
Institute for Policy Research
Northwestern University
Working Paper Series

WP-06-11

Correcting a Significance Test for Clustering

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Please do not quote or distribute without permission.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 trials 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 anticonservative conclusions about the precision and statistical significance of treatment effects. This working paper gives a simple correction to the *t*-statistic that would be computed if clustering were (incorrectly) ignored. The correction is a multiplicative factor depending on the total sample size, the cluster size, and the intraclass correlation *p*. The corrected *t*-statistic has a student's *t*-distribution with reduced degrees of freedom. The corrected statistic reduces to the *t*-statistic computed by ignoring clustering when $p = 0$. It reduces to the *t*-statistic computed using cluster means when $p = 1$. If $0 < p < 1$, it lies between these two, and the degrees of freedom are between those corresponding to these two extremes.

Correcting a Significance Test for Clustering

Field experiments often assign entire intact groups (such as sites, classrooms, or schools) to the same treatment group, with different intact groups assigned to different treatments. Because these intact groups correspond to 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 carry out a two stage analysis. That is, to compute mean scores on the outcome (and all other variables that may be involved in the analysis) and carry out the statistical analysis as if the site (cluster) means were the data. If all cluster sample sizes are equal, this approach provides exact tests for the treatment effect, but the tests may have lower statistical power than would be obtained by other approaches (see, e.g., Blair and Higgins, 1986). 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 trials is made when the data are analyzed as if the data were 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 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 it is desirable 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. What Works Clearinghouse reviewers found that, in the first areas they were investigating, the majority of the high quality studies involved assignment to treatments by clusters, but most of those studies did not account for clustering in their evaluation of the statistical significance of treatment effects. In this context, it would be desirable to be able to know how the conclusions about treatment effects might change if clustering were taken into account.

The purpose of this paper is to provide an analysis of the effects of 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. The derivations provide some insight into the properties of suggestions that have appeared in the literature for adjusting significance tests for the effects of clustering. Then we provide a generalization for unequal cluster 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 parameter) is known or that the intraclass correlation parameter 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_{ij}^{T} ($i = 1, \ldots, m^{T}; j = 1, \ldots, n_i^{T}$) and Y_{ij}^{C} ($i = 1, \ldots, m^{C}; j = 1, \ldots, n_i^{C}$) be the j^{th} observation in the i^{th} cluster in the treatment and control groups respectively, so that there are m^T clusters in the treatment group and m^C clusters in the control group, and a total of $M = m^T + m^C$ clusters with n observations each. Thus the sample size in the treatment group is

$$N^T = \sum_{i=1}^{m^T} n_i^T ,$$

the sample size in the control group is

$$N^C = \sum_{i=1}^{m^C} n_i^C ,$$

and the total sample size is $N = N^T + N^C$.

Let $\bar{Y}_{i\bullet}^{T}$ ($i = 1, \ldots, m^T$) and $\bar{Y}_{i\bullet}^{C}$ ($i = 1, \ldots, m^C$) be the means of the i^{th} cluster in the treatment and control groups, respectively, and let $\bar{Y}_{\bullet\bullet}^{T}$ and $\bar{Y}_{\bullet\bullet}^{C}$ be the overall (grand) means in the treatment and control groups, respectively. Define the (pooled) within-treatment group variance S^2 via

$$S^2 = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{n_i^T}(Y_{ij}^T - \bar{Y}_{\bullet\bullet}^T)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{n_i^C}(Y_{ij}^C - \bar{Y}_{\bullet\bullet}^C)^2}{N-2}. \qquad (1)$$

Suppose that observations within the treatment and control group clusters are normally distributed about cluster means μ_i^T and μ_i^C with a common within-cluster variance σ_W^2. That is

$$Y_{ij}^T \sim N(\mu_i^T, \sigma_W{}^2), i = 1, \ldots, m^T; j = 1, \ldots, n_i^T$$

and

$$Y_{ij}^C \sim N(\mu_i^C, \sigma_W{}^2) \; i = 1, \ldots, m^C; j = 1, \ldots, n_i^C.$$

Suppose further that the clusters are random effects (for example they are considered a sample from a population of clusters) so that the cluster means themselves have a normal sampling distribution with means $\mu_\bullet^T$ and $\mu_\bullet^C$ and common variance σ_B^2. That is

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

and

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

Note that in this formulation, σ_B^2 represents true variation of the population means of clusters over and above the variation in sample means that would be expected from variation in the sampling of observations within clusters.

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

The Intraclass Correlation

Note that there are three different within-treatment group standard deviations, σ_B, σ_W, and σ_T, the latter defined by

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2 .$$

In most educational data when clusters are schools, σ_B^2 is considerably smaller than σ_W^2. Obviously, if the between cluster variance σ_B^2 is small, then σ_T^2 will be very similar to σ_W^2. The relation between these three variances, in particular the fact that $\sigma_T \neq \sigma_W$ gives rise to the statistical effects of clustering.

A parameter that summarizes the relationship between the three variances (and therefore the clustering effect) is called the intraclass correlation ρ, which is defined by

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = \frac{\sigma_B^2}{\sigma_T^2} . \qquad (2)$$

The intraclass correlation is a measure of the effect of clustering in the data. If $\rho = 0$ then $\sigma_B^2 = 0$ and there is no clustering. If $\rho = 1$ then $\sigma_W^2 = 0$ and there is "complete clustering" in the sense that there is no *within*-cluster variability. Note that in this notation and throughout the paper, ρ is a population parameter, not a sample estimate.

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

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

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

$$t = \frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C)}{S}, \qquad (3)$$

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 $\rho = 0$), 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 $\rho \neq 0$) the test statistic has a different sampling distribution—one that depends on ρ, n, and m.

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 as a factor nested within treatments, analyses that use a hierarchical linear model including clusters as level two 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 (3) as a test statistic when the sample is a clustered sample.

Some Theory

The main result of this paper is a derivation of the sampling distribution of the t-statistic when $\rho \neq 0$, from which a modified test statistic is derived. However, it is useful to see why the t-statistic given in (3) does not have its nominal distribution under the null hypothesis when clustering is present. By definition, a statistic T has Student's t-distribution with v degrees of freedom if it can be written as

$$T = U / \sqrt{V / v}$$

where U is a standard normal and V is a chi-square with v degrees of freedom that is independent of U. Note that the degrees of freedom of the t-statistic are determined by the degrees of freedom of the chi-square in the denominator.

Distribution of the *t*-statistic When $\rho = 0$

When there is no clustering (that is when $\rho = 0$), the numerator of (3) has a normal distribution with standard deviation σ_T. In other words, when the null hypothesis is true

$$\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^{T} - \bar{Y}_{\bullet\bullet}^{C})/\boldsymbol{\sigma}_{\mathrm{T}}$$

has the standard normal distribution. Similarly, when there is no clustering (that is when $\rho = 0$), $(N - 2)S^2/\sigma_T^2$ 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

$$S/\sigma_T$$

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.

