


5-1-2013

Constructing a More Powerful Test in Two-Level Block Randomized Designs

Spyros Konstantopoulos
Michigan State University

Follow this and additional works at: <http://digitalcommons.wayne.edu/jmasm>

 Part of the [Applied Statistics Commons](#), [Social and Behavioral Sciences Commons](#), and the [Statistical Theory Commons](#)

Recommended Citation

Konstantopoulos, Spyros (2013) "Constructing a More Powerful Test in Two-Level Block Randomized Designs," *Journal of Modern Applied Statistical Methods*: Vol. 12: Iss. 1, Article 8.
Available at: <http://digitalcommons.wayne.edu/jmasm/vol12/iss1/8>

This Regular Article is brought to you for free and open access by the Open Access Journals at DigitalCommons@WayneState. It has been accepted for inclusion in Journal of Modern Applied Statistical Methods by an authorized administrator of DigitalCommons@WayneState.

Constructing a More Powerful Test in Two-Level Block Randomized Designs

Spyros Konstantopoulos
Michigan State University
East Lansing, MI

A more powerful test is proposed for the treatment effect in two-level block randomized designs where random assignment takes place at the first level. When clustering at the second level is assumed to be known, the proposed test produces higher estimates of power than the typical test.

Key words: Block randomized designs, statistical power, clustering, experiments.

Introduction

One important consideration when designing large-scale experiments is to ensure that the design is sensitive enough to detect the expected intervention effect. This task involves making decisions about sample sizes to ensure sufficient statistical power of the test of the treatment effect. Statistical power is defined as the probability of rejecting the null hypothesis when it is false. Extensive literature exists on the computation of statistical power (e.g., Cohen, 1988; Kraemer & Thiemann, 1987; Lipsey, 1990; Murphy & Myers, 2004). Much of the literature, however, involves the computation of power in studies that use simple random samples; hence nesting effects are typically not included in power computations.

In education and the social sciences, populations of interest have often multilevel structures, for example, students are nested within classrooms or schools. Because individuals within aggregate units are often more alike than individuals in different units,

this nesting produces an intraclass correlation structure which is often called clustering in the sampling literature (Kish, 1965). Clustering should be taken into account both in experimental design and statistical analysis. Treatment conditions in experiments may be assigned either to individuals (e.g., students), subgroups (e.g., classrooms) or entire groups (e.g., schools). When treatments are assigned to individuals or subgroups within entire groups the designs are called block randomized designs; these designs are also known as multisite experiments or multisite trials because each site runs a self-sufficient experiment.

In designs that involve clustering, the computation of statistical power is more complicated than in simple random samples designs. First, nested factors are usually assumed to have random effects, and hence, power computations should involve the variance components structures which are typically expressed via intraclass correlations of these random effects. In education for example, schools are clusters that are typically treated as random effects. Second, there is more than one sample size involved because there are units at different levels in the hierarchy. For example, in education where students are nested within schools the power of the test of the treatment effect depends not only on the number of students within a school, but also on the number of schools (Hedges & Hedberg, 2007; Raudenbush, 1997). The sample sizes at different levels may affect power estimates differently. Statistical theory for computing

Spyros Konstantopoulos is Associate Professor of measurement and quantitative methods at the department of counseling, educational psychology, and special education at the College of Education at Michigan State University. Website: <http://spyros.wiki.educ.msu.edu/>. Email him at: spyros@msu.edu.

power in two-level balanced designs has been documented (Barcikowski, 1981; Hedges & Hedberg, 2007; Raudenbush, 1997; Raudenbush & Liu, 2000). Hedges and Hedberg provided methods for computing statistical power in two-level balanced cluster randomized designs where the second level units (e.g., schools) are randomly assigned to a treatment and a control group. Raudenbush and Liu (2000) provided methods for power analysis in two-level balanced block randomized designs where the first level units (e.g., students within schools) are randomly assigned to a treatment or a control group within second level units. These methods are helpful for a priori power analysis during the designing phase of the experiment. Methods for power computations of tests of treatment effects in multi-level designs have also been discussed in the health sciences (Donner & Klar, 2000; Murray, 1998).

