DISCUSSION PAPER SERIES

IZA DP No. 3753

Incorporating Cost in Power Analysis for Three-Level Cluster Randomized Designs

Spyros Konstantopoulos

October 2008

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

Incorporating Cost in Power Analysis for Three-Level Cluster Randomized Designs

Spyros Konstantopoulos
Boston College
and IZA

Discussion Paper No. 3753
October 2008

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 3753
October 2008

ABSTRACT

Incorporating Cost in Power Analysis for Three-Level Cluster Randomized Designs

In experimental designs with nested structures entire groups (such as schools) are often assigned to treatment conditions. Key aspects of the design in these cluster randomized experiments include knowledge of the intraclass correlation structure and the sample sizes necessary to achieve adequate power to detect the treatment effect. However, the units at each level of the hierarchy have a cost associated with them and thus researchers need to decide on sample sizes given a certain budget, when designing their studies. This paper provides methods for computing power within an optimal design framework (that incorporates costs of units in all three levels) for three-level cluster randomized balanced designs with two levels of nesting. The optimal sample sizes are a function of the variances at each level and the cost of each unit. Overall, larger effect sizes, smaller intraclass correlations at the second and third level, and lower cost of level-3 and level-2 units result in higher estimates of power.

JEL Classification: I20

Keywords: experimental design, statistical power, optimal sampling

Corresponding author:

Spyros Konstantopoulos
Lynch School of Education
Boston College
Campion Hall, Room 336D
140 Commonwealth Avenue
Chestnut Hill, MA 02467
USA
E-mail: konstans@bc.edu

Many populations of interest in education and the social sciences have multilevel structures. For example, in education students are nested within classrooms, and classrooms are nested within schools. Experiments that involve nested population structures may assign treatment conditions to entire groups. In education, frequently, large-scale randomized experiments assign schools to treatment and control conditions and these designs are often called cluster or group randomized designs (see Bloom, 2005; Donner & Klar, 2000; Murray, 1998).

A critical issue in designing experiments is to ensure that the design has sufficient power to detect the intervention effects that are expected if the researchers' hypotheses were correct. There is an extensive literature on the computation of statistical power (e.g., Cohen, 1988; Lipsey, 1990, Murphy & Myors, 2004). Much of this literature however, involves the computation of power in studies that use simple random samples and thus clustering effects are not included in the power analysis. Software for computing statistical power in single-level designs has also become widely available recently (Borenstein, Rothstein, & Cohen, 2001).

Statistical theory for computing power in two-level designs has also been recently documented and statistical software for two-level balanced designs is currently available (e.g., Hedges & Hedberg, 2007; Murray, 1998; Raudenbush & Liu, 2000, 2001; Raudenbush, Spybrook, Liu, & Congdon, 2006). However, power analysis in nested designs entails challenges. First, nested factors are usually taken to have random effects, and hence, power computations usually involve the variance components structures (typically expressed via intraclass correlations) of these random effects. Second, there is not one sample size, but several sample sizes at each level of the hierarchy that may affect power differently. For example, in educational studies that assign treatments to schools, the power of the test of the treatment effect depends not only on the number of students within a classroom or a school, but on the number of classrooms or schools as well. Methods for power computations of tests of treatment effects in multi-level designs have also been discussed in the health sciences (e.g., Donner, 1984; Hsieh, 1988; Murray, 1998; Murray, Van Horn, Hawkins, & Arthur, 2006). For example, Murray and colleagues (2006) provided ways for analyzing data with complicated nested structures and discussed post-hoc power computations of tests of treatment effects within the ANCOVA framework.

In addition, a more recent study discussed methods for computing power in three-level balanced cluster randomized designs (Konstantopoulos, 2008). Many factors need to be taken into account when designing randomized experiments with a three-level structure. For instance, in three-level cluster randomized designs with two levels of clustering (second and third level) researchers need to take into account the clustering effects at both levels and consider trade-offs that involve sampling level-1, level-2, and level-3 units. In such designs maximizing the number of level-3 units in the sample has a larger impact on the power of the test of the treatment effect than maximizing the number of level1 or level-2 units (see Konstantopoulos, 2008). Also, clustering effects, often expressed via interclass correlations, affect the power estimates inversely.

In addition, the issue of optimal sampling of units at different levels of the hierarchy to maximize power is critical in designing multilevel experiments. Since larger units such as schools affect power much more than smaller units such as classrooms or students a researcher would be inclined to design large-scale experiments with numerous larger units and fewer smaller units. However, maximizing the number of larger units, such as schools, is more expensive than maximizing smaller units, such as classrooms or students. The researcher then faces the challenge of designing a cost-effective study that will optimize the power of the test of the treatment effect given the budget. This requires incorporating cost-related issues when maximizing power in cluster randomized designs (see Raudenbush, 1997). The present study discusses optimal design considerations that incorporate costs of sample sizes at different levels of the hierarchy when designing three-level cluster randomized designs with two levels of nesting. Specifically, I follow Cochran (1977) and Raundenbush (1997) and define cost functions that involve the cost ratios among level-1, level-2, and level-3 units, and then I determine the optimal number of level-1, level-2 (and level-3) units to maximize power, given the costs. Following Raudenbush and Liu (2000) I define optimal design, under specific assumptions, a design that results in the highest estimate of power for the treatment effect.