Distribution of the *t*-statistic When $\rho \neq 0$

When $\rho \neq 0$, neither the numerator nor the denominator of the *t*-statistic given in (3) has the same distribution as they do when $\rho = 0$. We now indicate how the distribution of the numerator and denominator are different when $\rho \neq 0$ and the cluster sample sizes n_i^T and n_i^C are all equal to n.

Distribution of the numerator of t when $\rho \neq 0$. Assuming cluster samples sizes are equal to n and $\rho \neq 0$, the numerator has a normal distribution with mean 0, but with a generally larger variance: $\sigma_T^2[1 + (n - 1)\rho]$. The factor $[1 + (n - 1)\rho]$ is Kish's (1965) design effect. In other words, when $\rho \neq 0$, and the null hypothesis is true

$$\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^{T} - \bar{Y}_{\bullet\bullet}^{C})/\boldsymbol{\sigma}_{\mathrm{T}}\sqrt{1+(n-1)\rho}$$

has the standard normal distribution.

Distribution of the denominator of t when $\rho \neq 0$. Assuming cluster samples sizes equal to n, and $\rho \neq 0$, the expected value of S^2 is no longer σ_T^2, but instead

$$E\left\{S^2\right\} = \sigma_W^2 + \left(\frac{N-2n}{N-2}\right)\sigma_B^2 = \sigma_T^2\left(1 - \frac{2(n-1)\rho}{N-2}\right).$$

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

$$\frac{hS^2}{\sigma_T^2\left(1 - \frac{2(n-1)\rho}{N-2}\right)}$$

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

$$h = \frac{[(N-2) - 2(n-1)\rho]^2}{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}. \qquad (4)$$

Taking the partial derivative of h with respect to ρ, we see that h is a decreasing function of ρ. If $\rho = 0$ and there is no clustering, $h = (N - 2)$ and S has the nominal degrees of freedom as expected. If $\rho = 1$ and there is complete clustering (no variability within clusters), then $h = (M - 2)$ as expected (because the only variability is that between the M clusters). If $0 < \rho < 1$, 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 $\rho \neq 0$, 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

$$\frac{S}{\sigma_T\sqrt{1 - \frac{2(n-1)\rho}{N-2}}}$$

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 $\rho \neq 0$ .The results in the previous section have three important implications for the *t*-statistic given in (3). First, the scale factors necessary to standardize the numerator and denominator of *t* no longer cancel, therefore the ratio (the *t*-statistic given in equation 3) no longer has the *t*-distribution. Second, the degrees of freedom of *S* (and therefore *t*) are no longer *(N – 2)*, but *h*. Third, the statistic

$$\frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^{T} - \bar{Y}_{\bullet\bullet}^{C})/\sigma_{T}\sqrt{1+(n-1)\rho}}{S/\sigma_{T}\sqrt{1-\frac{2(n-1)\rho}{N-2}}} = c\frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^{T} - \bar{Y}_{\bullet\bullet}^{C})}{S} = ct$$

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

$$c = \sqrt{\frac{(N-2)-2(n-1)\rho}{(N-2)[1+(n-1)\rho]}} \quad . \tag{5}$$

Thus the statistic

$$t_A = ct \tag{6}$$

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 $\rho = 0$ so that there is no clustering, then $c = 1$ and $h = N - 2$. That is, when $\rho = 0$, the test based on t_A reduces to the usual t-test ignoring clustering. When $\rho = 1$ and there is complete clustering, then $c = \sqrt{(M-2)/(N-2)}$ and $h = M - 2$. That is, when $\rho = 1$, and the test based on t_A reduces to a t-test computed using the cluster means. Note that throughout this section (and this paper), ρ is the population parameter, *not* a sample estimate.

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 $t(v, \alpha)$ is 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

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

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.

Accuracy of the Approximation

The method used to obtain the approximate sampling distribution of S (and therefore t_A) in this paper has proven quite accurate in other situations, such as the construction of tests in repeated measures analysis of variance (e.g., Geisser and Greenhouse, 1958). To verify that it is also accurate in this situation, an extensive simulation study was carried out. The number of clusters was varied from $m^T = m^C = m =$ 2 to $m = 20$ in each treatment group, and the sample size n per cluster was varied from $n = 2$ to $n = 100$. For each value of n and m, values of ρ from $\rho = 0.00$ (in which case the test ignoring clustering is correct) to $\rho = 0.40$ were examined. For each combination of n, m, and ρ examined, a total of 10,000 replications were generated, yielding a standard error of about 0.0022 for the simulated ejection rate of the test at the nominal 0.05 level.

The results suggested, as expected, that the unadjusted test provided poor control of Type I Errors, with actual rejection rates much higher than nominal. The results also suggested that the adjusted test based on t_A had actual significance levels that were indistinguishable (within the error of the simulation) from nominal. That is, the significance levels of the test based on t_A (and therefore the confidence intervals described in this paper) are quite accurate (provided, of course, that ρ is known). Additional simulations (not reported here) suggest that power calculations based on the noncentral t-distribution are also quite accurate.

Table 1 provides a selection of the empirical significance levels for the unadjusted t-test (ignoring the effects of clustering) and for the adjusted t-test based on t_A for selected values of n and $m^T = m^C = m$. The number of clusters varies from $m = 2$ to $m = 20$ in each treatment group, and the sample size n per cluster is varied from $n = 2$ in top panel of the table to $n = 100$ in the bottom panel of the table. For each value of n and m (that is,

for each panel of the table), there is a different value of ρ on each row, varying from ρ = 0.00 (in which case the test ignoring clustering is correct) to ρ = 0.40. For each value of n and ρ, the empirical rejection rates for the nominal 0.10, 0.05, and 0.01 level unadjusted t-tests that ignore the effects of clustering (that is, tests that reject the null hypothesis if $| t |$ exceeds the t critical value with $2nm - 2$ degrees of freedom) are given in columns 4, 5, and 6. The empirical rejection rates for the nominal 0.10, 0.05, and 0.01 level tests based on t_A (that is, tests that reject the null hypothesis if $| t_A |$ exceeds the t critical value with h degrees of freedom) are given in columns 7, 8, and 9. The value of the correction factor c and the adjusted degrees of freedom h are given in columns 10 and 11. (The details of the computations are given in the appendix.)

Examining the rejection rates for the unadjusted t-test (that ignores clustering) demonstrates why it is important to correct for its effects. For example with when the cluster sample size is relatively large (as it often is in educational studies where the clusters are schools), the effects of clustering on the unadjusted t-test are profound. For example, when m = 5, n = 20, and ρ = 0.10, the nominal 0.05 level test has an actual significance level of 0.253, 500% of the nominal level, and the nominal 0.01 level test has an actual significance level of 0.133 (1300% of its nominal level). The empirical rejection rates of the adjusted test are all quite near the nominal level.