Previous methods for power analysis in two-level balanced designs (e.g., students nested within schools) involved the computation of the non-centrality parameter of the non-central F - or t -distribution (Hedges & Hedberg, 2007; Raudenbush & Liu, 2000). Power is a function of the non-centrality parameter and of the degrees of freedom of the test, and higher values of these two factors correspond to higher values of statistical power. The non-centrality parameter is a function of clustering at the second level, which is typically expressed as an intraclass correlation, the number of level-1 and level-2 units, and the magnitude of the treatment effect. The degrees of freedom are a function of the number of the level-2 units.

Previous work has demonstrated that statistical power is an increasing function of the number of level-1 (e.g., students) and level-2 units (e.g., schools), and the magnitude of the treatment effect, but a decreasing function of the intraclass correlation (Hedges & Hedberg, 2007; Raudenbush, & Liu, 2000). Also, the number of level-2 units (e.g., schools) has a larger impact on power than the number of level-1 units (e.g., students). An implicit assumption in these methods, that are useful for a priori power analysis, is that the researcher has an idea about the value of the population intraclass correlation and the treatment effect in order to conduct the necessary power computations.

Hedges and Hedberg (2007) showed that in two-level balanced cluster randomized designs, where, for example, entire groups such as schools are randomly assigned to a treatment and a control condition, the power of the t -test has $2(\tilde{m} - 1)$ degrees of freedom (assuming no covariates), where \tilde{m} is the number of level-2 units (e.g., schools) assigned to each condition. When w covariates are included at the second level the degrees of freedom of the t -test are $2(\tilde{m} - 1) - w$. In such designs however, a more powerful test can be constructed when the intraclass correlation structure is assumed to be known (Blair & Higgins, 1986). Blair and Higgins showed that in two-level cluster randomized designs using an exact F -test with larger degrees of freedom is more powerful than that used by Hedges and Hedberg. Specifically, the test provided by Blair and Higgins had $2(\tilde{m}\tilde{n} - 1)$ degrees of freedom (assuming no covariates), where \tilde{n} is the number of level-1 units (e.g., students) within each level-2 unit (e.g., school).

As in two-level cluster randomized designs a test with larger degrees of freedom can also be constructed for two-level block randomized designs when the intraclass correlation (clustering) structure is assumed to be known. This test is more powerful than the typical test based on level-2 unit means because it preserves the degrees of freedom that are associated with level-1 units (e.g., students). Also, the test is exact, examines the same hypothesis about the treatment effect and has the same non-centrality parameter as the test presented by Raudenbush and Liu (2000) in the balanced case using the ANOVA framework.

The only difference between the two tests is in the degrees of freedom. The exact t -test for the treatment effect carried on level-2 unit (e.g., school) means assuming one treatment and one control group has $m - 1$ degrees of freedom when no covariates are included at any level, and $m - 1 - w$ degrees of freedom when w covariates are included at the second level, where m is the total number of level-2 units. Note that in this test the number of first level units is not taken into account in the degrees of the freedom of the test. However, following

Blair and Higgins (1986) a more powerful test for two-level block randomized designs can be constructed that includes the number of level-1 units in the degrees of freedom of the test. This article provides methods for constructing a more powerful test for treatment effects in two-level block randomized designs; these methods are useful for a priori power computations during the design phase of an experiment.

Methodology

A Two-Level Block Randomized Design

Following Graybill (1976) and Blair and Higgins (1986) consider a simple case of the general linear model in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (1)$$

where \mathbf{y} is a $N \times 1$ vector (N is the total number of observations), \mathbf{X} is a $N \times 2$ (assuming one treatment and one control group) design matrix for the regression coefficients, $\boldsymbol{\beta}$ is a 2×1 vector of the regression coefficients to be estimated (i.e., treatment and control means), and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of residuals that follows a multivariate normal distribution with a mean of zero and a variance matrix $\sigma^2 \mathbf{V}$, that is $\boldsymbol{\varepsilon} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{V})$, where σ^2 is the total variance in the outcome and is factored out of the variance covariance matrix \mathbf{V} . If the level-1 units are nested within level-2 units (the clusters), then matrix \mathbf{V} has elements that represent variances or intraclass correlations and ones within each cluster and zeroes between clusters. If these variances or intraclass correlations are assumed to be known, then matrix \mathbf{V} is positive definite and known (Graybill, 1976).