The paper is structured as follows. First, I define the intraclass correlations in three-level models with two levels of nesting. Second, I present the statistical model and provide an example for computing power in a three-level cluster randomized design. Then, I introduce cost functions that involve level-1, level-2, and level-3 units to maximize power. Finally, I summarize the usefulness of the methods and draw conclusions.

Clustering in Multilevel Designs

Suppose that a researcher samples level-3 units at the first stage, samples level-2 units within level-3 units at the second stage, and then samples level-1 units within level-2 units at the third stage. This is a three-stage cluster sample and the variance of the total population is the sum of the within-level-2 unit between-level-1 unit variance, σ_e^2; the within-level-3 unit between-level-2 unit variance, τ^2; and the between-level-3 unit variance, ω^2 (see Cochran, 1977; Lohr, 1999). That is, the total variance in the outcome is decomposed into three parts and is defined as $\sigma_T^2 = \sigma_e^2 + \tau^2 + \omega^2$. In such three-level designs two intraclass correlations are needed to describe the variance component structure. These are defined as the second level intraclass correlation:

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

and the third level intraclass correlation

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

where the subscripts 2 and 3 indicate the level of hierarchy.

The ANOVA Model

Consider a design where level-3 units are nested within treatment, and level-2 units are nested within level-3 units and treatment (Kirk, 1995), and both level-3 and level-2 units are random effects. A structural model for an outcome Y_{ijkl}, the l^{th} level-1 unit in the k^{th} level-2 unit in the j^{th} level-3 unit in the i^{th} treatment can be described in ACOVA notation as

$$Y_{ijkl} = \mu + \alpha_{\text{i}} + \beta_{\text{(i)j}} + \gamma_{\text{(ij)k}} + \varepsilon_{\text{(ijk)l}}, \tag{3}$$

where μ is the grand mean, α_i is the (fixed) effect of the i^{th} treatment (i = 1,2), and the last three terms represent level-3, level-2, and level-1 random effects respectively. Specifically,

$\beta_{(i)j}$ is the random effect of level-3 unit *j* (*j* = 1,…, *m*) within treatment *i*, $\gamma_{(ij)k}$ is the random effect of level-2 unit *k* (*k* = 1,…, *p*) within level-3 unit *j* within treatment *i*, and $\varepsilon_{(ijk)l}$ is the error term of level-1 unit *l* (*l* = 1,…, *n*) within level-2 unit *k,* within level-3 unit *j*, within treatment *i*. I assume that the level-1, level-2, and level-3 error terms are normally distributed with a mean of zero and residual variances σ_e^2, τ^2, and ω^2 respectively. For simplicity, I assume that there is one treatment and one control group and that the designs are balanced.

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

$$\text{H}_0\text{: } \alpha_1 = \alpha_2 \text{ or } \alpha_1 - \alpha_2 = 0 .$$

The researcher can test this hypothesis by carrying out the usual *t*-test. Following Konstantopoulos (2008) when the null hypothesis is false, the test statistic has a non-central *t*-distribution with *2m-2* degrees of freedom and non-centrality parameter λ (assuming no covariates). 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

$$\lambda = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{1+(n-1)\rho_2+(pn-1)\rho_3}}, \tag{4}$$

where *m* is the number of level-3 units in each condition (treatment or control group), *p* is the number of level-2 units within each level-3 unit, *n* is the number of level-1 units within each level-2 unit, and $\delta = \alpha_1 - \alpha_2 / \sigma_T$, where α_1 and α_2 are the treatment effect parameters from the ANOVA model (defined above) and σ_T is the population standard deviation.

The power of the two-tailed *t*-test at level α is

$$p_2 = 1 - \mathrm{H}\ [c(\alpha/2,\ 2m\text{-}2),\ (2m\text{-}2),\ \lambda_{\mathrm{A}}] + \mathrm{H}\ [\text{-}c(\alpha/2,\ 2m\text{-}2),\ (2m\text{-}2),\ \lambda_{\mathrm{A}}], \qquad (5)$$

where $c(\alpha,\ v)$ is the level α two-tailed critical value of the *t*-distribution with v degrees of freedom [$c(0.05,20) = 1.72$], 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. The test of the treatment effect and statistical power can also be computed using the *F*-statistic that has a non-central *F*-distribution with 1 degree of freedom in the numerator and $2m - 2$ degrees of freedom in the denominator and non-centrality parameter λ^2.

The ANCOVA Model

When covariates are included at each level the ANCOVA model is

$$Y_{ijkl} = \mu + \alpha_{\mathrm{A}i} + \boldsymbol{\theta}_I^T \mathbf{X}_{ijkl} + \boldsymbol{\theta}_C^T \mathbf{Z}_{ijk} + \boldsymbol{\theta}_S^T \boldsymbol{\Psi}_{ij} + \beta_{\mathrm{A(i)j}} + \gamma_{\mathrm{A(ij)k}} + \varepsilon_{A(\mathrm{ijk})\mathrm{l}}, \qquad (6)$$

where $\boldsymbol{\theta}_I^T = (\theta_{I1}, \ldots, \theta_{Ir})$ 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, and $\mathbf{W}_{ij}$ is a column vector of q level-3 covariates, and the last three terms represent residuals at the third, second, and first level respectively. The subscript A indicates adjustment due to covariate effects, that is, the level-2 and level-3 random effects are adjusted by level-2 and level-3 covariates respectively and the level-1 error term is adjusted by level-1 covariates. I assume that the covariates at each level are centered at their means to ensure that covariates explain variation in the outcome *only* at the level at which they are introduced. Note that although in practice covariates could slightly adjust the treatment effect, in principle, due to randomization the treatment effect should be unadjusted. I assume that the adjusted error terms first, second, and third level are normally distributed with a mean of zero and residual variances σ_{Re}^2, τ_R^2, and ω_R^2, respectively.