Relation to Previous Work

The sampling distribution of t_A derived in this paper provide some insight about other approaches to testing mean differences in clustered samples. Kish (1965) suggested multiplying S (or, equivalently, dividing the t-statistic) by the square root of the design

effect (the square root of [*1 + (n – 1)ρ*]) to remove the effect of clustering on the numerator of the *t*-statistic. The resulting statistic is

$$t_K = \frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C)}{S\sqrt{1+(n-1)\rho}}. \quad (7)$$

However because this statistic is does not correct for the fact that the scale factor necessary to standardize S 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

$$t_K = \frac{t_A}{\sqrt{1-\frac{2(n-1)\rho}{N-2}}}. \quad (8)$$

If $\rho \neq 0$ the denominator of (8) is less than one, so $t_K > t_A$. However note that

$$1-\frac{2(n-1)\rho}{N-2} = 1-\frac{\rho}{m}\left(1-\frac{m-1}{mn-1}\right) \quad (9)$$

so that the denominator of (8) will be quite close to 1 unless *m* is small and ρ is large. For example, if $\rho = 0.1$ $n = 50$ and $m = 2$, the denominator of (8) is about 0.975, but if $n = 50$ and $m = 10$, the denominator is 0.990. Therefore the sampling distribution of t_K is approximately a *t*-distribution with *h* degrees of freedom.

While many several authors have advocated the use of t_K (or its generalizations) as a test statistic, Hannan, et al. (1994) noted that "the choice of df for the adjusted test remains the subject of considerable debate"(p. 93). The results of this paper shed some light on the implication of various choices of degrees of freedom used for the selection of critical values. We can rewrite *h* in a somewhat more revealing form

$$h = (N-2)\left\{\frac{\left(1-\frac{2(n-1)\rho}{N-2}\right)^2}{1+(n-1)\rho\left\{\frac{[N-2(n-1)\rho-4]}{N-2}\right\}}\right\}.$$

This expression shows that h is essentially the nominal degrees of freedom ($N-2$) multiplied by a correction factor, which is the fraction in brackets. The numerator of this fraction is the square of the term given in (9) which is typically close to 1. The denominator of the fraction is almost (but not quite) the design effect $1 + (n-1)\rho$. Therefore the value of h should be similar, but not identical, to $(N-2)/[1 + (n-1)\rho]$. For small values of ρ, $(N-2)/[1 + (n-1)\rho]$ is usually smaller than h, but this need not be so for large values of ρ.

A test based on t_K with critical values based on (N – 2) degrees of freedom. Blair and Higgins (1986) have been interpreted as saying that critical values of t_K could be evaluated using the usual ($N-2$) degrees of freedom. (In fact, they made the claim not about t_K, but about a test statistic they derived based on generalized least squares analysis.) Let $t(v, \alpha)$ be the level α two-sided critical value for the t-distribution with v degrees of freedom. If $\rho \neq 0$, then $h < (N-2)$, so it follows that $t(N-2, \alpha) < t(h, \alpha)$. Moreover, $t_K > t_A$, so that, when the null hypothesis is true, the rejection rate of a test that rejects if $|t_K| > t(N-2, \alpha)$ must be greater than α. The two-tailed rejection rate of such a test is

$$2\left\{1-F\left[t(N-2,\alpha)\sqrt{1-\frac{2(n-1)\rho}{N-2}},\ h\right]\right\}, \qquad (10)$$

where $F(x, v)$ is the cumulative distribution function of the t-distribution with v degrees of freedom. Columns 7 and 8 of Table 2 show the degrees of freedom and the rejection rate

of the nominal α = 0.05 level test based on t_K using $(N - 2)$ degrees of freedom for m = 2, 5, and 10, n = 10, 25, and 100, and ρ = 0.10 and 0.20. These values show that the actual rejection rate of the test increases with ρ and n and decreases with m (as expected, since $N - 2 > h$). However unless the number m of clusters is small, the actual level of the test is only slightly larger than the nominal (not exceeding 0.07 for $n \leq 100$ and $\rho \leq 0.30$. However unless m is very small, the elevation in significance level is not large. In other words, the test is liberal, but only slightly so.

A test based on t_K with critical values based on $(M - 2)$ degrees of freedom. Donner and Klar (2000, p. 115) state that t_K, could be used with critical values based on $(M - 2)$ degrees of freedom. However because $h > (M - 2)$, it follows that $t(M - 2, \alpha) > t(h, \alpha)$. Although $t_K > t_A$, the rejection rate of a test that rejects if $|t_K| > t(M - 2, \alpha)$ is generally less (and often *much* less) than α. The two-tailed rejection rate of such a test is

$$2\left\{1 - F\left[t(M-2,\alpha)\sqrt{1-\frac{2(n-1)\rho}{N-2}},\ h \right]\right\}, \qquad (11)$$

where $F(x, v)$ is the cumulative distribution function of the t-distribution with v degrees of freedom. Columns 9 and 10 of Table 2 show the degrees of freedom and the rejection rate of the nominal α = 0.05 level test based on t_K using $(M - 2)$ degrees of freedom for m = 2, 3, 4, 5, and 10, n = 10, 25, and 100, and ρ = 0.10 and 0.20. These values show that the rejection rate is much lower than nominal when the number of clusters is small (as expected since $M - 2 << h$). In other words, the test is very conservative.

A test based on t_K with critical values based on $(N - 2)/[1 + (n - 1)\rho]$ degrees of freedom. Skinner, Holt, and Smith (1989) suggested that t_K could be used with critical

values based on degrees of freedom adjusted by the design effect, that is with $(N-2)/[1+(n-1)\rho]$ degrees of freedom. This test has a rejection rate of

$$2\left\{1-\mathrm{F}\left[\mathrm{t}((N-2)/[1+(n-1)\rho],\alpha)\sqrt{1-\frac{2(n-1)\rho}{N-2}},\ h\right]\right\}, \tag{12}$$

where F*(x, v)* is the cumulative distribution function of the *t*-distribution with *v* degrees of freedom. Columns 12 and 13 of Table 2 show the degrees of freedom and the rejection rate of the nominal $\alpha = 0.05$ level test based on t_K using $(N-2)/[1+(n-1)\rho]$ degrees of freedom, which is remarkably close to the nominal rejection rate for a wide variety of values of *n*, *m*, and ρ.

A test with degrees of freedom based on uncertainty of the estimate of ρ. A different approach to computing degrees of freedom was suggested by Hannan, Murray, Jacobs, and McGovern (1994) and elaborated by Blitstein, Hannan, Murray, and by Shadish (2005) and Blitstein, Murrray, Hannan, and Shadish (2005). They were concerned that external estimates of ρ were themselves subject to sampling uncertainty and wanted to take that into account in their choice of degrees of freedom. They noted that the if ρ was estimated from a total of *k* clusters each with a sample size of *n*, then the approximate variance of the analysis of variance estimate of ρ was given by Donner and Koval (1982) as

$$V(\hat{\rho}) = \frac{2(1-\rho)^2[1+(n-1)\rho]^2}{n(n-1)(k-1)}.$$

Hannan et al. noted that the equation could be solved for $(k-1)$, the degrees of freedom between groups used to estimate ρ. They denoted this estimate by *df**, namely

$$df^* = \frac{2(1-\rho)^2[1+(n-1)\rho]^2}{n(n-1)V(\hat{\rho})} \tag{13}$$

and suggested that df^* could be used as the degrees of freedom to compute critical values for a test based on t_K.

