Estimating effect size when there is clustering in one treatment group

Larry V. Hedges · Martyna Citkowicz

Published online: 26 November 2014
© Psychonomic Society, Inc. 2014

Abstract Some experimental designs involve clustering within only one treatment group. Such designs may involve group tutoring, therapy administered by multiple therapists, or interventions administered by clinics for the treatment group, whereas the control group receives no treatment. In such cases, the data analysis often proceeds as if there were no clustering within the treatment group. A consequence is that the actual significance level of the treatment effects is larger (i.e., actual p values are larger) than nominal. Additionally, biases will be introduced in estimates of the effect sizes and their variances, leading to inflated effects and underestimated variances when clustering in the treatment group is not taken into account. These consequences of clustering can seriously compromise the interpretation of study results. This article shows how information on the intraclass correlation can be used to obtain a correction for biases in the effect sizes and their variances, and also to obtain an adjustment to the significance test for the effects of clustering.

Keywords Effect size · Partial clustering · Significance tests · Meta-analysis

Randomized field experiments sometimes assign entire intact groups (such as schools, classes, wards, or hospitals) to treatments. Such designs are called *nested* or *cluster-randomized* designs, because the intact groups are nested within treatment conditions, and the groups may act as statistical clusters in the sense of cluster sampling (see, e.g., Donner & Klar, 2000). Such designs may be used for reasons of convenience (when it

is practically or politically difficult to assign individuals with the same cluster to different treatments) or to minimize contamination of different treatments within the same cluster. The purpose of such designs is typically to estimate the average effect of the treatment across groups (e.g., across schools). Although analyses could (and often do) attempt to model the variation across clusters, the cluster effects are taken to be random, and the purpose is to generalize treatment effects to a population of clusters from which the observed clusters are a sample.

Sometimes clustering arises within one treatment group (e.g., the active treatment), but not the other (e.g., a waitlist control). For example, when the treatment is administered to subgroups by agents such as teachers, tutors, or therapists and it is expected that the agents may not produce identical effects, then each subgroup may acquire a component of variance due to the agent with whom they worked. In such cases, the subgroups are statistical clusters whose effects need to be taken into account in the statistical analysis and in computing effect size estimates and their sampling variance. When such treatments are compared to a no-treatment control, the situation is one in which clustering occurs in the treatment group, but not in the control group. This is sometimes called a *partially clustered* (or *partially nested*) design. Designs such as that described above can be found in numerous fields, including education, psychology, and medicine.

Several authors have developed multilevel random- and mixed-effects models that take one-group clustering into account in order to test for individual-study significance (e.g., Bauer, Sterba, & Hallfors, 2008; Hoover, 2002; Lee & Thompson, 2005; Lohr, Schochet, & Sanders, 2014; Roberts & Roberts, 2005). However, too often the clustering is ignored, and the data are analyzed as if the observations are independent of one another. Pals et al. (2008) reviewed articles in six public health and behavioral health journals spanning the years 2002 to 2006, in order to examine how

L. V. Hedges · M. Citkowicz
Northwestern University, Evanston, IL, USA

M. Citkowicz (✉)
American Institutes for Research, San Mateo, CA, USA
e-mail: martyna.citkowicz@gmail.com

individually randomized trials with treatments administered in groups were analyzed. They found that 32 out of 34 articles ignored the group level entirely, analyzing the data only at the individual level. Analyses that incorrectly assume that the observations are independent may lead to wrong conclusions, because ignoring the extra between-cluster variation underestimates the error term, inflates the magnitude of the effect size, and can lead to inflated Type I errors in tests of hypotheses about treatment effects (see, e.g., Wampold & Serlin, 2000).

The effects of clustering in all treatment groups on effect size estimation were considered by Hedges (2007a, b, 2011). In 2007, he derived adjusted statistics that allow for the calculation of effect sizes (and their sampling variances) when the summary data come from a two-level cluster-randomized design (e.g., when students are clustered within classrooms). In 2011, Hedges extended his analysis to allow for correction in three-level designs (e.g., when students are clustered within classrooms, which are then clustered within schools). Both studies take clustering into account when it exists in both the treatment and control groups; however, we know of no development of effect size methods when there is clustering in only a single group. This problem was called to our attention by a colleague conducting a meta-analysis.

Meta-analyses involve a synthesis of statistical estimates from various studies being used to summarize a particular topic (Lipsey & Wilson, 2001). Summary statistics in each study are used to estimate effect sizes, which are then combined to compute tests of significance for the combined results and to describe the pattern of results on the given topic (Hedges & Olkin, 1985). However, if the magnitudes of the effect sizes included in the meta-analysis are inflated or their variances are underestimated, the overall combined mean effect and its variance will also be affected, which may lead to incorrect inferences. Although the multilevel random- and mixed-effects models that take one-group clustering into account lead to valid tests of individual-study significance, to our knowledge no account has described how to compute effect sizes or their sampling variances in partially clustered designs. The purpose of this report is to describe methods for computing effect size estimates and their variances when there is clustering in only one group and the analysis has not taken clustering into account.