Consider a two-level block randomized design where level-1 units such as students are randomly assigned to one treatment and one control condition within level-2 units such as schools (the blocks). Suppose that there are m level-2 units overall and that the total variance is $\sigma^2 = \sigma_e^2 + \tau^2$, where σ_e^2 is the level-1 variance and τ^2 is the level-2 variance; matrix \mathbf{V} then has the same structure as the matrix \mathbf{V}^* which is

block diagonal $\mathbf{V}^* = \mathbf{I}_m \otimes \{\mathbf{V}_j^*\}$ with m blocks (the total number of level-2 units in the sample), \mathbf{I} is the identity matrix, and \otimes is the kronecker product. The diagonal elements of matrix \mathbf{V}_j^* for cluster j are $v_{iij}^* = \sigma_e^2 + (1 + \vartheta)\tau^2$ for level-1 units that receive the treatment within the level-2 unit, and $v_{iicj}^* = \sigma_e^2 + \tau^2$ for level-1 units that do not receive the treatment within a level-2 unit, where $\vartheta = \tau_t^2 / \tau^2$ is the proportion of level-2 unit by treatment variance to the total level-2 variance ($0 \leq \vartheta \leq 1$) and subscripts i, j, t, c represent level-1, level-2 units, and treatment and control groups respectively. The off diagonal elements of matrix \mathbf{V}_j^* are $v_{ikj}^* = \tau^2$. If the intraclass correlation are defined as the proportion of the between level-2 unit variance to the total variance, namely $\rho = \tau^2 / \sigma^2$ and the total variance σ^2 is factored out from matrix \mathbf{V}^* matrix \mathbf{V} is constructed, which has ones in the main diagonal for level-1 units in the control group, $1 + \vartheta\rho$ in the main diagonal for level-1 units in the treatment group, ρ in the off diagonal between level-1 units within each level-2 unit and zeroes between level-2 units (see Appendix). If the intraclass correlation ρ and ϑ are known, which essentially means that the proportion of the level-2 unit by treatment variance to the total variance is known, then matrix \mathbf{V} is known.

To illustrate the structure of matrix \mathbf{V} consider a simple case where there are two schools and within each school two students are randomly assigned to a treatment and two students to a control group. Assuming the first two students receive treatment, $\mathbf{V}_j = \mathbf{V}_j^* / \sigma^2$ for school j is

$$\mathbf{V}_j = \begin{bmatrix} 1 + \vartheta\rho & \rho & \rho & \rho \\ \rho & 1 + \vartheta\rho & \rho & \rho \\ \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{bmatrix},$$

and \mathbf{V} is

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_j & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_j \end{bmatrix}$$

where $\mathbf{0}$ is a 2x2 matrix of zeros, namely $\mathbf{0} = [0, 0, 0, 0]$, expressed as a row vector. In this simple case when no covariates are included the matrix \mathbf{X} is

$$\mathbf{X}^T = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix},$$

and

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}.$$

Finally, the vectors \mathbf{y} (the outcome) and \mathbf{e} (the residuals of \mathbf{y}) are expressed as row vectors

$$\mathbf{y}^T = [y_1, \dots, y_8], \quad \mathbf{e}^T = [e_1, \dots, e_8].$$

According to Graybill (1976) when matrix \mathbf{V} is known the regression estimates of the general linear model are computed as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}, \quad (2)$$

the total variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{N-2} \mathbf{y}^T (\mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}) \mathbf{y} \quad (3)$$

and the variances of the regression estimates are computed as

$$\text{var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}. \quad (4)$$

Following Graybill (1976) and Blair and Higgins (1986) the test constructed for the hypothesis

$$\mathbf{H}\boldsymbol{\beta} = 0$$

is a general F -test

$$F = \frac{(\mathbf{H}\hat{\boldsymbol{\beta}} - 0)^T (\mathbf{H}^T (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{H})^{-1} (\mathbf{H}\hat{\boldsymbol{\beta}} - 0)^T}{\hat{\sigma}^2} / 1 \quad (5)$$

with 1 and $N - 2$ degrees of freedom (assuming one treatment and one control group). Matrix \mathbf{H} is a 1x2 design matrix that facilitates the contrast among the two treatment conditions and $\hat{\sigma}^2$ is defined in equation (3). Specifically, when there is one treatment and one control group and a researcher is interested in testing the equality between the two means, the vector of contrasts $\mathbf{H} = [1, -1]$, and the vector of coefficients is $\boldsymbol{\beta}^T = [\beta_1, \beta_2]$. Note that the proposed test can be used to test hypotheses for many general linear models including one-way, factorial ANOVA, and ANCOVA (Blair & Higgins, 1986; Graybill, 1976). When the null hypothesis is false the test follows a non-central F -distribution with a noncentrality parameter

$$\lambda^2 = \frac{(\mathbf{H}\boldsymbol{\beta} - 0)^T (\mathbf{H}^T (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{H})^{-1} (\mathbf{H}\boldsymbol{\beta} - 0)^T}{2\sigma^2} \quad (6)$$

and 1, $N - 2$ degrees of freedom. This test can be used for unbalanced or balanced data.