It is unclear how to evaluate the effectiveness of this estimate analytically. However we note that if the external estimate of ρ is very precise (as it may be if computed from sample surveys with large sample sizes), df^* can be very large, easily exceeding the total sample size of the study to which it is applied. In this case, Hannan et al. would suggest placing an upper bound on df^* of $N - 2$.

Studies of models with two levels of nesting. Murray, Hannan, and Baker (1996) used simulations to study the performance of analysis of variance F-tests in a model with two nested factors. They were particularly concerned with the effects of using critical values defined by numbers of degrees of freedom different than those conventionally used. They investigated a different sampling model than that in this paper (with two levels of clustering rather than the one level of clustering in this paper), but some of their results are similar. They found, as expected, that the rejection rate of an approximation to the usual F-test that ignores clustering (a test approximating the square of what is called in this paper the unadjusted t) is much higher than nominal, with rejection rates similar to those found in this paper. They also found that tests using the conventional degrees of freedom give close to nominal results. Given the differences in both the data generation model and the statistics computed, other results of theirs are more difficult to compare with those in this paper.

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" sample size might give most accurate results when substituted into the equal sample size formulas for rough approximations.

Distribution of the numerator of t when $\rho \neq 0$. The numerator of (3) has a normal distribution with mean 0 and variance $[1+(\tilde{n}-1)\rho]\sigma_T^2$, where $\tilde{n}$ is given by

$$\tilde{n} = \frac{N^C \sum_{i=1}^{m^T} \left(n_i^T\right)^2}{N^T N} + \frac{N^T \sum_{i=1}^{m^C} \left(n_i^C\right)^2}{N^C N}. \qquad (14)$$

In other words,

$$\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C)/\sigma_T\sqrt{1+(\tilde{n}-1)\rho},$$

has the standard normal distribution. Note that when all of the cluster sample sizes n_i^T and n_i^C are equal to n, then $\tilde{n} = n$.

Distribution of the denominator of t when $\rho \neq 0$. The expected value of S^2 is

$$E\left\{S^2\right\} = \sigma_T^2\left(1 - \frac{2(\bar{n}_U - 1)\rho}{N-2}\right),$$

where $\bar{n}_U$ is given by

$$\bar{n}_U = \frac{\sum_{i=1}^{m^T}\left(n_i^T\right)^2}{2N^T} + \frac{\sum_{i=1}^{m^C}\left(n_i^C\right)^2}{2N^C}. \qquad (15)$$

Note that when all of the cluster sample sizes n_i^T and n_i^C are equal to n, then $\bar{n}_U = n$. We show in the Appendix that

$$\frac{hS^2}{\sigma_T^2\left(1-\frac{2(\bar{n}_U-1)\rho}{N-2}\right)}$$

has the chi-square distribution with h degrees of freedom, where

$$h_U = \frac{\left[(N-2)-2(\bar{n}_U-1)\rho\right]^2}{(N-2)(1-\rho)^2 + A\rho^2 + 2(N-2\bar{n}_U)\rho(1-\rho)}, \tag{16}$$

where the auxiliary constant A is defined via $A = A^T + A^C$ and

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

and (17)

$$\boldsymbol{A^C} = \frac{\left(\boldsymbol{N^C}\right)^2 \sum_{i=1}^{\boldsymbol{m^C}} \left(\boldsymbol{n_i^C}\right)^2 + \left(\sum_{i=1}^{\boldsymbol{m^C}} \left(\boldsymbol{n_i^C}\right)^2\right)^2 - 2\boldsymbol{N^C} \sum_{i=1}^{\boldsymbol{m^C}} \left(\boldsymbol{n_i^C}\right)^3}{\left(\boldsymbol{N^C}\right)^2}.$$

Note that when the n_i^T and n_i^C are all equal to n, $A = n(N - 2n)$, $\bar{n}_U = n$, and (16) reduces to (4).

It follows that the statistic

$$\frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C)/\boldsymbol{\sigma}_T\sqrt{1+(\tilde{n}-1)\rho}}{S/\boldsymbol{\sigma}_T\sqrt{1-\frac{2(\bar{n}_U-1)\rho}{N-2}}} = c_U \frac{\sqrt{\tilde{N}}(\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C)}{S} = c_U t$$

has Student's t-distribution with h_U degrees of freedom, where c_U is a constant depending on $N, \tilde{n}, \bar{n}_U$, and ρ that absorbs the ratios of the scale factors in the numerator and the denominator, which is given by

$$c_U = \sqrt{\frac{(N-2)-2(\bar{n}_U-1)\rho}{(N-2)[1+(\tilde{n}-1)\rho]}}\,. \qquad (18)$$

Thus

$$t_{AU} = c_U t \qquad (19)$$

is a *t*-statistic adjusted for the effects of clustering in the case of unequal cluster sample sizes. When the cluster sample sizes n_i^T and n_i^C are all equal to n, c_U reduces to c, h_U reduces to h, and t_{AU} reduces to t_A.

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

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 $\rho \neq 0$ and the cluster size exceeds $n = 1$. 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 (or t_{AU})can be used to obtain confidence intervals that will have the correct probability content.

A $100(1 - \alpha)$ percent confidence interval for $\mu^T - \mu^C$ is given by

$$(\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C) - \mathrm{t}(\alpha,h)\,S/c\sqrt{\tilde{N}} \le \mu^T - \mu^C \le (\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C) + \mathrm{t}(\alpha,h)\,S/c\sqrt{\tilde{N}}\,, \qquad (20)$$

where c is the constant defined in (5) if cluster sample sizes are equal or (18) if they are unequal and $t(\alpha;\nu)$ 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 $t(\alpha, \nu) = 1.98$).

