.

Cost Considerations in Three-Level Block Randomized Designs:

Treatment Assigned at Middle Level

Spyros Konstantopoulos

Michigan State University

An important issue in designing experiments is to ensure that the design is sensitive enough to detect the intervention effects that are expected. This critical task involves making decisions about sample sizes to ensure sufficient statistical power of the test of the treatment effect. Power is defined as the probability of detecting the treatment effect when it exists. Many populations of interest in education and the social sciences have multilevel structure (e.g., students are nested within classrooms and classrooms are nested within schools. Because individuals within aggregate units are often more alike than individuals in different units, this nested structure produces an intraclass correlation structure that needs to be taken into account both in experimental design and statistical analysis.

Experiments that involve nested population structures may assign treatment conditions either to individuals such as students, subgroups such as classrooms or entire groups such as schools. Sometimes researchers are interested in implementing whole classroom interventions where treatments are assigned to entire subgroups such as classrooms (e.g., forms of formative assessment or differentiated instruction). In this case random assignment of subgroups to conditions takes place within blocks (e.g., schools) and the designs are called randomized block designs. Statistical theory for computing power in two-level designs has been well documented (e.g., Hedges & Hedberg, 2007; Raudenbush & Liu, 2000). Methods for computing power in three-level balanced cluster and block randomized designs have also been documented recently (Konstantopoulos, 2008a; 2008b).

In multilevel designs units at different levels of the hierarchy affect power differently. Because sampling units at different levels is under the control of the investigator the optimal sampling of units 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 (Raudenbush & Liu, 2000; Konstantopoulos, 2009). However, maximizing the number of larger units, such as schools, is typically more expensive than maximizing smaller units, such as classrooms or students. The challenge is to design a cost-effective study that will maximize the power of the test of the treatment effect given a certain budget. This requires incorporating cost-related issues when maximizing power in randomized block designs (Raudenbush & Liu, 2000). The present study discusses optimal design considerations that incorporate costs of sample sizes at different levels of the hierarchy when designing three-level randomized block designs where treatment is assigned at the middle level. For instance, consider a design where students are nested within classrooms and classrooms are randomly assigned to treatment conditions within schools. Raundenbush and Liu (2000) have discussed optimal deign for two-level randomized block designs. In this paper I extent their methods 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. Throughout the paper I discuss examples from education.

Three-Level Block Randomized Designs: Treatment Assigned at Middle Level

For simplicity, I assume that there is one treatment and one control group and that the design is balanced. In this design, level-3 units (blocks) 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 the discussion that follows, I assume that level-3 units, level-2 units, and the treatment by level-3 unit interaction are random effects.

The structural model in ANCOVA notation is

$$Y_{ijkl} = \mu + \alpha_{Aj} + \theta_1^T \mathbf{X}_{ijkl} + \theta_2^T \mathbf{Z}_{ijk} + \theta_3^T \Psi_i + \beta_{Ai} + \alpha\beta_{Aij} + \gamma_{A(ij)k} + \varepsilon_{A(ijk)l}, \quad (1)$$

where μ is the grand mean, α_{Aj} is the (fixed) effect of the j^{th} treatment (j = 1,2), $\theta_1^T = (\theta_{11}, \ldots, \theta_{1r})$ is a row vector of r level-1 covariate effects, $\theta_2^T = (\theta_{21}, \ldots, \theta_{2w})$ is a row vector of w level-2 covariate effects, $\theta_3^T = (\theta_{31}, \ldots, \theta_{3q})$ is a row vector of q level-3 covariate effects, $\mathbf{X}_{ijkl}$ is a column vector of r level-2 mean-centered level-1 covariates (e.g., student characteristics) in the k^{th} level-2 unit in the j^{th} treatment in the i^{th} level-3 unit, $\mathbf{Z}_{ijk}$ is a column vector of w level-3 mean-centered level-2 covariates (e.g., classroom or teacher characteristics) in the j^{th} treatment in the i^{th} level-3 unit, Ψ_i is a column vector of q level-3 covariates (e.g., school characteristics), and the last four terms are level-3, treatment by level-3 interaction, level-2, and level-1 random effects respectively. Specifically, β_{Ai} is the random effect of level-3 unit i (i = 1,…, m), $\alpha\beta_{Aij}$, is the treatment by level-3 interaction random effect, $\gamma_{A(ij)k}$ is the random effect of level-2 unit k (k = 1,…, p) within treatment j within level-3 unit i, and $\varepsilon_{A(ijk)l}$ is the error term of level-1 unit l (l = 1,…, n) within level-2 unit k, within treatment j, within level-3 unit i. The subscript A indicates adjustment due to covariate effects. In experiments, assuming randomization works, the treatment effect is orthogonal to the covariates and the error term, and the expected value of the adjustment of the treatment is zero. 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 residual level-1, level-2, and level-3 errors are normally distributed with a mean of zero and residual variances σ_{Re}^2, τ_R^2, and ω_R^2 respectively. The treatment by level-3 unit interaction random effect also follows a normal distribution with a mean of zero and a variance ω_{Rt}^2 (the subscript t indicates treatment).

