



Constructing a More Powerful Test in Three-Level Cluster Randomized Designs

Spyros Konstantopoulos

Faculty Associate, Institute for Policy Research
Assistant Professor of Human Development and Social Policy
Northwestern University

DRAFT

Please do not quote or distribute without permission.

Abstract

Experiments that involve nested structures may assign treatment conditions either to entire groups (such as schools), or subgroups (such as classrooms) or individuals (such as students). A key aspect of the design of such experiments includes knowledge of the intraclass correlation structure. This study provides methods for constructing a test for the treatment effect that is more powerful than the typical test based on level-3 unit means in three-level cluster randomized designs (with two levels of nesting). When the intraclass correlation structure at the second and third level is known the proposed test provides higher estimates of power because it preserves the degrees of freedom associated with the number of level-2 and level-1 units. The advantage in power estimates is more pronounced when the number of level-3 units is small.

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 in both experimental design and analysis.

Experiments that involve nested population structures, may assign treatment conditions either to individuals (such as students) or to entire groups (such as schools). Treatments are sometimes assigned to groups because the treatment is naturally administered to intact groups (such as a curriculum to a school or a management system to a firm), or because the assignment of treatments to groups is a matter of convenience (being much easier to implement than assignment to individuals). Designs that assign intact groups to treatments are often called cluster randomized, or group randomized, or hierarchical designs (see Bloom, 2005; Donner & Klar, 2000; Kirk, 1995; Murray, 1998). In other situations, treatments are assigned to entire subgroups such as classrooms (e.g., forms of formative assessment) or to individuals (e.g., different forms of computerized instruction to students within classrooms and schools). Such designs are called randomized block designs.

One of the most critical issues in designing experiments is to ensure that the design is sensitive enough to detect the intervention effects that are expected if the researchers' hypotheses were correct. This is, when planning experimental studies it is essential to ensure sufficient statistical power of the test of the treatment effect. Previous methods for power analysis in two-level balanced designs (e.g., students nested within schools) with one level of nesting (at the second level) involved the computation of the non-centrality parameter of the non-central F - or t -distribution (see Barcikowski, 1981; Hedges & Hedberg, 2007; Raudenbush & Liu, 2000). In

these designs the power is a function of the non-centrality parameter and 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 in turn is a function of the nesting effect 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.

As Barcikowski (1981) and Hedges and Hedberg (2007) showed in two-level 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 $2(m - 1)$ degrees of freedom (assuming no covariates), where m is the number of units (e.g. schools) assigned to each condition. When q covariates are included at the second level the degrees of freedom of the t -test are $2(m - 1) - q$. Previous work has also demonstrated that in such designs a more powerful test can be constructed when the intraclass correlation structure is known (see Blair & Higgins, 1986). Blair and Higgins showed that in two-level cluster randomized designs one can use an exact test with larger degrees of freedom that is more powerful than that used by Barcikowski, and Hedges and Hedberg. Specifically, the test provided by Blair and Higgins had $2(mn^* - 1)$ degrees of freedom (assuming no covariates), where n^* is the number of individuals (e.g. students) within each group (e.g. school). To illustrate the difference in the degrees of freedom consider the example where two schools are assigned to a treatment and two schools in a control group and each school has 50 students. The typical t -test for treatment effects carried out on school means (as provided in Barcikowski, and Hedges and Hedberg) would have two degrees of freedom, while a t -test that takes into account the observations within schools would have 198 degrees of freedom, which results in higher power of the test for the treatment effect as Blair and Higgins (1986) showed.