Example

The application that motivated this work was the synthesis of studies of the effectiveness of middle school mathematics curricula carried out as part of a broader program of syntheses by the US Institute of Education Sciences What Works Clearinghouse. This review encountered several studies that sampled and assigned to treatments at the cluster level, but ignored the clustering in their reports of significance tests. One of these studies was an evaluation of the Connected Mathematics curriculum reported by Ridgway, et al. (2002), which compared the achievement of $m^T = 18$ classrooms of 6th grade students who used connected mathematics with that of $m^C = 9$ classrooms in a comparison group that did not use connected mathematics. In this quasi-experimental design the clusters were classrooms. The cluster sizes were not identical but the average cluster size in the treatment groups was $N^T/m^T = 338/18 = 18.8$ and N^C/m^C $162/18 = 18.0$ in the control group. Here we treat the cluster sizes as if they were equal and choose $n = 18$ as a slightly conservative sample size and a total sample size of $N =$ 486. The mean difference between treatment and control groups is $\bar{Y}^T_{\bullet\bullet} - Y^C_{\bullet\bullet} = -1.5$, the pooled within-groups standard deviation $S = 2.436$. Although it was not a probability sample, the sample was drawn from sites located in all regions of the country (west, midwest, and east) and it was intended to be representative. The compendium of intraclass correlations computed from national probability samples (Hedberg, Santana,

and Hedges, 2004), gives a national value of $\rho = 0.264$ for sixth grade mathematics achievement (with a sampling standard error of 0.019).

The analysis carried out by the investigators ignored clustering. Comparing the mean of all of the students in the treatment with the mean of all of the students in the control group using a conventional *t*-test leads to an unadjusted *t* value of $t = 6.40$, which is highly statistically significant compared with a critical value based on $(N - 2) = 486 - 2 = 484$ degrees of freedom. Computing the constant *c* used to adjust the *t*-statistic using $\rho = 0.241$, we obtain

$$c = \sqrt{\frac{(486-2)-2(18-1)(0.264)}{(486-2)[1+(18-1)(0.264)]}} = 0.423 ,$$

so that the adjusted *t*-statistic $t_A = (0.423)(6.40) = 2.71$.

Computing the degrees of freedom *h*, we obtain

$$h = \frac{[(486-2)-2(18-1)0.264]^2}{(486-2)(1-0.264)^2 + 18(486-36)0.264^2 + 2(N-36)(0.264)(1-0.264)} = 225.29 .$$

Comparing 2.71 with the critical values of the *t*-distribution with 225.3 degrees of freedom, we obtain a two-tailed *p*-value of $p = 0.0073$. This value is still statistically significant, albeit less so than the (incorrect) test based on the unadjusted *p*-value.

By varying the value of ρ that is used to compute t_A, we can see how large a value of ρ would be required to yield a test statistic that was statistically insignificant at the $\alpha = 0.05$ significance level. We see that the test based on t_A would remain significant unless $\rho > 0.50$, a value that seems extremely implausible. Therefore it seems that, even though the statistical analysis used in this study did not take clustering into account, an analysis that did so would also have found statistically reliable evidence of treatment effects.

Note that the *t*-statistic adjusted only for the effect of clustering on the numerator would be quite similar, namely $t_K = 2.83$. The nominal degrees of freedom (486 – 2) divided by the design effect (5.49) yields an estimated degrees of freedom of 88.2, which is quite different from $h = 225.3$, but this difference in degrees of freedom has relatively small impact on the significance level. Evaluating the *p*-value of t_K using the *t*-distribution with 88.2 degrees of freedom we obtain $p = 0.0076$.

To compute the 95 percent confidence interval for the mean difference, note that $t(0.05; 225.3) = 1.971$. Then using (20) we obtain the 95 percent confidence interval

$$-2.59 = -1.5 - \frac{(1.971)(2.436)}{[0.423]\sqrt{108}} \leq \mu^T - \mu^C \leq -1.5 + \frac{(1.971)(2.436)}{[0.423]\sqrt{108}} = -0.41,$$

which is an interval of width 2.10. 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.

Power Considerations

The power of any statistical test and its power relative to alternative tests that might be used are major considerations. The *t*-test corrected for clustering 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). However, it is still useful to compare the power of this test with alternatives that could be used if more data were available.

Two of those alternatives are a *t*-test performed on cluster means (that is using the cluster as the unit of analysis)(see, e.g., Barcikowski, 1981) and a generalized least

squares (GLS) analysis computed using a known value of ρ to parameterize the error covariance matrix (see Blair and Higgins, 1986). Note that both of these alternatives require more information than the test given here, although they may be computed without complete reanalysis of the data. The analysis based on cluster means requires knowledge of either the cluster means or the mean difference between treatment and control groups and the standard deviations of the cluster means within treatment groups. The generalized least squares analysis proposed by Blair and Higgins also requires knowledge of cluster means (or their within-treatment group standard deviations), but in addition requires knowledge of the within-cluster standard deviations (or equivalent summary statistics).

These two tests provide a useful standard of comparison. The test based on cluster means is the most powerful exact test when ρ is unknown, while the test based on generalized least squares is the most powerful exact test when ρ is known..

The test statistic used in all three analyses (based on the results in this paper, and the two alternatives requiring more data) have noncentral *t*-distributions with the same noncentrality parameter,

$$\lambda = \frac{\sqrt{\tilde{N}}(\mu^T - \mu^C)}{\sigma_T}\sqrt{\frac{1}{1+(\tilde{n}-1)\rho}}, \qquad (21)$$

but different degrees of freedom [(N - 2), h, or (M – 2), respectively], when the null hypothesis is false. It is known that power is an increasing function of degrees of freedom for a fixed noncentrality parameter. Because the analysis based on generalized least squares that was suggested by Blair and Higgins has (N – 2) degrees of freedom and (N – 2) ≥ h ≥ (M – 2), it should provide the most powerful test if ρ is known and the raw data are available. Because the analysis based on group means has (M – 2) degrees of

freedom, 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 2 gives the power of each of the three tests in some illustrative situations when $\mu^T - \mu^C = 1.0\sigma_T$, 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 is unknown. That is, the power advantage of the GLS test and the adjusted *t*-test depends on having a known value of ρ. While the adjusted *t*-test is slightly less powerful than the GLS test, it is very nearly as powerful.

Using Results of This Paper

It is unusual for the value of ρ to be known exactly. However, the results in this paper may be useful even if exact values of ρ are not available. One major use of these results is to judge (when we cannot carry out a reanalysis) whether a plausible amount of clustering would dramatically change the findings reported from an analysis that ignored clustering. For this purpose it may be enough to know an upper bound on the value of ρ. Alternatively, it may be sufficient to have reasonable bounds for ρ in order to carry out a

sensitivity analysis to determine whether the results change qualitatively across the reasonable range of ρ values.

Because it is essential to know values of ρ for power and sample size computations in planning cluster randomized experiments, there have been systematic efforts to obtain information about reasonable values of ρ in realistic situations. For example chapter 5 of Donner and Klar (2000) provide references to many reports of intraclass correlations for health outcomes. Some information about reasonable values of ρ comes from cluster randomized trials that have been conducted. For example, Murray and Blitstein (2003) reported a summary of intraclass correlations obtained from 17 articles reporting cluster randomized trials in psychology and public health and Murray, Varnell, and Blitstein (2004) give references to 14 very recent studies that provide data on intraclass correlations for health related outcomes. Other information on reasonable values of ρ comes from sample surveys that use clustered sampling designs. For example Gulliford, Ukoumunne, and Chinn (1999) and Verma and Lee (1996) presented values of intraclass correlations based on surveys of health outcomes. Hedberg, Santana, and Hedges (2004) presented a compendium of several hundred intraclass correlations for reading and mathematics academic achievement computed from national probability samples at various grade levels. This later compendium provides national values for intraclass correlations as well as values for regions of the country and subsets of regions differing in level of urbanicity.