The article is organized into several parts. First, we define the model. Then we provide equations for the appropriate effect size computations and show the calculations via example. Finally, we provide an adjustment for the significance test. In the [Appendix](#), we provide the derivations for the statistics.

Model

Denote the j th observation in the i th cluster in the treatment by Y_{ij}^T ($i=1, \dots, m; j=1, \dots, n$), so that there are m clusters of

size n in the treatment group. Denote the i th observation in the control group by Y_i^C ($i=1, \dots, N^C$). Notice that there are *no* clusters in the control group. Thus, the sample size is $N^T = nm$ in the treatment group and N^C in the control group, and the total sample size is $N = N^T + N^C = nm + N^C$.

We denote the mean of the i th cluster in the treatment group by $\bar{Y}_{i\cdot}^T$ ($i=1, \dots, m$), and denote the overall (grand) means in the treatment and control groups by $\bar{Y}_{\cdot\cdot}^T$ and $\bar{Y}_{\cdot}^C$, respectively. We also define the total pooled within-group variance S_T^2 via

$$S_T^2 = \frac{\sum_{i=1}^m \sum_{j=1}^n \left(Y_{ij}^T - \bar{Y}_{\cdot\cdot}^T \right)^2 + \sum_{i=1}^{N^C} \left(Y_i^C - \bar{Y}_{\cdot}^C \right)^2}{N-2}.$$

Assume that the observations are normally distributed within the treatment group clusters with cluster means μ_i^T and within-cluster variance σ_W^2 . Assume that the observations are also normally distributed in the control group, with mean μ^C and variance σ_W^2 . That is,

$$Y_{ij}^T \sim N(\mu_i^T, \sigma_W^2), i = 1, \dots, m; j = 1, \dots, n, \\ Y_i^C \sim N(\mu^C, \sigma_W^2), i = 1, \dots, N^C.$$

Assume that the clusters in the treatment group have random effects (e.g., they are considered a sample from a population of clusters), so that the cluster means themselves are normally distributed with mean $\mu_{\bullet}^T$ and variance σ_B^2 . That is,

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

Thus, notice that the standard deviation in the control group is simply σ_W , whereas the standard deviation in the treatment group is a combination of σ_W and σ_B , or σ_T defined by

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

Note that the variance, σ_B^2 of the treatment clusters can be interpreted either as variation due to groups themselves or to variation (heterogeneity) in treatment effects. If it is due to variation in treatment effects, then some investigators might wish to model that variation as a function of explanatory variables (such as characteristics of therapists). Alternatively, they might be primarily concerned with the average treatment effect across clusters. This article deals with the situation like that found in over 90 % of the studies cited by Pals et al. (2008), in which the investigator has chosen not to model

between-treatment-cluster variation, so that the focus is implicitly on the average treatment effect across clusters.

A parameter that summarizes the relationship between the variances is called the *intraclass correlation coefficient* ρ , which is defined by

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} = \frac{\sigma_B^2}{\sigma_T^2}. \quad (1)$$

The intraclass correlation ρ can be used to obtain σ_W from σ_T , as $\sigma_W^2 = (1 - \rho)\sigma_T^2$.

Because intraclass correlations are used in planning cluster-randomized experiments, there have been considerable efforts to develop empirical evidence about intraclass correlations. One form of evidence is based on intraclass correlation estimates from experiments that have already been conducted (see, e.g., Baldwin et al., 2011; Schnurr et al., 2007). Another form of evidence is based on secondary analyses of sample surveys that use cluster-sampling designs (see, e.g., Hedges & Hedberg, 2007). In order to produce a reference guide for planning cluster-randomized trials in education, Hedges and Hedberg (2007) compiled a collection of intraclass correlation coefficients of academic achievement and provided extensive tables of ρ values in the Variance Almanac, available on the website for the University of Chicago's Center for Advancing Research and Communication. A third form of evidence is based on census data collections from administrative units, such as school districts or states (see, e.g., Bloom, Richburg-Hayes, & Black, 2007; Hedges & Hedberg, 2013).

Effect sizes

The most commonly used effect sizes in experimental educational and psychological research are standardized mean differences, defined as the difference between the treatment and control group means divided by (i.e., standardized by) a standard deviation. In designs in which there is statistical clustering, in at least one of the treatment groups, there is more than one standard deviation, leading to more than one possible definition of the standardized mean difference (see, e.g., Hedges, 2007b).

