Institute for Policy Research
Northwestern University
Working Paper Series

WP-07-13

Effect Sizes in Three Level Cluster Randomized Experiments

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Please do not quote or distribute without permission.

2040 Sheridan Rd. ◆ Evanston, IL 60208-4100 ◆ Tel: 847-491-3395 Fax: 847-491-9916
www.northwestern.edu/ipr, ◆ ipr@northwestern.edu

Abstract

Research designs involving cluster randomization are becoming increasingly important in educational and behavioral research. Many of these designs involve two levels of clustering or nesting (students within classes and classes within schools). Researchers would like to compute effect size indexes based on the standardized mean difference to compare the results of cluster randomized studies with other studies and to combine information across studies in meta-analyses. This paper addresses the problem of defining effect sizes in designs with two levels of clustering and computing estimates of those effect sizes and their standard errors from information that is likely to be reported in journal articles. Five effect sizes are defined corresponding to different standardizations. Estimators of each effect size index are also presented along with their sampling distributions (including standard errors).

Effect Sizes in Three Level Cluster Randomized Experiments

Educational experiments or quasi-experiments used to evaluate the effects of educational interventions, products or technologies often involve several sites (typically schools). One common design assigns entire *schools* to the same treatment group, with different schools assigned to different treatments. This design is often called a *cluster randomized* design because schools correspond to statistical clusters. Several analysis strategies for cluster randomized trials such as this are possible, 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).

Problems of representation of the results of cluster randomized trials (and the corresponding quasi-experiments) in the form of effect sizes and combining them across studies in meta-analyses have received less attention. Rooney and Murray (1996) called attention to the problem of effect size estimation for meta-analysis involving cluster randomized trials. They suggested that conventional estimates were not appropriate and noted that formulas for standard errors that assumed simple random sampling (rather than clustered sampling) were incorrect. Donner and Klar (2002) suggested that corrections for the effects of clustering should be introduced in meta-analyses of cluster randomized experiments. Hedges (in press) suggested definitions of several effect size parameters in cluster randomized experiments with one level of clustering or nesting (so called two-level designs) and gave their sampling distributions. Each of these papers cited above examined meta-analysis for studies using only one level of clustering (that is, two level designs).

The admittedly limited work on the effects of clustering on estimates of effect size and their combination in meta-analysis appears to have had little impact on the practice of meta-analysis. Laopaiboon (2003) reviewed the methods used in 25 published meta-analyses involving cluster randomized experiments, and found that only 3 used methods to account for clustering in their analysis. All of these three were meta-analyses of health care studies using binary outcomes. Of the six meta-analyses involving education, none used methods that addressed the impact of clustering.

The work reported in this paper was stimulated by problems faced by the US Department of Education's What Works Clearinghouse, 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 computing an estimate of effect size and its uncertainty. While some of these studies involved only one classroom per school, others involved multiple classrooms per school and thus the studies have three level designs involving a second level of clustering that may need to be taken into account.

While methods have now been developed for computing effect size estimates and their sampling properties in two level designs (e.g., Hedges, in press), less is known about methods for taking into account the two levels of clustering that are present in three

level designs. Three level designs are widespread in education (e.g., designs that assign schools with multiple classrooms within schools to treatments). 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 sampling plan that first sampling schools and then classrooms.

This paper has two purposes. One is to examine the problem of defining effect sizes for cluster randomized experiments with two levels of clustering or nesting (so-called three-level designs). The second is to examine how to estimate these effect sizes and obtain standard errors for them from statistics that are typically given in reports of research (that is, without a reanalysis of the raw data).

Although we use the terms "schools" and "classrooms" to characterize the stages of clustering, this is merely a matter of convenient and readily understandable terminology. The results of this paper apply equally well to any situation in which there is a two-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] but treatments are assigned to clusters (e.g., schools).

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

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

the pooled within-schools variance S_{WS}^2 via

$$S_{WS}^2 = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\sum_{k=1}^{n_{ij}^T}(Y_{ijk}^T - \bar{Y}_{i\bullet\bullet}^T)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\sum_{k=1}^{n_{ij}^C}(Y_{ijk}^C - \bar{Y}_{i\bullet\bullet}^C)^2}{N-M}, \quad (2)$$

and the (pooled) within-classroom sample variance S_{WC}^2 via

$$S_{WC}^2 = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\sum_{k=1}^{n_{ij}^T}(Y_{ijk}^T - \bar{Y}_{ij\bullet}^T)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\sum_{k=1}^{n_{ij}^C}(Y_{ijk}^C - \bar{Y}_{ij\bullet}^C)^2}{N-P}. \quad (3)$$

Define the between schools but within treatment groups variance S_{BS}^2 via

$$S_{BS}^2 = \frac{\sum_{j=1}^{m^T}(\bar{Y}_{i\bullet\bullet}^T - \bar{Y}_{\bullet\bullet\bullet}^T)^2 + \sum_{j=1}^{m^C}(\bar{Y}_{i\bullet\bullet}^C - \bar{Y}_{\bullet\bullet\bullet}^C)^2}{M-2}, \quad (4)$$

and the (pooled) between classrooms but within treatment groups and schools variance S_{BC}^2 via

$$S_{BC}^2 = \frac{\sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}(Y_{ij\bullet}^T - \bar{Y}_{i\bullet\bullet}^T)^2 + \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}(Y_{ij\bullet}^C - \bar{Y}_{i\bullet\bullet}^C)^2}{P-M}. \quad (5)$$

Suppose that observations within the j^{th} classroom in the i^{th} school within the treatment and control group groups are normally distributed about 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

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

Suppose further that the 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

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

and

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

Finally suppose that the 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

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

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 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

$$\sigma_{WT}^2 = \sigma_{BS}^2 + \sigma_{BC}^2 + \sigma_{WC}^2 .$$

In most educational achievement data when clusters are schools and 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_{WT}^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

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

Similarly, define the classroom level intraclass correlation ρ_C by

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

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

Effect Sizes

The effect sizes typically used in educational and psychological research are standardized mean differences, defined as the ratio of a difference between treatment and control group means to a standard deviation (see, e.g., Hedges, 1981). Effect sizes represent the magnitude of treatment effects in a metric that is intended to be interpretable in the same way across different studies and thus facilitates comparability and synthesis of experimental findings across studies. For this reason, effect sizes have been mandated in reports of educational and psychological research by both the American Psychological Association and the American Educational Research Association.

In single level designs the notion of standardized mean difference is often unambiguous: There is only one possibility because there is only one within-treatment-groups standard deviation. In multi-site designs and multi-level designs such as cluster randomized trials, there are several possible standardized mean differences. The possibilities in two level designs were discussed by Hedges (in press). In this section we clarify the possible effect size definitions in three-level designs.

The alternative possibilities for the standard deviation lead to different possible definitions for the population effect size in three-level designs. The choice of one of these effect sizes should be determined on the basis of the inference of interest to the researcher. If the effect size measures are to be used in meta-analysis, an important inference goal may be to estimate parameters that are comparable with those that can be estimated in other studies. In such cases, the standard deviation may be chosen to be the same kind of standard deviation used in other studies to which this study will be compared. We focus on three effect sizes that seem likely to be the most useful (meaning the most widely used), but indicate two others that may be useful in some circumstances.

One effect size standardizes the mean difference by the total standard deviation within-treatment groups. It is of the form

$$\delta_{WT} = \frac{\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C}}{\sigma_{WT}} . \tag{8}$$

The effect size δ_{WT} might be of interest in a meta-analysis where the other studies have three-level designs involving multiple schools and classrooms or studies that sample from a broader population but do not include schools or classes as clusters in the sampling design (this would typically imply that they used an individual, rather than a cluster, assignment strategy). In such cases, δ_{WT} might be the effect size that is most comparable with the effect sizes in other studies.

If $\sigma_{BC}^2 + \sigma_{WC}^2 \neq 0$ (and hence $\rho_S \neq 1$), a second effect size can be defined that standardizes the mean difference by the standard deviation within-schools. It is of the form

$$\delta_{WS} = \frac{\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C}}{\sqrt{\sigma_{BC}^2 + \sigma_{WC}^2}} = \frac{\delta_{WT}}{\sqrt{1-\rho_S}} . \tag{9}$$

The effect size δ_{WS} might be of interest in a situation where most of the studies of interest had two level designs involving multiple schools. In such studies δ_{WS} may (implicitly or explicitly) be the effect size estimated and hence be most comparable with the effect size estimates of the other studies.

If $\sigma_{WC} \neq 0$ (and hence $\rho_S + \rho_C \neq 1$), a third effect size can be defined that standardizes the mean difference by the within-classroom standard deviation. It has the form

$$\delta_{WC} = \frac{\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C}}{\sigma_{WC}} = \frac{\delta_{WT}}{\sqrt{1-\rho_S-\rho_C}} = \frac{\delta_{WS}\sqrt{1-\rho_S}}{\sqrt{1-\rho_S-\rho_C}} . \tag{10}$$

The effect size δ_{WC} might be of interest, for example, in a meta-analysis where the other studies to which the current study is compared are typically studies with a single school. In such studies δ_{WC} may (implicitly) be the effect size estimated and hence δ_{WC} might be the effect size most comparable with the effect size estimates in other studies.

Two other effect sizes that standardize the mean difference by between classroom and between school standard deviations may be of interest (although less frequently). If $\sigma_{BS} \neq 0$ (and hence $\rho_S \neq 0$), a fourth possible effect size would be

$$\delta_{BS} = \frac{\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C}}{\sigma_{BS}} = \frac{\delta_{WT}}{\sqrt{\rho_S}}. \qquad (11)$$

The effect size δ_{BS} may also be of interest in a meta-analysis where the studies being compared are typically multi-level studies that have been analyzed by using school means as the unit of analysis. This effect size is less likely to be of general interest, but it might be of interest in cases where the outcome variable on which the treatment effect is defined is conceptually at the level of schools and the individual observations are of interest only because the average defines an aggregate property.

