Institute for Policy Research
Northwestern University
Working Paper Series

WP-07-14

Correcting a Significance Test for Clustering in Designs With Two Levels of Nesting

Larry V. Hedges
Faculty Fellow, Institute for Policy Research
Board of Trustees Professor of Statistics and Social Policy
Northwestern University

DRAFT
Please do not quote or distribute without permission.

Abstract

A common mistake in analysis of cluster randomized experiments is to ignore the effect of clustering and analyze the data as if each treatment group were a simple random sample. This typically leads to an overstatement of the precision of results and anti-conservative conclusions about precision and statistical significance of treatment effects. This paper gives a simple correction to the *t*-statistic that would be computed if clustering were (incorrectly) ignored in an experiment with two levels of nesting (e.g., classrooms and schools). The correction is a multiplicative factor depending on the number of clusters and subclusters, the subcluster sample size, the subcluster size, and the cluster and subcluster intraclass correlations ρ_S and ρ_C. The corrected *t*-statistic has Student's *t*-distribution with reduced degrees of freedom. The corrected statistic reduces to the *t*-statistic computed by ignoring clustering when $\rho_S = \rho_C = 0$. It reduces to the *t*-statistic computed using cluster means when $\rho_S = 1$. If ρ_S and ρ_C are between 0 and 1, the adjusted *t*-statistic lies between these two and the degrees of freedom are in between those corresponding to these two extremes.

Note: This material is based upon work supported by the National Science Foundation under Grant No. 0129365 and IES under Grant No. R305U040003.

Correcting a Significance Test for Clustering in Designs With Two Levels of Nesting

Experiments in educational research often assign entire intact groups (such as schools or classrooms) to the same treatment group, with different intact groups assigned to different treatments. Because these intact groups correspond to statistical clusters, this design is often called a group randomized or *cluster randomized* design. Several analysis strategies for cluster randomized trials are possible, but the simplest is to use the cluster as the unit of analysis. This analysis involves computing mean scores on the outcome (and all other variables that may be involved in the analysis) and carrying out the statistical analysis as if the cluster means were the data. If all cluster sample sizes are equal, this approach provides exact tests for the treatment effect, but more flexible and informative analyses are also available, including analyses of variance using clusters as a nested factor (see, e.g., Hopkins, 1982) and analyses involving hierarchical linear models (see e.g., Raudenbush and Bryk, 2002). For general discussions of the design and analyses of cluster randomized experiments see Raudenbush and Bryk (2002), Donner and Klar (2000), Klar and Donner (2001), Murray (1998), or Murray, Varnell, & Blitstein (2004).

A common mistake in analysis of cluster randomized experiments in education is to analyze the data as if it were based on a simple random sample and assignment was carried out at the level of individuals. This typically leads to an overstatement of the precision of results and consequently to anti-conservative conclusions about precision and statistical significance of treatment effects (see, e.g., Murray, Hannan, and Baker, 1996). This analysis can also yield misleading estimates of effect sizes and incorrect estimates of their sampling uncertainty. If the raw data were available, then reanalysis using more appropriate analytic methods is usually desirable.

In some cases, however, the raw data is not available but one wants to be able to interpret the findings of a research report that improperly ignored clustering in the analysis. This problem often arises in reviewing the findings of studies carried out by other investigators. In particular, this problem has arisen in the work of the What Works Clearinghouse, a US Institute of Education Sciences funded project whose mission is to evaluate, compare, and synthesize evidence of effectiveness of educational programs, products, practices, and policies. The What Works Clearinghouse reviewers found that the majority of the high quality studies they were examining involved assignment of treatment by schools, which led to clustering that needed to be taken into account in assessing the uncertainty of the treatment effect (e.g., by computing confidence intervals) or in testing its statistical significance. While some of these studies sampled students directly within schools (at least roughly approximating a simple random sample within schools), most studies sampled students by first sampling classrooms within schools and thus there is a second level of clustering (nesting) that may need to be taken into account. Moreover, most of the statistical analyses in these studies did not attempt to take clustering into account. In this context, it would be desirable to be able to know how the conclusions about treatment effects might change if both levels of clustering were taken into account.

Another way to conceive the issue is in terms of survey sampling theory. In experiments that assign schools to treatments, treatment effects are just differences

between independent treatment group means. The variance of the treatment group means depends on the sampling design. If students are sampled by first selecting schools and then selecting classrooms within schools and then students within classrooms, the sampling design is a three-stage cluster sample with schools as clusters and classrooms as subclusters. Each stage of cluster sampling adds to the design effect (inflates the variance) of the treatment group mean. Ignoring these design effect (which is the equivalent to assuming that the sampling design is a simple random sample of students form the total population) leads to an underestimate of the variance of the treatment group means and therefore an underestimate of the variance of the treatment effect.

Designs involving two levels of clustering are widespread in education (e.g., designs that assign schools with multiple classrooms within schools to treatments). While methods are available to adjust for the effects of one level of clustering on simple tests of significance (e.g., Hedges, in press), less is known about methods for taking two levels of clustering into account. Such methods are likely to have wide application in education for two reasons. The first reason is the increasing prevalence of educational experiments that assign treatments to schools in order to avoid cross contamination of different treatments within the same school. The second reason is the practical fact that, since students are nested within classrooms and classrooms are nested within schools, it is easier to sample students by using a multistage cluster sampling plan that first samples schools and then classrooms. Such designs are therefore widely used in quasi-experiments as well as experiments.

Although we use the terms "schools" and "classrooms" to characterize the stages of clustering, this is merely a matter of convenience and readily understandable terminology. The results of this paper apply equally well to any situation in which there is a three-stage sampling design [where individual units are sampled by first sampling clusters (e.g., schools) and then sampling subclusters (e.g., classrooms) within the clusters, and finally sampling individual units (e.g., students) within the subclusters] and treatments are assigned to clusters (e.g., schools).

The purpose of this paper is to provide an analysis of the effects of two-levels of nesting (clustering) on significance tests and confidence intervals for treatment effects. First we derive the sampling distribution of the *t*-statistic under a clustered sampling model with equal cluster sample sizes. Then we provide and evaluate some simpler approximate methods for adjusting significance tests for the effects of clustering. Next we consider whether acceptable corrections may be obtained by adjusting for only one of the levels of nesting. Then we provide a generalization for unequal cluster (and subcluster) sample sizes. This research provides a simple correction that may be applied to a statistical test that was computed (incorrectly) ignoring the clustering of individuals within groups. The correction requires that a bound on the amount of clustering (in the form of an upper bound on the intraclass correlation parameters) is known or that the intraclass correlation parameters can be imputed for sensitivity analysis. We then derive confidence intervals for the mean difference based on the corrected test statistic. Finally we consider the power of the corrected test.

Model and Notation

Let $Y_{ijk}^{\ T}$ ($i = 1, \ldots, m^T$; $j = 1, \ldots, p_i^T$; $k = 1, \ldots, n_{ij}^{\ T}$) and $Y_{ijk}^{\ C}$ ($i = 1, \ldots, m^C$; $j = 1, \ldots, p_i^C$; $k = 1, \ldots, n_{ij}^{\ C}$) be the k^{th} observation in the j^{th} classroom in the i^{th} school in the treatment and control groups respectively. Thus, in the treatment group, there are m^T

schools, the i^{th} school has p_i^T classrooms, and the j^{th} classroom in the i^{th} school has n_{ij}^T observations. Similarly, in the control group, there are m^C schools, the i^{th} school has p_i^C classrooms, and the j^{th} classroom in the i^{th} school has n_{ij}^C observations. Thus there is a total of $M = m^T + m^C$ schools, a total of

$$P = \sum_{i=1}^{m^T} p_i^T + \sum_{i=1}^{m^C} p_i^C$$

classrooms, and a total of

$$N = N^T + N^C = \sum_{i=1}^{m^T} \sum_{j=1}^{p_i^T} n_{ij}^T + \sum_{i=1}^{m^C} \sum_{j=1}^{p_i^C} n_{ij}^C$$

observations overall.