This holds for three-level cluster randomized designs as well. Consider a three-level design where nesting occurs at the second and third levels (e.g., classrooms and schools) and the three-level units (e.g., schools) are randomly assigned to treatment and control conditions. The exact t -test for the treatment effect carried on level-3 (school) means assuming one treatment and one control group and no covariates has $2(m - 1)$ degrees of freedom, and $2(m - 1) - q$ degrees of freedom when q covariates are included at the third level (see Konstantopoulos, in press). As in two-level cluster randomized designs a test with larger degrees of freedom can also be constructed. This test is more powerful because it preserves the degrees of freedom that are associated with subgroups (e.g., classrooms) and individual observations (e.g., students within classrooms). This is also an exact test that examines the same hypothesis about the treatment effect, and has the same non-centrality parameter as that presented by Konstantopoulos (in press). The only difference between the two tests is in the degrees of freedom. Both tests are useful for a priori power computations during the design phase of the experiment.

A Three-level Cluster Randomized Design with Two Levels of Nesting

For simplicity, in the following sections I discuss balanced three-level cluster randomized designs with two levels of nesting (at the second and third level) where there is one treatment and one control group. That is, level-3 units are randomly assigned to one treatment and one control group. I focus on the power of treatment contrasts, not omnibus (multiple-degree of freedom) treatment effects, because in my experience, multilevel designs are chosen to ensure the power of particular treatment contrasts. Even when several treatments are being compared, there is typically one contrast that is most important and the design is chosen to ensure adequate sensitivity for that contrast. I represent the number of level-3 units within each condition by m , the number of level-2 units within each level-3 unit by p , and the number of level-1 units within

each level-2 unit by n . Then, the sample size for the treatment and the control group is $N_t = N_c = mpn$ and the total sample size is $N = N_t + N_c = 2mpn$.

The Analytic Model

In this design, level-3 units are nested within treatment conditions, and level-2 and level-1 units are nested within level-3 units and treatment conditions (see Kirk, 1995). I assume that both level-2 and level-3 units are random effects and for simplicity I assume that no covariates are included in the model. Then, a structural model for a student outcome Y_{ijkl} , the l^{th} student in the k^{th} classroom in the j^{th} school in the i^{th} treatment can be described as

$$Y_{ijkl} = \alpha_i + \beta_{(i)j} + \gamma_{(ij)k} + \varepsilon_{(ijk)l}, \quad (1)$$

where α_i is the (fixed) effect of the i^{th} treatment ($i = 1, 2$), $\beta_{(i)j}$ is the random effect of school j ($j = 1, \dots, m$) within treatment i , $\gamma_{(ij)k}$ is the random effect of classroom k ($k = 1, \dots, p$) within school j within treatment i , and $\varepsilon_{(ijk)l}$ is the error term of student l ($l = 1, \dots, n$) within classroom k , within school j , within treatment i . I assume that the student, classroom, and school-level random effects have variances σ_e^2 , τ^2 , and ω^2 respectively. I also assume that the random effects at different levels are orthogonal to each other.

Nesting Effects

Consider a three-level design that follows a three-stage cluster sampling (see Cochran, 1977). The nesting effect in such designs occurs at the second and third level (e.g., classrooms and schools) and can be defined via two intraclass correlations. The total variance in the outcome is decomposed into three components: the within level-2 and between level-1 units variance, σ_e^2 , the between level-2 and within level-3 units variance, τ^2 , and the between level-3 units variance,

ω^2 . Then, the total variance in the outcome is defined as $\sigma^2 = \sigma_e^2 + \tau^2 + \omega^2$. Hence, in such three-level designs two intraclass correlations are defined:

$$\rho_2 = \frac{\tau^2}{\sigma^2} \tag{2}$$

at the second level and

$$\rho_3 = \frac{\omega^2}{\sigma^2} \tag{3}$$

at the third level (and the subscripts 2 and 3 indicate the second and third level respectively).

The General Linear Model

Following Graybill (1976) and Blair and Higgins (1986) the model presented above in equation 1 can be expressed as a general linear model in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

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 that need to be estimated (e.g., treatment and control means), and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of the residuals that follows a multivariate normal distribution with a mean of zero and a variance matrix $\sigma^2 \mathbf{V}$, that is $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{V})$. The matrix \mathbf{V} is positive definite and known, since its elements are either intraclass correlations (which are assumed known) or ones and zeroes, and σ^2 is the total variance of an individual observation given by $\sigma^2 = \sigma_e^2 + \tau^2 + \omega^2$.