The objective in this case is to examine the statistical significance of the treatment effect adjusted by covariates, which means to test the hypothesis

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

Note that in this case δ and the intraclass correlations are adjusted. Specifically, the numerator of δ remains unchanged (because of orthogonality between the treatment and the covariates), whilst the denominator changes (because the total variance is now residual variance). The intraclass correlations are also adjusted. The second level intraclass correlation is now defined as

$$\rho_{A2} = \frac{\tau_R^2}{\sigma_{RT}^2}$$

and the third level intraclass correlation is defined as

$$\rho_{A3} = \frac{\omega_R^2}{\sigma_{RT}^2},$$

where subscript A indicates adjustment and subscript R indicates residual variance. When the null hypothesis is false, the test statistic has the non-central *t*-distribution with *2m-q-2* degrees of freedom (where *q* is the number of level-3 covariates) and non-centrality parameter λ_A. Following Konstantopoulos (2008) the non-centrality parameter is defined now as

$$\lambda_A = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{\eta_1 + \left(n\eta_2 - \eta_1\right)\rho_2 + \left(pn\eta_3 - \eta_1\right)\rho_3}}, \qquad (7)$$

where

$$\eta_3 = \omega_R^2 / \omega^2,\ \eta_2 = \tau_R^2 / \tau^2,\ \eta_1 = \sigma_{Re}^2 / \sigma_e^2, \qquad (8)$$

(see also Hedges & Hedberg, 2007; Murray, 1998). The η's indicate the proportion of the variances at each level of the hierarchy that is still unexplained. For example when η_e= 0.50, this indicates that the variance at the student level decreased by 50 percent due to the inclusion of covariates. Alternatively, the η s can be defined as a function of the proportion of variance explained (R^2) at each level, that is, $\eta_s = 1 - R_s^2$, $\eta_c = 1 - R_c^2$, $\eta_e = 1 - R_e^2$. The power of the two-tailed *t*-test at level α is

$$p_2 = 1 - \mathrm{H}\left[c(\alpha/2, 2m\text{-}q\text{-}2), (2m\text{-}q\text{-}2), \lambda_\mathrm{A}\right] + \mathrm{H}\left[-c(\alpha/2, 2m\text{-}q\text{-}2), (2m\text{-}q\text{-}2), \lambda_\mathrm{A}\right]. \qquad (9)$$

The test of the treatment effect and statistical power can also be computed using the F-statistic that 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.

The power is a function of the non-centrality parameter and the degrees of freedom, and larger values of those factors result in higher power. As equations 4 and 7 indicate the non-centrality parameter becomes larger as the level-3 units within a condition and the effect size parameter become larger. The intraclass correlations are inversely related to power, that is, larger clustering effects result in lower power. Level-3 units affect power via the degrees of freedom as well, that is, larger numbers of level-3 units result in higher power.

To illustrate the computation of power consider an example from education. Suppose the effect size is δ = 0.2, there are m = 20 schools per condition with p = 3 classrooms per school and n = 20 students per classroom, and the intraclass correlations are ρ_3 = 0.1 and ρ_2 = 0.05 (and no covariates at any level). To compute the power of a two-tailed test at significance level 0.05, first we compute

$$\lambda = \sqrt{\frac{20 \times 3 \times 20}{2}}\left(0.2\right)\sqrt{\frac{1}{1 + \left(3 \times 20 - 1\right)0.1 + \left(20 - 1\right)0.05}} = 1.75,$$

then we compute the critical value of the *t*-distribution with 2(20) – 2 = 38 degrees of freedom as *c*(0.05*/2, 2(20) – 2*) = 2.02, and then we use equation 5 to compute the power as

$$1 - \mathrm{H}\,[2.02, 38, 1.75] + \mathrm{H}\;[-2.02, 38, 1.75] = 1 - 0.60 = 0.40,$$

that is, the power in this example is 0.40.

Determinants of Power in Three-Level Cluster Randomized Designs

As equations 5 and 9 indicate mainly three factors impact power: the effect size parameter, the degrees of freedom, and the non-centrality parameter. The degrees of freedom are a function of the number of level-3 units and hence larger number of level-3 units result in higher power. Larger effect sizes result in higher power estimates as well. In addition, as equations 4 and 7 suggest the non-centrality parameter is affected by the number of level-1, level-2, and level-3 units and the intraclass correlations at the second and third level. Of course covariates also affect power as indicated in equation 9.