Let $\overline{Y}_{i\bullet\bullet}^T$ ($i = 1, \ldots, m^T$) and $\overline{Y}_{i\bullet\bullet}^C$ ($i = 1, \ldots, m^C$) be the means of the i^{th} school in the treatment and control groups, respectively, let $\overline{Y}_{ij\bullet}^T$ ($i = 1, \ldots, m^T$; $j = 1, \ldots, k$) and $\overline{Y}_{ij\bullet}^C$ ($i = 1, \ldots, m^C$; $j = 1, \ldots, k$) be the means of the j^{th} class in the i^{th} school in the treatment and control groups, respectively, and let $\overline{Y}_{\bullet\bullet\bullet}^T$ and $\overline{Y}_{\bullet\bullet\bullet}^C$ be the overall means in the treatment and control groups, respectively. Define the (pooled) within treatment groups variance S^2 via

$$S^2 = \frac{\sum_{i=1}^{m^T} \sum_{j=1}^{p_i^T} \sum_{k=1}^{n_{ij}^T} (Y_{ijk}^T - \overline{Y}_{\bullet\bullet\bullet}^T)^2 + \sum_{i=1}^{m^C} \sum_{j=1}^{p_i^C} \sum_{k=1}^{n_{ij}^C} (Y_{ijk}^C - \overline{Y}_{\bullet\bullet\bullet}^C)^2}{N-2}. \quad (1)$$

Suppose that observations within the j^{th} subcluster (classroom) in the i^{th} cluster (school) within the treatment and control group groups are normally distributed about cluster (classroom) means μ_{ij}^T and μ_{ij}^C with a common within-cluster variance σ_{WC}^2. That is

$$Y_{ijk}^T \sim N(\mu_{ij}^T, \sigma_{WC}^2),\ i = 1, \ldots, m^T;\ j = 1, \ldots, p_i^T;\ k = 1, \ldots, n_{ij}^T$$

And (2)

$$Y_{ijk}^C \sim N(\mu_{ij}^C, \sigma_{WC}^2),\ i = 1, \ldots, m^C; j = 1, \ldots, p_i^C;\ k = 1, \ldots, n_{ij}^C.$$

Suppose further that the subcluster (classroom) means are random effects (for example they are considered a sample from a population of means) so that the class means themselves have a normal distribution about the school means $\mu_{i\bullet}^T$ and $\mu_{i\bullet}^C$ and common variance σ_{BC}^2. That is

$$\mu_{ij}^T \sim N(\mu_{i\bullet}^T, \sigma_{BC}^2),\ i = 1, \ldots, m^T; j = 1, ..., p_i^T$$

and (3)

$$\mu_{ij}^C \sim N(\mu_{i\bullet}^C, \sigma_{BC}^2),\ i = 1, \ldots, m^C; j = 1, ..., p_i^C.$$

Finally suppose that the cluster (school) means $\mu_{i\bullet}^T$ and $\mu_{i\bullet}^C$ are also normally distributed about the treatment and control group means $\mu_{\bullet\bullet}^T$ and $\mu_{\bullet\bullet}^C$ with common variance σ_{BS}^2. That is

$$\mu_{i\bullet}^T \sim N(\mu_{\bullet\bullet}^T, \sigma_{BS}^2),\ i = 1, \ldots, m^T$$

and (4)

$$\mu_{i\bullet}^{C} \sim N(\mu_{\bullet\bullet}^{C}, {\sigma_{BS}}^2),\ i = 1, \ldots, m^{C}.$$

Note that in this formulation, σ_{BC}^2 represents true variation of the population means of classrooms over and above the variation in sample means that would be expected from variation in the sampling of observations into classroom. Similarly, σ_{BS}^2 represents the true variation in school means, over and above the variation in sample means that would be expected from variation dues to the sampling of observations into schools.

These assumptions correspond to the usual assumptions that would be made in the analysis of a multi-site trial by a three-level hierarchical linear models analysis, an analysis of variance (with treatment as a fixed effect and schools and classrooms as nested random effects), or a *t*-test using the school means in treatment and control group as the unit of analysis.

Intraclass Correlations

In principle there are several different within-treatment group variances in a design with two levels of nesting (a three level design). We have already defined the within-classroom, between-classroom, and between-school, variances σ_{WC}^2, σ_{BC}^2, and σ_{BS}^2. There is also the total variance within treatment groups σ_{WT}^2 defined via

$$\sigma_T^2 = \sigma_{BS}^2 + \sigma_{BC}^2 + \sigma_{WC}^2 . \quad (5)$$

In most educational achievement data when clusters are schools and subclusters are classrooms, σ_{BS}^2 and σ_{BC}^2 are considerably smaller than σ_{WC}^2. Obviously, if the between school and classroom variances σ_{BS}^2 and σ_{BC}^2 are small, then σ_T^2 will be very similar to σ_{WC}^2.

In two-level models (e.g., those with schools and students as levels), the relation between variances associated with the two levels is characterized by an index called the intraclass correlation. In three-level models, two indices are necessary to characterize the relationship between these variances, and they are generalizations of the intraclass correlation. Define the school-level intraclass correlation ρ_S by

$$\rho_S = \frac{\sigma_{BS}^2}{\sigma_{BS}^2 + \sigma_{BC}^2 + \sigma_{WC}^2} = \frac{\sigma_{BS}^2}{\sigma_T^2} . \quad (6)$$

Similarly, define the classroom level intraclass correlation ρ_C by

$$\rho_C = \frac{\sigma_{BC}^2}{\sigma_{BS}^2 + \sigma_{BC}^2 + \sigma_{WC}^2} = \frac{\sigma_{BC}^2}{\sigma_T^2} . \quad (7)$$

These intraclass correlations can be used to obtain one of these variances from any of the others, since ${\sigma_{BS}}^2 = \rho_S{\sigma_T}^2$, ${\sigma_{BC}}^2 = \rho_C{\sigma_T}^2$, and ${\sigma_{WC}}^2 = (1 - \rho_S - \rho_C){\sigma_T}^2$.

Hypothesis Testing

The object of the statistical analysis may be to test the statistical significance of the intervention effect, that is, to test the hypothesis of no treatment effect

H_0: ${\mu_{\bullet\bullet}}^T = {\mu_{\bullet\bullet}}^C$.

The Test Statistic Ignoring Clustering

Suppose that the researcher wishes to test the hypothesis and carries out the usual *t*- or *F*-test. The *t*-test involves computing the test statistic

$$t = \frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet\bullet}^{T} - \bar{Y}_{\bullet\bullet\bullet}^{C})}{S} , \quad (8)$$

where S is the usual pooled within treatment group standard deviation defined in (1) and

$$\tilde{N} = \frac{N^T N^C}{N^T + N^C}.$$

The F-test statistic from a one-way analysis of variance ignoring clustering is of course $F = t^2$. If there is no clustering (that is, if $\rho_S = \rho_C = 0$), the test statistic t has Student's t-distribution with $N - 2$ degrees of freedom when the null hypothesis is true. If there is clustering (that is if either $\rho_S \neq 0$ or $\rho_C \neq 0$) the test statistic has a different sampling distribution—one that depends on ρ_S and ρ_C.

Note that this t-test (or the corresponding F-test) would not be computed if the analyst was properly addressing the clustered nature of the sample. As we noted above, other analyses that would be appropriate include analyses that include the clusters and subclusters as factors nested within treatments, analyses that use a hierarchical linear model including subclusters and clusters as level 2 and level 3 units, or use cluster means as the units of analysis. However, the objective of this paper is not to examine these analyses but to examine the effects of using (8) as a test statistic when the sample is a clustered sample.

When there is no clustering (that is when $\rho_S = \rho_C = 0$), the numerator of (8) has a normal distribution with standard deviation σ_T. In other words, when the null hypothesis is true

$$\sqrt{\tilde{N}}(\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet})/\sigma_T$$

has the standard normal distribution. Similarly, when there is no clustering (that is when $\rho_S = \rho_C = 0$), $(N - 2)S^2/\sigma_T^2$ is distributed as a chi-square with $(N - 2)$ degrees of freedom so that S^2 is distributed as σ_T^2 times a chi-square with $(N - 2)$ degrees of freedom. In other words S/σ_T is distributed as the square root of a chi-square with $(N - 2)$ degrees of freedom divided by its degrees of freedom. Note that the scale factor σ_T, which occurs in both the numerator and the denominator, cancels so that the ratio, t, is scale free. Because the numerator has the standard normal distribution and the denominator is the square root of the ratio of a chi-square with $(N - 2)$ degrees of freedom to its degrees of freedom that is independent of the numerator, the ratio in (8) has (by definition) Student's t-distribution with $(N - 2)$ degrees of freedom.

The Impact of Clustering

When there is clustering (either $\rho_S \neq 0$ or $\rho_C \neq 0$), neither the numerator nor the denominator of the t-statistic given in (8) has the same distribution as they do when either $\rho_S = \rho_C = 0$. We now indicate how the distribution of the numerator and denominator are different when $\rho_S \neq 0$ or $\rho_C \neq 0$ in the balanced design where the cluster sample sizes p_i^T and p_i^C are all equal to p and the subcluster sample sizes n_{ij}^T and n_{ij}^C are all equal to n.

Assuming that the design is balanced, the numerator has a normal distribution with mean 0, but with a generally larger variance: $\sigma_T^2[1 + (pn - 1)\rho s + (n - 1)\rho c]$. The factor $[1 + (pn - 1)\rho s + (n - 1)\rho c]$ is a generalization of Kish's (1965) design effect for two levels of nesting. In other words, when ρ_S or $\rho_C \neq 0$, and the null hypothesis is true

$$\sqrt{\tilde{N}}(\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet})/\boldsymbol{\sigma}_T\sqrt{1 + (pn-1)\rho_S + (n-1)\rho_C}$$

has the standard normal distribution.