The matrix \mathbf{V} has the same structure as the matrix \mathbf{V}^* which is block diagonal $\mathbf{V}^* = \mathbf{I}_{2m} \otimes \{\mathbf{V}_j^*\}$ with $2m$ blocks (the total number of level-3 units in the sample), where \mathbf{I} is the

identity matrix, $\mathbf{V}_j^* = \sigma_e^2 \mathbf{I}_{n_j} + \mathbf{I}_p \otimes \tau^2 \mathbf{J}_{n_k} + \omega^2 \mathbf{J}_{n_j}$, $\mathbf{J}_{n_k} = \mathbf{1}_{n_k} \mathbf{1}_{n_k}^T$, $\mathbf{J}_{n_j} = \mathbf{1}_{n_j} \mathbf{1}_{n_j}^T$, n_j , n_k are the number of level-1 units in level-3 unit j and level-2 unit k respectively, and \otimes is the Kronecker product. The diagonal elements of \mathbf{V}^* are $\text{var}(y_i) = \sigma^2 = \sigma_e^2 + \tau^2 + \omega^2$, the non-diagonal elements for level-1 units in the same level-2 and level-3 unit are $\text{cov}(y_i, y_j) = \tau^2 + \omega^2$, and the non-diagonal elements for level-1 units in different level-2 units, but in the same level-3 unit are $\text{cov}(y_i, y_j) = \omega^2$. The non-diagonal elements of \mathbf{V}^* can be expressed in terms of the intraclass correlations ρ_2, ρ_3 and the total variance σ^2 , since $\sigma^2 \rho_3 = \omega^2$, and $\sigma^2(\rho_2 + \rho_3) = \tau^2 + \omega^2$. Factoring out the total variance σ^2 from matrix \mathbf{V}^* produces matrix \mathbf{V} with ones in the diagonal, zeros in the non-diagonal elements between level-3 unit blocks, ρ_3 in the non-diagonal elements for level-1 units in the same level-3 unit but different level-2 units, and $\rho_2 + \rho_3$ in the non-diagonal elements for level-1 units in the same level-2 and level-3 units. If the intraclass correlations are known, then matrix \mathbf{V} is known.

Consider a simple case where there are two schools, each school has two classrooms, and each classroom has two students. Then $\mathbf{V}_j = \mathbf{V}_j^* / \sigma^2$ is

$$\mathbf{V}_j = \begin{bmatrix} 1 & \rho_2 + \rho_3 & \rho_3 & \rho_3 \\ \rho_2 + \rho_3 & 1 & \rho_3 & \rho_3 \\ \rho_3 & \rho_3 & 1 & \rho_2 + \rho_3 \\ \rho_3 & \rho_3 & \rho_2 + \rho_3 & 1 \end{bmatrix},$$

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

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

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

$$\mathbf{y}^T = [y_1, \dots, y_6], \quad \mathbf{e}^T = [e_1, \dots, e_6].$$

Following Graybill when matrix \mathbf{V} is known the total variance is estimated as

$$\hat{\sigma}^2 = \frac{1}{N-2} \mathbf{y}^T (\mathbf{V}^{-1} - \mathbf{V}^{-1} \mathbf{X} (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{V}^{-1}) \mathbf{y}. \quad (5)$$

Estimation

In balanced three-level cluster randomized designs the estimate of the treatment effect is the difference in the means between the treatment and the control group namely $\bar{Y}_{1\dots} - \bar{Y}_{2\dots}$ (assuming no covariates). The standard error of the estimate of the treatment effect is defined as

$$SE(\bar{Y}_{1\dots} - \bar{Y}_{2\dots}) = \sqrt{\frac{2}{mpn} \sqrt{pn\omega^2 + n\tau^2 + \sigma_e^2}} = \sqrt{\frac{2}{mpn}} \sigma \sqrt{1 + (n-1)\rho_2 + (pn-1)\rho_3}. \quad (6)$$

Hypothesis Testing

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

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

where α_i is the mean of the treatment or control group.