Several interesting findings emerge from the results above. First, the power increases as the effect size increases. Consider an example in education where there are *n* = 20 students per classroom, *p* = 3 classrooms per school, and *m* = 15 schools per treatment condition (and no covariates). When the school and classroom intraclass correlations are respectively ρ_3 = 0.1 and ρ_2 = 0.05, and the effect size is δ = 0.25, the power of a two-tailed *t*-test is 0.45. If the effect size is twice as large however, δ = 0.5, and everything else remains unchanged the power is 0.95, more than two times as large. Second, the power decreases as the intraclass correlations increase. In the previous example, when the effect size is δ = 0.5, but the intraclass correlations at the school and classroom level are increased respectively to ρ_3 = 0.2 and ρ_2 = 0.1 (e.g., twice as large as in the previous case), the power of a two-tailed *t*-test is 0.76 (an absolute decrease of 19 percent from 0.95).

Third, the number of level-1, level-2 and level-3 units also affects power. However, the number of classrooms has a larger impact on power than the number of students per classroom, but the number of schools influences power much more than the number of classrooms per school or the number of students per classroom. For example, suppose that the total sample size is the same but the total number of schools is increased from 30 to 40 (*m* = 20 per condition), there are *p* = 3 classrooms per school, *n* = 15 students per classroom, the intraclass correlations at the school and at the classroom level are respectively ρ_s = 0.2 and ρ_c = 0.1, and the effect size is δ = 0.5 standard deviations. The power of a two-tailed *t*-test in this case is 0.87 (a 11 percent increase from 0.76 above). In fact, as Konstantopoulos (2008) showed when the number of level-3 units becomes vary large the power tends to 1.

The Effect of Covariates on Power Estimates

As one would expect, the power is higher when covariates are included in the model. Overall, the larger the proportion of variance explained at each level, the higher the power, other things being equal. As Cook (2005) argues covariates with considerable predictive power are critical for reducing the number of larger units (such as schools) needed, and for making the study less expensive or affordable given a fixed budget (see also Bloom, Richburg-Hayes, & Black, 2007). However, covariates do not affect power exactly the same. For example, in education, a school-aggregate measure of prior achievement is typically a very useful covariate, and recent work has shown that school-level pretest measures are at least as effective in increasing power as student-level pretest measures in two-level cluster randomized designs (see Bloom et al., 2007; Hedges & Hedberg, 2007). Similarly, in three-level cluster randomized designs the level-3 covariates can influence power potentially as much or more than the level-2 or level-1 covariates (assuming covariates at different levels explain the same proportion of the variance at that level). Consider the following example in education. Suppose that the ANCOVA model includes only one covariate at a time, each time the covariate is at a different level, and that the covariate explains 50 percent of the variance at that level. When only a first level covariate is included in the model and explains 50 percent of the variance at the first level the second part of equation 7 becomes

$$\sqrt{\frac{1}{0.5(1-\rho_2-\rho_3)+n\rho_2+pn\rho_3}}\,. \tag{10}$$

When only a second level covariate is included in the model and explains 50 percent of the variance at the second level the second part of equation 7 becomes

$$\sqrt{\frac{1}{(1-\rho_2-\rho_3)+0.5n\rho_2+pn\rho_3}}, \tag{11}$$

and when only a third level covariate is included in the model and explains 50 percent of the variance at the third level the second part of equation 7 becomes

$$\sqrt{\frac{1}{(1-\rho_2-\rho_3)+n\rho_2+0.5pn\rho_3}} \; . \tag{12}$$

In this example, the first level covariate adjusts the term $(1-\rho_2-\rho_3)$ by one-half, the second level covariate adjusts the term $n\rho_2$ by one-half, and the third level covariate adjusts the term $pn\rho_3$ by one-half. In education, typically, the clustering effect at the second level is smaller than that in the third level, and the third level intraclass correlation is typically between 0.1 and 02 (see Hedges & Hedberg, 2007; Nye, Konstantopoulos, & Hedges, 2004). When there are at least 10 students per classroom and at least 2 classrooms per school, the terms $n\rho_2$ and especially $pn\rho_3$ are more likely to be larger than 1, whilst the term $(1-\rho_2-\rho_3)$ is smaller than 1 (assuming the clustering effects are not exactly zero). With educational data the assumption $pn\rho_3 > n\rho_2$ is likely to hold, and the adjustment in equations 10 to 12 is typically larger when the third level covariate is included in the model. As a result, the non-centrality parameter becomes larger and hence the power becomes larger.

For instance, suppose that there are *n* = 20 students per classroom, *p* = 3 classrooms per school, and *m* = 15 schools per treatment condition and only one covariate at the first level that explains 50 percent of the first level variance. When the school and classroom intraclass correlations are respectively ρ_3 = 0.15 and ρ_2 = 0.10, and the effect size is δ = 0.25, the power of a two-tailed *t*-test is 0.33. When only one covariate is included at the second level and explains 50 percent of the second level variance, and everything else is unchanged, the power is 0.35. However, when only one third level covariate is included in the model and explains 50 percent of the third level variance, and everything else is unchanged, the power is 0.48 (much larger than 0.33 or 0.35). Of course different values of the parameters that affect power will produce different values of power, but overall level-1 and level-2 covariates affect power similarly whereas level-3 covariates seem to have a higher impact on power under certain assumptions and using achievement data. However, the level-3 covariates, q, are included in the computation of the degrees of freedom of the test and, hence, it is preferable to include a small number of covariates with high explanatory power at the third level. These results hold for two-level cluster randomized designs. That is, assuming achievement data, level-2 covariates seem to have a higher impact on power than level-1 covariates.