Here we focus on designs in which there is clustering in only one group (the treatment group). Two effect sizes are possible, depending on whether or not one wants to include between-cluster-within-treatment group variance in the standard deviation.

In some cases, the clustering in the treatment group is caused by the administration of the treatment. For example, if the treatment is administered by agents (e.g., therapists) that have different effects from one another, additional variance might be created that would not occur in the control group. In

this case, one might argue that the “natural” standard deviation of the outcome is one that does not include the extra outcome variance induced by the treatment clustering. This would be consistent with the position that the appropriate standard deviation to use in defining the effect size is the control group standard deviation. In this case (if $\sigma_W \neq 0$, and hence $\rho \neq 1$), one effect size parameter has the form

$$\delta_W = \frac{\mu_*^T - \mu^C}{\sigma_W}. \quad (2)$$

In other cases, it may be reasonable to think that the clustering is the “natural” situation. (In this case, it may be easier to think of the group without clustering as the “treatment” group, but the labels “treatment” and “control” are entirely arbitrary.) For example, consider a treatment that is usually delivered in a small group setting by agents leading groups, and an experiment to evaluate whether the treatment was just as effective if administered in a single large group. Because the small-group administration of the treatment is the standard administration, it might be sensible to include the additional variance associated with clustering in the standard deviation used to define the effect size. This leads to a second effect size parameter of the form

$$\delta_T = \frac{\mu_*^T - \mu^C}{\sqrt{\sigma_B^2 + \sigma_W^2}} = \frac{\mu_*^T - \mu^C}{\sigma_T}. \quad (3)$$

Note that if the intraclass correlation ρ is known, it is possible to obtain δ_W from δ_T (and vice versa), because

$$\delta_W = \delta_T / \sqrt{1 - \rho}.$$

Estimating the effect sizes

It is our experience that the information reported in studies will often involve S_T (e.g., when the clustering within the treatment group was ignored in the statistical analysis). The actual data might be a standard deviation or a t or F statistic from an analysis that ignored clustering. Because we believe that the information most likely to be reported will involve S_T , we focus first on estimating δ_T using S_T and converting this estimate to an estimate of δ_W , if that estimate is desired. However, occasionally, only S_W is reported and not S_T (see, e.g., Chaplin & Capizzano, 2006). Consequently, we also

address estimation of δ_W and δ_T when only S_W is reported.

Estimating δ_T when S_T is reported

One might expect that

$$d_{Naive} = \frac{\bar{Y}_{..}^T - \bar{Y}_{..}^C}{S_T} \quad (4)$$

would be an unbiased, or nearly unbiased, estimate of δ_T . However, this d_{Naive} has a small but systematic bias (which does not disappear even in large samples), due to the fact that S_T^2 consistently underestimates σ_T^2 when there is clustering in the treatment group.

The intraclass correlation coefficient ρ may be used to obtain an estimate of δ_T that takes clustering in the treatment group into account. A direct argument shows that a consistent

estimator of δ_T is

$$d_T = \left(\frac{\bar{Y}_{..}^T - \bar{Y}_{..}^C}{S_T} \right) \sqrt{1 - \frac{(N^C + n - 2)\rho}{N - 2}}. \quad (5)$$

The second factor (under the radical sign) arises as a correction for the bias in S_T^2 as an estimate of σ_T^2 . Note that when $\rho = 0$, so that there is no clustering, the correction factor is 1 and $d_T = d_{Naive}$. The expression for d_T makes it clear that $|d_T| \leq |d_{Naive}|$ and the difference between d_T and d_{Naive} is an increasing function of ρ , but also that the value of d_T is often very similar to that of d_{Naive} . For example, if $N^C = 10$, $m = 2$, and $n = 5$, so that $N^C = N^T = 10$, the ratio d_T/d_{Naive} ranges from .98 for $\rho = .05$ to .91 for $\rho = .25$. Note that if there is no clustering, so that $\rho = 0$, d_T given in Eq. 5 reduces to the usual standardized mean difference.

The estimator d_T is approximately normally distributed with variance

$$\begin{aligned} v_T &= \frac{N}{N^T N^C} \left[1 + \left(\frac{n N^C}{N} - 1 \right) \rho \right] + \frac{[(N-2)(1-\rho)^2 + (N^T - n)n\rho^2 + 2(N^T - n)(1-\rho)\rho] \delta_T^2}{2[(N-2)(1-\rho) + (N^T - n)\rho]^2} \\ &= \frac{N}{N^T N^C} \left[1 + \left(\frac{n N^C}{N} - 1 \right) \rho \right] + \frac{\delta_T^2}{2h}, \end{aligned} \quad (6)$$

where h is the effective degrees of freedom of S_T^2 , given by

$$h = \frac{[(N-2)(1-\rho) + (N^T - n)\rho]^2}{(N-2)(1-\rho)^2 + (N^T - n)n\rho^2 + 2(N^T - n)(1-\rho)\rho}. \quad (7)$$