Suppose that the researcher wants to test the hypothesis and carries out the usual t -test.

This involves computing the test statistic

$$t = \frac{\sqrt{\frac{mpn}{2}} (\bar{Y}_{1\dots} - \bar{Y}_{2\dots})}{S},$$

where $\bar{Y}_{i\dots}$ is the mean of the i^{th} treatment group and S is the estimated standard error of the mean difference. When the null hypothesis is false, the test statistic t has the non-central t -distribution with a non-centrality parameter λ . 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

$$\lambda = \sqrt{\frac{mpn}{2}} \delta \sqrt{\frac{1}{1+(n-1)\rho_2+(pn-1)\rho_3}}. \quad (7)$$

(see Konstantopoulos, in press). The degrees of freedom of the t -test based in the level-3 unit means are $2(m-1)$. The power of the one-tailed t -test at level α is $p_1 = 1 - H[c(\alpha, 2(m-1)), 2(m-1), \lambda]$, where $c(\alpha, \nu)$ is the level α one-tailed critical value of the t -distribution with ν degrees of freedom [e.g., $c(0.05, 20) = 1.72$], and $H(x, \nu, \lambda)$ is the cumulative distribution function of the non-central t -distribution with ν degrees of freedom and non-centrality parameter λ . The power of the two-tailed t -test at level α is $p_2 = 1 - H [c(\alpha/2, 2(m-1)), 2(m-1), \lambda] + H [-c(\alpha/2, 2(m-1)), 2(m-1), \lambda]$. 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 $2(m-1)$ degrees of freedom in the denominator and non-centrality parameter λ^2 .

However, when the intraclass correlation structure is known a more powerful F - or t -test can be constructed. The non-centrality parameter of the t -test is the same as that reported in equation 7 above. However, this test has larger degrees of freedom, since σ in equation 6 above is estimated by $\hat{\sigma}$ in equation 5. Because the degrees of freedom associated with $\hat{\sigma}$ are $N-2$,

the degrees of freedom of this t -test are $2(mp_n - 1)$ assuming one treatment and one control and no covariates (see also Blair & Higgins, 1986). The power of the one-tailed t -test at level α is

$$p_1 = 1 - H[c(\alpha, 2(mp_n - 1)), 2(mp_n - 1), \lambda], \quad (8)$$

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

$$p_2 = 1 - H[c(\alpha/2, 2(mp_n - 1)), 2(mp_n - 1), \lambda] + H[-c(\alpha/2, 2(mp_n - 1)), 2(mp_n - 1), \lambda]. \quad (9)$$

Equivalently, the F -statistic has a non-central F -distribution with 1 degree of freedom in the numerator and $2(mp_n - 1)$ degrees of freedom in the denominator and a non-centrality parameter λ^2 . Specifically the numerator mean square (treatment) of the F -statistic assuming no covariates is

$$MST = \frac{mp_n / 2}{1 + (n-1)\rho_2 + (pn-1)\rho_3} (\bar{Y}_{1\dots} - \bar{Y}_{2\dots})^2 / 1,$$

and has 1 degree of freedom.

To compute the denominator mean square (error) of the F -statistic, one needs first to compute the necessary sums of squares of all three sources of variation: the within level-2 unit (classroom) sums of squares (SS_{wc}), the between level-2 unit sums of squares (SS_{bc}), and the between level-3 unit (schools) sums of squares (SS_{bs}). The within level-2 unit sums of squares assuming no covariates are

$$SS_{wc} = (n-1) \sum_{i=1}^2 \sum_{j=1}^m \sum_{k=1}^p S_{ijk}$$

where

$$S_{ijk} = \sum_{l=1}^n (Y_{ijkl} - \bar{Y}_{ijk.})^2 / (n-1).$$

The expected value of SS_{wc} is

$$E(SS_{wc}) = 2mp(n-1)\sigma^2(1 - \rho_2 - \rho_3).$$

Similarly, the within level-3 unit between level-2 units sums of squares assuming no covariates are