For example, Bryk and Schneider (2000) study the level of school trust as an important predictor of school functioning. Trust level in a school is measured as an aggregate of reported trust scores of individual students or teachers, but only the school means are meaningful as measures of trust as a property of the school. Thus an intervention that was targeted at increasing school trust might well measure trust in a school as the mean of trust scores across students, and evaluate the treatment effect as the difference in means between all students in schools assigned to the treatment group and those in schools assigned to the control group. In this case, only σ_{BS} represents a standard deviation of a meaningful aggregate (school average trust), while σ_{BC} and σ_{WC} do not, and therefore only δ_{BS} seems conceptually meaningful as an effect size.

If $\sigma_{BC} \neq 0$ (and hence $\rho_C \neq 0$), a fifth possible effect size would be

$$\delta_{BC} = \frac{\mu_{\bullet\bullet}^{T} - \mu_{\bullet\bullet}^{C}}{\sigma_{BC}} = \frac{\delta_{WT}}{\sqrt{\rho_C}}. \qquad (12)$$

The effect size δ_{BC} may also be of interest in a meta-analysis where the studies being compared are typically multi-level studies that have been analyzed by using classroom means as the unit of analysis or the outcome is conceptually defined as a classroom-level property. This effect size might be of interest in cases where the outcome variable on which the treatment effect is defined conceptually at the level of classrooms and the individual observations are of interest only because the classroom average defines an aggregate property.

For example, consider classroom climate as an important determinant of instructional effectiveness (see, e.g., Dunkin and Biddle, 1974). Classroom climate is measured as an aggregate of reports by students, but only classroom means are meaningful as measures of classroom climate. An intervention that was intended to improve classroom climate might well measure the intervention effect via the difference between the means of classrooms assigned to the intervention group and those assigned to the control group. In this case both σ_{BC} and σ_{BS} represent standard deviations of a meaningful aggregate score (while σ_{WC} does not). Therefore either δ_{BS} or δ_{BC} would be conceptually meaningful effect sizes, but δ_{BC} might be preferred if many studies were carried out within single schools.

Because ρ_S and ρ_C will often be rather small, δ_{WS} and δ_{WC} will often be similar in magnitude to δ_{WT}. For the same reason, δ_{BS} and δ_{BC} will typically be much larger (in the same population) than δ_{WS}, δ_{WC}, or δ_{WT}. Note also that if all of the effect sizes are defined (that is, if $0 < \rho_S < 1$, $0 < \rho_C < 1$, and $\rho_S + \rho_C < 1$), any one of these effect sizes may be

obtained algebraically from any of the others, provided ρ_C and ρ_S are known. In particular, equations (8) through (11) can be solved for δ_{WT} in terms of any of the other effect sizes, ρ_S, and ρ_C. These same equations, in turn can be used to define any of the five effect sizes in terms of δ_{WT}.

Estimates of Effect Sizes: Equal Sample Sizes

In this section we present estimates of the effect sizes and their approximate sampling distributions when all of the schools have the same number, p, of classrooms, that is when $p_i^T = p$, $i = 1, \ldots, m^T$ and $p_i^C = p$, $i = 1, \ldots, m^C$, and all the classroom sample sizes are equal to n, that is when $n_{ij}^T = n$, $i = 1, \ldots, m^T$; $j = 1, \ldots, p$ and $n_{ij}^C = n$, $i = 1, \ldots, m^C$; $j = 1, \ldots, p$ In this case $N^T = npm^T$, $N^C = npm^C$, and $N = N^T + N^C = np(m^T + m^C) = nM$. Derivations and the details of the small sample distribution of the effect size estimators are given in the Appendix.

We present results explicitly for the case of equal school and classroom sample sizes for two reasons. The first reason is that most designs attempt to achieve equal sample sizes but specific (realized) school and classroom sample sizes are rarely reported, so that the equal sample size formulas will be of most practical use. The second reason is that the results become considerably more complicated when sample sizes are unequal—sufficiently complicated that it is difficult to obtain much insight from examining the formulas in the unequal sample size case.

Estimation of δ_{WC}

We start with estimation of δ_{WC}, which is the most straightforward. If $\rho_S + \rho_C \neq 1$, then δ_{WC} is defined and the estimate

$$d_{WC} = \frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{S_{WC}} \tag{13}$$

is a consistent estimator of δ_{WC}. The estimator d_{WC} is approximately normally distributed about δ_{WC} with variance

$$V\{d_{WC}\} = \frac{1 + (pn-1)\rho_S + (n-1)\rho_C}{\tilde{m}pn(1-\rho_S-\rho_C)} + \frac{\delta^2_{WC}}{2(N-Mp)}, \tag{14}$$

where

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

An estimate v_{WC} of the variance of d_{WC} can be computed by substituting the consistent estimate d_{WC} for δ_{WC} in equation (14) above.

The leading term of the variance in equation (14) arises from uncertainty in the mean difference. Note that it is $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]/(1 - \rho_S - \rho_C)]$ as large as would be expected if there were no clustering in the sample (that is if $\rho_S = \rho_C = 0$). Thus $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]/(1 - \rho_S - \rho_C)]$ is a kind of variance inflation factor for the variance of the effect size estimate d_{WC}.

Note that if either $\rho_S = 0$ or $\rho_C = 0$, there is no clustering at one level and the three-level design logically reduces to a two-level design. If $\rho_S = 0$, the expression (13) for d_{WC} and expression (14) for its variance reduce to those given in equations (7) and (8) in Hedges (in press) for the corresponding effect size estimate (called there d_W) and its variance in a two-level design with clusters of size n and intraclass correlation $\rho = \rho_C$. Similarly, if $\rho_C = 0$, expressions (13) and (14) reduce to the estimate given in equation (7)

and its variance given in equation (8) of Hedges (in press) with cluster size np and intraclass correlation $\rho = \rho_S$. If $\rho_S = \rho_C = 0$ so that there is no clustering by classroom or by school, the design is logically equivalent to a one-level design. In this case the expression (13) for d_{WC} and expression (14) for its variance reduce to the usual expressions for the standardized mean difference and its variance (see, e.g., Hedges, 1981).

Estimation of δ_{WS}

If $\rho_S \neq 1$,then δ_{WS} is defined and an estimate of δ_{WS} can also be obtained from the mean difference and S_{WS} if the intraclass correlations ρ_S and ρ_C are known or can be imputed. A direct argument shows that a consistent estimator of δ_{WS} is

$$d_{WS} = \left(\frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{S_{WS}} \right) \sqrt{1 - \frac{(n-1)\rho_C}{(1-\rho_S)(np-1)}} \,. \qquad (15)$$

The estimate d_{WS} is normally distributed in large samples with variance

$$V\{d_{WS}\} = \frac{1 + (pn-1)\rho_S + (n-1)\rho_C}{\tilde{m}pn(1-\rho_S)} + \delta^2_{WS} \left(\frac{(pn-1)\bar{\rho}^2 + 2n(p-1)\bar{\rho}\rho_C + n^2(p-1)\rho_C^2}{2M\left[(pn-1)^2(1-\rho_S)^2 - (pn-1)(n-1)\rho_C(1-\rho_S)\right]} \right) \,, \qquad (16)$$

where $\bar{\rho} = (1 - \rho_S - \rho_C)$. An estimate v_{WS} of the variance of d_{WS} can be computed by substituting the consistent estimate d_{WS} for δ_{WS} in equation (16) above.

The leading term of the variance in equation (16) arises from uncertainty in the mean difference. Note that it is $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]/(1 - \rho_S)]$ as large as would be expected if there were no clustering in the sample (that is if $\rho_S = \rho_C = 0$). Thus the term $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]/(1 - \rho_S)]$ is a kind of variance inflation factor for the variance of the effect size estimate d_{WS}.

Note that if $\rho_C = 0$ so that there is no clustering effect by classrooms, the three level design is logically equivalent to a two-level design. In this case, the expression (15) for d_{WS} and expression (16) for its variance reduce to those given in equations (7) and (8) in Hedges (in press) for the corresponding effect size estimate (called there d_W) and its variance in a two-level design with clusters of size pn and intraclass correlation $\rho = \rho_S$. If $\rho_S = \rho_C = 0$ so that there is no clustering by classroom or by school, the design is logically equivalent to a one-level design. In this case the expression (15) for d_{WS} and expression (16) for its variance reduce to those for the standardized mean difference and its variance in a one-level design (see, e.g., Hedges, 1981).

Estimation of δ_{WT}

An estimate of δ_{WT} can also be obtained from the mean difference and S_{WT} if the intraclass correlations ρ_S and ρ_C are known or can be imputed. A direct argument shows that a consistent estimator of δ_{WT} is

$$d_{WT} = \left(\frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{S_{WT}} \right) \sqrt{1 - \frac{2(pn-1)\rho_S + 2(n-1)\rho_C}{N-2}} \,. \qquad (17)$$

The estimate d_{WT} is normally distributed in large samples with variance

$$V\{d_{WT}\} = \frac{1+(pn-1)\rho_S+(n-1)\rho_C}{\tilde{m}pn} + \delta_{WT}^2\left(\frac{pn\hat{N}\rho_S^2+n\breve{N}\rho_C^2+(N-2)\bar{\rho}^2+2n\hat{N}\rho_S\rho_C+2\hat{N}\rho_S\bar{\rho}+2\breve{N}\rho_C\bar{\rho}}{2(N-2)\left[(N-2)-2(pn-1)\rho_S-2(n-1)\rho_C\right]}\right), \quad (18)$$

where $\hat{N} = (N - 2pn)$, $\breve{N} = (N - 2n)$, and $\bar{\rho} = 1 - \rho_S - \rho_C$. An estimate v_{WT} of the variance of d_{WT} can be computed by substituting the consistent estimate d_{WT} for δ_{WT} in equation (18) above.