Computing Power

When the null hypothesis is false the test statistic that examines the significance of the treatment effect has approximately a non-central t-distribution with m-q-1 degrees of freedom and a non-centrality parameter λ_A. The non-centrality parameter is defined as the expected value of the estimate of the treatment effect divided by the square root of the variance of the estimate of the treatment effect, namely

$$\lambda_A = \sqrt{\frac{mpn}{2}}\delta\sqrt{\frac{1}{\eta_1 + (m\eta_2 - \eta_1)\rho_2 + (pn\vartheta_{R3}\eta_3 - \eta_1)\rho_3}}. \quad (2)$$

where

$$\eta_1 = \sigma_{Re}^2 / \sigma_e^2, \eta_2 = \tau_R^2 / \tau^2, \eta_3 = \omega_R^2 / \omega^2, \ \rho_2 = \frac{\tau^2}{\sigma_T^2}, \ \rho_3 = \frac{\omega^2}{\sigma_T^2}, \ \sigma_T^2 = \sigma_e^2 + \tau^2 + \omega^2, \quad (3)$$

$\vartheta_{R3} = \omega_{Rt}^2 / \omega_R^2$ is the proportion of the treatment by level-3 unit interaction random effect residual variance to the total between level-3 units residual variance ($0 \le \vartheta_{R3} \le 1$), and δ is the effect size parameter (the standardized mean difference) that is independent of sample size and describes the magnitude of the treatment effect. The η's indicate the proportion of the variances at each level of the hierarchy that is still unexplained (percentage of residual variation). For example when $\eta_1 = 0.25$, this indicates that the variance at the first level decreased by 75 percent due to the inclusion of covariates (Konstantopoulos, 2008a). The ρ's are intraclass correlations and represent the clustering at the middle and top levels. The effect size parameter is expressed in (total population) standard deviation units, namely $\delta = \alpha_1 - \alpha_2 / \sigma_T$. When no covariates are included the non-centrality parameter reduces to

$$\lambda = \sqrt{\frac{mpn}{2}}\delta \sqrt{\frac{1}{1+(n-1)\rho_2 + (pn\vartheta_3 - 1)\rho_3}}, \tag{4}$$

where $\vartheta_3 = \omega_t^2 / \omega^2$ is the proportion of the treatment by level-3 unit interaction random effect variance to the total between level-3 units variance and $0 \le \vartheta_3 \le 1$. The power of the two-tailed *t*-test at level α is

$$p_2 = 1 - H[c(\alpha/2, m-q-1), (m-q-1), \lambda_A] + H[-c(m-q-1), (m-q-1), \lambda_A]. \tag{5}$$

where $c(\alpha, v)$ is the level α one-tailed critical value of the *t*-distribution with v degrees of freedom [e.g., $c(0.05, 20) = 1.72$], and $H(x, \nu, \lambda)$ is the cumulative distribution function of the non-central *t*-distribution with ν degrees of freedom and non-centrality parameter λ_A. In the case of no covariates the degrees of freedom for the *t*-test are slightly changed, and the non-centrality parameter $\lambda_A = \lambda$.