Incorporating Cost in Three-Level Cluster Randomized Designs

In three-level balanced designs the researcher needs to choose three samples sizes: the number of level-1 units within level-2 units, the number of level-2 units within level-3 units, and the number of level-3 units. Because of budget constraints however, the choice of sampling of each unit at each level is affected by the cost of the units. In survey methods there is a long tradition of optimum sampling in both stages of two-stage cluster designs (see Cochran 1977; Lohr, 1999). In psychology, methodologists have discussed optimal allocation and power analysis in generalizability studies and measurement designs with budget constraints (see Marcoulides, 1993, 1997). In addition, psychology methodologists have discussed optimal allocation methods for many aspects of experimental designs such as the number of individuals to different treatment levels, and the number of measurements within individuals (Allison, Allison, Faith, Paultre, & Pi-Sunyer, 1997; McClelland, 1997). In education, statisticians have provided methods for optimal allocation in two-level cluster and block randomized designs with equal and unequal costs per unit of randomization (Raudenbush & Liu, 1997, 2000; Liu, 2003). Below I present methods for optimal allocation in three-level randomized designs with two levels of nesting. The methods resemble optimum sampling for two-stage sampling (see Cochran, 1977) and optimal design for two-level cases as discussed by Raudenbush (1997), and Raudenbush and Liu (2000). For simplicity I discuss balanced designs.

Although level-3 units affect power more than level-2 or level-1 units in practice it may be too expensive to have numerous level-3 units (e.g., schools) in the sample. In contrast, it may be less expensive to add level-2 (e.g., classrooms) or level-1 (e.g., students) units in the sample. Hence, given the budget constraints, the researcher needs to configure the best allocation of resources possible to optimize power. This suggests that the researcher needs to incorporate the costs of level-1, level-2 and level-3 units in the design phase of the study. Following Raudenbush (1997) and Raudenbush and Liu (2000) consider a linear cost function for the total cost of the study

$$TC = MpnC_1 + MpC_2 + MC_3 \tag{13}$$

where TC is the total cost for all units in all levels, M = 2*m* is the total number of level-3 units, C_1 is the cost of each level-1 unit, C_2 is the cost of each level-2 unit, and C_3 is the cost of each level-3 unit, and all other terms have been defined earlier. It follows that

$$M = 2m = \frac{TC}{pnC_1 + pC_2 + C_3} . \qquad (14)$$

Now, suppose that the total cost as well as the cost for each unit at each level is fixed. Suppose also for simplicity that the cost for the units in the treatment and the control group is the same. The objective then is to determine the optimal number of level-1 and level-2 units that maximizes power. As Raudenbush (1997) argued choosing the optimal samples size within larger units (or clusters) informs decisions about the total number of larger units (or clusters) that need to be included in the sample. To achieve that, one needs to maximize the non-centrality parameter λ in equations 4 and 7 with respect to n and p (see Raudenbush & Liu, 2000). In the case of no covariates at any level when we substitute equation 14 in equation 4 λ becomes

$$\lambda = \sqrt{\frac{Mpn}{4}}\delta\sqrt{\frac{1}{1+(n-1)\rho_2+(pn-1)\rho_3}} = \sqrt{\frac{TCpn}{4(pnC_1+pC_2+C_3)}}\delta\sqrt{\frac{1}{1+(n-1)\rho_2+(pn-1)\rho_3}} . \qquad (15)$$

The task at hand is to maximize the above equation with respect to n (number of level-1 units in each level-2 unit) and p (number of level-2 units in each level-3 unit). The maximization produces an optimal n of

$$n_{opt} = \sqrt{\frac{C_2}{C_1}}\sqrt{\frac{1-\rho_2-\rho_3}{\rho_2}} , \qquad (16)$$

and an optimal p of

$$p_{opt} = \sqrt{\frac{C_3}{C_2}}\sqrt{\frac{\rho_2}{\rho_3}}\,. \tag{17}$$

The total number of level 3 clusters is then determined as

$$M = \frac{TC}{p_{opt}n_{opt}C_1 + p_{opt}C_2 + C_3}\,. \tag{18}$$

Note that in the two-level case with one level of nesting at the second level equation 16 becomes

$$\tilde{n}_{opt} = \sqrt{\frac{C_2}{C_1}}\sqrt{\frac{1-\rho_2}{\rho_2}}$$

which replicates the results by Raudenbush (1997) and Cochran (1977). The same logic holds when covariates are included in the model. However, in the case of covariates equation 13 becomes

$$TC = MpnC_1^{*} + MpC_2^{*} + MC_3^{*}, \tag{19}$$

where the asterisk indicates that the cost of level-1, level-2, and level-3 units changes when covariates are measured. Suppose again that the total cost is fixed. When we incorporate equation 19 in equation 7 λ_A becomes

$$\lambda = \sqrt{\frac{Mpn}{4}}\delta\sqrt{\frac{1}{\eta_1 + (n\eta_2 - \eta_1)\rho_2 + (pn\eta_3 - \eta_1)\rho_3}} = \sqrt{\frac{TCpn}{4\left(pnC_1^* + pC_2^* + C_3^*\right)}}\delta\sqrt{\frac{1}{\eta_1 + (n\eta_2 - \eta_1)\rho_2 + (pn\eta_3 - \eta_1)\rho_3}}. \tag{20}$$