Assuming a balanced design, the expected value of S^2 is no longer σ_T^2, but instead

$$E\left\{S^2\right\} = \sigma_{WC}^2 + \left(\frac{N-2pn}{N-2}\right)\sigma_{BS}^2 + \left(\frac{N-2n}{N-2}\right)\sigma_{BC}^2 = \sigma_T^2\left(1 - \frac{2(pn-1)\rho_S + 2(n-1)\rho_C}{N-2}\right).$$

Thus the scale factor necessary to standardize S is not σ_T. We show in the Appendix that

$$\frac{hS^2}{\sigma_T^2\left(1 - \frac{2(pn-1)\rho_S + 2(n-1)\rho_C}{N-2}\right)}$$

has, to an excellent approximation, the chi-square distribution with h degrees of freedom, where

$$h = \frac{\left[N - 2 - 2(pn-1)\rho_S - 2(n-1)\rho_C\right]^2}{pn\widehat{N}\rho_S^2 + n\breve{N}\rho_C^2 + (N-2)\overline{\rho}^2 + 2n\widehat{N}\rho_S\rho_C + 2\widehat{N}\rho_S\overline{\rho} + 2\breve{N}\rho_C\overline{\rho}}, \tag{9}$$

where $\widehat{N} = (N - 2pn)$, $\breve{N} = (N - 2n)$, and $\overline{\rho} = 1 - \rho_S - \rho_C$.

Taking the partial derivative of h with respect to ρ_S or ρ_C, we see that h is a decreasing function of ρ_S and ρ_C. If $\rho_S = \rho_C = 0$ and there is no clustering, $h = (N - 2)$ and S has the nominal degrees of freedom as expected. If $\rho_S = 1$ (so that $\rho_C = 0$) and there is complete clustering by school (no variability within clusters), then $h = (M - 2)$ as expected (because the only variability is that between the M clusters). If $\rho_C = 1$ (so that $\rho_S = 0$) and there is complete clustering by classroom (no variability within subclusters or between clusters), then $h = (Mp - 2)$ as expected (because the only variability is that between the Mp subclusters). If $0 < \rho_S < 1$ and $0 < \rho_C < 1$, then h is between $(M - 2)$ and $(N - 2)$ and its value reflects the effective degrees of freedom in S.

These results imply that when either $\rho_S \neq 0$ or $\rho_C \neq 0$, S/σ_T is no longer distributed as the square root of a chi-square with $(N - 2)$ degrees of freedom divided by its degrees of freedom, but

$$\frac{S}{\sigma_T\sqrt{1 - \frac{2(pn-1)\rho_S + 2(n-1)\rho_C}{N-2}}}$$

is distributed as the square root of a chi-square with h degrees of freedom divided by its degrees of freedom.

The Sampling Distribution of the *t*-Statistic When Either $\rho_S \neq 0$ or $\rho_C \neq 0$

The results in the previous section imply that when either $\rho_S \neq 0$ or $\rho_C \neq 0$, the statistic

$$\frac{\sqrt{\tilde{N}}(\overline{Y}_{\bullet\bullet\bullet}^T - \overline{Y}_{\bullet\bullet\bullet}^C)/\sigma_T\sqrt{1 + (pn-1)\rho_S + (n-1)\rho_C}}{S/\sigma_T\sqrt{1 - \frac{2(pn-1)\rho_S + 2(n-1)\rho_C}{N-2}}} = c\frac{\sqrt{\tilde{N}}(\overline{Y}_{\bullet\bullet\bullet}^T - \overline{Y}_{\bullet\bullet\bullet}^C)}{S} = ct$$

has the t-distribution with h degrees of freedom, where c is a constant depending on N, p, n, ρ_S, and ρ_C that absorbs the ratios of the scale factors in numerator and denominator, which given by

$$c = \sqrt{\frac{N - 2 - 2(pn-1)\rho_S - 2(n-1)\rho_C}{(N-2)\left[1 + (pn-1)\rho_S + (n-1)\rho_C\right]}} \tag{10}$$

Thus the statistic

$$t_A = ct \tag{11}$$

has the *t*-distribution with *h* degrees of freedom and can be thought of as a *t*-statistic adjusted for both for clustering effects on the mean difference and on the standard deviation.

Thus a two-sided test of the null hypothesis of equal group means consists of rejecting H_0 if $| t_A |$ exceeds the 100α percent two-tailed critical value of the *t*-distribution with *h* degrees of freedom. The one sided test rejects H_0 on the positive side if t_A exceeds the 100α percent one-tailed critical value of the *t*-distribution with *h* degrees of freedom.

Note that if $\rho_S = 0$ and $\rho_C = 0$ so that there is no clustering, then $c = 1$ and $h = N - 2$. That is, when $\rho_S = 0$ and $\rho_C = 0$, the test based on t_A reduces to the usual *t*-test ignoring clustering. When $\rho_S = 1$ and $\rho_C = 0$ and there is complete clustering by school, then $c = \sqrt{(M-2)/(N-2)}$ and $h = M - 2$. That is, when $\rho_S = 1$ and $\rho_C = 0$, and the test based on t_A reduces to a *t*-test computed using the cluster (school) means. Note that when $\rho_S = 0$ and $\rho_C = 1$, $c = \sqrt{(Mp-2)/(N-2)}$ and $h = Mp - 2$, so that the test based on t_A reduces to a *t*-test computed using the subcluster (classroom) means.

The sampling distribution of t_A is not exact, but it is based on theory that yields a very good approximation (see, e.g., Welch, 1949; Welch, 1956; Gaylor and Hopper 1969) and is widely used in other settings to construct tests in complex analyses of variance, such as unbalanced between-subjects designs and repeated measures designs (see, e.g., Geisser and Greenhouse, 1958). Extensive simulation experiments in connection with two-level designs found the rejection rates of the corresponding test to be indistinguishable from nominal (see Hedges, in press). Our simulation results in three level designs (not reported here) also confirm that rejection rates do not appear to differ from nominal.

One immediate application of the results in this paper is to study the rejection rate of the unadjusted *t*-test. While it is well known that the unadjusted *t*-test has a rejection rate that is often much higher than nominal (see, e.g., Murray, Hannan, and Baker, 1996), previous studies have relied on simulation to study this test. The sampling distribution of t_A provides an analytic expression for the rejection rates of the unadjusted *t*-test under the cluster sampling model. Let $t(v, \alpha)$ be the level α two-sided critical value for the *t*-distribution with *v* degrees of freedom. Then the usual unadjusted *t*-test rejects if $|t| > t(N - 2, \alpha)$. Because $t_A = ct$ has the *t*-distribution with *h* degrees of freedom under the null hypothesis, the rejection rate of the unadjusted test is

$$2\left\{1 - F\left[ct((N-2),\alpha),\ h\right]\right\}, \qquad (12)$$

where $F[x, v]$ is the cumulative distribution function of the *t*-distribution with *v* degrees of freedom. Computations with this expression (not reported in this paper) are very consistent with the empirical rejection rates obtained in our simulations.

Relation to Previous Work

The properties of significance tests in designs with two-levels of nesting were discussed by Murray, Hannan, and Baker (1996). In one part of their paper, they provided results of Monte Carlo studies of rejection rates of the naïve test that ignored clustering (the test based on the statistic F_{ind} with degrees of freedom ddf_{ind} in their notation). The rejection rates computed using the methods in this paper agree well with their results. Table 1 gives the values computed using the methods in this paper and the results given in Table 1 of Murray, Hannan, and Baker (1996) for F_{ind} with degrees of freedom ddf_{ind}. All of these results based on this paper are within two standard errors of

the empirical proportion obtained in the simulation, and all but one are within one standard error.

The sampling distribution of t_A derived in this paper provide some insight about other approaches to testing mean differences in clustered samples. For designs with a single level of clustering, Kish (1965) suggested multiplying S (or, equivalently, dividing the t-statistic) by the square root of the design effect to remove the effect of clustering on the numerator of the t-statistic. The generalization of that suggestion would be to divide the t-statistic by the square root of $[1 + (pn - 1)\rho s + (n - 1)\rho c]$, yielding the statistic is

$$t_K = \frac{\sqrt{\tilde{N}}(\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet})}{S\sqrt{1+(pn-1)\rho_S+(n-1)\rho_C}}.$$

However because this statistic is does not correct for the fact that the scale factor necessary to standardize S_{WT} is not σ_T, the sampling distribution of t_K is not a t-distribution but a constant times a t-distribution with h degrees of freedom, namely

$$t_K = \frac{t_A}{\sqrt{1-\frac{2(pn-1)\rho_S+2(n-1)\rho_C}{N-2}}}. \qquad (13)$$

If $\rho_S \neq 0$ or $\rho_C \neq 0$ the denominator of (13) is less than one, so $t_K > t_A$. However note that the denominator of (13) will be quite close to 1 unless m is small and ρ_S is large. For example, if $\rho_S = 0.25$, $\rho_C = 0.15$, $n = 30$, $p = 3$ and $m = 2$, the denominator of (13) is about 0.925, but if $n = 30$, $p = 3$, and $m = 10$, the denominator is 0.986. Therefore the sampling distribution of t_K is approximately a t-distribution with h degrees of freedom.

