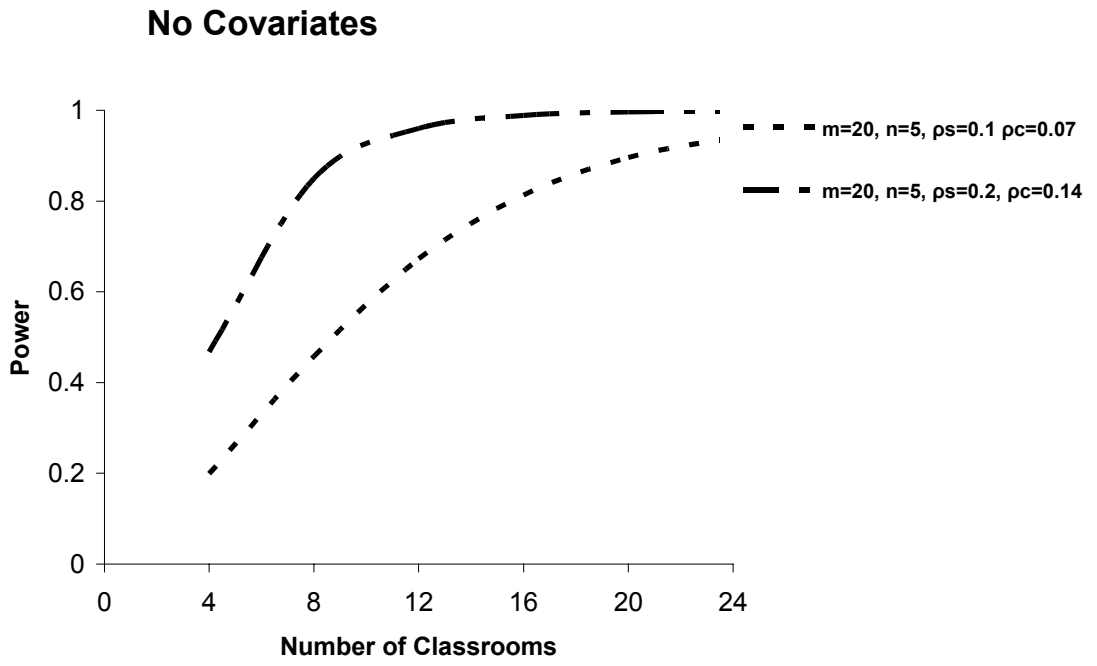

Figure 4. The power of the between-school variance of the treatment effect in design one as a function of the number of classrooms per school holding constant the number of students within classrooms and the number of schools.

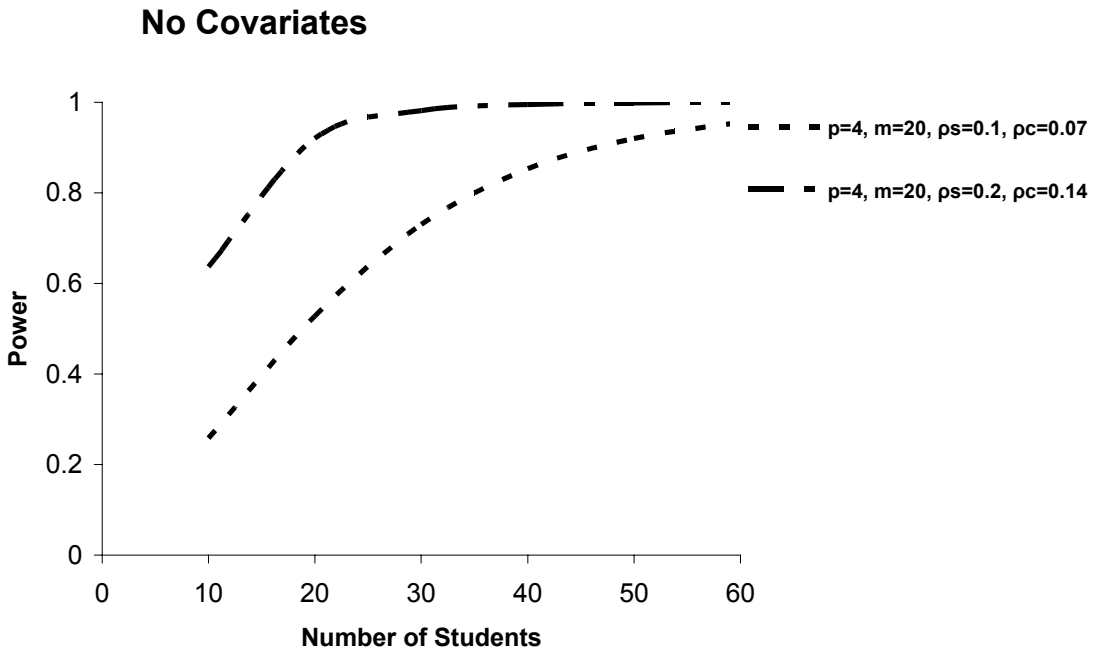

Figure 5. The power of the between-school variance of the treatment effect in design two as a function of the number of students within classrooms holding constant the number of classrooms per school and the number of schools.

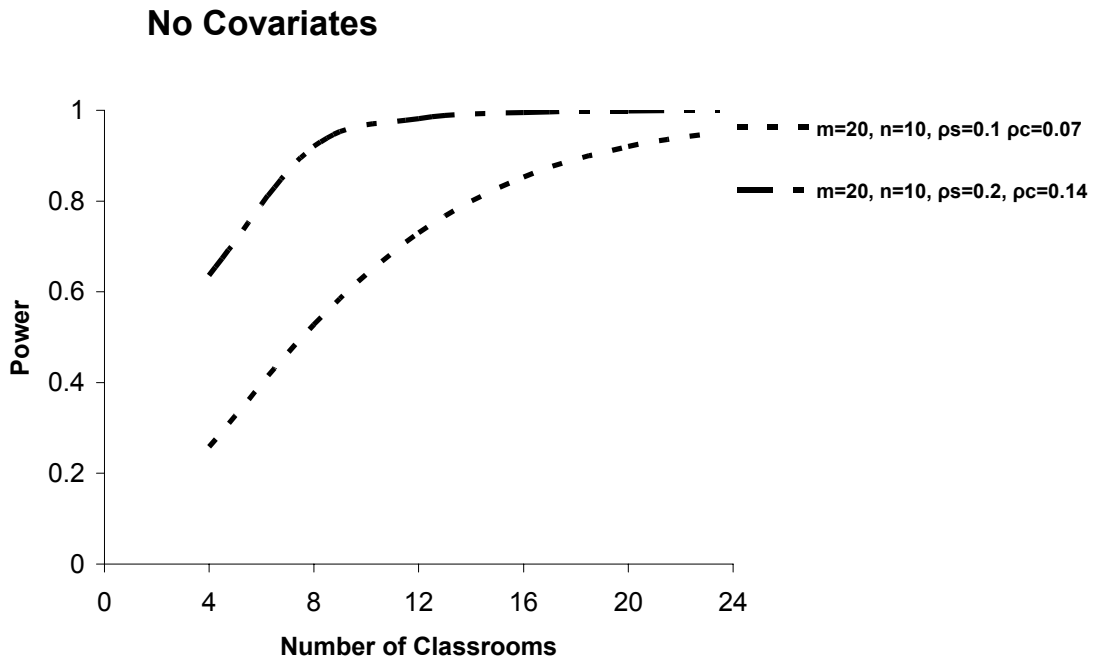

Figure 6. The power of the between-school variance of the treatment effect in design two as a function of the number of classrooms per school holding constant the number of students within classrooms and the number of schools.