The leading term of the variance in equation (18) arises from uncertainty in the mean difference. Note that this leading term is $[1 + (np - 1)\rho_S + (n - 1)\rho_C]$ as large as would be expected if there were no clustering in the sample (that is if $\rho_S = \rho_C = 0$). The term $[1 + (np - 1)\rho_S + (n - 1)\rho_C]$ is a generalization to two stages of clustering of the variance inflation factor mentioned by Donner (1981) for the variance of means in clustered samples (with one stage of clustering) and also corresponds to a variance inflation factor for the effect size estimate d_{WT}.

Note that if $\rho_C = 0$ so that there is no clustering by classrooms, or if $\rho_S = 0$ so that there is no clustering by schools, the three-level design is logically equivalent to a two-level design. If $\rho_S = 0$, the expression (17) for d_{WT} and expression (18) for its variance reduce to those given in equations (15) and (16) in Hedges (in press) for the corresponding effect size estimate and its variance in a two-level design with clusters of size n and intraclass correlation $\rho = \rho_C$. Similarly, if $\rho_C = 0$, expressions (17) and (18) reduce to the estimate given in equation (15) and its variance given in equation (16) of Hedges (in press) for the corresponding effect size in a two-level design with cluster size np and intraclass correlation $\rho = \rho_S$.

Estimation of δ_{BC}

If $\rho_C \neq 0$ (so that δ_{BC} is defined), S_{BC} may be used, along with the intraclass correlations ρ_S and ρ_C, to obtain an estimate of δ_{BC}. A direct argument shows that

$$d_{BC} = \frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{S_{BC}}\sqrt{\frac{1-\rho_S+(n-1)\rho_C}{n\rho_C}} \quad (19)$$

is a consistent estimate of δ_{BC}. This estimate is normally distributed in large samples with variance

$$V\{d_{BC}\} = \frac{1-\rho_S+(n-1)\rho_C}{\tilde{m}pn\rho_C} + \frac{[1-\rho_S+(n-1)\rho_C]\delta_{BC}^2}{2M(p-1)n\rho_C}. \quad (20)$$

An estimate v_{BC} of the variance of d_{BC} can be computed by substituting the consistent estimate d_{BC} for δ_{BC} in equation (20) above. Note that the presence of ρ_C in the denominator of the estimator and its variance is possible since the effect size parameter only exists if $\rho_C \neq 0$.

The variance in equation (20) is $[1 - \rho_S + (n - 1)\rho_C]/n\rho_C]$ as large as the variance of the standardized mean difference computed from an analysis using class means as the unit of analysis if the data had come from a simple random sample. Thus $[1 - \rho_S + (n - 1)\rho_C]/n\rho_C]$ is a kind of variance inflation factor for the variance of effect size estimates like d_{BC} compared to this alternative effect size estimate.

Note that if $\rho_S = 0$ so that there is no clustering by school, the expression (19) for d_{BC} and expression (20) for its variance reduce to those given in equations (11) and (12)

in Hedges (in press) for the corresponding effect size estimate (called there d_{B2}) and its variance in a two-level design with clusters of size n and intraclass correlation $\rho = \rho_C$.

Estimation of δ_{BS}

If $\rho_S \neq 0$ (so that δ_{BS} is defined), S_{BS} may be used, along with the intraclass correlations ρ_S and ρ_C, to obtain an estimate of δ_{BS}. A direct argument shows that

$$d_{BS} = \frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{S_{BS}} \sqrt{\frac{1-(pn-1)\rho_S + (n-1)\rho_C}{pn\rho_S}} \tag{21}$$

is a consistent estimate of δ_{BS}. This estimate is normally distributed in large samples with variance

$$V\{d_{BS}\} = \frac{1+(pn-1)\rho_S + (n-1)\rho_C}{\tilde{m}pn\rho_S} + \frac{[1+(pn-1)\rho_S + (n-1)\rho_C]\delta^2_{BS}}{2(M-2)pn\rho_S}. \tag{22}$$

An estimate v_{BS} of the variance of d_{BS} can be computed by substituting the consistent estimate d_{BS} for δ_{BS} in equation (22) above. Note that the presence of ρ_S in the denominator of the estimator and its variance is possible since the effect size parameter only exists if $\rho_S \neq 0$.

The variance in equation (22) is $[1 + (pn - 1)\rho_S + (n - 1)\rho_C]/pn\rho_S]$ as large as the variance of the standardized mean difference computed from a simple random sample. Thus $[1 - \rho_S + (n - 1)\rho_C]/pn\rho_S]$ is a kind of variance inflation factor for the variance of effect size estimates like d_{BS} compared to this alternative effect size estimate.

Note that if $\rho_C = 0$ so that there is no clustering by classroom, expressions (21) and (22) reduce to the corresponding effect size estimate (called there d_{B2}) given in equation (11) and its variance given in equation (12) of Hedges (in press) in a two-level design with cluster size np and intraclass correlation $\rho = \rho_S$.

Estimation of Effect Size: Unequal Sample Sizes

When classroom or school sample sizes are unequal, expressions for the effect size estimates are more complex and expressions for their variances are much more complex. In this section we give expressions for several of the effect size estimates and their sampling distributions. These expressions may be of use in cases where the number of classrooms and students within classrooms are markedly unequal and they are given explicitly in reports of research. They also provide some in sight into what single "compromise" value of p and n might give most accurate results when substituted in to equal sample size formulas.

Estimation of δ_{WC}

If $\rho_S + \rho_C \neq 1$, then the estimate d_{WC} of δ_{WC} when sample sizes are unequal is identical to that in the equal sample size case given in equation (13), and it is approximately normally distributed. However in the case of unequal sample sizes, the variance is given by

$$V\{d_{WC}\} = \frac{1+(p_U-1)\rho_S + (n_U-1)\rho_C}{\tilde{N}(1-\rho_S-\rho_C)} + \frac{\delta^2_{WC}}{2(N-P)}, \tag{23}$$

where

$$p_U = \frac{N^C \sum_{i=1}^{m^T} \left(\sum_{j=1}^{p_i^T} n_{ij}^T \right)^2}{NN^T} + \frac{N^T \sum_{i=1}^{m^C} \left(\sum_{j=1}^{p_i^C} n_{ij}^C \right)^2}{NN^C}, \quad (24)$$

$$n_U = \frac{N^C \sum_{i=1}^{m^T} \sum_{j=1}^{p_i^T} \left(n_{ij}^T \right)^2}{NN^T} + \frac{N^T \sum_{i=1}^{m^C} \sum_{j=1}^{p_i^C} \left(n_{ij}^C \right)^2}{NN^C}, \quad (25)$$

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

and

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

is the total number of classrooms. Note that if if all the n_i^T and n_i^C are equal to n, then $n_U = n$, and that if all the p_i^T and p_i^C are equal to p, then $p_U = pn$, and $P = Mp$, so that (23) reduces to (14). As in the case of the equal sample size formulas an estimate v_{WC} of the variance of d_{WC} can be computed by substituting the consistent estimate d_{WC} for δ_{WC} in equation (23) above.

Estimation of δ_{WS}

If $\rho_S \neq 1$,then δ_{WS} is defined and an estimate d_{WS} is the unequal sample size case becomes

$$d_{WS} = \left(\frac{\bar{Y}_{\bullet\bullet\bullet}^T - \bar{Y}_{\bullet\bullet\bullet}^C}{S_{WS}} \right) \sqrt{1 - \frac{(\tilde{n} - 1)\rho_C}{(1 - \rho_S)(N_S - 1)}}, \quad (26)$$

where

$$M\tilde{n} = M(\tilde{n}^T + \tilde{n}^C) = \sum_{i=1}^{M^T} \left(\sum_{j=1}^{p_i^T} \left(n_{ij}^T \right)^2 \Big/ n_{i+}^T \right) + \sum_{i=1}^{M^C} \left(\sum_{j=1}^{p_i^C} \left(n_{ij}^C \right)^2 \Big/ n_{i+}^C \right), \quad (27)$$

$$n_{i+}^T = \sum_{j=1}^{p_i^T} n_{ij}^T,$$

$$n_{i+}^C = \sum_{j=1}^{p_i^C} n_{ij}^C,$$

and $N_S = N/M$ is the average sample size per school. Note that if all of the n_{ij}^T and n_{ij}^C are equal to n and all the p_j^T and p_j^C are equal to p, then $\tilde{n} = n$, $N_S = pn$, and (26) reduces to (15).