One might wish to avoid the computation of h by using a simpler approximation for the degrees of freedom that is used to obtain a critical value for the test using t_K. Obvious possibilities for degrees of freedom include the degrees of freedom based on the number of individuals, namely $(N - 2)$; degrees of freedom based on the number of schools, namely $(M - 2)$; and the effective degrees of freedom reduced by the design effect, namely $(N - 2)/[1 + (pn - 1)\rho s + (n - 1)\rho c]$. Table 2 shows the actual rejection rates for two-sided tests at the $\alpha = 0.05$ significance level for the naïve test that ignores clustering and for tests using the statistic t_K with critical values based on $(N - 2)$, $(M - 2)$, and $(N - 2)/[1 + (pn - 1)\rho s + (n - 1)\rho c]$ degrees of freedom for plausible situations. The eighth column of the table, which gives the results of the naïve test ignoring clustering, shows that the effects of two levels of clustering can be profound. It shows that the actual rejection rates for the 5 percent test under the null hypothesis are as large as 70 percent. Note that the test based on statistic t_K using $(N - 2)$ degrees of freedom is liberal, rejecting more often than its nominal rate of 5 percent, particularly when the number M of clusters is small. The test based on statistic t_K using $(M - 2)$ degrees of freedom is conservative, rejecting less often than its nominal rate of 5 percent, and is very conservative when the number M of clusters is small. In contrast, the test based on statistic t_K using $(N - 2)/[1 + (pn - 1)\rho s + (n - 1)\rho c]$ degrees of freedom is sometimes slightly liberal, sometimes slightly conservative, but generally has a level very close to the nominal 5 percent.

Unequal Cluster Sample Sizes

When cluster sample sizes are unequal, the expression for the sampling distribution of the t-test statistic from clustered samples and is considerably more complex. In this section we give the sampling distribution of the usual t-statistic and a

statistic that is adjusted for the effects of clustering when cluster sample sizes are not equal. These expressions may be of use when cluster sample sizes are unequal and are reported explicitly. They also give some insight about what single "compromise" value of p or n might give most accurate results when substituted into the equal sample size formulas for rough approximations.

The expressions are quite complex when subcluster sample sizes are unequal. Consequently we provide expressions for the adjusted t-statistic and its degrees of freedom when the subcluster sizes are equal, but the cluster sizes are unequal. Then we give expressions when the subcluster sample sizes are unequal.

Unequal Cluster (School) Sample Sizes but Equal Subcluster (Classroom) Sizes

In this section we consider the case when the subcluster (classroom) sample sizes are equal or nearly so, but clusters differ in the number of subclusters (e.g., schools have different numbers of classrooms). That is we assume that the subcluster sample sizes n_{ij}^T and n_{ij}^C are all equal to n, but the number of treatment and control group clusters (m^T and m^C) may differ and the number of subclusters within each treatment and control group clusters (p_i^T and p_i^C) may also differ.

This situation is of interest for several reasons. First, as a practical matter, schools that are sampled in research studies have different numbers of classrooms, but the classroom sample sizes are equal or approximately equal (see, e.g., Ridgeway, et al., 2000). Second, the adjustment to the t-statistic and the degrees of freedom depend much more on cluster (school) sample sizes than on subcluster (classroom) sample sizes. Therefore adjustment for unequal classroom sample sizes is a second order correction to both test statistic and degrees of freedom, so treating the subcluster sample sizes as equal when they are not quite equal has relatively little effect. Third, the subcluster sample sizes are much less likely to be reported than the cluster sample sizes, so these expressions are more likely to be of practical use. Finally, the expressions for the adjustment and the degrees of freedom are much simpler when subcluster sample sizes are equal.

When the number of clusters is unequal, the adjusted t-statistic that is a generalization of (11) becomes

$$t_{AU} = c_U t \tag{14}$$

where the adjustment constant c_U is given by

$$c_U = \sqrt{\frac{(N-2) - 2(\bar{p}_U n - 1)\rho_S - 2(n-1)\rho_C}{(N-2)\left[1 + (\tilde{p}_U n - 1)\rho_S + (n-1)\rho_C\right]}}, \tag{15}$$

where

$$\bar{p}_U = \frac{n\sum_{i=1}^{m^T}\left(p_i^T\right)^2}{2N^T} + \frac{n\sum_{i=1}^{m^C}\left(p_i^C\right)^2}{2N^C} \tag{16}$$

and

$$\tilde{p}_U = \frac{N^C n\sum_{i=1}^{m^T}\left(p_i^T\right)^2}{N^T N} + \frac{N^T n\sum_{i=1}^{m^C}\left(p_i^C\right)^2}{N^C N}. \tag{17}$$

Note that if all the p_i^T and p_i^C are equal to p, then $\bar{p}_U = p$, $\tilde{p}_U = p$, and expression (15) for c_U reduces to expression (10) for c.

The statistic t_{AU} has Student's *t*-distribution with *h* degrees of freedom, where h_U is given by

$$h_U = \frac{\left[N - 2 - 2(\overline{p}_U n - 1)\rho_S - 2(n-1)\rho_C\right]^2}{A\rho_S^2 + n\breve{N}\rho_C^2 + (N-2)\overline{\rho}^2 + 2n\hat{N}_U\rho_S\rho_C + 2\hat{N}_U\rho_S\overline{\rho} + 2\breve{N}\rho_C\overline{\rho}} \tag{18}$$

where $\hat{N}_U = (N - 2\overline{p}_U n)$, $\breve{N} = (N - 2n)$, and $\overline{\rho} = 1 - \rho_S - \rho_C$ and the auxiliary constant *A* is defined via $A = A^T + A^C$ and

$$A^T = \frac{n^2\left(N^T\right)^2\sum_{i=1}^{m^T}\left(p_i^T\right)^2 + n^4\left(\sum_{i=1}^{m^T}\left(p_i^T\right)^2\right)^2 - 2n^3N^T\sum_{i=1}^{m^T}\left(p_i^T\right)^3}{\left(N^T\right)^2},$$

$$A^C = \frac{n^2\left(N^C\right)^2\sum_{i=1}^{m^C}\left(p_i^C\right)^2 + n^4\left(\sum_{i=1}^{m^C}\left(p_i^C\right)^2\right)^2 - 2n^3N^C\sum_{i=1}^{m^C}\left(p_i^C\right)^3}{\left(N^C\right)^2}, \tag{19}$$

$$P^T = \sum_{i=1}^{m^T} p_i^T \ ,$$

and

$$P^C = \sum_{i=1}^{m^C} p_i^C \ .$$

Note that when the $p_i{}^T$ and $p_i{}^C$ are all equal to *p*, then $\overline{p}_U = p$, $A = pn(N - 2pn)$, expression (18) for h_U reduces to expression (9) for *h*.

Unequal Subcluster (Classroom) Sample Sizes

The exact expression for the degrees of freedom *h* is quite complex when subcluster (classroom) sample sizes are unequal. The complexity of the expression is not unexpected. The denominator of *h* is the variance of a linear combination of three correlated variance component estimates, and the variances and covariances of these variance component estimates are themselves quite complex in unbalanced designs with two nested factors (see e.g., Searle, 1971, pp. 475 - 477). To obtain reasonably compact expressions, it is useful to definite several auxiliary constants, which are given in Table 3.

When the sample size in the subclusters is unequal, the adjusted *t*-statistic that is a generalization of (11) becomes

$t_{AU} = c_U t$

where the adjustment constant c_U is given by

$$c_U = \sqrt{\frac{(N-2) - 2(k_1 - 1)\rho_S - 2(k_3 - 1)\rho_C}{(N-2)\left[1 + (\tilde{k}_1 - 1)\rho_S + (\tilde{k}_3 - 1)\rho_C\right]}} \tag{20}$$

where $k_1 = k_1{}^T + k_1{}^C$, $k_3 = k_3{}^T + k_3{}^C$,

$$\tilde{k}_1 = \frac{N^C k_1^T + N^T k_1^C}{N^T + N^C},$$

$$\tilde{k}_3 = \frac{N^C k_3^T + N^T k_3^C}{N^T + N^C},$$

and the auxiliary constants k_1^T , k_1^C, k_3^T , and k_3^C are defined in Table 2. Note that if all the p_i^T and p_i^C are equal to p and if all the n_{ij}^T and n_{ij}^C are equal to n, then $k_1 = pn$ and $k_3 = n$, and expression (20) for c_U reduces to expression (10) for c.