Note that the term $[1 + (nN^C/N - 1)\rho]$ is a design effect or variance inflation factor that describes the ratio of the variance of the mean difference when one group is clustered to that when neither group is clustered. Moreover, observe that in a balanced design in which $N^T = N^C$, the design effect becomes $[1 + (n/2 - 1)\rho]$. Comparing the design effect just given for clustering in one group with that when there is clustering in both groups, $[1 + (n - 1)\rho]$, we see that clustering in one group has the same impact on the variance of the mean difference as clustering in both groups, with cluster sizes half as large (i.e., clusters of size $n/2$ vs. n). When clustering is ignored, the variance is computed as

$$v_{Naive} = \frac{N}{N^T N^C} + \frac{d_{Naive}^2}{2(N^T + N^C - 2)}. \quad (8)$$

Although the value of d_T is often quite similar to the value of d_{Naive} , the variance of d_T is typically considerably larger than the value of the variance that would be computed if clustering was ignored. Note that if $\rho = 0$, so that there is no clustering, Expression 6 reduces to Expression 8 and $v_T = v_{Naive}$. For example, if $N^C = 10$, $m = 2$, and $n = 5$, so that $N^C = N^T = 10$, the ratio v_T/v_{Naive} ranges from 1.07 for $\rho = .05$ to 1.36 for $\rho = .25$. In the latter case, v_{Naive} underestimates the variance of d_T by more than a third. Note that when there is no clustering so that $\rho = 0$, then $h = N - 2$, and the variance given in Eq. 6 reduces to the variance ignoring clustering given in Eq. 8.

An approximately unbiased estimate of δ_T (analogous to that given by Hedges, 1981) can be obtained by multiplying

d_T by

$$J(h) = 1 - \frac{3}{4h-1}, \quad (9)$$

where h is the effective degrees of freedom of S_T^2 given in Expression 7. Thus, the approximately unbiased estimator of δ_T is

$$g_T = J(h)d_T, \quad (10)$$

which is an analogue of Hedges's g in designs without clustering, and which has variance $[J(h)]^2 v_T$.

Estimating δ_W when S_T is reported

If $\rho \neq 1$ (so that $\sigma_W \neq 0$) and an estimate of δ_W is desired, then the estimate d_T can be converted into an estimate of δ_W by using the intraclass correlation; namely,

$$\frac{d_T}{\sqrt{1-\rho}} \quad (11)$$

estimates δ_W and has variance $v_W = v_T/(1-\rho)$. Similarly, an approximately unbiased estimator of δ_W is

$$\frac{g_T}{\sqrt{1-\rho}} = \frac{J(h)d_T}{\sqrt{1-\rho}}, \quad (12)$$

and has variance $[J(h)]^2 v_W = [J(h)]^2 v_T/(1-\rho)$.

Estimating δ_W when S_W is reported

If $\rho \neq 1$ (so that $\sigma_W \neq 0$), the effect size estimator

$$d_W = \frac{\bar{Y}_{..}^T - \bar{Y}_{..}^C}{S_W}. \quad (13)$$

is a consistent estimator of δ_W , where S_W is the standard deviation of the control group. It is approximately normally distributed with variance

$$v_W = \frac{N}{N^T N^C} \left(\frac{1 + \left(\frac{nN^C}{N} - 1 \right) \rho}{1-\rho} \right) + \frac{\delta_W^2}{2(N^C-1)}. \quad (14)$$

Note that when there is no clustering, so that $\rho = 0$, the variance given in Eq. 14 reduces to the variance of an effect size standardized by the control group standard deviation when clustering is ignored. An approximately unbiased estimator of δ_W is

$$g_W = J(N^C - 1)d_W, \quad (15)$$

which is an analogue of Hedges's g , and which has variance $[J(N^C - 1)]^2 v_W$.

Estimating δ_T when S_W is reported

If $\rho \neq 1$ (so that $\sigma_W \neq 0$) and an estimate of δ_T is desired, then the estimate d_W can be converted into an estimate of δ_T by using the intraclass correlation; namely,

$$d_W \sqrt{1-\rho} \quad (16)$$

estimates δ_T and has variance $v_W(1-\rho)$. Similarly, an approximately unbiased estimator of δ_T is

$$g_W \sqrt{1-\rho} = J(N^C - 1)d_W \sqrt{1-\rho} \quad (17)$$

and has variance $[J(N^C - 1)]^2 v_T(1-\rho)$.