The estimate d_{WS} is normally distributed in large samples with variance

$$V\{d_{WS}\} = \frac{1+(p_U-1)\rho_S+(n_U-1)\rho_C}{\tilde{N}(1-\rho_S)} + \delta_{WS}^2\left(\frac{(N_S-1)\overline{\rho}^2+2(N_S-\tilde{n}_1)\overline{\rho}\rho_C+A\rho_C^2}{2M\left[(N_S-1)^2(1-\rho_S)^2-(N_S-1)(\tilde{n}_1-1)\rho_C(1-\rho_S)\right]}\right), \quad (28)$$

where the auxiliary constant $A = (A^T + A^C)/M$ and A^T and A^C are given by

$$A^T = \sum_{i=1}^{m^T}\left(\left[\sum_{j=1}^{p_i^T}\left(n_{ij}^T\right)^2\right]^2 \Big/ \left(n_{i+}^T\right)^2\right) + \sum_{i=1}^{m^T}\sum_{j=1}^{p_i^T}\left(n_{ij}^T\right)^2 \Big/ N + \sum_{i=1}^{m^T}\left(\sum_{j=1}^{p_i^T}\left(n_{ij}^T\right)^3 \Big/ n_{i+}^T\right)^2$$

and (29)

$$A^C = \sum_{i=1}^{m^C}\left(\left[\sum_{j=1}^{p_i^C}\left(n_{ij}^C\right)^2\right]^2 \Big/ \left(n_{i+}^C\right)^2\right) + \sum_{i=1}^{m^C}\sum_{j=1}^{p_i^C}\left(n_{ij}^C\right)^2 \Big/ N + \sum_{i=1}^{m^C}\left(\sum_{j=1}^{p_i^C}\left(n_{ij}^C\right)^3 \Big/ n_{i+}^C\right)^2 .$$

Note that if all of the n_{ij}^T and n_{ij}^C are equal to n and all the p_j^T and p_j^C are equal to p, then $\tilde{n}_1 = n$, $P = pn$, $A = n^2(p-1)$, and (28) reduces to (16).

The second term of (28) is rather complex and involves detailed sample size information that may not be available. A direct argument shows that the second term of (28) is always smaller than $\delta_{WS}^2/2[P-M]$, therefore a slightly conservative overestimate of the variance of d_{WS} is obtained from

$$V\{d_{WS}\} \approx \frac{1+(p_U-1)\rho_S+(n_U-1)\rho_C}{\tilde{N}(1-\rho_S)} + \frac{\delta_{WS}^2}{2(P-M)} . \quad (30)$$

As in the case of the equal sample size formulas an estimate v_{WS} of the variance of d_{WS} can be computed by substituting the consistent estimate d_{WS} for δ_{WS} in equation (28) or (30) above.

Estimation of δ_{WT}

A consistent estimate d_{WT} of δ_{WT} in the case of unequal sample sizes is

$$d_{WT} = \left(\frac{\overline{Y}_{\bullet\bullet\bullet}^T - \overline{Y}_{\bullet\bullet\bullet}^C}{S_{WT}}\right)\sqrt{1-\frac{2(p_U-1)\rho_S+2(n_U-1)\rho_C}{N-2}} , \quad (31)$$

where p_U is given by (23) and n_U is given by (24) above. Note that if all of the n_{ij}^T and n_{ij}^C are equal to n and all the p_j^T and p_j^C are equal to p, (29) reduced to (16).

The exact expression for the variance of d_{WT} is quite complex when sample sizes are unequal. The simplest expression that we have been able to obtain for the variance of d_{WT} is about a page and a half in length. The complexity of the expression is not unexpected. It is quite similar to that of the variance component estimates from which it is derived, which are also quite complex in unbalanced designs with two nested factors (see e.g., Searle, 1971, pp. 473 - 474).

Note however that the variance consists of two terms. The first term arises from the uncertainty of the mean difference, and it typically much larger than the second term, which arises because of the uncertainty of the standard deviation. Because the leading

term almost completely determines the variance, we suggest a simple approximation to the variance of d_{WT} when sample sizes are unequal that makes use of the exact first term and an approximation to the second term. That approximation is

$$V\{d_{WT}\} = \frac{1+(p_U-1)\rho_S+(n_U-1)\rho_C}{\tilde{N}} + \delta_{WT}^2\left(\frac{p_U\widehat{N}_U\rho_S^2+n_U\breve{N}_U\rho_C^2+(N-2)\overline{\rho}^2+2n_U\widehat{N}_U\rho_S\rho_C+2\widehat{N}_U\rho_S\overline{\rho}+2\breve{N}_U\rho_C\overline{\rho}}{2(N-2)\left[(N-2)-2(p_U-1)\rho_S-2(n_U-1)\rho_C\right]}\right) \quad (32)$$

where $\widehat{N}_U = (N - 2p_U)$, $\breve{N}_U = (N - 2n_U)$, and $\overline{\rho} = 1 - \rho_S - \rho_C$. Note that if if all the n_i^T and n_i^C are equal to n, then $n_U = n$, and that if all the p_i^T and p_i^C are equal to p, then $p_U = pn$, and (32) reduces to (18).

However the approximation above for the variance of d_{WT} is still rather complex and a simpler conservative approximation might be desired. A direct argument shows that the second term of (32) is always smaller than $\delta_{WT}^2/2[M - 2]$, therefore a slightly conservative overestimate of the variance of d_{WT} is obtained from

$$V\{d_{WT}\} \approx \frac{1+(p_U-1)\rho_S+(n_U-1)\rho_C}{\tilde{N}} + \frac{\delta_{WT}^2}{2(M-2)}. \quad (33)$$

As in the case of the equal sample size formulas an estimate v_{WT} of the variance of d_{WT} can be computed by substituting the consistent estimate d_{WT} for δ_{WT} in equation (32) or (33) above.

Estimation of δ_{BC} and δ_{BS}

There is more than one way to generalize the estimator d_{BC} to the case of unequal school and classroom sample sizes. One possibility is to use the means of the classroom means in the treatment and control group in the numerator and standard deviation of the classroom means in the denominator. This corresponds to using the classroom means as the unit of analysis. Another possibility for the numerator is to use the grand means in the treatment and control groups. Similarly, there are multiple possibilities for the denominator, such as the mean square between classrooms. When school and classroom sample sizes are identical, then all of these approaches are equivalent in the sense that the effect size estimates are identical. When the cluster sample sizes are not identical, the resulting estimators (and their sampling distributions) are not the same.

Similarly, there is more than one way to generalize the estimator d_{BS} to the case of unequal school and classroom sample sizes. One possibility is to use the means of the school means in the treatment and control group in the numerator and standard deviation of the school means in the denominator. This corresponds to using the school means as the unit of analysis. Another possibility for the numerator is to use the grand means in the treatment and control groups. Similarly, there are multiple possibilities for the denominator, such as the mean square between schools. When school and classroom sample sizes are identical, then all of these approaches are equivalent in the sense that the effect size estimates are identical. When the school and classroom sample sizes are not identical, the resulting estimators (and their sampling distributions) are not the same.

Because there are so many possibilities, because some of them are rather complex, and because the information necessary to use them (i.e., means, standard deviations, and sample sizes for each classroom and school) is frequently not reported in research reports, we do not give the estimates or their sampling distributions in this paper. A sensible

approach if these estimators are needed is to compute the variance estimates for the equal sample size cases, substituting the equal sample size formulas.

Confidence Intervals for δ_{WC}, δ_{WS}, δ_{WT}, δ_{BC}, and δ_{BS}

The results in this paper can also be used to compute confidence intervals for effect sizes. If δ is any one of the effect sizes mentioned, d is a corresponding estimate, and v_d is the estimated variance of d, then a $100(1 - \alpha)$ percent confidence interval for δ based on d and v_d is given by

$$d - c_{\alpha/2}\sqrt{v_d} \le \delta \le d + c_{\alpha/2}\sqrt{v_d}, \quad (34)$$

where $c_{\alpha/2}$ is the $100(1 - \alpha/2)$ percent point of the standard normal distribution (e.g., 1.96 for $\alpha/2 = 0.05/2 = 0.025$).

Applications in Meta-analysis

The statistical results in this paper should be useful in deciding what effect sizes are desirable in a three level experiment or quasi-experiment. They should also be useful for finding ways to compute effect size estimates and their variances from data that may be reported. We illustrate applications in some examples in the sections that follow.

Obtaining Values of Intraclass Correlations

Intraclass correlations are needed for the methods described in this paper are often not reported, particularly when these effect sizes are calculated retrospectively in meta-analyses. However, because plausible values of ρ are essential 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. 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. Hedges and Hedberg (2007) presented a compendium of intraclass correlations computed from surveys of academic achievement.

While most of these surveys have focused in intraclass correlations at a single level of nesting, they do provide some empirical information and guidelines about patterns of intraclass correlations at lower, versus higher, levels of nesting. We know of two sources of information about classroom level intraclass correlations from reasonably large samples. One is project STAR, which collected data on Kindergarten through third grade students in 79 schools in Tennessee. Nye, Konstantopoulos, and Hedges (2004) reported variance components from a three level model that permits the computation of school and classroom level intraclass correlations (our ρ_S and ρ_C). The second source of intraclass correlation data is based on analyses of the National Assessment of Educational Progress (NAEP). Konstantopoulos (personal communication) used NAEP to estimate school and classroom level intraclass correlations in reading and mathematics achievement for the 1992 and 1996 fourth grade NAEP assessments. The values of intraclass correlations from these sources are given in Table 1. The values of school level intraclass correlations (ρ_S values) are very consistent with those obtained with those obtained by Hedges and Hedberg (2007) using data from other surveys. The classroom

level intraclass correlations (ρ_C values) fall in a range of 0.08 to 0.14. The school level intraclass correlations (ρ_S values) range from 0.11 to 0.18 in project STAR, but are a little larger in NAEP, possibly because of its broader (nationally representative) sample of schools. While these values may be useful in suggesting a plausible range of classroom level intraclass correlations, more data on classroom intraclass correlations is clearly needed.

Computing Effect Sizes When Individuals are (Incorrectly) the Unit of Analysis

The results given in this paper can be used to produce effect size estimates and their variances from studies that incorrectly analyze cluster randomized trials as if individuals were randomized. The required means, standard deviations, and sample sizes cannot always be extracted from what may be reported, but often it is possible to extract the information to compute at least one of the effect sizes discussed in this paper.