When the null hypothesis is true, the statistic t_{AU} has Student's t-distribution with h degrees of freedom, where h_U is given by

$$h_U = \frac{\left[N - 2 - 2(k_1 - 1)\rho_S - 2(k_3 - 1)\rho_C\right]^2}{(N-2)\overline{\rho}^2 + B\rho_S^2 + C\rho_C^2 + 2D\rho_S\rho_C + 2E\rho_S\overline{\rho} + 2F\rho_C\overline{\rho}} \qquad (21)$$

where $\overline{\rho} = 1 - \rho_S - \rho_C$, and $B = B^T + B^C$, $C = C^T + C^C$, $D = D^T + D^C$, $E = E^T + E^C$, and $F = F^T + F^C$ are defined below. In the definition below, the T and C superscripts denoting the Treatment and Control groups are omitted for simplicity. Thus, the definition below gives the value of the constants B, C, D, E, and F within each treatment group (B^T, C^T, etc.) in terms of auxiliary constants k_1 to k_9 given in Table 2:

$B = [k_1(N + k_1) - 2k_9/N]$,

$C = \{2k_3[N(k_{12} - k_3)^2 + k_3(N - k_{12})^2] + 2(N - k_3)^2(2k_7 + Nk_3 - 2k_5)$

$- 4(N - k_3)(k_{12} - k_3)(k_7 + Nk_3 - k_5) + 4(N - k_3)(N - k_{12})(k_5 - k_7 - k_4/N)$

$+ 4(N - k_{12})(k_{12} - k_3)k_4/N\}/\{(N - k_{12})^2\}$,

$D = [k_3(N + k_1) - 2k_8/N]$,

$E = [N - k_1]$

$F = [N - k_3]$.

Note that when the p_i^T and p_i^C are all equal to p, and all the the n_{ij}^T and n_{ij}^T are equal to n then expression (21) for h_U reduces to expression (9) for h.

Confidence Intervals

Confidence intervals based on the standard error of the mean difference and using the critical values used in the test based on t assuming simple random sampling will not be accurate when either $\rho_S \neq 0$ and $p > 1$ or $\rho_C \neq 0$ and $n > 1$. That is, the actual probability content of these confidence intervals will usually be smaller than nominal (the confidence intervals will be too short). The corrected t-statistic t_A can be used to obtain confidence intervals that will have the correct probability content.

A $100(1 - \alpha)$ percent confidence interval for the treatment effect $\mu_{\bullet\bullet}^T - \mu_{\bullet\bullet}^C$ is given by

$$(\bar{Y}_{\bullet\bullet\bullet}^T - \bar{Y}_{\bullet\bullet\bullet}^C) - \mathrm{t}(\alpha, h)\, S / c\sqrt{\tilde{N}} \le \mu_{\bullet\bullet}^T - \mu_{\bullet\bullet}^C \le (\bar{Y}_{\bullet\bullet\bullet}^T - \bar{Y}_{\bullet\bullet\bullet}^C) + \mathrm{t}(\alpha, h)\, S / c\sqrt{\tilde{N}} \ , \qquad (22)$$

where c is the constant defined in (10) if the cluster and subcluster sample sizes, respectively are equal or the constant c_U defined in (15) or (20) if they are unequal and $\mathrm{t}(\alpha;\nu)$ is the 100α percent two-sided critical value of the t-distribution with ν degrees of freedom (e.g., if $\alpha = 0.05$ and $\nu = 120$, then $\mathrm{t}(\alpha, \nu) = 1.98$).

Example

An evaluation of the connected mathematics curriculum reported by Ridgway, et al. (2002) compared the achievement of $p^T = 2$ classrooms of 6th grade students who used connected mathematics in each of $m^T = 9$ schools with that of $p^C = 1$ classroom in each of

$m^C = 9$ schools in a comparison group that did not use connected mathematics. In this quasi-experimental design the clusters were schools and the subclusters were classrooms. The class sizes were not identical but the average class size in the treatment group was $N^T/m^T = 338/18 = 18.8$ and $N^C/m^C = 162/18 = 18$ in the control group. The exact sizes of all the classes were not reported, but here we treat the subcluster sizes as if they were equal and choose $n = 18$ as a slightly conservative sample size. The mean difference between treatment and control groups is $\bar{Y}^T_{\bullet\bullet\bullet} - Y^C_{\bullet\bullet\bullet} = -1.5$, the pooled within-groups standard deviation $S_{WT} = 2.436$. This evaluation involved sites in all regions of the country and it was intended to be nationally representative. Ridgeway et al. did not give an estimate of the intraclass correlation based on their sample. Hedges and Hedberg (2007) provide an estimate of the school level grade 6 intraclass correlation in mathematics achievement for the nation as a whole (based on a national probability sample) of 0.264. Therefore for this example we assume that the intraclass correlation at the school level is $\rho_S = 0.264$ and that the classroom level intraclass correlation is about two thirds as large, namely $\rho_C = 0.176$.

The analysis carried out by the investigators ignored clustering. Comparing the mean of all of the students in the treatment group with the mean of all of the students in the control group using a conventional *t*-test leads to an unadjusted *t* value of $t = 6.399$, which is highly statistically significant compared with a critical value based on $(N - 2) = 500 - 2 = 498$ degrees of freedom or $486 - 2 = 484$ degrees of freedom using our slightly conservative assumption that classrooms had an equal sample size of $n = 18$.

To determine what impact clustering may have had on the statistical significance of these findings we compute the adjusted *t*-test. We start by computing $\bar{p}_U$ using (16) and $\tilde{p}_U$ from (17) we obtain $\bar{p}_U = 1.5$ and $\tilde{p}_U = 1.33$. Inserting these values into the expression (15) for c_U yields $c_U = 0.309$ and a *t*-statistic adjusted for clustering of $t_{AU} = 1.976$, which is much smaller than the unadjusted *t*-statistic. To compute the degrees of freedom for the adjusted test, we first compute the auxiliary constant A using (19) and obtain $A = 12{,}960$, then we insert this value of A along with $\widehat{N} = 432$ and $\breve{N} = 450$ into (18) to obtain $h_U = 96.02$. Comparing the value of the adjusted statistic, $t_{AU} = 1.976$, with Student's *t*-distribution with $h = 96.02$ degrees of freedom, we see that the two-tailed *p*-value is $p = 0.051$. Thus a conventional interpretation would be that the result is not quite statistically significant at the 5 percent level. A 95 per cent confidence interval for $\mu_{\bullet\bullet}{}^T - \mu_{\bullet\bullet}{}^C$ computed from (22) is given by

$$-3.007 \le \mu_{\bullet\bullet}{}^T - \mu_{\bullet\bullet}{}^C \le 0.007,$$

which has width 3.014, and as expected from the outcome of the significance test, contains zero. Comparing this to the confidence interval that would be computed ignoring clustering, (-1.96 to -1.04) which has width 0.92, we see that the confidence interval which ignores clustering is considerably (and erroneously) narrower than that using t_A, which takes clustering into account.

This example illustrates that a finding that implies treatment effects that may seem very reliably different from zero when the analysis ignores clustering may be equivocal when clustering is taken into account. The adjustment used in this example involves assumptions about intraclass correlations that may not be exactly correct. It should be viewed more as a sensitivity analysis than as a sharp estimate of actual significance values. (For example, if the value of ρ_S was decreased to $\rho_S = 0.25$, the

adjusted *t*-test would yield a *p*-value less than 0.05.) However the assumptions made in this example are likely to be more plausible than the assumption that $\rho_S = \rho_C = 0$ that corresponds to the idea that clustering can be safely ignored.

This example also illustrates that when the sampling design in an experiment involves a three stage sample with two levels of clustering (nesting), such as sampling students by first selecting schools, then classrooms within schools, then students within classrooms, it is important to include all of the levels of nesting in adjustments for clustering. If we had ignored the clustering at the classroom level (or equivalently assumed that $\rho_C = 0$) and continued to assume that $\rho_S = 0.264$, then we would have calculated a value of $c_U = 0.371$ and an adjusted *t*-statistic of $t_{AU} = 2.372$ with $h = 165.87$ degrees of freedom and a *p*-value of $p = 0.019$. Thus we would have concluded that that the treatment effect was still reliably different from zero, even after adjusting for clustering at the school level.