When we maximize equation 20 with respect to n and p we obtain an optimal n of

$$n_{opt}^* = \sqrt{\frac{C_2^*}{C_1^*}}\sqrt{\frac{\eta_1(1-\rho_2-\rho_3)}{\eta_2\rho_2}}, \tag{21}$$

and an optimal p of

$$p_{opt}^* = \sqrt{\frac{C_3^*}{C_2^*}}\sqrt{\frac{\eta_2\rho_2}{\eta_3\rho_3}}. \tag{22}$$

In this case the total number of level-3 units is determined as

$$M^* = \frac{TC}{p_{opt}^* n_{opt}^* C_1^* + p_{opt}^* C_2^* + C_3^*}. \tag{23}$$

The last step involves the computation of the power of the test for the treatment effect. To compute power one needs to include the optimal values of n, p, and M in the computation of the non-centrality parameter (and the degrees of freedom).

Computing the Optimal Number of Level-1 and Level-2 Units and Power

To illustrate the usefulness of the methods presented above I consider a simple example where the total cost TC = 1000, and the cost of level-1 units C_1 = 1 (see e.g., Raudenbush, 1997; Raudenbush & Liu, 2000). The optimal n, and p, and the power for multiple values of the cost ratios, for multiple effect sizes (expressed in standard deviation units), and intraclass correlations are reported in Table 1 (assuming no covariates at any level). Specifically, Table 1 shows how sample sizes at each level and power are affected when level-3 units become much more expensive than level-2 units. Several findings emerged from Table 1. First, as level-3 units become much more expensive than level-2 units, the number of level-3 units becomes smaller and the number of level-2 units becomes larger (as equations 17 and 22 suggest). For example, when level-3 units are five times as costly as level-2 units and level-2 units are two times as costly as level-1 units, the intraclass correlations at the second the third level are respectively ρ_3 = 0.03 and ρ_2 = 0.02, the optimal number of level-1 units within level-2 units is n = 10, the number of level-2 units within level-3 units is p = 2, and the number of level-3 units is 32. In this example when the effect size is δ = 0.3, the power is 0.79. However, when the cost ratio of level-3 to level-2 units is four times larger (C_3/C_2 = 20) the optimal number of level-2 units within level-3 units p = 4 and the number of level-3 units is 12. The power is also affected differently in this case and it is much smaller, 0.49.

Insert Table 1 Here

Second, the larger the intraclass correlations at the second and third level, the smaller the number of level-1 units (as equations 16 and 21 suggest) and the larger the number of level-3 units. In the previous example, when level-3 units are five times as costly as level-2 units and level-2 units are two times as costly as level-1 units, and the intraclass correlations at the second the third level are respectively ρ_3 = 0.08 and ρ_2 = 0.12, the optimal number of level-1 units within level-2 units is n = 4, the optimal number of level-2 units within level-3 units is p = 2, and the number of level-3 units is 46. When the effect size is δ = 0.3, the power is 0.50.

In addition, equations 17 and 22 indicate when the intraclass correlation at the second level becomes larger relative to the intraclass correlation at the third level the optimal p becomes larger. In the previous example, when level-3 units are five times as costly as level-2 units and level-2 units are two times as costly as level-1 units, and the intraclass correlations at the second the third level are respectively ρ_2 = 0.20 and ρ_3 = 0.05, the optimal number of level-1 units within level-2 units is n = 3, the optimal number of level-2 units within level-3 units is p = 4, and the number of level-3 units is 32. When the effect size is δ = 0.3, the power is 0.53.

As expected, larger effect sizes, smaller intraclass correlations at the second and third level, and lower cost of level-3 and level-2 units result in higher estimates of power of the test of the treatment effect (see last column of Table 1). The magnitude of the effect size has a large impact on power and effect sizes larger than or equal to 0.8 standard deviations can produce optimal power estimates (e.g., 0.80) even when clustering effects and cost ratios are large. For example, when level-3 units are 20 times as costly as level-2 units and level-2 units are two times as costly as level-1 units, the intraclass correlations at the second the third level are respectively ρ_3 = 0.20 and ρ_2 = 0.1, the optimal number of level-1 units within level-2 units is n = 4, the optimal number of level-2 units within level-3 units is p = 3, and the number of level-3 units is 17. If the effect size is δ = 0.8, the power is 0.83, slightly larger than the typical threshold of 0.80.

Table 2 summarizes power estimates when the cost of level-2 units becomes increasingly large with respect to level-1 units and the cost of level-3 units to level-2 units remains constant. As Table 2 and equations 16 and 21 indicate when level-2 units are much more expensive than level-1 units, the optimal n becomes larger, other things being equal. For example, when level-3 units are five times as costly as level-2 units and level-2 units are also five times as costly as level-1 units, the intraclass correlations at the second the third level are respectively ρ_3 = 0.03 and ρ_2 = 0.02, the optimal number of level-1 units within level-2 units is n = 15, the optimal number of level-2 units within level-3 units is p = 2, and the number of level-3 units is 16. However, when the cost ratio of level-2 to level-1 units is four times larger (C_2/C_1 = 20), and everything else remains unchanged, the optimal number of level-1 units within level-2 units n = 31 and the number of level-3 units is 5. Note that in Tables 1 and 2 different values of cost ratios, intraclass correlations, and effect sizes will provide different estimates of power. However, overall the computations follow the same pattern.