Suppose it is decided that the effect size δ_{WT} is appropriate because most other studies both assign and sample individually from a clustered population. Suppose that the data in a study are analyzed by ignoring clustering, then the test statistic is likely to be either

$$t = \sqrt{\frac{N^T N^C}{N^T + N^C}}\left(\frac{\overline{Y}^T_{\bullet\bullet\bullet} - \overline{Y}^C_{\bullet\bullet\bullet}}{S_{WT}}\right)$$

or

$$F = \left(\frac{N^T N^C}{N^T + N^C}\right)\left(\frac{\overline{Y}^T_{\bullet\bullet\bullet} - \overline{Y}^C_{\bullet\bullet\bullet}}{S_{WT}}\right)^2.$$

Either can be solved for

$$\left(\frac{\overline{Y}^T_{\bullet\bullet\bullet} - \overline{Y}^C_{\bullet\bullet\bullet}}{S_{WT}}\right),$$

which can then be inserted into equation (18) or (31) along with ρ_S and ρ_C to obtain d_{WT}. This estimate of d_{WT} can then be inserted into equation (19) or (32) to obtain v_{WT}, an estimate of the variance of d_{WT}.

Alternatively, suppose it is decided that the effect size δ_{WC} is appropriate because most other studies involve only a single site. We may begin by computing d_{WT} and v_{WT} as before. Because we want an estimate of δ_{WC}, not δ_{WT}, we use the fact given in equation (10) that

$$\delta_{WC} = \frac{\delta_{WT}}{\sqrt{1-\rho_S-\rho_C}}$$

and therefore

$$\frac{d_{WT}}{\sqrt{1-\rho_S-\rho_C}} \tag{35}$$

is an estimate of δ_{WC} with a variance of

$$\frac{v_{WT}}{1-\rho_S-\rho_C}. \tag{36}$$

Similarly, we might decide that δ_{WS} was a more appropriate effects size because most other studies involve schools that have several classrooms and nesting of students within classrooms is not taken into account in most of them. We may begin by

computing d_{WT} and v_{WT} as before. Because we want an estimate of δ_{WS}, not δ_{WT}, we use the fact given in equation (9) that

$$\delta_{WS} = \frac{\delta_{WT}}{\sqrt{1-\rho_S}}$$

and therefore

$$\frac{d_{WT}}{\sqrt{1-\rho_S}} \tag{37}$$

is an estimate of δ_{WS} with a variance of

$$\frac{v_{WT}}{1-\rho_S}. \tag{38}$$

If we decided that δ_{BC} or δ_{BS} was a more appropriate effect size, then we could still begin by computing d_{WT} and v_{WT}, but then use (12), namely

$$\delta_{BC} = \delta_{WT} \Big/ \sqrt{\rho}_C$$

to obtain

$$d_{WT} \Big/ \sqrt{\rho}_C \tag{39}$$

as an estimate of δ_{BC} with variance

$$v_{WT}/\rho_C \tag{40}$$

or use (11), namely

$$\delta_{BS} = \delta_{WT} \Big/ \sqrt{\rho}_S$$

to obtain

$$d_{WT} \Big/ \sqrt{\rho}_S \tag{41}$$

as an estimate of δ_{BS} with variance

$$v_{WT}/\rho_S. \tag{42}$$

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 $p^C = 1$ classrooms 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 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 cluster 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 $\bar{Y}^T_{\bullet\bullet\bullet} - Y^C_{\bullet\bullet\bullet} = 1.9$, the pooled within-groups standard deviation $S_{WT} = 12.37$. 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 with a standard error of 0.019. 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 ignored clustering and compared the mean of all of the students in the treatment with the mean of all of the students in the control group via a *t*-test, the *t*-value obtained was $t = 3.488$. This leads to a value of the standardized mean difference of

$$\frac{\bar{Y}^{T}_{\bullet\bullet\bullet} - \bar{Y}^{C}_{\bullet\bullet\bullet}}{S_{WT}} = 0.1536 ,$$

which is not an estimate of any of the three effect sizes considered here. If an estimate of the effect size δ_{WT} is desired, and we had imputed a school level intraclass correlation of $\rho_S = 0.264$ and a classroom level intraclass correlation of $\rho_S = 0.176$, then we use equation (31) to obtain

$d_{WT} = (0.1536)(0.9811) = 0.1507.$

The effect size estimate is very close to the original standardized mean difference because the amount of clustering in this case is rather small. However even this small amount of clustering has a substantial effect on the variance of the effect size estimate. The variance of the standardized mean difference ignoring clustering is

$$\frac{324+162}{324*162} + \frac{0.1536^2}{2(324+162-2)} = 0.009259 .$$

However, computing the variance of d_{WT} using equation (32) with $\rho_S = 0.264$ and $\rho_S = 0.176$, we obtain a variance estimate of 0.093295, which is more than 10 times the variance ignoring clustering. It is worth noting that if we had used the conservative approximation for v_{WT} given in (33), we would have obtained a variance estimate of 0.093895, which differs from the computation given (32) by less than 1 percent. A 95 percent confidence interval for δ_T is given by

$$-0.4480 = 0.1507 - 1.96\sqrt{0.093295} \le \delta_T \le 0.1507 + 1.96\sqrt{0.093295} = 0.7494 .$$

If clustering had been ignored in computing the variance of d_{WT}, the confidence interval for the population effect size would have been -0.0382 to 0.3395.

If we wanted to estimate δ_{WC}, then an estimate of δ_{WC} given by expression (35) is

$$\frac{0.1507}{\sqrt{1-0.264-0.176}} = 0.2014 ,$$

with variance given by expression (36) as

$0.093295/(1 - 0.264 - 0.176) = 0.166598.$

If an estimate of δ_{WS} was wanted, then an estimate of δ_{WS} could be computed from expression (37) as

$$\frac{0.1507}{\sqrt{1-0.264}} = 0.1757 ,$$

with variance given by expression (38) as

$0.093295/(1 - 0.264) = 0.126759.$

If we decided that δ_{BC} was a more appropriate effect size, an estimate could be computed from (39) as

$$0.1507 \big/ \sqrt{0.176} = 0.3592 ,$$

with variance given by expression (40) as

$0.093295/0.176 = 0.53008.$

If we decided that δ_{BS} was a more appropriate effect size, an estimate could be computed from (41) as

$0.1507/\sqrt{0.264} = 0.2933$,

with variance computed from (42) as

0.093295/0.264 = 0.353389.

Is it Really Necessary to Take *Both* Levels of Clustering Into Account?

One might be tempted to think that accounting for clustering at one the highest level (e.g. schools) would be sufficient to account for most of the effects of clustering at both levels. Indeed there is a widely held belief among some researchers that only the highest level of clustering matters in analysis, and that lower levels of clustering can safely be omitted if they are not explicitly part of the analysis. Another belief that is sometimes cited is that any level of the design can be ignored if it is not explicitly included in the analysis. If this were the case the results of this paper might be theoretically interesting, but have few practical implications. It is clear from (15) and (17) that the effect size estimates d_{WS} and d_{WT} do not depend strongly on the intraclass correlation (at least within plausible ranges of intraclass correlations that are likely to be encountered in educational research). However, the analytic results presented here show that, in situations where there is a plausible amount of clustering at two levels, omitting either level of clustering from variance computations can lead to substantial biases in the variance of effect size estimators.

Table 2 shows some calculations of the variance of the effect size estimate d_{WT} and its variance in a balanced design with $m = 10$, $n = 20$, $\rho_S = 0.15$, $\rho_C = 0.10$, and $d_{WT} = (\bar{T}^T_{\bullet\bullet\bullet} - \bar{T}^C_{\bullet\bullet\bullet})/S_{WT} = 0.15$ (fairly typical values). Omitting the classroom (lowest level) of clustering leads to underestimates of the variance when the number of classrooms (lower level clusters) per school is small, underestimation that become less serious as the number of classrooms per school increases. For example, omitting the classroom level of clustering from variance computations leads to underestimation of the variance of d_{WT} by 22% when $p = 2$, but by only 11% when $p = 5$. Because the number of classrooms per school is typically small in educational experiments, we judge that variance is likely to be seriously underestimated when the classroom level is ignored in computing variances of effect size estimates in three level designs.

The table shows that omitting the school (highest level) of clustering leads to larger underestimates of variance when the number of classrooms (lower level clusters) per school is small, which become even more serious when the number of classrooms becomes larger. Omitting the school level of clustering from variance computations leads to underestimation of the variance of d_{WT} by about 67% when $p = 2$, but by 84% when $p = 5$.

One potential substitute for computing variances using the formulas for two levels of clustering is to compute the variances assuming two levels of clustering, but use a composite intraclass correlation that "includes" both between-school and between-class components. Table 2 shows that using only the school level of clustering to compute the variance of d_{WT} but using $\rho = \rho_S + \rho_C$ in place of the school level intraclass correlation performs poorly, leading to overestimates of the variance of from 23% (for $p = 2$) to 45% (for $p = 5$). However, using only the school level of clustering to compute the variance of d_{WT} but using $\rho = \rho_S + \rho_C / p$ in place of the school level intraclass correlation performs much better, leading to overestimates of the variance of less than 1% for all p.

Conclusions

This paper has provided definitions of five different effect sizes that can be estimated in three level studies using cluster randomization. Methods of estimation are provided for each effect size, and the sampling variances are also given. The sampling distribution of each estimator is shown to be a constant times a noncentral *t*-distribution and simple normal approximations are given in each case. Because these approximations have been extensively studied in the context of simpler effect size estimates and power analysis, there is reason to believe that they are reasonably accurate unless sample sizes are quite small (which is unlikely in cluster randomized designs). Simulation studies (reported in other studies) evaluating the accuracy of these approximations confirm expectations.