Power Considerations

In evaluating any statistical test, it is useful to know its power relative to alternative tests that might be used. The corrected *t*-test presented in this paper is likely to be used in situations where there is no obvious alternative (that is in situations where only a data summary such as a *t*-statistic computed ignoring clustering is available). Yet it is still useful to know something about the power of this test compared with that of the alternatives that could be used if more data were available.

Two alternatives that require more information than the test given here, but which may be computed without complete reanalysis of the data, are a *t*-test performed on cluster (school) means (that is using the school as the unit of analysis) and a generalized least squares (GLS) analysis computed using known values of ρ_S and ρ_C to parameterize the error covariance matrix. Blair and Higgins (1986) give the two level version of the test based on GLS, but its extension to three levels is straightforward. These two tests provide useful standards of comparison because the test based on cluster (school) means is the most powerful exact test when both ρ_S and ρ_C are unknown, while the test based on generalized least squares is the most powerful exact test when both ρ_S and ρ_C are known.

When the null hypothesis is false (and the design is balanced), the test statistic used in all three analyses (the one based on the results in this paper, and the two alternatives requiring more data) have noncentral *t*-distributions with the same noncentrality parameter,

$$\lambda = \frac{\sqrt{\tilde{N}}\left(\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C}\right)}{\sigma_T} \sqrt{\frac{1}{1+(pn-1)\rho_S + (n-1)\rho_C}}, \tag{23}$$

but different degrees of freedom [$(N-2)$, h, or $(M-2)$, respectively]. Because the power is an increasing function of degrees of freedom for a fixed noncentrality parameter the relative power of these three tests is therefore determined by the degrees of freedom. Because the analysis based on generalized least squares has $(N-2)$ degrees of freedom and $(N-2) \geq h \geq (M-2)$, it will provide the most powerful test if ρ is known and the raw data are available. Because the analysis based on school means has $(M-2)$ degrees of freedom and $(M-2) \leq h \leq (N-2)$, it should always provide the least powerful of the three tests. Because the test based on t_A has h degrees of freedom, it should have power in between the other two tests. However, because the dependence of the power function on degrees of freedom for a fixed noncentrality parameter) is slight when degrees of

freedom are 30 or more, the difference in the power of these three tests need not be substantial.

Table 4 gives the power of each of the three tests in some illustrative situations when $\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C} = 1.0\sigma_T$, and the last column is the ratio of the power of the test proposed here to that of the test based on generalized least squares. This table illustrates that when the number of clusters is small, the adjusted *t*-test is considerably more powerful than the test using cluster means as the unit of analysis, but the power advantage decreases as the number of clusters increases. However it is important to remember that the test based on cluster means is the most powerful test if ρ_S and ρ_C are unknown. That is, the power advantage of the GLS test and the adjusted *t*-test depends on having known values of ρ_S and ρ_C. While the adjusted *t*-test is slightly less powerful than the GLS test, it is very nearly as powerful.

Conclusions

Cluster randomized trials are important in education and the social and policy sciences, but these trials are often improperly analyzed by ignoring the effects of clustering on significance tests. It is obviously desirable that these trials should be analyzed using more appropriate statistical methods (such as multilevel statistical methods). However, when conclusions must be drawn from published reports (using *t*- or *F*-tests that ignore clustering), corrected significance levels and confidence intervals can be obtained if the intraclass correlations are known or plausible values can be imputed. Such procedures provide reasonably accurate significance levels and are suitable for bounds on the results.

The theory given in this paper can also be used to study alternative suggestions for adjusting *t*-tests for clustering. Such analyses show that a test based on Kish's statistic t_K gives quite conservative results when critical values are obtained using degrees of freedom based strictly on the number of clusters. A test based on t_K has rejection rates that are generally close to nominal (but not always strictly conservative) when critical values are obtained using degrees of freedom adjusted for the design effect involving both levels of clustering.

When using the adjustments to test statistics given in this paper, it is important to adjust for both levels of clustering. Ignoring one of the levels of nesting (clustering) in computing the adjusted *t*-statistic or t_K can result in substantial inflation of significance levels.

This paper considered only the simplest analyses for treatment effects under a sampling model with two levels of nesting. Educational experiments sometimes involve the use of covariates at one or more levels of the design to increase precision. The generalization of the methods used in this paper to more complex designs and more complex analyses would be desirable to provide methods for dealing with such cases.

Appendix

Derivations with the Equal Cluster and Subcluster Sample Sizes

Under the model the sampling distribution of the numerator of (8) is normal with mean $\sqrt{\tilde{N}}(\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C})$ and variance $\sigma_W^2 + pn\sigma_{BS}^2 + n\sigma_{BC}^2 = \sigma_T^2[1 + (pn - 1)\rho_S + (n - 1)\rho_C]$. The square of the denominator of (8), can be written as

$$S^2 = \frac{SSBS + SSBC + SSWC}{N - 2}, \tag{24}$$

where *SSBS* is the pooled sum of squares between cluster (school) means within treatment groups, *SSBC* is the pooled sum of squares between subcluster (classroom) means within schools and treatment groups, and *SSWC* is the pooled sum of squares within subclusters (classrooms). Therefore $SSWC/\sigma_{WC}^2$ has the chi-squared distribution with $(N - Mp)$ degrees of freedom, where $M = m^T + m^C$. Similarly

$$\frac{SSBC}{\sigma_{WC}^2 + n\sigma_{BC}^2} \tag{25}$$

has the chi-squared distribution with $(Mp - M)$ degrees of freedom and

$$\frac{SSBS}{\sigma_{WC}^2 + n\sigma_{BC}^2 + pn\sigma_{BS}^2} \tag{26}$$

has the chi-squared distribution with $(M - 2)$ degrees of freedom.

Thus S^2 is a linear combination of independent chi-squares. To obtain the sampling distribution of S^2, we use a result of Box (1954), which gives the sampling distribution of quadratic forms in normal variables in terms of the first two cumulants of the quadratic form. Theorem 3.1 in Box (1954) implies that S^2 is distributed to an excellent approximation as a constant g times chi-square with h degrees of freedom, where g and h are given by

$$g = \frac{\mathrm{V}\{S^2\}}{2\mathrm{E}\{S^2\}} \tag{27}$$

and

$$h = \frac{2\left(\mathrm{E}\{S^2\}\right)^2}{\mathrm{V}\{S^2\}}, \tag{28}$$

where $\mathrm{E}\{X\}$ and $\mathrm{V}\{X\}$ are the expected value and the variance of X. Therefore we have that $S^2/gh = S^2/\mathrm{E}\{S^2\}$ is distributed as a chi-square with h degrees of freedom divided by h.

By the definition of the noncentral t-distribution (see, e.g., Johnson and Kotz, 1970), it follows that

$$\frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^{T} - \bar{Y}_{\bullet\bullet}^{C})/\sigma_T\sqrt{1 + (pn-1)\rho_S(n-1)\rho_C}}{S/\sigma_T\sqrt{E\left\{S^2\right\}}} = ct$$

has the noncentral t-distribution with h degrees of freedom and noncentrality parameter

$$\lambda = \frac{\sqrt{\tilde{N}}(\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C})}{\sigma_T\sqrt{1 + (pn-1)\rho_S + (n-1)\rho_C}},$$

where c is given by

$$c = \frac{\sqrt{E\left\{S^2\right\}/\sigma_T^2}}{\sqrt{1+(n-1)\rho}} \tag{29}$$

and h is given by (28). When $\mu^T - \mu^C = 0$ (and therefore $\lambda = 0$), the distribution is a central t-distribution with h degrees of freedom.

It follows from (24), and standard theory for expected mean squares in hierarchical designs (see, e.g., Kirk, 1995) that

$$\mathrm{E}\left\{S^2\right\} = \sigma_{WC}^2 + \left(\frac{N-2n}{N-2}\right)\sigma_{BC}^2 + \left(\frac{N-2pn}{N-2}\right)\sigma_{BS}^2$$

and

$$\mathrm{V}\{S^2\} = \frac{2\left[\, pn\widehat{N}\sigma_{BS}^4 + n\breve{N}\sigma_{BC}^4 + (N-2)\sigma_{WC}^4 + 2n\widehat{N}\sigma_{BS}^2\sigma_{BC}^2 + 2\widehat{N}\sigma_{BS}^2\sigma_{WC}^2 + 2\breve{N}\sigma_{BC}^2\sigma_{WC}^2 \,\right]}{(N-2)^2},$$

where $\widehat{N} = (N - 2pn)$, $\breve{N} = (N - 2n)$, and $\overline{\rho} = 1 - \rho_S - \rho_C$. Inserting these values for the mean and variance of S^2 into (27) and (28), using the fact that $\rho_S\sigma_T^2 = \sigma_{BS}^2$, $\rho_C\sigma_T^2 = \sigma_{BC}^2$ and $(1 - \rho_S - \rho_C)\sigma_T^2 = \sigma_{WC}^2$, and simplifying gives the values we obtain for c given in (10) and h given in (9).