It is particularly important to use external values of ρ with considerable caution, because the value of ρ has substantial influence on the results of analyses. In particular, it would be difficult to justify the use of the methods described in this paper using estimates

of ρ obtained from small samples (small numbers of clusters) because those estimates are likely to be subject to considerable sampling error. Similarly, it would be difficult to justify the use of external estimates of ρ, even from large sample sizes if those estimates were not based on a similar sampling strategy, with similar populations, and similar outcome measures. If raw data were available, reanalysis would be preferable to the use of methods described in this paper. However, to the extent that external values of ρ can be justified, the methods suggested in this paper can provide a means of adjusting results of a study that ignored clustering for the probable effects of clustering as an alternative to discarding the results of that study altogether.

Conclusion

Cluster randomized trials are increasingly important in education and the social and policy sciences. However these trials are often improperly analyzed by ignoring the effects of clustering on significance tests. Reanalysis using more appropriate methods (such as multilevel statistical methods) is obviously desirable. 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 correlation is 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.

This paper considered only *t*-tests under a sampling model with one level of clustering. Educational experiments sometimes involve additional levels of clustering that would be desirable to include in the statistical analyses (such as classrooms within schools of schools within school districts). The generalization of the methods used in this paper to more designs with additional levels of nesting and more complex analyses would be desirable.

Appendix

Derivations with the Equal Cluster Sample Sizes

Under the model the sampling distribution of the numerator of (3) is normal with mean $\sqrt{\tilde{N}}(\mu^T - \mu^C)$ and variance $\sigma_W^2 + n\sigma_B^2 = \sigma_T^2[1 + (n - 1)\rho]$. The square of the denominator of (3), can be written as

$$S^2 = \frac{SSBC + SSWC}{N - 2}, \tag{22}$$

where *SSBC* is the pooled sum of squares between cluster means within treatment groups, and *SSWC* is the pooled sum of squares within clusters. Therefore $SSWC/\sigma_W^2$ has a chi-squared distribution with $(N - M)$ degrees of freedom, where $M = m^T + m^C$. Similarly

$$\frac{SSBC}{\boldsymbol{\sigma}_W^2 + n\boldsymbol{\sigma}_B^2} \tag{23}$$

has a 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

$$g = \frac{\mathrm{V}\{S^2\}}{2\mathrm{E}\{S^2\}} \tag{24}$$

and

$$h = \frac{2\left(\mathrm{E}\{S^2\}\right)^2}{\mathrm{V}\{S^2\}}, \tag{25}$$

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

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

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

$$\lambda = \frac{\sqrt{\tilde{N}}(\mu^{T} - \mu^{C})}{\sigma_{T}\sqrt{1+(n-1)\rho}},$$

where c is given by

$$c = \frac{\sqrt{E\left\{S^2\right\}/\sigma_{T}^2}}{\sqrt{1+(n-1)\rho}} \tag{26}$$

and h is given by (25). 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 (22), and standard theory for expected mean squares in hierarchical designs (see, e.g., Kirk, 1995) that

$$\mathrm{E}\left\{S^2\right\} = \sigma_{W}^2 + \left(\frac{N-2n}{N-2}\right)\sigma_{B}^2$$

and

$$\mathrm{V}\{S^2\} = \frac{2(N-2)\sigma_{W}^4 + 2n(N-2n)\sigma_{B}^4 + 4(N-2n)\sigma_{B}^2\sigma_{W}^2}{(N-2)^2}.$$

Inserting these values for the mean and variance of S^2 into (26) and (25), using the fact that $\rho\sigma_T^2 = \sigma_B^2$ and $(1-\rho)\sigma_T^2 = \sigma_W^2$, and simplifying gives the values we obtain for c given in (4) and h given in (5).

Unequal Cluster Sample Sizes

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

$$\mathsf{V}\left\{\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C\right\} = \left(\frac{N^T N^C}{N^T + N^C}\right)^{-1}\left(\sigma_W^2 + \tilde{n}\sigma_B^2\right) \tag{27}$$

where $\tilde{n}$ is defined in (14). Therefore the sampling distribution of the numerator of (3) is normal with mean $\sqrt{\tilde{N}}(\mu^T - \mu^C)$ and variance $\sigma_W^2 + \tilde{n}\,\sigma_B^2 = \sigma_T^2[1 + (\tilde{n} - 1)\rho]$.

The expected value and variance of S^2 can be calculated from the analysis of variance between clusters within the treatment groups. When cluster sample sizes are unequal, the between and within cluster sums of squares are still independent, and the within cluster sum of squares has a chi-square distribution, but if $\rho \neq 1$, 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 of S^2, use the fact that

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

where $SSBC^T$ and $SSWC^T$ and $SSBC^C$ and $SSWC^C$ are the sums of squares between and within clusters in the treatment and control groups, respectively. Using the expected

values of the *SSBC*'s and *SSWC*'s given, for example, in equations 77 and 78 on page 70 of Searle, Casella, and McCulloch (1992), we obtain

$$E\left\{S^2\right\} = \sigma_W^2 + \frac{(N - 2\overline{n}_U)\sigma_B^2}{N-2} = \sigma_T^2\left(1 - \frac{2(\overline{n}_U - 1)\rho}{N-2}\right), \tag{28}$$

where $\overline{n}_U$ is given in (15). To obtain the variance of S^2, we use the between clusters variance component estimates in the treatment and control groups

$$\left(\hat{\sigma}_B^T\right)^2 = \frac{MSBC^T - MSWC^T}{n_U^T}$$

and

$$\left(\hat{\sigma}_B^C\right)^2 = \frac{MSBC^C - MSWC^C}{n_U^C},$$

where

$$\boldsymbol{n_U^T} = \frac{\left(\boldsymbol{N^T}\right)^2 - \sum_{i=1}^{\boldsymbol{m^T}}\left(\boldsymbol{n_i^T}\right)^2}{\boldsymbol{N^T}(\boldsymbol{m^T} - 1)},$$

and

$$\boldsymbol{n_U^C} = \frac{\left(\boldsymbol{N^C}\right)^2 - \sum_{i=1}^{\boldsymbol{m^C}}\left(\boldsymbol{n_i^C}\right)^2}{\boldsymbol{N^C}(\boldsymbol{m^C} - 1)}.$$

Then write S^2 as a function of between and within cluster variance components

$$S^2 = \frac{(N^T - m^T)\left(\hat{\sigma}_W^T\right)^2 + (N^C - m^C)\left(\hat{\sigma}_W^C\right)^2 + (m^T - 1)n_U^T\left(\hat{\sigma}_B^T\right)^2 + (m^C - 1)n_U^C\left(\hat{\sigma}_B^C\right)^2}{N-2}.$$