$$SS_{bc} = n(p-1) \sum_{i=1}^2 \sum_{j=1}^m S_{ij}$$

where

$$S_{ij} = \sum_{k=1}^p (\bar{Y}_{ijk.} - \bar{Y}_{ij..})^2 / (p-1).$$

The expected value of SS_{bc} is

$$E(SS_{bc}) = 2m(p-1)\sigma^2(1 + (n-1)\rho_2 - \rho_3).$$

Finally, the between level-3 units sums of squares assuming no covariates are

$$SS_{bs} = np(m-1) \sum_{i=1}^2 S_i$$

where

$$S_i = \sum_{j=1}^m (\bar{Y}_{ij..} - \bar{Y}_{i...})^2 / (m-1).$$

The expected value of SS_{bs} is

$$E(SS_{bs}) = 2(m-1)\sigma^2(1 + (n-1)\rho_2 + (pn-1)\rho_3).$$

The sums of square error using information from all three sources of variation is

$$SSE = \left(\frac{SS_{wc}}{1 - \rho_2 - \rho_3} + \frac{SS_{bc}}{1 - (n-1)\rho_2 - \rho_3} + \frac{SS_{bs}}{1 - (n-1)\rho_2 - (pn-1)\rho_3} \right)$$

and it follows that the denominator mean square (error) of the F -statistic assuming no covariates is

$$MSE = \frac{SSE}{N-2} = \left(\frac{SS_{wc}}{1 - \rho_2 - \rho_3} + \frac{SS_{bc}}{1 - (n-1)\rho_2 - \rho_3} + \frac{SS_{bs}}{1 - (n-1)\rho_2 - (pn-1)\rho_3} \right) / (N-2). \quad (10)$$

Notice that the degrees of freedom of the mean square error above are the sum of the degrees of freedom of all three different sums of squares which is $2mpn - 2mp + 2mp - 2m + 2m - 2 = 2(mpn - 1)$. This indicates that the F -statistic has 1 and $2(mpn - 1)$ degrees of freedom and the t -statistic has $2(mpn - 1)$ degrees of freedom. Note that although the numerator mean square (treatment) is unchanged, the denominator mean square error can take different forms. That is, the denominator mean square error can use information from one or more of the three sources of variation. Hence, different F - or t -tests can be constructed that will have different degrees of freedom (due to the denominator mean square error used each time) and hence will provide different estimates of power. For example, the exact F -test in three-level cluster randomized designs with two levels of nesting carried out on level-3 unit means uses the denominator mean square error

$$MSE_s = SSE_s / 2(m - 1) = \left(\frac{SS_{bs}}{1 - (n - 1)\rho_2 - (pn - 1)\rho_3} \right) / 2(m - 1) \quad (11)$$

assuming no covariates and one treatment and one control group. The F -statistic has 1 and $2(m - 1)$ degrees of freedom and the t -statistic has $2(m - 1)$ degrees of freedom (see Konstantopoulos, in press). Since the non-centrality parameter is unchanged the F - or t -statistic that uses equation 10 to compute the denominator mean square error is more powerful than the statistic that uses equation 11. In addition, the F - or t -tests that use other combinations of sums of squares to compute the denominator mean square error will always have smaller power (due to smaller degrees of freedom) than that computed using equation 9.

Including Covariates

The three-level model discussed in the previous sections can be modified to include q level-3 covariates, w level-2 covariates, and r level-1 covariates. I assume that the covariates at levels 1 and 2 are centered around their means respectively (group-mean centering). This ensures that predictors explain variation in the outcome *only* at the level at which they are introduced.