Unequal Cluster Sample Sizes

When cluster sample sizes are unequal but samples sizes in subclasses are equal, expressions for the expressions for the constant c and degrees of freedom h are more complex. A direct argument leads to

$$\mathrm{V}\left\{\overline{Y}_{\bullet\bullet\bullet}^T - \overline{Y}_{\bullet\bullet\bullet}^C\right\} = \left(\frac{N^T N^C}{N^T + N^C}\right)^{-1}\left(\sigma_{WC}^2 + n\sigma_{BC}^2 + \tilde{p}_U n\sigma_{BS}^2\right) \tag{30}$$

where $\tilde{p}_U$ is defined in (17). Therefore the sampling distribution of the numerator of (8) is normal with mean $\sqrt{\tilde{N}}(\mu_{\bullet\bullet}^T - \mu_{\bullet\bullet}^C)$ and variance $\sigma_W^2 + \tilde{n}\,\sigma_B^2 = \sigma_T^2[1 + (\tilde{p}_U n - 1)\rho S + (n - 1)\rho_C]$.

The expected value and variance of S^2 can be calculated from the analysis of variance between clusters, between subclusters, and within clusters within the treatment groups. When cluster sample sizes are unequal, the sums of squares are still independent, and the within cluster sum of squares has a chi-square distribution, but if $\rho_S \neq 1$, the between cluster sum of squares does not have a chi-square distribution. However because S^2 is a quadratic form, Box's theorem can be used to obtain the distribution of S^2.

To obtain the expected value and variance of S^2, use the fact that

$$S^2 = \frac{SSBS^T + SSBS^C + SSBC^T + SSBC^C + SSWC^T + SSWC^C}{N-2},$$

where $SSBS^T$, $SSBC^T$ and $SSBS^T$, $SSWC^T$ and $SSBS^C$, $SSBC^C$ and $SSWC^C$ are the sums of squares between schools, between classes, and within classes in the treatment and control groups, respectively. When subcluster sample sizes are equal, it is easiest to do this in two steps. Start by computing the sum of squares within schools in the treatment and control groups via $SSWS^T = SSBC^T + SSWC^T$ and $SSWS^C = SSBC^C + SSWC^C$. Because the classroom sample sizes are equal, this computation is straightforward and follows exactly from results for the two-level model given in Hedges (2007). Then S^2 can be written as

$$S^2 = \frac{SSBS^T + SSBS^C + SSWS^T + SSWS^C}{N-2}.$$

Because $SSBS^T$ and $SSBS^C$ are functions of the school means in the treatment and control groups, and they are independent of $SSWS^T$ and $SSWS^C$, the mean and variance of S^2 follow exactly from the results for the unequal sample size case for the two-level model given in Hedges (2007) with clusters of size np_i^T or np_i^C, respectively.

When the subcluster sample sizes are unequal, we compute S^2 as

$$S^2 = \frac{SST^T + SST^C}{N-2},$$

where SST^T and SST^C are the sums of squares about the treatment and control group means, respectively. Each treatment group can be viewed as a design with two nested factors. The mean and variance of SST^T and SST^C are calculated separately from results on the estimation of variance components in unbalanced designs with two nested factors (see, e.g., Searle, 1971, pages 474 – 477). Specifically, for either group,

$$SST = (N-1)\hat{\sigma}^2_{WC} + (N-k_3)\hat{\sigma}^2_{BC} + (N-k_1)\hat{\sigma}^2_{BS}.$$

Using results on the variances and covariances of $\hat{\sigma}^2_{WC}, \hat{\sigma}^2_{BC},$ and $\hat{\sigma}^2_{BS}$ (see, e.g., Searle, 1971, pages 474 – 477), the mean and variance of S^2 are obtained from the mean and variance of SST^T and SST^C. Inserting these values for the mean and variance of S^2 into (29) and (28), and simplifying gives the values we obtain for c_U given in (20) and h_U given in (21).

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 on "Statistical power with group mean as the unit of analysis." *Journal of Educational Statistics*, *11*, 161-169.

Blitstein, J. L., Hannan, P. J., Murray, D. M., & Shadish, W. R. (2005). Increasing degrees of freedom in existing group randomized trials through the use of external estimates of intraclass correlation: The df* approach. *Evaluation Review*, *29*, 241-267.

Blitstein, J. L., Murray, D. M., Hannan, P. J., & Shadish, W. R. (2005). Increasing degrees of freedom in future group randomized trials through the use of external estimates of intraclass correlation: The df* approach. *Evaluation Review*, *29*, 268-286.

Box, G. E. P. (1954). Some theorems on quadratic forms applied to the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, *25*, 290-302.

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

Donner, A. & Koval, J.J. (1982). Design considerations in the estimation of intraclass correlations. *Annals of Human Genetics*, *46*, 271-277.

Gaylord, D. W. & Hopper, F. N. (1969). Estimating degrees of freedom for linear combinations of mean squares by Satterthwaite's formula. *Technometrics*, *11*, 691-706.

Geisser, S. & Greenhouse, S. W. (1958). An extension of Box's results on the use of the *F* distribution in multivariate analysis. *Annals of Mathematical Statistics*, *29*, 885-891.

Guilliford, M. C., Ukoumunne, O. C., & Chinn, S. (1999). Components of variance and intraclass correlations for the design of community-based surveys and intervention studies. Data from the Health Survey for England 1994. *American Journal of Epidemiology*, *149*, 876-883.

Hannan, P. J., Murray, D. M., Jacobs, D. R., & McGovern, P. G. (1994). Parameters to aid in the design and analysis of community trials: Intraclass correlations from the Minnesota heart health program. *Epidemiology*, *5*, 88-95.

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

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, *32*, 151-179.

Hopkins, K. D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal*, *19*, 5-18.

Johnson, N. L. & Kotz, S. (1970). *Distributions in statistics-Continuous univariate distributions-2*. New York: John Wiley.

Kirk, R. (1995). *Experimental design*. Belmont, CA: Brooks Cole.

Klar, N. & Donner, A. (2001). Current and future challenges in the design and analysis of cluster randomization trials. *Statistics in Medicine*, *20*, 3729-3740.

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

Murray, D. M. (1998). *Design and analysis of group-randomized trials*. New York: Oxford University Press.

Murray, D. M. & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials, *Evaluation Review*, *27*, 79-103.

Murray, D. M., Hannan, P. J., & Baker, W. L. (1996). A Monte Carlo study of alternative responses to intraclass correlation in community trials. *Evaluation Review*, *20*, 313-337.

Murray, D. M., Varnell, S. P., & Blitstein, J. L. (2004). Design and analysis of group-randomized trials: A review of recent methodological developments. *American Journal of Public Health*, *94*, 423-432.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models*. Thousand Oaks, CA: Sage Publications.

Ridgeway, J. E., Zawgewski, J. S., Hoover, M. N., & Lambdin, D. V. (2002). Student attainment in connected mathematics curriculum. Pages 193-224 in S. L. Senk & D. R. Thompson (Eds.) *Standards-based school mathematics curricula: What are they? What do students learn?* Mahwah, NJ: Erlbaum.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley.

Skinner, C. J., Holt, D. & Smith, T. M. F. (1989). *The analysis of complex surveys*. New York: Wiley.

Verma, V. & Lee, T. (1996). An analysis of sampling errors for demographic and health surveys. *International Statistical Review*, *64*, 265-294.

Welch, B. L. (1949). Further notes on Mrs. Aspin's tables and on certain approximations to the tabled function. *Biometrika*, *36*, 293-296.

Welch, B. L. (1956). On linear combinations of several variances. *Journal of the American Statistical Association*, *51*, 132-148.