Clustering

The test proposed herein assumes that matrix \mathbf{V} is known. This implies that the variance of the treatment effect across level-2 units is known. Eventually, this translates to knowing the intraclass correlation ρ and ϑ , which means knowing the proportion of level-2 unit by treatment variance to the total level-2 variance. Typically values of population parameters are not likely to be known exactly. A more realistic assumption for a priori power analysis is that there is a range of intraclass correlations which will most likely capture the real value of the population parameter. Hedges & Hedberg (2007) provided a comprehensive collection of intraclass correlations for achievement data based on national representative samples of students. Specifically, they gave an array of plausible values of intraclass correlations for achievement outcomes using recent large-scale studies that surveyed national probability samples of elementary and secondary students in America. This compilation of intraclass correlations is useful for planning

two-level designs. The values of intraclass correlations ranged typically between 0.10 and 0.20 for typical samples and were smaller than 0.10 for more homogeneous samples (e.g., low-achieving schools). Evidence from two-level models of the National Assessment of Educational Progress (NAEP) trend data and Project STAR data (Konstantopoulos & Hedges, 2008; Nye, Hedges & Konstantopoulos, 2000) also points to intraclass correlations between 0.10 and 0.20. Finally, evidence from Project STAR has suggested that the between school variance of the small class effect is typically less than 50 percent of the between school variance.

The ANOVA Model

The proposed test presented in equation (5) is a general test that can be used in both unbalanced and balanced designs. To simplify computations, how the proposed test can be used in simple two-level balanced block randomized designs is now discussed. Using the ANOVA framework the noncentrality parameter of a test can be computed, which facilitates power computations. In this model, level-1 units are randomly assigned to treatment and control groups within level-2 units. The number of level-2 units are represented by m , and the number of level-1 units within each condition by n . The assumption is that there is one treatment and one control group and hence the total sample size is $N = 2mn$. At this point the model does not include any covariates. A structural model for a student outcome Y_{ijk} , the k^{th} level-1 unit in the j^{th} treatment in the i^{th} level-2 unit can then be described as

$$Y_{ijk} = \alpha_j + \beta_i + \alpha\beta_{ij} + \varepsilon_{ijk}, \quad (7)$$

where α_j is the (fixed) effect of the j^{th} treatment ($j = 1, 2$) within level-2 unit i , β_i is the random effect of level-2 unit i ($i = 1, \dots, m$), $\alpha\beta_{ij}$ is the treatment by level-2 unit interaction random effect, and ε_{ijk} is the error term of student k ($k = 1, \dots, n$) within treatment j , within level-2 unit i . The level-1, level-2 and treatment by level-2 unit random effects have variances $\sigma_e^2, \tau^2, \tau_i^2$

respectively. The random effects at different levels are orthogonal to each other.

The objective is to examine the statistical significance of the treatment effect, meaning to test the hypothesis:

$$H_0: \alpha_1 = \alpha_2 \text{ or } \alpha_1 - \alpha_2 = 0.$$

Suppose that a researcher wants to test the hypothesis and carries out the usual t - or F -test. When the null hypothesis is false, the test statistic F has the non-central F -distribution with a non-centrality parameter λ^2 . In the balanced case 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

$$\lambda^2 = \delta^2 \frac{mn}{2} \frac{1}{1+(n\vartheta-1)\rho} \quad (8)$$

(Hedges & Hedberg, 2007; Raudenbush & Liu, 2000). This F -test is based on the level-2 unit means and hence the degrees of freedom of the denominator of the test are $m - 1$ assuming no covariates at the second level. The power of the F -test at level α is

$$p = 1 - H[c(\alpha, l, m-1), 1, m-1, \lambda^2], \quad (9)$$

where $c(\alpha, l, v)$ is the level α critical value of the F -distribution with l, v degrees of freedom (e.g., $c(0.05, 1, 20) = 4.35$) and $H(x, 1, v, \lambda)$ is the cumulative distribution function of the non-central F -distribution with l, v degrees of freedom and non-centrality parameter λ^2 . Equivalently, the test of the treatment effect and statistical power can also be computed using the t -statistic that has a non-central t -distribution with $m - 1$ degrees of freedom and a non-centrality parameter λ (the square root of equation (8)).