The analytic work shows that while clustering has only a small effect on the estimates of effect size involving between-student standard deviations, it can have a substantial effect on the variance of these effect size estimates. The example provided illustrates that small, but plausible, amounts of clustering can have a very large effect on the variance of effect sizes and therefore on confidence intervals in practical situations. This implies that ignoring clustering can have serious effects in meta-analysis, leading to serious underestimates of uncertainty in effect size estimates. Moreover, when the sampling design involves two levels of clustering (such as schools and classrooms), it appears that simply ignoring one of the levels of clustering in computing the variances of effect size estimates can also lead to marked underestimates of the variance of effect size estimates.

The results given in this paper can be used to estimate the effect sizes (and their variances) in cluster randomized trials that have been improperly analyzed by ignoring clustering, provided an intraclass correlation is known or can be imputed. The effect size estimates can then be used in meta-analyses along with any other effect size estimates of the same conceptual parameter, using the variances of the estimates to compute weights in the usual way, and using those weights in fixed or mixed effects analyses.

The results given in this paper require that a value of the intraclass correlation parameters ρ_S and ρ_C be known or imputed for sensitivity analysis. In some cases external data about ρ_S and ρ_C may be available (e.g., from previous studies or compendia such as that of Hedges and Hedberg, 2007). It is important to use external values of ρ_S and ρ_C with considerable caution, because the values of ρ_S and ρ_C have 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_S and ρ_C 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_S and ρ_C, even from large sample sizes if those estimates were not based on a similar sampling strategy, with similar populations, and similar outcome measures. However, making no correction for the effects of clustering at all corresponds to assuming that $\rho_C = \rho_S = 0$. The assumption that $\rho_S = \rho_C = 0$ is often very far from the case and thus it may introduce more serious biases in the computation of variances than using values of ρ that are slightly in error.

A major practical problem is that while there is rather extensive reference data on school level intraclass correlations (ρ_S), the data on classroom level intraclass correlations (ρ_C) is less extensive. Such data, either computed from surveys or from reports of

experiments themselves, is badly needed to improve the meta-analysis of studies with multiple levels of clustering. Moreover, given the connection between effect sizes and statistical power, we speculate that our findings about the variance of effect size estimates imply that computations of statistical power in three level designs by omitting levels of clustering (that is, compute power in three level designs as if there were only two levels) will badly underestimate power in plausible situations. If this is so, information on classroom level intraclass correlations may be as crucial for planning of educational experiments that assign schools to treatments as well as for computing effect sizes.

Appendix: Derivation of Sampling Distributions of Effect Size Estimates

The sampling distribution of the effect size estimates proposed in this paper all follow from the same theorem, which was proven in Hedges (in press) and is restated below.

Theorem: Suppose that $Y \sim \mathrm{N}(\mu, a\sigma^2/\tilde{N})$ and that S^2 is a quadratic form in normal variables that is independent of Y, so that $\mathrm{E}\{S^2\} = b\sigma^2$, and $\mathrm{V}\{S^2\} = 2c\sigma^4$, where a, b, c, and $\tilde{N}$ are constants. Then

$$T = \sqrt{\frac{\tilde{N}b}{a}}\left(\frac{Y}{S}\right)$$

has approximately the noncentral t-distribution with b^2/c degrees of freedom and noncentrality parameter

$$\theta = \sqrt{\frac{\tilde{N}b}{a}}\left(\frac{\mu}{\sigma}\right) = \sqrt{\frac{\tilde{N}b}{a}}\delta,$$

where $\delta = \mu/\sigma$. Consequently

$$D = \frac{Y\sqrt{b}}{S} = T\sqrt{\frac{a}{\tilde{N}}} \tag{A1}$$

is a consistent estimate of the effect size δ with approximate variance

$$\frac{a}{\tilde{N}} + \frac{c\delta^2}{2b}. \tag{A2}$$

The theorem can be applied to obtain the sampling distribution of each of the effect size estimators given in this paper, using some elementary facts. In each case we apply the theorem with $Y = \bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}$, $\mu = \mu^T_{\bullet\bullet} - \mu^C_{\bullet\bullet}$, and $\tilde{N} = N^T N^C/(N^T + N^C)$, but with different definitions of S and σ (which imply different choices of the constants a, b, and c). In each case, we use the fact that the expected value of the mean difference is given by

$$\mathrm{E}\left\{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}\right\} = \mu^T_{\bullet\bullet} - \mu^C_{\bullet\bullet}$$

However the variance of variance of $\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}$ and the mean and variance of various choices of S require different derivations in the balanced (equal cluster sample size) and unbalanced (unequal cluster sample size) cases.

Equal Cluster Sample Sizes

In the case of equal cluster sample sizes, a direct argument gives the variance of the mean difference as

$$\begin{aligned}\mathrm{V}\left\{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}\right\} &= \left(\tilde{N}\right)^{-1}\left(\sigma^2_{WC} + n\sigma^2_{BC} + np\sigma^2_{BS}\right) \\ &= \left(\tilde{N}\right)^{-1}\sigma^2_{WT}\left[1 + (n-1)\rho_C + (pn-1)\rho_S\right]\end{aligned}. \tag{A3}$$

We also use the moments of S_{BS}^2, S_{BC}^2, S_{WC}^2, S_{WS}^2, and S_{WT}^2, which are derived from their relation to sums of squares in the analysis of variance (see, e.g., Snedecor, 1956). The design described here has a treatment factor with schools (nested within treatments) and classrooms (nested within schools and treatments) as nested factors. Using the expected mean squares from a design with one crossed factor and two nested

factors, where both nested factors are considered random effects (see, e.g., Kirk, 1995, p.488) we have

$$E\{MS_{BS}\} = \sigma_{WC}^2 + n\sigma_C^2 + pn\sigma_S^2 = \sigma_{WT}^2[\,1 + (n-1)\rho_C + (pn-1)\,\rho_S],$$
$$E\{MS_{BC}\} = \sigma_{WC}^2 + n\sigma_C^2 = \sigma_{WT}^2[1 + (n-1)\rho_C - \rho_S],$$

and

$$E\{MS_{WC}\} = \sigma_{WC}^2 = \sigma_{WT}^2[1 - \rho_C - \rho_S].$$

A standard result from analysis of variance of this design, is that the sums of squares divided by their expectations are independently distributed as chi-squares. In particular,

$$SS_{BS} / \sigma_{WT}^2[\,1 + (n-1)\rho_C + (pn-1)\,\rho_S] \sim \chi_{(M-2)}^2,$$
$$SS_{BC} / \sigma_{WT}^2[\,1 + (n-1)\rho_C - \rho_S] \sim \chi_{M(p-1)}^2,$$

and

$$SS_{WC} / \sigma_{WT}^2[\,1 - \rho_C - \rho_S] \sim \chi_{Mp(n-1)}^2.$$

Therefore, because the expected value of a chi-square equals its degrees of freedom, it follows that the expectations of the sums of squares are

$$E\{SS_{BS}\} = \sigma_{WT}^2(M-2)\,[\,1 + (n-1)\rho_C + (pn-1)\,\rho_S],$$
$$E\{SS_{BC}\} = \sigma_{WT}^2 M(p-1)[\,1 + (n-1)\rho_C - \rho_S],$$

and

$$E\{SS_{WC}\} = \sigma_{WT}^2 Mp(n-1)[\,1 - \rho_C - \rho_S].$$

Because the variance of a chi-square equals twice its degrees of freedom, it follows that the variances of the sums of squares are

$$V\{SS_{BS}\} = 2\sigma_{WT}^4(M-2)\,[\,1 + (n-1)\rho_C + (pn-1)\,\rho_S]^2,$$
$$V\{SS_{BC}\} = 2\sigma_{WT}^4 M(p-1)[\,1 + (n-1)\rho_C - \rho_S]^2,$$

and

$$V\{SS_{WC}\} = 2\sigma_{WT}^4 Mp(n-1)[\,1 - \rho_C - \rho_S]^2.$$

Using these results, we obtain the expected value and variance of the five sample variances involved in the expressions for the effect size estimates described in this paper. Because $S_{WC}^2 = SS_{WC}/(N - Mp)$, it follows that the expected value and variance of S_{WC}^2 are

$$E\left\{S_{WC}^2\right\} = \sigma_{WT}^2\left(1 - \rho_C - \rho_S\right) \tag{A4}$$

and

$$V\left\{S_{WC}^2\right\} = \frac{2\sigma_{WT}^4\left(1 - \rho_C - \rho_S\right)^2}{N - Mp} = \frac{2\sigma_{WC}^4}{N - Mp}. \tag{A5}$$

Because $S_{BC}^2 = SS_{BC}/M\,(p-1)\,n$, it follows that the expected value and variance of S_{BC}^2 are

$$E\left\{S_{BC}^2\right\} = \sigma_{WT}^2\left[1 + (n-1)\rho_C\right]/n \tag{A6}$$

and

$$V\left\{S_{BC}^2\right\} = \frac{2\sigma_{WT}^4\left[1 + (n-1)\rho_C\right]^2}{\left(Mp - M\right)n^2}. \tag{A7}$$