When covariates are included in the model the level-1 (student), level-2 (classroom), and level-3 (school) variances are defined as $\sigma_{Re}^2, \tau_R^2, \omega_R^2$ respectively, and $\sigma_{RT}^2 = \sigma_{Re}^2 + \tau_R^2 + \omega_R^2$ (and R indicates residual variances because of the adjustment for the effect of covariates). Two parameters (analogous to the unadjusted level-2 and level-3 intraclass correlations) summarize the associations between these residual variances: The adjusted level-2 intraclass correlation

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

and the adjusted level-3 intraclass correlation

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

where the subscripts A and R indicate adjustment due to covariates and residual variance. Then, the estimate of the treatment effect is the difference between the covariate adjusted treatment and the control group means namely $\bar{Y}_{A1...} - \bar{Y}_{A2...}$. The standard error of the estimate of the treatment effect is now defined as

$$SE(\bar{Y}_{A1...} - \bar{Y}_{A2...}) = \sqrt{\frac{2}{mpn}} \sqrt{pn\omega_R^2 + n\tau_R^2 + \sigma_{Re}^2} = \sqrt{\frac{2}{mpn}} \sigma_{RT} \sqrt{1 + (n-1)\rho_{A2} + (pn-1)\rho_{A3}}.$$

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

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

This involves computing the test statistic

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

where $\bar{Y}_{Ai\dots}$ is the adjusted mean of the i^{th} treatment group and S_A is the estimated covariate adjusted standard error of the covariate adjusted mean difference. When the null hypothesis is false, the test statistic t_A has the non-central t -distribution with a non-centrality parameter

$$\lambda_A = \sqrt{\frac{mpn}{2}} \delta \sqrt{\frac{1}{\eta_1 + (n\eta_2 - \eta_1)\rho_2 + (pm\eta_3 - \eta_1)\rho_3}}. \quad (12)$$

where

$$\eta_3 = \omega_R^2 / \omega^2, \eta_2 = \tau_R^2 / \tau^2, \eta_1 = \sigma_{Re}^2 / \sigma_e^2.$$

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 student level decreased by 75 percent due to the inclusion of covariates such as pre-treatment measures. The degrees of freedom of the t -test are $2(mpn - 1) - q - w - r$. The power of the one-tailed t -test at level α is

$$p_1 = 1 - H[c(\alpha, 2(mpn-1)-q-w-r), 2(mpn-1)-q-w-r, \lambda_A], \quad (13)$$

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

$$p_2 = 1 - H [c(\alpha/2, 2(mpn-1)-q-w-r), 2(mpn-1)-q-w-r, \lambda_A] + H [-c(\alpha/2, 2(mpn-1)-q-w-r), 2(mpn-1)-q-w-r, \lambda_A]. \quad (14)$$

Equivalently, the F -statistic has a non-central F -distribution with 1 degree of freedom in the numerator and $2(mpn - 1) - q - w - r$ degrees of freedom in the denominator and non-centrality parameter λ_A^2 . In two-level designs (e.g. students nested within schools) when r and q covariates are included at the first and second level respectively, the degrees of freedom of the t -test are $2(mn^* - 1) - q - r$, where $n^* = pn$.

Computational Example

This section provides power comparisons between two t -tests: the t -test carried out on level-3 unit means (with $2(m-1) - q$ degrees of freedom) and the t -test outlined in this paper (with $2(mpn-1) - q - w - r$ degrees of freedom). The power computations are presented in Table 1. I compute power (assuming one treatment and one control group) for two-tailed t -tests at the 0.05 significance level for an effect size $\delta = 0.5$, assuming no covariates are included in the model, and values of the intraclass correlations at the second level $\rho_2 = 0.07$, and at the third level $\rho_3 = 0.10$. To select plausible values of intraclass correlations I consulted Hedges & Hedberg (2007) and Nye, Konstantopoulos, and Hedges (2004).