Example

Kubany et al. (2004) conducted a study on cognitive trauma therapy for battered women (CTT-BW) with posttraumatic stress disorder (PTSD). They administered cognitive therapy to women assigned to one of two groups: the immediate therapy group, who received cognitive trauma therapy right away, and the delayed therapy group, who received therapy only six weeks after initial assessment. The women were assessed on various measures at four different time-points in order to assess treatment effects. We focused on the results of the Clinician-Administered PTSD Scale (CAPS) at the second time point, which compares the women two weeks after the immediate therapy group has completed CTT-BW and directly before the delayed therapy group has received CTT-BW. (The results that we use to conduct our calculations for this example can be found in columns 3 and 8 of Table 3 of that article: "Posttherapy for the immediate therapy group and Pretherapy 2 for the delayed therapy group.") That time point is particularly useful for illustrating the method because the immediate therapy group contains $m = 7$ therapist clusters, whereas the delayed therapy group does not contain any clustering. The authors do not take this one-group clustering into account,

thus, their results may be inflated due to the omission of between-cluster treatment variance. Note that our goal is not to criticize the authors for not accounting for clustering in the treatment group, since the method presented here was not previously available, but rather to use the article for illustrating the new method and describe the difference in results.

The control (delayed therapy) group contained a total of $N^C=40$ women, whereas the treatment (immediate therapy) group consisted of $N^T=45$ women. With $m=7$ therapists, the average cluster size in the treatment groups was $N^T/m=45/7=6.4$. The exact sizes of the clusters were not reported, so here we treat the clusters as equal. Moreover, we chose a slightly conservative sample size by rounding down to $n=6$. That resulted in a total treatment group sample size of $N^T=nm=6 \times 7=42$ and a total sample size of $N=N^T+N^C=42+40=82$. The means (and standard deviations) for the treatment and control groups reported in Table 3 of Kubany et al. (2004) are 15.8 (14.4) and 71.9 (23.8), respectively. Finally, in order to adjust the results for clustering, an intraclass correlation coefficient is required. Schnurr et al. (2007), reported a ρ of .05 for therapists in the context of cognitive behavioral therapy for PTSD in women.

Let us first examine how the results for the effect size change when one accounts for clustering. Note that here we are interested in estimating d_T when information to compute S_T is reported (i.e., Section “Estimating δ_T when S^T is reported” above). With the mean difference between the treatment and control groups $\bar{Y}_{..}^T - \bar{Y}_{..}^C = 15.8 - 71.9 = -56.1$, and the pooled within-group standard deviation $S_T=19.555$, the standardized difference between means (SMD) effect size that does not account for clustering is -2.869 . Given an intraclass

correlation of $\rho = .05$, the SMD that accounts for treatment group clustering may be calculated using Expression 5:

$$d_T = -2.869 \sqrt{1 - \frac{(40+6-2)(.05)}{82-2}} = (-2.869)(0.986) \\ = -2.829.$$

With a correction factor that is so close to 1 (.986), the adjustment is minimal; the adjustment results in a 1.4 % increase in effect size (note that the effect size is adjusted up toward 0). The variance of the effect size, on the other hand, is affected a bit more. When not accounting for clustering, the variance is

$$v_{Naive} = \frac{82}{(42)(40)} + \frac{(-2.869)^2}{2(42+40-2)} = 0.100.$$

Accounting for treatment group clustering, the variance is

$$v_T = \frac{82}{(42)(40)} \left[1 + \left(\frac{(6)(40)}{82} - 1 \right) (0.05) \right] + \frac{(-2.829)^2}{2(79.475)} \\ = 0.104,$$

where $h=79.475$ degrees of freedom is calculated using Expression 7; that is,

$$h = \frac{[(82-2)(1-0.05) + (42-6)(0.05)]^2}{(82-2)(1-0.05)^2 + (42-6)(6)(0.05)^2 + 2(42-6)(1-0.05)(0.05)} = 79.475.$$

That results in a 3.6 % increase in variability of effects. The confidence interval for the population effect is also affected. A 95 % confidence interval for d_T is given by

$$-2.829 - 1.96\sqrt{0.104} \leq d_T \leq -2.829 + 1.96\sqrt{0.104} \\ = -3.461 \leq d_T \leq -2.197.$$

With a confidence interval of $[-3.489, -2.248]$ for the population effect ignoring clustering, the difference in lengths is minimal (0.022, or 1.8 %).