Optimal Sampling

In three-level 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 randomized block designs with equal and unequal costs per unit of randomization (Raudenbush & Liu, 1997, 2000; Liu, 2003). A recent study provided methods for optimal allocation in three-level cluster randomized designs where for example level-3 units such as schools are randomly assigned to conditions (Konstantopoulos, 2009). Below I present methods for optimal allocation in three-level randomized block designs with two levels of nesting where treatment is assigned at the middle level. The methods resemble optimum sampling for two-stage sampling (see Cochran, 1977), and optimal design for two-level cases (Raudenbush, 1997; Raudenbush & Liu, 2000) or three-level cluster randomized designs (Konstantopoulos, 2009). 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 and Liu (2000) and Konstantopoulos (2009) consider a linear cost function for the total cost of the study

$$TC = 2mpnC_1 + 2mpC_2 + mC_3 \qquad (6)$$

where TC is the total cost for all units in all levels, *m* is the total number of level-3 units, *p* is the number of level-2 units within a treatment condition, *n* is the number of level-1 units within a level-2 unit, 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.

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, level-2 and eventually level-3 units that maximizes power. Choosing the optimal sample 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 equation 6 with respect to n and p (Konstantopoulos, 2009; Raudenbush & Liu, 2000). When one substitutes equation *m* from equation 6 to equation 2 the noncentrality parameter λ_A becomes

$$\lambda_A = \sqrt{\frac{TCpn}{2\left(2pnC_1 + 2pC_2 + C_3\right)}}\delta\sqrt{\frac{1}{\eta_1 + \left(n\eta_2 - \eta_1\right)\rho_2 + \left(pn\vartheta_{R3}\eta_3 - \eta_1\right)\rho_3}}. \qquad (7)$$

When we maximize equation 7 with respect to n and p we obtain

$$n_{opt} = \sqrt{\frac{C_2}{C_1}}\sqrt{\frac{\eta_1\left(1-\rho_2-\rho_3\right)}{\eta_2\rho_2}},\ p_{opt} = \sqrt{\frac{C_3}{2C_2}}\sqrt{\frac{\eta_2\rho_2}{\eta_3\vartheta_{R3}\rho_3}}. \qquad (8)$$

(15)

Then the total number of level-3 units is determined as

$$m = \frac{TC}{2p_{opt}n_{opt}C_1 + 2p_{opt}C_2 + C_3}. \qquad (9)$$

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). When covariates are not included in the model the optimal n and p are respectively

$$n_{opt} = \sqrt{\frac{C_2}{C_1}}\sqrt{\frac{\left(1-\rho_2-\rho_3\right)}{\rho_2}},\ p_{opt} = \sqrt{\frac{C_3}{2C_2}}\sqrt{\frac{\rho_2}{\vartheta_3\rho_3}}. \qquad (10)$$

Power Computations within Optimal Design

To illustrate the usefulness of the methods presented above I followed Raudenbush & Liu (2000) and Konstantopoulos (2009) and I consider a simple example where the total cost TC = 1000, and the cost of level-1 units $C_1 = 1$. 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 are more expensive than level-2 units and the level-2 to level-1 cost ratio is fixed. First, as level-3 units become more expensive than level-2 units, the number of level-3 units becomes smaller and the number of level-2 units becomes larger. For example, when level-3 units are 5 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 = 0.06 and ρ_2 = 0.04, the optimal number of level-1 units within level-2 units is n = 7, the number of level-2 units within treatments within level-3 units is p = 3, and the number of level-3 units is 15. In this example when the effect size is δ = 0.3, the power is 0.85. However, when the cost ratio of level-3 to level-2 units is twice as large (C_3/C_2 = 10) the optimal number of level-2 units within treatments within level-3 units p = 5 and the number of level-3 units is 10. The power is also affected differently in this case and it is slightly smaller, 0.82. Other things being equal when level-3 units become much more expensive than level-2 units and level-2 units become much more expensive than level-1 units the power becomes much smaller. Second, the larger the intraclass correlations at the second and third level, the smaller the number of level-1 units and the larger the number of level-3 units. Other things being constant the power also becomes smaller when intraclass correlations become larger. 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).

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 indicates when level-2 units are much more expensive than level-1 units, the optimal n becomes larger, and m becomes smaller 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_2 = 0.04 and ρ_3 = 0.06, the optimal number of level-1 units within level-2 units is n = 11, the optimal number of level-2 units within treatments within level-3 units is p = 3, and the number of level-3 units is m = 8. However, when the cost ratio of level-2 to level-1 units is two times larger (C_2/C_1 = 10), and everything else remains unchanged, the optimal number of level-1 units within level-2 units n = 15 and the number of level-3 units is m = 5. Other things being equal the power becomes smaller as cost ratios increase (see last column of Table 2).