When the intraclass correlation structure is assumed known, however, a more powerful F - or t -test can be constructed (see equation (5)). In the balanced case the non-centrality parameter of the test is the same as that reported in

equation (8). However, this test has larger degrees of freedom, because σ in equation (5) is estimated by $\hat{\sigma}$ in equation (3). Because the degrees of freedom associated with $\hat{\sigma}$ are $N - 2$, the degrees of freedom of the denominator of the proposed test are $N - 2 = 2(mn - 1)$ assuming one treatment and one control and no covariates. The power of the F -test at level α is

$$p = 1 - H[c(\alpha, 1, 2(mn - 1)), 2(mn - 1), \lambda^2]. \quad (10)$$

Equivalently, the t -statistic has a non-central t -distribution with $2(mn - 1)$ degrees of freedom and a non-centrality parameter λ .

The ANCOVA Model

When covariates are included at the first and second level the linear model is

$$Y_{ijk} = \alpha_{Aj} + \beta_{Ai} + \boldsymbol{\theta}_1^T \mathbf{X}_{ijk} + \boldsymbol{\theta}_2^T \mathbf{Z}_i + \alpha \beta_{Aij} + \varepsilon_{A(ij)k} \quad (11)$$

where $\boldsymbol{\theta}_1^T = (\theta_{11}, \dots, \theta_{1r})$ is a row vector of r first-level covariate effects, $\boldsymbol{\theta}_2^T = (\theta_{21}, \dots, \theta_{2w})$ is a row vector of w second-level covariate effects, \mathbf{X}_{ijk} is a column vector of r first-level covariates (e.g., student characteristics) in the j^{th} treatment in the i^{th} second level unit, \mathbf{Z}_i is a column vector of w second-level covariates (e.g., school characteristics); all other terms have been previously defined. The subscript A indicates that both the treatment and the random effects are adjusted by the covariates in the model. In principle however, assuming randomization is successful, the treatment effect is orthogonal to the covariates and the error term and the expected value of the adjustment is zero. The first and second level random effects are adjusted by first and second level covariates respectively. The first level covariates are centered around their second level unit means and therefore they do not explain variance of the random effects at the second level (i.e., group-mean centering). Centering also ensures orthogonality among predictors at the first and second level. All first level covariates are treated as fixed at the second level. When covariates are

included in the model the level-1 and level-2 residual variances are defined as σ_{Re}^2, τ_R^2 respectively, and the residual total variance is $\sigma_{RT}^2 = \sigma_{Re}^2 + \tau_R^2$ (and R indicates residual variances because of the adjustment for the effect of covariates). The adjusted level-2 intraclass correlation is defined then as

$$\rho_{A2} = \frac{\tau_R^2}{\sigma_{RT}^2}. \quad (12)$$

Covariates are useful when conducting power analysis because they typically increase the power of the test for the treatment effect. Specifically, covariates that are significantly associated with the outcome typically explain some proportion of the variance in the outcome, which in turn results in a reduction of the unconditional intraclass correlations and the standard errors of the treatment effects. In experimental studies this indicates that the F - or the t -tests for the treatment effects will have higher values when covariates are included in the model because the treatment effect remains virtually unchanged due to the fundamental principle of randomization, which assumes independence between treatment effects and covariates. That is, a researcher can achieve optimal power estimates (e.g., 0.80) without having to increase sample size. In fact, as Cook (2005) argues covariates with considerable predictive power are important for reducing the number of larger units such as schools needed, and for making the study less expensive or affordable given a fixed budget. Powerful covariates at the first level, when modeling achievement data, include previous achievement and socioeconomic status (Hedges & Hedberg, 2007). Powerful covariates at the second level include school aggregate measures of achievement or socioeconomic status.