If one wanted to estimate an unbiased estimate of δ_T , then one would multiply d_T by $J(h)$, resulting in an analogue of Hedges's g :

$$g_T = \left(1 - \frac{3}{4(79.475)-1} \right) (-2.829) = (0.991)(-2.829) \\ = -2.802.$$

Note that g_T may also be converted into an estimate of g_W by using the intraclass correlation:

$$g_W = \frac{-2.802}{\sqrt{1-0.05}} = 2.875.$$

Similarly, an estimate d_W may be calculated as

$$d_W = \frac{-2.829}{\sqrt{1-0.05}} = -2.903$$

with variance

$$v_W = \frac{0.104}{1-0.05} = 0.109.$$

How large are the effects of clustering in one group? The effects of clustering in our example were quite modest, which might suggest that the results are not worth the extra effort of correcting the estimates for treatment group clustering. However, even though the correction is modest in this example, it need not be in other (plausible) situations. The magnitude of corrections for clustering on both d_T and its variance depend on two things: the size of the intraclass correlation ρ and the size of the clusters n . If either of these is large, then the effects of clustering will be large. In some areas of research, either of them can be large.

For example, in education, the intraclass correlation of academic achievement test scores at the school level is .22 (on average across all school grades; Hedges & Hedberg, 2007). That is much larger than the ρ of .05 for therapists presented in the example. It is also easy to imagine large cluster sizes, for example when the treatment is administered by clinics or social service centers, there may be few clusters but large sample sizes associated with them.

Suppose that the actual intraclass correlation in the example above was $\rho = .22$ (a representative value from education), rather than $\rho = .05$. The effect size d_T decreases by 6.24 % in absolute magnitude (from -2.869 to -2.690), the variance v_T increases by 21.5 % (from .100 to .122), the degrees of freedom h decrease by 13.5 % (from 80 to 69.18), and the confidence interval length increases by 10.2 % (from 1.241 to 1.368).

A similar difference would result if the clusters had been larger. In our example, suppose that, instead of $m = 7$ clusters of size $n = 6$, there were $m = 2$ clusters of size $n = 21$, for the same total sample size of $N^T = 42$ and $N^C = 40$ (and $\rho = .05$ once again). In this case, the adjustment for clustering still produces a negligible impact on the estimate, decreasing d_T by 1.9 % in absolute magnitude (from -2.869 to -2.815), increasing the variance v_T by 21.3 % (from .1002 to .1216), decreasing the effective degrees of freedom by 1.5 % (from 80 to 78.84), and increasing the length of the confidence interval by

10.2 % (from 1.241 to 1.367). One might still argue that the change in effect size estimate is modest, but the changes in the variance and length of the confidence intervals certainly are not. To explore these potential changes further, let us examine the effects of clustering for various potential values of n and ρ .

Figure 1 shows plots of the effects on the estimates d_T , v_T , and h as ρ and n increase. The results display the effect ratio, defined as the adjusted estimate divided by the naïve estimate, plotted on the y -axis and either ρ (Fig. 1a) or n (Fig. 1b) plotted on the x -axis. Note that we only examine a balanced design here, in which $m = 4$ and $N^C = N^T = nm$. In Fig. 1a $n = 10$, and in Fig. 1b $\rho = .20$, an intraclass correlation often found in education research.

First, notice that, in Fig. 1a, when $\rho = 0$, the effect ratio is 1 for all three estimates. That is because, in that case, no excess between-group variability exists (just as in the case when clustering does not exist in the study), so the estimates that account for clustering are identical to those that do not account for clustering. However, the ratio diverges from 1 as ρ increases. The changes in the estimates are small for the effect size and moderate for the degrees of freedom, with d_T decreasing by only 13 % and h decreasing by 51 % when $\rho = .40$.

Fig. 1 Plots of the effects on the estimates d_T , v_T , and h

v_T , on the other hand, obtains a 2.5-fold increase by the time that $\rho = .40$. Figure 1b displays a similar pattern. The changes in the effect size and degrees of freedom are small as compared to the changes in the variance as n increases (although note that the decrease in degrees of freedom is by no means negligible); by the time $n = 100$, there is a more than nine-fold increase in the variance, whereas d_T and h decrease by 6 % and 66 %, respectively. Nonetheless, keep in mind that even small decreases in the effect size may affect meta-analyses' results.

Recall that, in meta-analyses, effects are combined in order to summarize a particular set of research findings. If a portion of the effect sizes are adjusted toward zero in order to account for treatment group clustering, the combined mean effect will also decrease in size. Moreover, with substantially larger variances about the effects, the effect sizes will be less precise and more variable about the overall mean effect. Depending on the magnitudes of both, the meta-analytic results may change considerably.

Adjusting the significance test for clustering in only one group

Research studies that involve clustering in one treatment group often ignore that clustering in their statistical analyses (Pals et al., 2008). This is important because it leads to computed levels of statistical significance that are smaller (and can be *much* smaller) than the actual level of statistical significance (see, e.g., Hedges, 2007a). That is, the p values computed for tests of the null hypothesis of no treatment effects are too small when clustering is ignored. In reviewing such research, it is sometimes useful to determine the approximate actual level of statistical significance for such studies (to more accurately state the findings of individual studies). Such an approach is used by the US Institute of Education Sciences (IES) What Works Clearinghouse in its reviews.