Insert Table 2 Here

When covariates are included in the model the results are overall similar to those reported above. Note however that, as equation 23 indicates the number of level-3 units becomes larger as the number of level-1 and level-2 units becomes smaller, other things being equal. As equations 21 and 22 suggest, in order to decrease the optimal n and p, other things being equal, one would want to include in the model level-1 predictors that explain more variance at the first level than level-2 predictors at the second level, and level-2 predictors to explain more variance at the second level than level-3 predictors at the third level. That is, within the optimal design framework, level-1 covariates are more important than level-2 covariates in minimizing the number of level-1 units. In addition, level-2 covariates are more important than level-3 covariates in minimizing the number of level-1 units. And minimizing level-1 and level-2 units will result in maximizing level-3 units (see equation 23). For example, suppose that the level-3 units are five times as costly as level-2 units, the level-2 units are two times as costly as level-1 units, and the intraclass correlations at the second the third level are respectively ρ_3 = 0.07 and ρ_2 = 0.10. Also, suppose that a level-1 covariate explains 80 percent of the variance at the first level, a level-2 covariate explains 40 percent of the variance at the second level, and a level-3 covariate explains 20 percent of the variance at the third level. Then, the optimal number of level-1 units within level-2 units is n = 3, the optimal number of level-2 units within level-3 units is p = 2, and the number of level-3 units is 56.

However, when the covariates explain the same proportion of the variance and everything else remains unchanged the optimal number of level-1 units within level-2 units is n = 5, the optimal number of level-2 units within level-3 units is p = 2, and the number of level-3 units is 44. Finally, when a level-1 covariate explains 20 percent of the variance at the first level, a level-2 covariate explains 40 percent of the variance at the second level, and a level-3 covariate explains 80 percent of the variance at the third level, and everything else remains unchanged the optimal number of level-1 units within level-2 units is n = 6, the optimal number of level-2 units within level-3 units is p = 3, and the number of level-3 units is 29. However, note that in these examples the power estimates are almost identical since first the non-centrality parameter is affected differently by the proportion of variance explained by the covariates (as discussed in previous sections) and the degrees of freedom become smaller as the level-3 units become smaller. That is, in the examples above, when the effect size is δ = 0.4, the power is 0.97 or 0.98.

Conclusion

In education three-level experimental designs are becoming increasingly common, and frequently such designs assign randomly entire clusters to a treatment and a control group. In these large-scale cluster randomized studies the researcher faces the challenge of obtaining sufficient power of the test of the treatment effect given budget constraints. That is, the researcher needs to incorporate the costs associated with recruiting samples at each level of the hierarchy and collecting data in the power computations (see Raudenbush, 1997). The present study provided methods for computing power of tests of treatment effects (within an optimal design framework) in three-level cluster randomized designs where nesting occurs at the second and at the third level.

Several findings emerged from this study. First, as in two-level designs the number of level-3 units impacts power more than the number of level-2 or level-1 units, and the number of level-2 units influences power more than the number of level-1 units. In addition, the number of level-3 units impacts power via the degrees of freedom of the *t*- or *F*-test. Second, the clustering at the second and third level affects power inversely. Third, larger effect sizes affect power positively. Fourth, useful covariates affect power positively. In addition, in education (e.g., achievement data), it appears that the level-3 covariates have a larger impact on power than level-2 covariates.

Results from optimal allocation methods suggested that as level-3 units become much more expensive than level-2 units, the researcher should sample a larger number of level-2 units within level-3 units. Similarly, as level-2 units become much more expensive than level-1 units, the researcher should sample a larger number of level-1 units within level-2 units. However, when the cost of level-3 units is not much higher than the cost of other units, sampling more level-3 units is recommended because it results in higher power. Also, larger clustering effects result in a smaller optimal number of level-1 units within level-2 units and a larger number of level-3 units (other things being equal). When the clustering effect at the second level is smaller than that in the third level, the optimal number of level-2 units within level-3 units decreases and as a result the number of level-3 units increases.

Covariates also affect optimal design computations. Specifically, level-1 covariates seem more important than level-2 covariates in maximizing the number of level-3 units. In addition, level-2 covariates seem more important than level-3 covariates in maximizing the number of level-3 units. However, power is affected differently via the non-centrality parameter and the degrees of freedom.

The methods provided here apply to both experimental designs and any non-experimental studies that involve nesting and estimate group differences in an outcome (assuming trivial correlations between observed covariates and treatment). The logic of power computations remains the same and one can compute the power of a test that examines a group difference using the results presented in this study.

One potential limitation of this study is that it provided methods for optimal design assuming balanced designs. Although researchers aim to design balanced experimental studies, imbalance often takes place in practice or sometimes by design (e.g., studies about class or school size effects). In principle, the results of the present paper should apply approximately to unbalanced designs when treatment and control groups (or level-2 and level-3 units) have similar sample sizes. When imbalance is extreme among groups however, the use of the harmonic mean is recommended to compute power (see Cohen, 1998). In addition, the present study assumed that the cost in experimental and control groups is the same, which does not always hold. Liu (2003) discussed cases where the cost is unequal for treatment and control units.

References

Allison, D. B., Allison, R. L., Faith, M. S., Paultre, F., & Pi-Sunyer, F. X. (1997). Power and money: Designing statistically powerful studies while minimizing financial costs. *Psychological Methods, 2*, 20-33.