Table 1 shows that the proposed test statistic has large power advantages when the number of level-3 units (schools) is small, compared to the test statistic carried out on level-3 unit means. For example, when two schools are assigned to a treatment and a control condition the power of the proposed test statistic is consistently twice as large. As the number of level-3 units (schools) increases however, the difference in power between the two tests decreases, and when the number of level-3 units becomes infinitely large the two tests provide identical estimates of power. Nonetheless, in practice this test is always more powerful since typically randomized experiments include a relatively small number of level-3 units. This replicates the results presented by Blair and Higgins for two-level designs.

Conclusion

Three-level designs are increasingly common in educational research. Experiments that involve multiple schools with multiple classrooms in each school are inherently three-level designs with two levels of nesting (assuming a three-stage sampling scheme). The present study proposed a more powerful test for treatment effects in three-level cluster randomized designs

where nesting occurs at two levels (e.g., classrooms and schools). This test statistic is more powerful than the typical test based on level-3 unit means because it preserves the degrees of freedom that are associated with level-2 and level-1 units. The proposed test assumes that the nesting effects (expressed as intraclass correlations) are known, which does not always hold. In education however, when the outcome is achievement, there is evidence from a large-scale randomized experiment (Project STAR) that the school-level intraclass correlation is on average 0.16 and the classroom-level intraclass correlation is on average 0.11 (see Nye et al., 2004). Analyses using NAEP data have provided comparable estimates. The increased use of three-level models in educational research should help with providing additional estimates of intraclass correlations. Note that the knowledge of the intraclass correlations is also necessary for the typical test based on level-3 unit means. Nonetheless, when educated guesses of the intraclass correlations are available the proposed test provides higher estimates of power, especially when the number of level-3 units is smaller.

References

- Barcikowski, R. S. (1981). Statistical power with 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.
- Bloom, H. S. (2005). *Learning more from social experiments*. New York: Russell Sage.
- Cochran, W. G. (1977). *Sampling techniques*. New York: Wiley.
- 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: 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.
- Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing.
- Konstantopoulos, S. (in press). The Power of the Test for Treatment Effects in Three-Level Cluster Randomized Designs. *Journal for Research on Educational Effectiveness*.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.
- Nye, B, Konstantopoulos, S., & Hedges, V. E. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trails. *Psychological Methods, 5*, 199-213.

Power Analysis in Three-Level Designs

Table 1. Power comparisons between a t-test based on means and the proposed t-test

Number of Level-3 Units	Number of Level-2 Units	Number of Level-1 Units	Power/ Means	Power/ Observations	df/ Means	df/ Observations
2	2	10	0.11	0.22	2	78
2	2	20	0.12	0.24	2	158
2	2	30	0.12	0.25	2	238
2	4	10	0.13	0.27	2	158
2	4	20	0.14	0.29	2	318
2	4	30	0.14	0.29	2	478
2	6	10	0.14	0.29	2	238
2	6	20	0.14	0.30	2	478
2	6	30	0.14	0.31	2	718
4	2	10	0.30	0.39	6	158
4	2	20	0.33	0.43	6	318
4	2	30	0.34	0.45	6	478
4	4	10	0.36	0.47	6	318
4	4	20	0.38	0.51	6	638
4	4	30	0.39	0.52	6	958
4	6	10	0.39	0.51	6	478
4	6	20	0.41	0.54	6	958
4	6	30	0.41	0.54	6	1438
6	2	10	0.46	0.54	10	238
6	2	20	0.51	0.59	10	478
6	2	30	0.53	0.61	10	718
6	4	10	0.56	0.64	10	478
6	4	20	0.59	0.68	10	958
6	4	30	0.60	0.69	10	1438
6	6	10	0.60	0.68	10	718
6	6	20	0.62	0.71	10	1438
6	6	30	0.63	0.72	10	2158
8	2	10	0.60	0.66	14	318
8	2	20	0.65	0.72	14	638
8	2	30	0.67	0.74	14	958
8	4	10	0.71	0.77	14	638
8	4	20	0.74	0.80	14	1278
8	4	30	0.75	0.81	14	1918
8	6	10	0.75	0.81	14	958
8	6	20	0.77	0.83	14	1918
8	6	30	0.78	0.83	14	2878