The objective of the significance test is to determine the statistical significance of the treatment effect—that is, to test the null hypothesis

$$H_0 : \mu_*^T = \mu^C.$$

To carry out the usual t test, when clustering is ignored, the Student's t would be computed as

$$t_{Naive} = \left(\frac{\bar{Y}_{..}^T - \bar{Y}_{..}^C}{S_T} \right) \sqrt{\frac{N^T N^C}{N^T + N^C}}, \quad (18)$$

with $N^T + N^C - 2$ degrees of freedom. Taking clustering into account, the t value must be adjusted by multiplying t_{Naive} by

the square root of factor f ,

$$f = \frac{(N-2) - (N^C + n-2)\rho}{(N-2) \left[1 + \left(\frac{nN^C}{N} - 1 \right) \rho \right]}, \quad (19)$$

which leads to the following expression for

$$t_A = \left(\frac{\bar{Y}_{..}^T - \bar{Y}_{..}^C}{S_T} \right) \sqrt{\frac{N^T N^C}{N^T + N^C}} \sqrt{\frac{t_A = t_{Naive} \sqrt{f}}{(N-2) \left[1 + \left(\frac{nN^C}{N} - 1 \right) \rho \right]}}. \quad (20)$$

Note that t_A has a t distribution with h degrees of freedom when the null hypothesis is true, where h is defined in Expression 7. If there is no clustering so that $\rho = 0$, then $f = 1$, $h = N - 2$, and t_A reduces to the usual test statistic ignoring clustering.

One application of the results in this article is to study the rejection rate of the unadjusted t test when there is clustering. It is well known that the unadjusted t test has a rejection rate that is often much higher than nominal. 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 $t(\nu, \alpha)$ denote the level α two-sided critical value for the t distribution with ν degrees of freedom. The usual unadjusted t test rejects if $|t| > t(N-2, \alpha)$. Because $t_A = t_{Naive} \sqrt{f}$ has a t distribution with h degrees of freedom under the null hypothesis, the rejection rate of the unadjusted test is

$$2 \left\{ 1 - F \left[\sqrt{f} t(N-2, \alpha), h \right] \right\}, \quad (21)$$

where $F[x, \nu]$ is the cumulative distribution function of the t distribution with ν degrees of freedom. Table 1 provides the actual rejection rates of the $\alpha = .05$ level two-sided significance test for various values of $N^T = N^C$, n , m , and ρ .

This table illustrates that the actual significance level of the nominal 5 % test is higher than 5 % when $\rho > 0$. Moreover, when either ρ or n is large, the actual significance level can be *much* higher than nominal. For example, when the cluster size $n = 100$, the actual rejection rate of the unadjusted test when the null hypothesis is true is .30 if $\rho = .05$, .43 when $\rho = .10$, and .52 when $\rho = .15$. Therefore, when $\rho = .15$ and $n = 100$, the unadjusted test with a nominal 5 % significance level rejects the (true) null hypothesis more than 50 % of the time. Even with rather small values of ρ (like $\rho = .05$), the behavior of the

Table 1 Actual significance level of a nominal .05 significance test when there is clustering in the treatment group

ρ	m	n	N^C	$N-2$	h	$\sqrt{f}$	Actual p
.05	2	5	10	18	17.9	.947	.06
.05	2	10	20	38	37.7	.896	.08
.05	2	20	40	78	76.0	.815	.11
.05	2	25	50	98	96.4	.782	.12
.05	2	50	100	198	191.5	.661	.19
.05	2	75	150	298	283.6	.584	.25
.05	2	100	200	398	372.8	.528	.30
.05	2	200	400	798	703.0	.402	.43
.05	2	500	1,000	1,998	1,493.9	.268	.60
.10	5	5	25	48	47.0	.905	.08
.10	5	10	50	98	93.9	.820	.11
.10	5	20	100	198	181.7	.704	.17
.10	5	25	125	248	223.0	.661	.19
.10	5	50	250	498	406.4	.526	.30
.10	5	75	375	748	558.8	.450	.38
.10	5	100	500	998	687.5	.399	.43
.10	5	200	1,000	1,998	1,049.2	.294	.56
.10	5	500	2,500	4,998	1,532.0	.191	.71
.15	5	5	25	48	45.6	.863	.09
.15	5	10	50	98	88.6	.755	.14
.15	5	20	100	198	163.0	.622	.22
.15	5	25	125	248	195.4	.578	.26
.15	5	50	250	498	323.2	.445	.38
.15	5	75	375	748	412.7	.375	.46
.15	5	100	500	998	478.8	.330	.52
.15	5	200	1,000	1,998	630.1	.240	.64
.15	5	500	2,500	4,998	777.1	.154	.76

$N^C = N^T = nm$, where N^C and N^T represent the sample sizes in the control and treatment groups, respectively, and m is the number of clusters of size n in the treatment group. N denotes the total sample size $N = N^C + N^T$, $N-2$ represents the degrees of freedom when clustering is ignored, h is the corrected degrees of freedom, and $\sqrt{f}$ is the correction factor by which t_{Naive} is multiplied to adjust for one-group clustering ($t_A = t_{Naive}\sqrt{f}$)

unadjusted test is unacceptable for all but the smallest cluster sizes.