In the computations I used estimates that may be plausible for educational data given results from previous work, but they may not apply well to other kinds of data. The estimates for cost ratios for units at different levels illustrate the logic and applicability of incorporating cost in the model, but are not necessarily plausible values of cost ratios. Note that different values of cost ratios, intraclass correlations, and effect sizes will provide different estimates of power. However, overall the computations follow the same pattern.

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.

Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.

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

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

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

Konstantopoulos, S. (2008b). The power of the test for treatment effects in three-level block randomized designs. *Journal of Research on Educational Effectiveness, 1,* 265-288.

Konstantopoulos, S. (2009). Incorporating Cost in Power Analysis for Three-Level Cluster Randomized Designs. *Evaluation Review, 33,* 335-357.

Liu, XF (2003). Statistical power and optimum sample allocation ratio for treatment and control having unequal costs per unit randomization. *Journal of Educational and Behavioral Statistics, 28,* 231-248.

Lohr, S. L. (1998). *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.

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.

Table 1. Power Computations that Incorporate Cost: No Covariates

C_3/C_2	C_2/C_1	Level-2 ICC	Level-3 ICC	Effect Size	Optimal n	Optimal p	m	θ	Power
5	2	0.04	0.06	0.20	7	3	15	0.15	0.51
5	2	0.04	0.06	0.30	7	3	15	0.15	0.85
5	2	0.04	0.06	0.40	7	3	15	0.15	0.98
5	2	0.04	0.06	0.50	7	3	15	0.15	1.00
5	2	0.08	0.12	0.20	4	3	19	0.15	0.42
5	2	0.08	0.12	0.30	4	3	19	0.15	0.75
5	2	0.08	0.12	0.40	4	3	19	0.15	0.94
5	2	0.08	0.12	0.50	4	3	19	0.15	0.99
10	2	0.04	0.06	0.20	7	5	10	0.15	0.49
10	2	0.04	0.06	0.30	7	5	10	0.15	0.82
10	2	0.04	0.06	0.40	7	5	10	0.15	0.97
10	2	0.04	0.06	0.50	7	5	10	0.15	1.00
10	2	0.08	0.12	0.20	4	5	12	0.15	0.38
10	2	0.08	0.12	0.30	4	5	12	0.15	0.69
10	2	0.08	0.12	0.40	4	5	12	0.15	0.91
10	2	0.08	0.12	0.50	4	5	12	0.15	0.98

Note: ICC = Intraclass Correlation

Table 2. Power Computations that Incorporate Cost: No Covariates

C_3/C_2	C_2/C_1	Level-2 ICC	Level-3 ICC	Effect Size	Optimal n	Optimal p	m	θ	Power
5	5	0.04	0.06	0.20	11	3	8	0.15	0.34
5	5	0.04	0.06	0.30	11	3	8	0.15	0.64
5	5	0.04	0.06	0.40	11	3	8	0.15	0.87
5	5	0.04	0.06	0.50	11	3	8	0.15	0.97
5	5	0.08	0.12	0.20	7	3	9	0.15	0.26
5	5	0.08	0.12	0.30	7	3	9	0.15	0.49
5	5	0.08	0.12	0.40	7	3	9	0.15	0.73
5	5	0.08	0.12	0.50	7	3	9	0.15	0.90
5	10	0.04	0.06	0.20	15	3	5	0.15	0.22
5	10	0.04	0.06	0.30	15	3	5	0.15	0.42
5	10	0.04	0.06	0.40	15	3	5	0.15	0.64
5	10	0.04	0.06	0.50	15	3	5	0.15	0.82
5	10	0.08	0.12	0.20	10	3	5	0.15	0.19
5	10	0.08	0.12	0.30	10	3	5	0.15	0.36
5	10	0.08	0.12	0.40	10	3	5	0.15	0.55
5	10	0.08	0.12	0.50	10	3	5	0.15	0.73

Note: ICC = Intraclass Correlation