Because $S_{BS}^2 = SS_{BS}/(M-2)np$, it follows that the expected value and variance of S_{BS}^2 are

$$E\left\{S_{BS}^2\right\} = \sigma_{WT}^2\left[1 + (n-1)\rho_C + (pn-1)\rho_S\right]/np \tag{A8}$$

and

$$V\left\{S_{BS}^2\right\} = \frac{2\sigma_{WT}^4\left[1+(n-1)\rho_C+(pn-1)\rho_S\ \right]^2}{(M-2)n^2p^2}. \tag{A9}$$

The expectations and variances of these sums of squares permit computation of the expectations and variances of the composite sums of squares SS_{WS} and SS_{WT} and the variances S_{WS}^2 and S_{WT}^2. Because $S_{WS}^2 = (SS_{WC} + SS_{BC})/(N - M)$, it follows that the expected value and variance of S_{WS}^2 are

$$E\left\{S_{WS}^2\right\} = \sigma_{WT}^2\left[1-\rho_S-\frac{(n-1)\rho_C}{pn-1}\right] = \sigma_{WS}^2\left[1-\frac{(n-1)\rho_C}{(1-\rho_S)(pn-1)}\right] \tag{A10}$$

and

$$\begin{aligned} V\left\{S_{WS}^2\right\} &= \frac{2\sigma_{WT}^4\left[(pn-1)\overline{\rho}^2+2n(p-1)\overline{\rho}\rho_C+n^2(p-1)\rho_C^2\right]}{M(pn-1)^2} \\ &= \frac{2\sigma_{WS}^4\left[(pn-1)\overline{\rho}^2+2n(p-1)\overline{\rho}\rho_C+n^2(p-1)\rho_C^2\right]}{M(pn-1)^2(1-\rho_S)^2}. \end{aligned} \tag{A11}$$

Because $S_{WT}^2 = (SS_{WC} + SS_{BC} + SS_{BS})/(N - 2)$, it follows that the expected value and variance of S_{WT}^2 are

$$E\left\{S_{WT}^2\right\} = \sigma_{WT}^2\left(1-\frac{2(pn-1)\rho_S-2(n-1)\rho_C}{N-2}\right) \tag{A12}$$

and

$$V\left\{S_{WT}^2\right\} = \frac{2\sigma_{WT}^4\left(pn\widehat{N}\rho_S^2+n\breve{N}\rho_C^2+(N-2)\overline{\rho}^2+2n\widehat{N}\rho_S\rho_C+2\widehat{N}\rho_S\overline{\rho}+2\breve{N}\rho_C\overline{\rho}\right)}{(N-2)^2}, \tag{A13}$$

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

The distribution of d_{WC}. In this case we apply the theorem with $\sigma^2 = \sigma_{WC}^2$ and $S^2 = S_{WC}^2$. Here

$$a = \frac{\sigma_{WC}^2+n\sigma_{BC}^2+np\sigma_{BS}^2}{\sigma_{WC}^2} = \frac{1+(n-1)\rho_C+(pn-1)\rho_S}{1-\rho_C-\rho_S}.$$

Because the expected value of S_{WC}^2 is σ_{WC}^2it follows that $b = 1$. Because the variance of S_{WC}^2 is $2\sigma_{WC}^4/(N - Mp)$, it follows that $c = 1/(N - Mp)$. Substituting the expressions for a, b, and c into (A1) and (A2), and simplifying, gives the results in expressions (13) and (14). Since S^2 involves only a single chi-square, it follows that the t-statistic corresponding to d_{WC} has exactly the noncentral t-distribution with $(N - Mp)$ degrees of freedom.

The distribution of d_{WS}. In this case we apply the theorem with $\sigma^2 = \sigma_{WS}^2 = \sigma_{WC}^2 + \sigma_{BS}^2$ and $S^2 = S_{WS}^2$. Here

$$a = \frac{\sigma_{WC}^2+n\sigma_{BC}^2+np\sigma_{BS}^2}{\sigma_{WS}^2} = \frac{1+(n-1)\rho_C+(pn-1)\rho_S}{1-\rho_S},$$

$$b = \frac{E\left\{S_{WS}^2\right\}}{\sigma_{WS}^2} = 1 - \frac{(n-1)\rho_C}{(1-\rho_S)(pn-1)},$$

and

$$c = \frac{V\left\{S_{WS}^2\right\}}{2\sigma_{WS}^4} = \frac{(pn-1)\bar{\rho}^2 + n(p-1)\bar{\rho}\rho_C + n^2(p-1)\rho_C^2}{M(pn-1)(1-\rho_S)^2}.$$

Substituting the expressions for *a*, *b*, and *c* into (A1) and (A2), and simplifying, gives the results in expressions (15) and (16).

The distribution of d_{WT}. In this case we apply the theorem with $\sigma^2 = \sigma_{WT}^2$ and $S^2 = S_{WT}^2$. Here

$$a = \frac{\sigma_{WC}^2 + n\sigma_{BC}^2 + pn\sigma_{BS}^2}{\sigma_{WT}^2} = 1 + (n-1)\rho_C + (pn-1)\rho_S.$$

Using the expected value of S_{WT}^2 given in (A12), we compute

$$b = \frac{E\left\{S_{WT}^2\right\}}{\sigma_{WT}^2} = 1 - \frac{2(pn-1)\rho_S + 2(n-1)\rho_C}{N-2},$$

and using the variance of S_{WT}^2 given in (A13), we compute

$$c = \frac{V\left\{S_{WT}^2\right\}}{2\sigma_{WT}^4} = \frac{pn\tilde{N}\rho_S^2 + n\breve{N}\rho_C^2 + (N-2)\bar{\rho}^2 + 2n\hat{N}\rho_S\rho_C + 2\hat{N}\rho_S\bar{\rho} + 2\breve{N}\rho_C\bar{\rho}}{(N-2)^2}.$$

Substituting the expressions for *a*, *b*, and *c* into (A1) and (A2), and simplifying, gives the results in expressions (17) and (18).

The distribution of d_{BC}. In this case we apply the theorem with $\sigma^2 = \sigma_{BC}^2$ and $S^2 = S_{BC}^2$. Here

$$a = \frac{\sigma_{WC}^2 + n\sigma_{BC}^2 + pn\sigma_{BS}^2}{\sigma_{BC}^2} = \frac{1 + (n-1)\rho_C + (pn-1)\rho_S}{1-\rho_C},$$

$$b = \frac{E\left\{S_{BC}^2\right\}}{\sigma_{BC}^2} = \frac{1 - \rho_S - (n-1)\rho_C}{n\rho_C},$$

and

$$c = \frac{V\left\{S_{BC}^2\right\}}{\sigma_{BC}^2} = \frac{\left[1 - \rho_S + (n-1)\rho_C\right]^2}{M(p-1)n^2\rho_C^2}.$$

Substituting the expressions for *a*, *b*, and *c* into (A1) and (A2), and simplifying, gives the results in expressions (19) and (20). Since S^2 involves only a single chi-square, it follows that the *t*-statistic corresponding to d_{BC} has exactly the noncentral *t*-distribution with $M(p - 1)$ degrees of freedom.

The distribution of d_{BS}. In this case we apply the theorem with $\sigma^2 = \sigma_{BS}^2$ and $S^2 = S_{BS}^2$. Here

$$a = \frac{\sigma_{WC}^2 + n\sigma_{BC}^2 + pn\sigma_{BS}^2}{\sigma_{BS}^2} = \frac{1+(n-1)\rho_C + (pn-1)\rho_S}{1-\rho_S},$$

$$b = \frac{E\left\{S_{BS}^2\right\}}{\sigma_{BS}^2} = \frac{1-(n-1)\rho_C - (pn-1)\rho_S}{pn\rho_S},$$

and

$$c = \frac{V\left\{S_{BS}^2\right\}}{2\sigma_{BS}^4} = \frac{\left[1-(pn-1)\rho_S + (n-1)\rho_C\right]^2}{(M-2)p^2n^2\rho_S^2}.$$

Substituting the expressions for *a*, *b*, and *c* into (A1)and (A2), and simplifying, gives the results in expressions (21) and (22). Since S^2 involves only a single chi-square, it follows that the *t*-statistic corresponding to d_{BS} has exactly the noncentral *t*-distribution with M – 2 degrees of freedom.

Unequal Cluster Sample Sizes

When cluster sample sizes are unequal, expressions for the effect size estimators and their variances are more complex. We first derive the variance of the mean differences. A direct argument leads to

$$V\left\{\bar{Y}_{\bullet\bullet\bullet}^T - \bar{Y}_{\bullet\bullet\bullet}^C\right\} = \left(\tilde{N}\right)^{-1}\left[1+(p_U-1)\rho_S + (n_U-1)\rho_C\right] \qquad \text{(A14)}$$

where p_U and n_U are defined by (24) and (25) respectively. The expected value and variance of $S_{WS}^2$and S_{WT}^2 can be calculated from the analysis of variance across classrooms within schools and across schools within the treatment groups. When cluster sample sizes are unequal, the between school, between classroom, and within classroom sums of squares are still independent, and the within classroom sum of squares has a chi-square distribution, but if school and classroom sample sizes are unequal, the between school and between classroom sums of squares do not, in general, have chi-square distributions. However because both of these sum of squares are quadratic forms, the methods used in this paper apply and the distribution of effect size estimates can be obtained.

The distribution of d_{WC}. In this case, we apply the theorem with $\sigma^2 = \sigma_{WC}^2$ and $S^2 = S_{WC}^2$. Here

$$a = \frac{\sigma_{WC}^2 + p_U\sigma_{BS}^2 + n_U\sigma_{BC}^2}{\sigma_{WC}^2} = \frac{1+(pn-1)\rho_S + (n-1)\rho_C}{1-\rho_S-\rho_C}.$$

Because S_{WC}^2 is the same whether school or classroom sample sizes are equal or not, it follows that b = 1and $c = 1/(N - N_S)$. Substituting the expressions for *a*, *b*, and *c* into (A1) and (A2), and simplifying, gives the results in expressions (13) and (23). Since S^2 involves only a single chi-square, it follows that the *t*-statistic corresponding to d_{WC} has exactly the noncentral *t*-distribution. with $(N - N_S)$ degrees of freedom.