Table 1

					Theoretical Results	Empirical Results	
m	p	N^a	ρ_s	ρ_c	Rejection Rate From Equation 12	Empirical Rejection Rate	Empirical SE^b
2	2	200	0.0008	0.0002	0.055	0.059	0.004
2	2	200	0.0400	0.0100	0.277	0.282	0.008
2	8	800	0.0008	0.0002	0.069	0.069	0.004
2	8	800	0.0400	0.0100	0.522	0.516	0.009
8	2	800	0.0008	0.0002	0.055	0.062	0.004
8	2	800	0.0400	0.0100	0.274	0.276	0.008

Note: Empirical results are from the first three rows of Table 1 in Murray, Hannan, and Baker (1996).

a. $n = 25$ in this table

b. $SE = \sqrt{p(1-p)/3200}$

Table 2
The actual significance level of four nominal significance level $\alpha = 0.05$ significance tests: the naïve test (ignoring clustering) and tests using t_K computed from critical values based on $(N-2)$, $(N-2)/DEF^a$, and $(M-2)$ degrees of freedom

						Naïve Test		t_K with $N-2$ df		t_K with $(N-2)/DEF$ df		t_K with $M-2$ df	
m	p	n	N	DEF^a	h	$N-2$	Actual p	t_A/t_K	Actual p	$(N-2)/DEF$	Actual p	$M-2$	Actual p
$\rho_S = 0.25$ $\rho_C = 0.15$													
2	2	30	240	20.1	40.1	238	0.690	0.9250	0.076	11.8	0.048	2	<.001
5	2	30	600	20.1	84.7	598	0.673	0.9709	0.060	29.8	0.050	8	0.028
10	2	30	1200	20.1	163.8	1198	0.667	0.9856	0.055	59.6	0.050	18	0.040
20	2	30	2400	20.1	323.2	2398	0.665	0.9928	0.052	119.3	0.050	38	0.045
2	3	30	360	27.6	50.0	358	0.731	0.9285	0.074	13.0	0.048	2	<.001
5	3	30	900	27.6	103.0	898	0.718	0.9721	0.059	32.5	0.050	8	0.027
10	3	30	1800	27.6	198.0	1798	0.713	0.9862	0.055	65.1	0.050	18	0.040
20	3	30	3600	27.6	389.3	3598	0.711	0.9931	0.052	130.4	0.050	38	0.045
2	5	30	600	42.6	62.3	598	0.781	0.9313	0.072	14.0	0.050	2	<.001
5	5	30	1500	42.6	124.5	1498	0.771	0.9732	0.059	35.2	0.050	8	0.027
10	5	30	3000	42.6	237.4	2998	0.767	0.9867	0.054	70.4	0.050	18	0.039
20	5	30	6000	42.6	465.2	5998	0.766	0.9934	0.052	140.8	0.050	38	0.045
$\rho_S = 0.25$ $\rho_C = 0.25$													
2	2	30	240	23	27.4	238	0.714	0.9184	0.082	10.3	0.051	2	0.001
5	2	30	600	23	61.2	598	0.695	0.9684	0.062	26.0	0.051	8	0.029
10	2	30	1200	23	120.1	1198	0.689	0.9843	0.056	52.1	0.051	18	0.041
20	2	30	2400	23	238.4	2398	0.686	0.9922	0.053	104.3	0.050	38	0.046

2	3	30	360	30.5	36.3	358	0.747	0.9241	0.077	11.7	0.049	2	<.001
5	3	30	900	30.5	78.7	898	0.732	0.9705	0.061	29.4	0.051	8	0.028
10	3	30	1800	30.5	153.2	1798	0.727	0.9854	0.055	59.0	0.050	18	0.040
20	3	30	3600	30.5	303.0	3598	0.725	0.9927	0.053	118.0	0.050	38	0.045
2	5	30	600	45.5	48.8	598	0.789	0.9287	0.074	13.1	0.050	2	<.001
5	5	30	1500	45.5	101.9	1498	0.779	0.9722	0.059	32.9	0.050	8	0.027
10	5	30	3000	45.5	196.3	2998	0.775	0.9862	0.055	65.9	0.050	18	0.040
20	5	30	6000	45.5	386.4	5998	0.773	0.9931	0.052	131.8	0.050	38	0.045
$\rho_S = 0.15$ $\rho_C = 0.25$													
2	2	30	240	17.1	45.5	238	0.660	0.9455	0.069	13.9	0.047	2	<.001
5	2	30	600	17.1	99.9	598	0.645	0.9787	0.057	35.0	0.049	8	0.026
10	2	30	1200	17.1	194.3	1198	0.640	0.9894	0.054	70.1	0.050	18	0.039
20	2	30	2400	17.1	384.1	2398	0.638	0.9947	0.052	140.2	0.050	38	0.045
2	3	30	360	21.6	63.4	358	0.692	0.9510	0.066	16.6	0.048	2	<.001
5	3	30	900	21.6	137.0	898	0.681	0.9807	0.056	41.6	0.050	8	0.025
10	3	30	1800	21.6	264.9	1798	0.677	0.9904	0.053	83.2	0.050	18	0.038
20	3	30	3600	21.6	522.1	3598	0.675	0.9952	0.052	166.6	0.050	38	0.044
2	5	30	600	30.6	92.7	598	0.737	0.9553	0.064	19.5	0.049	2	<.001
5	5	30	1500	30.6	194.6	1498	0.729	0.9824	0.055	49.0	0.050	8	0.025
10	5	30	3000	30.6	373.1	2998	0.726	0.9912	0.053	98.0	0.050	18	0.038
20	5	30	6000	30.6	732.2	5998	0.724	0.9956	0.051	196.0	0.050	38	0.044

a. *DEF* is Kish's design effect, $DEF = [1 + (pn - 1)\rho s + (n - 1)\rho c]$.

Table 3
Auxiliary constants for computing the adjusted test statistic and its degrees of freedom when subcluster sample sizes are unequal

$k_1 = \frac{1}{N}\sum_{i=1}^{m} n_{i\bullet}^2$	$k_{12} = \sum_{i=1}^{m}\left(\sum_{j=1}^{p_i} \frac{n_{ij}^2}{n_{i\bullet}}\right)$	$k_3 = \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{p_i} n_{ij}^2$
$k_4 = \sum_{i=1}^{m}\sum_{j=1}^{p_i} n_{ij}^3$	$k_5 = \sum_{i=1}^{m}\left(\sum_{j=1}^{p_i} \frac{n_{ij}^3}{n_{i\bullet}}\right)$	$k_6 = \sum_{i=1}^{m}\left[\frac{\left(\sum_{j=1}^{p_i} n_{ij}^2\right)^2}{n_{i\bullet}}\right]$
$k_7 = \sum_{i=1}^{m}\frac{\left(\sum_{j=1}^{p_i} n_{ij}^2\right)^2}{n_{i\bullet}^2}$	$k_8 = \sum_{i=1}^{m} n_{i\bullet}\left(\sum_{j=1}^{p_i} n_{ij}^2\right)$	$k_9 = \sum_{i=1}^{m} n_{i\bullet}^3$

Note: The superscripts T and C for treatment and control group are omitted, but each of these constants must be computed within each treatment group to obtain the k_i^T used in (20) and (21).

Table 4
Power of the Adjusted *t*-test Based on t_A, GLS, and the Test Based on Cluster Means, along with the Ratio of the Power of the Adjusted Test to that Based on GLS when $\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C} = 1.0\sigma_T$

			GLS Test		Test based on t_A		Test Based on Cluster means		Power
m	*p*	*N*	Power	*df*	Power	*df*	Power	*df*	Ratio
$\rho_S = 0.25, \rho_C = 0.15$									
2	2	160	0.399	158	0.388	41.9	0.176	2	0.97
2	3	240	0.432	238	0.422	50.9	0.188	2	0.98
2	5	400	0.462	398	0.453	62.3	0.198	2	0.98
3	2	240	0.552	238	0.541	54.5	0.363	4	0.98
3	3	360	0.592	358	0.582	66.0	0.391	4	0.98
3	5	600	0.628	598	0.619	80.2	0.417	4	0.99
5	2	400	0.772	398	0.764	83.3	0.662	8	0.99
5	3	600	0.809	598	0.803	101.0	0.702	8	0.99
5	5	1000	0.840	998	0.835	122.2	0.736	8	0.99
$\rho_S = 0.25, \rho_C = 0.25$									
2	2	160	0.358	158	0.344	32.1	0.161	2	0.96
2	3	240	0.399	238	0.386	39.9	0.176	2	0.97
2	5	400	0.439	398	0.427	51.2	0.190	2	0.97
3	2	240	0.500	238	0.485	41.6	0.327	4	0.97
3	3	360	0.552	358	0.539	52.2	0.362	4	0.98
3	5	600	0.600	598	0.589	66.7	0.396	4	0.98
5	2	400	0.717	398	0.706	63.3	0.606	8	0.98
5	3	600	0.771	598	0.762	80.0	0.660	8	0.99
5	5	1000	0.816	998	0.809	102.2	0.709	8	0.99
$\rho_S = 0.15, \rho_C = 0.25$									
2	2	160	0.454	158	0.445	50.8	0.197	2	0.98
2	3	240	0.524	238	0.515	66.7	0.222	2	0.98
2	5	400	0.594	398	0.587	93.6	0.250	2	0.99
3	2	240	0.620	238	0.611	66.0	0.412	4	0.99
3	3	360	0.697	358	0.690	88.1	0.472	4	0.99
3	5	600	0.769	598	0.763	123.7	0.535	4	0.99
5	2	400	0.834	398	0.828	100.2	0.730	8	0.99
5	3	600	0.893	598	0.889	134.9	0.801	8	1.00
5	5	1000	0.936	998	0.934	189.7	0.861	8	1.00

Note: $n = 20$.