In a balanced design within the ANCOVA framework the objective is to examine the statistical significance of the treatment effect net of the possible effects of covariates, namely to test the hypothesis

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

which involves the computation of the typical t or F -test statistic. When the null hypothesis is false, the F -test statistic has the non-central F -distribution with a non-centrality parameter

$$\lambda_A^2 = \delta^2 \frac{mn}{2} \frac{1}{\eta_1 + (n\eta_2\vartheta - \eta_1)\rho}, \quad (13)$$

where

$$\eta_2 = \tau_R^2 / \tau^2, \eta_1 = \sigma_{Re}^2 / \sigma_e^2 \quad (14)$$

(Hedges & Hedberg 2007; Murray, 1998). 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% due to the inclusion of covariates such as pre-treatment measures. The degrees of freedom of the F -test are $1, 2(mn - 1) - w - r$. The power of the F -test at level α is

$$p = 1 - H \left[\begin{array}{c} c(\alpha, 1, 2(mn - 1) - w - r), \\ 1, 2(mn - 1) - w - r, \lambda_a^2 \end{array} \right] \quad (15)$$

Equivalently, the t -statistic has a non-central t -distribution with $2(mn - 1) - w - r$ degrees of freedom and a non-centrality parameter λ_A (square root of equation (13)).

Results

Computational Example

Power comparisons between two t -tests are now discussed: the typical t -test carried out on level-2 unit means with $m - 1 - w$ degrees of freedom and the proposed t -test with $2(mn - 1) - w - r$ degrees of freedom. The power computations are presented in Tables 1 and 2 and apply to balanced designs. For this exercise power is computed assuming one treatment and one control group for two-tailed t -tests, or equivalently an F -test, at the 0.05 significance level assuming no covariates in the model. In Table 1 the effect size parameter is $\delta = 0.25$, and in Table 2 the effect size parameter is $\delta = 0.40$. Three values of intraclass correlations were used: 0.05, 0.10, and 0.20. These values have

been reported in previous work as typical values for homogeneous and more heterogeneous samples (Hedges & Hedberg, 2007; Raudenbush & Liu, 2000). Results from Project STAR have also indicated that $\vartheta \leq 0.50$. The first step in the power analysis is to compute the noncentrality parameter. Suppose that there are a total of $m = 6$ schools, $n = 15$ students in each condition (30 students total per school) within each school and that $\delta = 0.40$, $\rho = 0.10$, and $\vartheta = 0.50$. The noncentrality parameter using equation (8) is

$$\lambda = 0.4^2 \frac{6*15}{2} \frac{1}{1 + (15*0.5 - 1)*0.10} = 4.36$$

and the degrees of freedom are $6 - 1 = 5$ for the test using the level-2 means and $2*(6*15 - 1) = 178$ for the proposed test. Using equation (10) the power is 0.39 and using equation (11) the power is 0.55 (see seventh row in Table 2). The functions for the noncentral F - or t -test are available in mainstream packages such as SPSS (the functions are Ncdf.F or Ncdf.T), SAS (using the cumulative distribution functions, CDF, of the F - or t -distribution), S-Plus (the functions are pf or ptnoncent) or R (the functions are pf or pt).

The first column of Table 1 shows the number of level-2 units in the sample. The second column shows the number of level-1 units within each condition within each level-2 unit. The third and fourth columns show values of ρ and ϑ , and columns five and six show the degrees of freedom for each test. Finally, columns seven and eight show power values for each test. The number of level-2 units ranges from 6 to 12, and the number of level-1 units per condition per level-2 unit ranges from 15 to 30. Results from Table 1 suggest that the power of the proposed test is always higher than the power of the typical test based on level-2 unit means. The difference in power is more pronounced when the number of level-2 units is smaller, the number of level-1 units is larger, and ρ , ϑ are small. For example, when the total number of level-2 units $m = 6$, the number of level-1 units $n = 30$ in each condition per level-2 unit, $\delta = 0.25$, $\rho = 0.05$, and $\vartheta = 0.25$

the power is 0.54 for the proposed test and 0.39 for the typical test that uses level-2 unit means. The difference in power becomes smaller however, as the number of level-2 units increases.