Bloom, H. S. (2005). *Learning more from social experiments*. New York: Russell Sage.

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

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

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

Cook, T. D. (2005). Emergent principles for the design, implementation, and analysis of cluster-based experiments in social science. *The Annals of the American Academy of Political and Social Science, 599,* 176-198.

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

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

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

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. (2008). The power of the test for treatment effects in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness, 1,* 66-88.

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

Lohr, S. L. (1999). *Sampling: Design and analysis.* Duxbury Press.

Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. *Journal of Educational Statistics, 18*, 197-206.

Marcoulides, G. A. (1997). Optimizing measurement designs with budget constraints: The variable cost case. *Educational and Psychological Measurement, 57,* 800-812.

McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods, 2,* 3-19.

Murphy, K. R., & Myors, B. (2004). *Statistical power analysis: A simple and general model fortraditional 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., 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, Konstantopoulos, S., & Hedges, L. V. (2000). 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., & 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.

Raudenbush, S. W., Spybrook, J., Liu, X., & Congdon, R. (2006). *Optimal design for longitudinal and multi-level research*. Software documentation.

Table 1. Power Computations that Incorporate Cost: No Covariates

Cost ratio: C_3/C_2	Cost Ratio: C_2/C_1	Second Level ICC	Third Level ICC	Effect Size	Optimal n	Optimal p	M = 2m	Power
5	2	0.02	0.03	0.20	10	2	32	0.46
5	2	0.02	0.03	0.30	10	2	32	0.79
5	2	0.02	0.03	0.40	10	2	32	0.96
5	2	0.04	0.06	0.20	7	2	39	0.36
5	2	0.04	0.06	0.30	7	2	39	0.67
5	2	0.04	0.06	0.40	7	2	39	0.89
5	2	0.08	0.12	0.20	4	2	46	0.26
5	2	0.08	0.12	0.30	4	2	46	0.50
5	2	0.08	0.12	0.40	4	2	46	0.74
10	2	0.02	0.03	0.20	10	3	20	0.37
10	2	0.02	0.03	0.30	10	3	20	0.68
10	2	0.02	0.03	0.40	10	3	20	0.90
10	2	0.04	0.06	0.20	7	3	24	0.28
10	2	0.04	0.06	0.30	7	3	24	0.54
10	2	0.04	0.06	0.40	7	3	24	0.78
10	2	0.08	0.12	0.20	4	3	27	0.19
10	2	0.08	0.12	0.30	4	3	27	0.37
10	2	0.08	0.12	0.40	4	3	27	0.58
20	2	0.02	0.03	0.20	10	4	12	0.25
20	2	0.02	0.03	0.30	10	4	12	0.49
20	2	0.02	0.03	0.40	10	4	12	0.73
20	2	0.04	0.06	0.20	7	4	14	0.19
20	2	0.04	0.06	0.30	7	4	14	0.37
20	2	0.04	0.06	0.40	7	4	14	0.58
20	2	0.08	0.12	0.20	4	4	16	0.14
20	2	0.08	0.12	0.30	4	4	16	0.25
20	2	0.08	0.12	0.40	4	4	16	0.40

Note: ICC = Intraclass Correlation

Table 2. Power Computations that Incorporate Cost: No Covariates

Cost ratio: C_3/C_2	Cost Ratio: C_2/C_1	Second Level ICC	Third Level ICC	Effect Size	Optimal n	Optimal p	M = 2m	Power
5	5	0.02	0.03	0.20	15	2	16	0.29
5	5	0.02	0.03	0.30	15	2	16	0.55
5	5	0.02	0.03	0.40	15	2	16	0.79
5	5	0.04	0.06	0.20	11	2	19	0.22
5	5	0.04	0.06	0.30	11	2	19	0.43
5	5	0.04	0.06	0.40	11	2	19	0.66
5	5	0.08	0.12	0.20	7	2	21	0.15
5	5	0.08	0.12	0.30	7	2	21	0.29
5	5	0.08	0.12	0.40	7	2	21	0.46
5	10	0.02	0.03	0.20	22	2	9	0.18
5	10	0.02	0.03	0.30	22	2	9	0.35
5	10	0.02	0.03	0.40	22	2	9	0.55
5	10	0.04	0.06	0.20	15	2	10	0.13
5	10	0.04	0.06	0.30	15	2	10	0.24
5	10	0.04	0.06	0.40	15	2	10	0.39
5	10	0.08	0.12	0.20	10	2	12	0.11
5	10	0.08	0.12	0.30	10	2	12	0.18
5	10	0.08	0.12	0.40	10	2	12	0.29
5	20	0.02	0.03	0.20	31	2	5	0.11
5	20	0.02	0.03	0.30	31	2	5	0.17
5	20	0.02	0.03	0.40	31	2	5	0.27
5	20	0.04	0.06	0.20	21	2	6	0.09
5	20	0.04	0.06	0.30	21	2	6	0.15
5	20	0.04	0.06	0.40	21	2	6	0.22
5	20	0.08	0.12	0.20	14	2	6	0.07
5	20	0.08	0.12	0.30	14	2	6	0.10
5	20	0.08	0.12	0.40	14	2	6	0.14

Note: ICC = Intraclass Correlation