Therefore the variance of S^2 is given by

$$
\begin{aligned}
(N-2)^2\,\mathrm{V}\left\{S^2\right\} &= (N^T-m^T)^2\,\mathrm{V}\left\{\left(\hat{\sigma}_W^T\right)^2\right\}+(N^C-m^C)^2\,\mathrm{V}\left\{\left(\hat{\sigma}_W^C\right)^2\right\} \\
&\quad+\left[\left(m^T-1\right)n_U^T\right]^2\,\mathrm{V}\left\{\left(\hat{\sigma}_B^T\right)^2\right\}+\left[\left(m^C-1\right)n_U^C\right]^2\,\mathrm{V}\left\{\left(\hat{\sigma}_B^C\right)^2\right\} \\
&\quad+2(N^T-m^T)(m^T-1)n_U^T\mathrm{Cov}\left\{\left(\hat{\sigma}_W^T\right)^2,\left(\hat{\sigma}_B^T\right)^2\right\} \\
&\quad+2(N^C-m^C)(m^C-1)n_U^C\mathrm{Cov}\left\{\left(\hat{\sigma}_W^C\right)^2,\left(\hat{\sigma}_B^C\right)^2\right\},
\end{aligned}
\tag{29}
$$

where $\mathrm{Cov}\{X, Y\}$ is the covariance between X and Y. Using expressions (95) and (102) for the variances of the variance component estimates and expression (96) for the covariance term from pages 74 and 75 of Searle, Casella, and McCulloch (1992), and simplifying yields

$$
\mathrm{V}\left\{S^2\right\}=\frac{2\sigma_W^4}{N-2}+\frac{2(N-2\bar{n}_U)\sigma_B^2\sigma_W^2}{(N-2)^2}+\frac{2A\sigma_B^4}{(N-2)^2}, \tag{30}
$$

where $A = A^T + A^C$ defined in (17). Using expression (28) for the expected value and expression (29) for the variance of S^2, inserting these values for the mean and variance of S^2 into (26) and (25), using the fact that $\rho\sigma_T^2 = \sigma_B^2$ and $(1-\rho)\sigma_T^2 = \sigma_W^2$, and simplifying gives the values we obtain for c_U given in (18) and h_U given in (16).

Simulations and Power Computations

The simulation was carried out by a Compac Visual FORTRAN program on a Dell Pentium IV computer. Because the *t*-statistic is invariant under affine transformations of the data, we lose no generality in assuming that (under the null hypothesis) $\mu_i^{\mathrm{T}} = \mu_i^{\mathrm{C}} = 0$ and $\sigma_{\mathrm{W}} = 1$. By the definition of ρ, $\sigma_{\mathrm{B}}^2 = \rho\sigma_{\mathrm{W}}^2/(1-\rho)$. The data analyzed in this simulation was generated based on 10,000 replications for each particular combination of n, m, and ρ values. For each replication, two vectors of standard normal

deviates were generated using IMSL subroutine RRNOR. The first, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_{2nm})$ had $2nm$ values and the second, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{2m})$ had $2m$ values. The elements of the vector $\boldsymbol{\varepsilon}$ were the individual level sampling errors. Because $\sigma_W^2 = 1$, the elements of the vector $\boldsymbol{\alpha}$ were transformed to yield the group effects with variance σ_B^2 by multiplying each element of $\boldsymbol{\alpha}$ by $\sqrt{\rho/(1-\rho)}$. The Y^T and Y^C values were then created via

$$Y^T_{ij} = \alpha_i + \varepsilon_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, mn$$

and

$$Y^C_{ij} = \alpha_i + \varepsilon_{ij},\ i = m + 1, \ldots, 2m;\ j = mn + 1, \ldots, 2mn.$$

The data generated by the program was checked to determine that the observed within- and between-cluster variances, and the intraclass correlations generated were as expected. The *t*-statistic, adjusted *t*-statistic, and value of *h* were computed for each replication. The critical values of Student's *t*-distribution were evaluated using IMSL subroutine TIN and the empirical rejection rates were the proportion of replications in which $|t|$ exceeded the two-tailed critical value of the *t*-distribution with $(N - 2)$ degrees of freedom or the proportion of replications in which $|t_A|$ exceeded the two-tailed critical value of the *t*-distribution with *h* degrees of freedom. All power computations used subroutine TNDF to compute the distribution function of the noncentral *t*-distribution.

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.

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.

Hedberg, E. C. Santana, R., & Hedges, L. V. (2004). The variance structure of academic achievement in America. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego.

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.

Table 1

Actual Significance Levels of Tests Based on Unadjusted t and Adjusted t (t_A) with Nominal Significance Levels of $\alpha = 0.10$, 0.05, and 0.01

			Unadjusted t-test Nominal Significance Level			Adjusted t-test using t_A Nominal Significance Level				
n	m	ρ	0.1	0.05	0.01	0.1	0.05	0.01	c	h
2	2	0.00	0.103	0.052	0.010	0.103	0.052	0.010	1.000	6.0
2	2	0.05	0.105	0.051	0.011	0.096	0.046	0.009	0.968	6.0
2	2	0.10	0.113	0.062	0.016	0.097	0.051	0.011	0.937	5.9
2	2	0.20	0.134	0.070	0.017	0.091	0.046	0.010	0.882	5.8
2	2	0.30	0.164	0.095	0.024	0.103	0.050	0.010	0.832	5.5
2	2	0.40	0.194	0.118	0.033	0.104	0.049	0.009	0.787	5.0
20	5	0.00	0.102	0.051	0.010	0.102	0.051	0.010	1.000	198.0
20	5	0.05	0.245	0.167	0.070	0.103	0.051	0.010	0.713	190.5
20	5	0.10	0.338	0.253	0.133	0.100	0.048	0.010	0.582	170.5
20	5	0.20	0.455	0.372	0.240	0.094	0.048	0.009	0.448	118.5
20	5	0.30	0.541	0.465	0.337	0.096	0.050	0.010	0.375	77.0
20	5	0.40	0.585	0.513	0.391	0.095	0.046	0.009	0.328	50.6
2	20	0.00	0.103	0.047	0.009	0.103	0.047	0.009	1.000	78.0
2	20	0.05	0.116	0.060	0.012	0.107	0.054	0.010	0.975	77.8
2	20	0.10	0.117	0.059	0.012	0.099	0.048	0.009	0.952	77.2
2	20	0.20	0.135	0.073	0.020	0.100	0.050	0.011	0.911	75.0
2	20	0.30	0.150	0.089	0.025	0.102	0.052	0.011	0.874	71.5
2	20	0.40	0.166	0.097	0.030	0.098	0.049	0.010	0.841	67.1
100	2	0.00	0.100	0.050	0.011	0.100	0.050	0.011	1.000	398.0
100	2	0.05	0.511	0.437	0.303	0.102	0.050	0.011	0.405	351.8
100	2	0.10	0.626	0.560	0.445	0.095	0.048	0.010	0.295	256.2
100	2	0.20	0.732	0.684	0.589	0.096	0.048	0.008	0.208	114.8
100	2	0.30	0.784	0.746	0.670	0.100	0.050	0.010	0.166	55.1
100	2	0.40	0.820	0.786	0.724	0.105	0.052	0.009	0.140	29.6

Note: The standard error of the simulated rejection rates of t_A are 0.0030, 0.0022, and 0.0010 for 0.10, 0.05, and 0.01 nominal significance levels, respectively.