The structure of Table 2 is identical to that in Table 1. As expected, because the effect size is larger, power estimates in Table 2 are larger. Again, the power of the proposed test is always higher than the power of the typical test based on level-2 unit means. As in Table 1, the difference in power is more pronounced when the number of level-2 units is smaller, the number of level-1 units is larger, and ρ , ϑ are smaller. For example, when the total number of level-2 units $m = 6$, the number of level-1 units $n = 30$ in each condition per level-2 unit, $\delta = 0.40$, $\rho = 0.10$, and $\vartheta = 0.25$ the power is 0.84 for the proposed test and 0.66 for the typical test based on level-2 unit means. The difference in power becomes smaller as the number of level-2 units becomes larger. Overall, power is positively affected by the effect size and the number of level-1 and level-2 units, and negatively affected by ρ , ϑ , which suggests that the larger the between level-2 unit variance of the treatment effect the smaller the power, other things being equal.

These findings replicate the results presented by Blair and Higgins for two-level cluster randomized designs. The power estimates of the proposed test will always be larger than those obtained by the test based on the level-2 unit means, and the difference in power is larger when the number of the level-2 units is smaller, the number of level-1 units is larger and the between level-2 unit variance of the treatment effect is smaller. However, as the number of level-2 units increases, the difference in power between the two tests decreases, and when the number of level-2 units becomes infinitely large the two tests provide almost identical estimates of power.

Conclusion

This study proposed a more powerful test for treatment effects in two-level block randomized designs where, for example, students within schools are randomly assigned to a treatment and a control group. The proposed test statistic is

more powerful than the typical test based on level-2 unit means because it preserves the degrees of freedom that are associated both with level-2 and level-1 units. The test can be used to compute power both in unbalanced and balanced designs. However, this study focused on the balanced case. The assumption of the proposed test is that the between level-2 unit variance of the treatment effect is known, that is, ρ , ϑ are known.

In education, when the outcome is achievement, there is evidence that the level-2 intraclass correlation ranges typically from 0.10 to 0.20, and it is less than 0.10 for more homogeneous samples. There is also some evidence that the between level-2 unit variance of the treatment effect is typically less than 50 percent of the between level-2 unit variance. As with some statistical procedures a limitation of the current test is that the information used to compute power is not always known exactly. Nonetheless, for a priori power analysis some knowledge of clustering and effect sizes is always necessary for computing power of the typical test based on level-2 unit means (Raudenbush & Liu, 2000).

It is important to stress that the methods for a priori power computations provided are intended to serve simply as useful guides for experimental designs; the sample sizes proposed, although informative, should be treated as approximate and not exact (Kraemer & Thieman, 1987). The results of the methods presented are accurate as long as the assumptions about the model and the tests, as well as the estimates of effect sizes and intraclass correlations, are accurate. Regardless, assuming educated or accurate guesses for the information used to compute the power in the proposed test produce higher estimates of power than in the typical test, especially when the number of level-2 units and the intraclass correlations are small. The findings of this study are useful because in education and the social sciences many times researchers focus on homogeneous groups (e.g., minorities, disadvantaged students). In addition, sampling fewer level-2 units (e.g., schools) is cost-effective because it reduces the cost of the study overall without compromising statistical power.

A MORE POWERFUL TEST IN TWO-LEVEL BLOCK RANDOMIZED DESIGNS

Table 1: Power Comparisons between a F-test Based on Level-2 Unit Means
and the Proposed F-test: Effect Size is 0.25

Number of Level-2 Units	Number of Level-1 Units	Intraclass Correlation	Theta	df/Level-2 Unit Means	df/ All Observations	Power/Level-2 Unit Means	Power/ All Observations
6	15	0.05	0.25	5	178	0.25	0.35
6	30	0.05	0.25	5	358	0.39	0.54
6	15	0.05	0.50	5	178	0.22	0.31
6	30	0.05	0.50	5	358	0.32	0.44
6	15	0.10	0.25	5	178	0.23	0.31
6	30	0.10	0.25	5	358	0.32	0.45
6	15	0.10	0.50	5	178	0.19	0.25
6	30	0.10	0.50	5	358	0.24	0.33
6	15	0.20	0.25	5	178	0.20	0.27
6	30	0.20	0.25	5	358	0.25	0.34
6	15	0.20	0.50	5	178	0.15	0.20
6	30	0.20	0.50	5	358	0.17	0.23
12	15	0.05	0.25	11	358	0.53	0.60
12	30	0.05	0.25	11	718	0.76	0.83
12	15	0.05	0.50	11	358	0.47	0.54
12	30	0.05	0.50	11	718	0.65	0.73
12	15	0.10	0.25	11	358	0.48	0.55
12	30	0.10	0.25	11	718	0.66	0.74
12	15	0.10	0.50	11	358	0.39	0.45
12	30	0.10	0.50	11	718	0.50	0.58
12	15	0.20	0.25	11	358	0.41	0.48
12	30	0.20	0.25	11	718	0.52	0.60
12	15	0.20	0.50	11	358	0.30	0.34
12	30	0.20	0.50	11	718	0.35	0.40