The distribution of d_{WS}. In this case we apply the theorem with $\sigma^2 = \sigma_{WS}^2 = \sigma_{WC}^2 + \sigma_{BS}^2$ and $S^2 = S_{WS}^2$. Here

$$a = \frac{\sigma_{WC}^2 + n_U\sigma_{BC}^2 + p_U\sigma_{BS}^2}{\sigma_{WS}^2} = \frac{1+(n_U-1)\rho_C + (p_U-1)\rho_S}{1-\rho_S}.$$

We obtain the expected value and variance of S_{WS}^2 by using the fact that

$$S_{WS}^2 = \frac{SSB^T + SSW^T + SSB^C + SSW^C}{N-2},$$

where $SSBC^T$ and $SSWC^T$ and $SSBC^C$ and $SSWC^C$ are the sums of squares between and within classes in the treatment and control groups, respectively. Using the expected values of the *SSBC*'s and *SSWC*'s given, for example, on page 429 of Searle, Casella, and McCulloch (1992), we obtain (in our notation)

$$E\left\{S_{WS}^2\right\} = \sigma_{WC}^2 + \frac{(N-M\tilde{n})\sigma_{BC}^2}{(N-M)},$$

so that

$$b = \frac{E\left\{S_{WS}^2\right\}}{\sigma_{WS}^2} = 1 - \frac{(\tilde{n}-1)\rho_C}{(N_S-1)(1-\rho_S)}.$$

Because the between clusters variance component estimates in the treatment and control groups are

$$\left(\hat{\sigma}_{BC}^T\right)^2 = \frac{MSBC^T - MSWC^T}{\tilde{n}^T}$$

and

$$\left(\hat{\sigma}_{BC}^C\right)^2 = \frac{MSBC^C - MSWC^C}{\tilde{n}^C},$$

it follows that S_{WS}^2 can be written as a function of between and within cluster variance components

$$S_{WS}^2 = \frac{(N^T - M^T)\left(\hat{\sigma}_{WC}^T\right)^2 + (N^C - M^C)\left(\hat{\sigma}_{WC}^C\right)^2 + B^T\left(\hat{\sigma}_{BC}^T\right)^2 + B^C\left(\hat{\sigma}_{BC}^C\right)^2}{N-M},$$

where $B^T = N^T - \tilde{n}^T/M^T$ and $B^C = N^C - \tilde{n}^C/M^C$. Therefore the variance of S_{WS}^2 is given by

$$\begin{aligned}(N-M)^2\,\mathrm{V}\left\{S_{WS}^2\right\} &= (N^T - M^T)^2\mathrm{V}\left\{\left(\hat{\sigma}_{WC}^T\right)^2\right\} + (N^C - M^C)^2\mathrm{V}\left\{\left(\hat{\sigma}_{WC}^C\right)^2\right\} \\ &+ \left(B^T\right)^2\mathrm{V}\left\{\left(\hat{\sigma}_{BC}^T\right)^2\right\} + \left(B^C\right)^2\mathrm{V}\left\{\left(\hat{\sigma}_{BC}^C\right)^2\right\} \\ &+ 2(N^T - M^T)B^T\mathrm{Cov}\left(\left(\hat{\sigma}_{WC}^T\right)^2, \left(\hat{\sigma}_{BC}^T\right)^2\right) + 2(N^C - M^C)B^C\mathrm{Cov}\left(\left(\hat{\sigma}_{WC}^C\right)^2, \left(\hat{\sigma}_{BC}^C\right)^2\right)\end{aligned}.$$

Using the expressions for the variances and covariances of the variance component estimates on page 430 of Searle, Casella, and McCulloch (1992), simplifying, and substituting into the formula for *c* yields

$$c = \frac{V\left\{S_{WS}^2\right\}}{2\sigma_{WS}^4} = \frac{(pn-1)\bar{\rho}^2 + n(p-1)\bar{\rho}\rho_C + n^2(p-1)\rho_C^2}{M(pn-1)(1-\rho_S)^2}.$$

Substituting the expressions for *a*, *b*, and *c* into (A1) and (A2), and simplifying, gives the results in expressions (26) and (28).

The distribution of d_{WT}. In this case, we apply the theorem with $\sigma^2 = \sigma_{WT}^2$ and $S^2 = S_{WT}^2$. We obtain the expected value and variance of S_{WT}^2 by using the fact that

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

where $SSBS^T$, $SSBC^T$, and $SSWC^T$ 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. Using the expected values of the *SSBS*'s, *SSBC*'s, and *SSWC*'s given, for example, on page 429 of Searle, Casella, and McCulloch (1992), we obtain the expected value and variance of S_{WT}^2. Here

$$a = \frac{\sigma_{WC}^2 + p_U\sigma_{BS}^2 + n_U\sigma_{BC}^2}{\sigma_{WT}^2} = 1 + (p_U - 1)\rho_S + (n_U - 1)\rho_C$$

and

$$b = \frac{E\left\{S_{WT}^2\right\}}{\sigma_{WT}^2} = 1 - \frac{2(p_U - 1)\rho_S + 2(n_U - 1)\rho_C}{N-2}.$$

The exact variance of S^2 is quite complex, therefore we substitute p_U for pn and n_U for n into expression (A13) to obtain the approximate variance of S_{WT}^2, use this approximate variance to obtain the constant c in the second term of the variance of d_{WT}. Substituting the expressions for a, b, and c into (A1) and (A2), and simplifying, gives the results in expressions (31) and (32).

References

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.

Bryk, A. & Schneider, B. (2000). *Trust in schools*. New York: Russell Sage Foundation.

Donner, N., Birkett, N., & Buck, C. (1981). Randomization by cluster. *American Journal of Epidemiology*, 114, 906-914.

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

Donner, A. & Klar, N. (2002). Issues in the meta-analysis of meta-analysis of cluster randomized trials. *Statistics in Medicine*, *21*, 1971-2980.

Dunkin, M. J. & Biddle, B. J. (1974). *The study of teaching*. New York: Holt, Rinehardt, and Winston.

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.

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. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107-128.

Hedges, L. V. (in press). Effect sizes in cluster randomized designs. *Journal of Educational and Behavioral Statistics*.

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

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

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

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

Laopaiboon, M. (2003). Meta-analyses involving cluster randomization trials: A review of published literature in health care. *Statistical Methods in Medical Research*, *12*, 515-530.

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., 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.

Nye, B., Hedges, L. V., & Konstantopoulos, S. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

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

Rooney, B. L. & Murray, D. M. (1996). A meta-analysis of smoking prevention programs after adjustment for errors in the unit of analysis. *Health Education Quarterly*, *23*, 48-64.

Searle, S. R. (1971). *Linear models*. New York: John Wiley.

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

Snedecor, G. W. (1956). *Statistical methods applied to experiments in agriculture and biology*. Ames, IA: Iowa State University Press.

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

Table 1
School and classroom level intraclass correlations estimated from STAR and NAEP

	Mathematics			Reading		
Grade	ρ_S	ρ_C	ρ_C/ρ_S	ρ_S	ρ_C	ρ_C/ρ_S
K (STAR)	0.17	0.13	0.76	0.17	0.11	0.69
1 (STAR)	0.18	0.13	0.74	0.18	0.08	0.46
2 (STAR)	0.16	0.12	0.74	0.15	0.10	0.64
3 (STAR)	0.15	0.11	0.74	0.11	0.09	0.77
4 (NAEP 1992)	0.24	0.14	0.59	0.22	0.12	0.53
4 (NAEP 1996)	0.29	0.13	0.45	0.24	0.09	0.40

Note: The intraclass correlations in STAR are computed from data in Nye, Konstantopoulos, And Hedges (2004).

Table 2
The effect of ignoring levels of clustering on the variance of d_{WT}

			Variance Computed				Estimated variance / True variance			
p	d_{WT}	v_{WT}	Ignoring Classes	Ignoring Schools	Ignoring Classes $\rho = \rho_S + \rho_C$	Ignoring Schools $\rho = \rho_S + \rho_C / p$	Ignoring Classes	Ignoring Schools	Ignoring Classes $\rho = \rho_S + \rho_C$	Ignoring Schools $\rho = \rho_S + \rho_C / p$
1	0.1482	0.0576	0.0385	0.0290	0.0576	0.0576	0.670	0.504	1.000	1.000
2	0.1485	0.0438	0.0343	0.0145	0.0538	0.0440	0.783	0.332	1.229	1.006
3	0.1486	0.0392	0.0329	0.0097	0.0525	0.0394	0.838	0.247	1.341	1.006
4	0.1487	0.0369	0.0321	0.0073	0.0519	0.0371	0.871	0.197	1.407	1.005
5	0.1487	0.0355	0.0317	0.0058	0.0515	0.0357	0.893	0.163	1.451	1.004
8	0.1488	0.0335	0.0311	0.0036	0.0510	0.0336	0.929	0.108	1.524	1.003
16	0.1488	0.0317	0.0305	0.0018	0.0505	0.0318	0.963	0.057	1.591	1.002

Note: In this example, $m = 10$, $n = 20$, $\rho_S = 0.15$, $\rho_C = 0.10$, and $(\bar{T}^T_{\bullet\bullet\bullet} - \bar{T}^C_{\bullet\bullet\bullet})/S_{WT} = 0.15$