Table 2

Rejection rates of tests based on t_K using critical values based on different degrees of freedom

						Test Based on t_K with (N - 2) df		Test Based on t_K with (M - 2) df		Test Based on t_K with (N - 2)/*DEF* df	
n	m	c	*DEF*[a]	d[b]	h	N - 2	Rejection Rate	M - 2	Rejection Rate	(N - 2)/*DEF*	Rejection Rate
ρ =0.1											
10	2	0.708	1.90	0.976	36.0	38	0.055	2	0.0002	20.0	0.049
25	2	0.529	3.40	0.975	86.1	98	0.056	2	0.0001	28.8	0.049
100	2	0.295	10.90	0.975	256.2	398	0.056	2	0.0000	36.5	0.049
10	3	0.714	1.90	0.984	54.3	58	0.054	4	0.0085	30.5	0.049
25	3	0.533	3.40	0.984	125.9	148	0.054	4	0.0072	43.5	0.049
100	3	0.298	10.90	0.983	349.6	598	0.054	4	0.0067	54.9	0.049
10	4	0.717	1.90	0.988	72.6	78	0.053	6	0.0181	41.1	0.050
25	4	0.536	3.40	0.988	166.0	198	0.053	6	0.0167	58.2	0.050
100	4	0.299	10.90	0.988	447.1	798	0.053	6	0.0161	73.2	0.050
10	5	0.719	1.90	0.991	90.9	98	0.052	8	0.0247	51.6	0.050
25	5	0.537	3.40	0.990	206.2	248	0.052	8	0.0234	72.9	0.050
100	5	0.3	10.90	0.990	546.0	998	0.052	8	0.0228	91.6	0.050
10	10	0.722	1.90	0.995	182.6	198	0.051	18	0.0379	104.2	0.050
25	10	0.54	3.40	0.995	407.5	498	0.051	18	0.0372	146.5	0.050
100	10	0.301	10.90	0.995	1045.7	1998	0.051	18	0.0368	183.3	0.050
ρ = 0.2											
10	2	0.569	2.80	0.951	30.6	38	0.061	2	0.0003	13.6	0.049

25	2	0.394	5.80	0.950	60.7	98	0.062	2	0.0001	16.9	0.049
100	2	0.208	20.80	0.949	114.8	398	0.063	2	0.0001	19.1	0.049
10	3	0.579	2.80	0.968	44.9	58	0.057	4	0.0101	20.7	0.049
25	3	0.402	5.80	0.967	84.5	148	0.058	4	0.0087	25.5	0.050
100	3	0.212	20.80	0.966	147.7	598	0.058	4	0.0081	28.8	0.050
10	4	0.584	2.80	0.977	59.4	78	0.055	6	0.0201	27.9	0.050
25	4	0.405	5.80	0.975	109.3	198	0.056	6	0.0187	34.1	0.050
100	4	0.214	20.80	0.975	185.4	798	0.056	6	0.0181	38.4	0.050
10	5	0.587	2.80	0.981	74.1	98	0.054	8	0.0266	35.0	0.050
25	5	0.407	5.80	0.980	134.4	248	0.055	8	0.0254	42.8	0.050
100	5	0.215	20.80	0.980	224.3	998	0.055	8	0.0248	48.0	0.050
10	10	0.592	2.80	0.991	147.4	198	0.052	18	0.0391	70.7	0.050
25	10	0.411	5.80	0.990	261.3	498	0.052	18	0.0384	85.9	0.050
100	10	0.217	20.80	0.990	423.6	1998	0.052	18	0.0381	96.1	0.050

a. DEF = $1 + (n - 1)\rho$, Kish's design effect.

b. $d = \sqrt{1 - \frac{2(n-1)\rho}{N-2}}$

Table 3
Power of the Adjusted *t*-test Based on t_A, GLS, and the Test Based on Cluster Means with the Ratio of the Power of the Test Based on t_A to that Based on GLS when $\mu^T - \mu^C = 1.0\sigma_T$

		GLS Test		Adjusted t-test		Test on Cluster Means		Power
n	*m*	Power	*df*	Power	*h*	Power	*df*	Ratio[a]
$\rho = 0.10$								
10	2	0.609	38	0.607	36.0	0.265	2	1
25	2	0.766	98	0.765	86.1	0.336	2	1
100	2	0.856	398	0.855	256.2	0.393	2	1
10	3	0.789	58	0.788	54.3	0.566	4	1
25	3	0.910	148	0.909	125.9	0.703	4	1
100	3	0.959	598	0.959	349.6	0.790	4	1
10	4	0.893	78	0.893	72.6	0.771	6	1
25	4	0.968	198	0.968	166.0	0.889	6	1
100	4	0.990	798	0.990	447.1	0.943	6	1
10	5	0.949	98	0.948	90.9	0.887	8	1
25	5	0.990	248	0.989	206.2	0.962	8	1
100	5	0.998	998	0.998	546.0	0.986	8	1
10	10	0.999	198	0.999	182.6	0.998	18	1
25	10	1	498	1	407.5	1	18	1
100	10	1	1998	1	1045.7	1	18	1
$\rho = 0.20$								
10	2	0.453	38	0.449	30.6	0.201	2	0.99
25	2	0.538	98	0.533	60.7	0.230	2	0.99
100	2	0.590	398	0.585	114.8	0.248	2	0.99
10	3	0.624	58	0.620	44.9	0.424	4	0.99
25	3	0.714	148	0.710	84.5	0.490	4	0.99
100	3	0.765	598	0.761	147.7	0.531	4	0.99
10	4	0.752	78	0.748	59.4	0.609	6	1
25	4	0.832	198	0.829	109.3	0.689	6	1
100	4	0.872	798	0.870	185.4	0.734	6	1
10	5	0.841	98	0.839	74.1	0.745	8	1
25	5	0.905	248	0.903	134.4	0.819	8	1
100	5	0.934	998	0.932	224.3	0.858	8	1
10	10	0.988	198	0.987	147.4	0.979	18	1
25	10	0.996	498	0.996	261.3	0.992	18	1
100	10	0.998	1998	0.998	423.6	0.996	18	1

a. The ratio of the power of the adjusted *t*-test to that of the GLS test, rounded to two decimal places