SPYROS KONSTANTOPOULOS

Table 2: Power Comparisons between a F-test Based on Level-2 Unit Means and the Proposed F-test: Effect Size is 0.4

Number of Level-2 Units	Number of Level-1 Units	Intraclass Correlation	Theta	df/Level-2 Unit Means	df/ All Observations	Power/Level-2 Unit Means	Power/ All Observations
6	15	0.05	0.25	5	178	0.53	0.71
6	30	0.05	0.25	5	358	0.75	0.91
6	15	0.05	0.50	5	178	0.47	0.64
6	30	0.05	0.50	5	358	0.65	0.83
6	15	0.10	0.25	5	178	0.48	0.66
6	30	0.10	0.25	5	358	0.66	0.84
6	15	0.10	0.50	5	178	0.39	0.55
6	30	0.10	0.50	5	358	0.51	0.69
6	15	0.20	0.25	5	178	0.42	0.57
6	30	0.20	0.25	5	358	0.52	0.70
6	15	0.20	0.50	5	178	0.30	0.42
6	30	0.20	0.50	5	358	0.35	0.49
12	15	0.05	0.25	11	358	0.90	0.94
12	30	0.05	0.25	11	718	0.99	1.00
12	15	0.05	0.50	11	358	0.85	0.91
12	30	0.05	0.50	11	718	0.96	0.98
12	15	0.10	0.25	11	358	0.86	0.92
12	30	0.10	0.25	11	718	0.97	0.99
12	15	0.10	0.50	11	358	0.77	0.84
12	30	0.10	0.50	11	718	0.88	0.93
12	15	0.20	0.25	11	358	0.79	0.86
12	30	0.20	0.25	11	718	0.90	0.94
12	15	0.20	0.50	11	358	0.65	0.70
12	30	0.20	0.50	11	718	0.71	0.79

A MORE POWERFUL TEST IN TWO-LEVEL BLOCK RANDOMIZED DESIGNS

References

Barcikowski, R. S. (1981). Statistical power with a group mean as the unit of analysis. *Journal of Educational Statistics*, 6, 267-285.

Blair, R. C., & Higgins, J. J. (1986). Comment of statistical power with group mean as the unit of analysis. *Journal of Educational Statistics*, 11, 161-169.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed.). New York, NY: 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., & Klar, N. (2000). *Design and analysis of cluster randomization trials in health research*. London: Arnold.

Graybill, F. A. (1976). *Theory and application of the linear model*. Boston, MA: Duxbury Press.

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

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

Konstantopoulos, S., & Hedges, L. V. (2008). How Large an effect can we expect from school reforms? *Teachers College Record*, 110, 1613-1640.

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

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

Murphy, K. R., & Myers, 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, NY: Oxford University Press.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2000). 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. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2, 173-185.

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

Appendix

The diagonal elements of matrix \mathbf{V}_j^* for cluster j are $v_{iij}^* = \sigma_e^2 + (1 + \vartheta)\tau^2$ for level-1 units that receive the treatment within the level-2 unit, and $v_{iicj}^* = \sigma_e^2 + \tau^2$ for level-1 units that do not receive the treatment within a level-2 unit. The off diagonal elements of matrix \mathbf{V}_j^* are $v_{ikj}^* = \tau^2$. The structure of the block diagonal matrix \mathbf{V}^* is

$$\mathbf{V}^* = \begin{bmatrix} \mathbf{A}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_2 & \cdots & \mathbf{0} \\ \vdots & & & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{A}_m \end{bmatrix}$$

assuming m level-2 units, where \mathbf{A}_j is a $2n \times 2n$ matrix

$$\mathbf{A}_j = \begin{bmatrix} \sigma^2(1 + \vartheta\rho) & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2(1 + \vartheta\rho) & \cdots & \sigma^2\rho \\ \vdots & & & \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{bmatrix}$$

assuming n level-1 units per condition.