Example

In the Kubany et al. (2004) example of cognitive trauma therapy for battered women with PTSD described above, we can test the null hypothesis that CTT-BW is no more useful than no therapy, or that $H_0: \mu_{Immediate} = \mu_{Delayed}$. Ignoring clustering, t is

$$t_{Naive} = \left(\frac{15.8-71.9}{19.555} \right) \sqrt{\frac{(42)(40)}{42+40}} = -12.985,$$

with $df=80$ and a highly significant p value of $p < .0001$. Adjusting the naive t value for clustering in the treatment group, t_A is

$$t_A = -12.985 \sqrt{\frac{(82-2)-(40+6-2)(0.05)}{(82-2) \left[1 + \left(\frac{(6)(40)}{82} - 1 \right) (0.05) \right]}}$$

$$= (-12.985)(0.942) = -12.230,$$

and $h=79.475$ from above. With the square root of the adjustment factor f being so close to 1 (.942), notice that the absolute magnitude of t decreases by only 5.8 %. The degrees of freedom essentially do not change as well (from 80 to 79.475), leading again to a highly significant p value of $p < .0001$.

Remember that these minimal changes in the estimates are due to the small intraclass correlation of .05 for therapists. Once again, imagine that $\rho = .22$, the average intraclass correlation for schools. In this case, the absolute magnitude of t_A decreases by 21.4 % (from -12.985 to -10.202), and h

Fig. 2 Plots of the effects on the estimates t_A and h

decreases by 13.5 % (from 80 to 69.177). Although this still yields a highly significant p value, one can see that changing just one of the data characteristics can make a substantial impact on the estimates. Let us examine how the t value, degrees of freedom, and resulting p value change depending on how some of the other data characteristics change.

Figure 2 displays plots of the effects on the estimates t_A and h as ρ and n increase. As in Fig. 1, the effect ratio is plotted on the y -axis, and either ρ (Fig. 2a) or n (Fig. 2b) is plotted on the x -axis. We again consider a balanced design with $m=4$, $N^C = N^T = nm$, and $n=10$ in Fig. 2a, and $\rho = .20$ in Fig. 2b.

Figure 2a and b show that the t statistic and degrees of freedom obtain similar patterns of adjustment (i.e., a decrease in the effect ratio slope) as ρ and n increase, albeit t_A is consistently more adjusted than h (except for the crossover that occurs at a ρ of about .35 in Fig. 2a). The results suggest that increasing ρ can easily make a significant t value nonsignificant, because when $\rho > 0$, one requires a much larger sample size in order to obtain significance than if one did not have clustering in the treatment group. This is partially due to the fact that t_A is adjusted more than d_T (see Fig. 1), with the absolute magnitude adjustment ranging from 0 % to 46 % in t_A and from 0 % to 13 % in d_T as ρ increases from 0 to .40. It is also due to the substantial decrease in degrees of freedom (51 % at $\rho = .40$), since fewer degrees of freedom mean less power to detect an effect. For example, in Fig. 2a, when $\rho = .05$, the p value is about .05, but once $\rho = .10$, the p value increases to .07 and is no longer significant. Figure 2b shows that, in addition to having a $\rho > 0$, increasing n can add to the change estimates, with d_T and h decreasing by 72 % and 66 %, respectively, by the time $n=100$.

Unequal cluster sizes

Previous sections of this article have involved the assumption that each cluster has the same number n of individuals. Although this will often be true (at least approximately), it need not always be true. The same principles used to derive the effect size estimate, its variance, and the adjustment to the significance test when cluster sizes are equal can be used to derive the corresponding statistics when the cluster sizes are unequal. Suppose that n_i individuals are in the i th cluster in the treatment group, and that the n_i values need not be equal.

Estimation of δ_T from S_T

A direct argument shows that a consistent estimator of δ_T is given by

$$d_T = \left(\frac{\bar{Y}_{..}^T - \bar{Y}_{..}^C}{S_T} \right) \sqrt{1 - \frac{(N^C + \tilde{n} - 2)\rho}{N - 2}}, \quad (22)$$

where $\tilde{n}$ is an “average” cluster size, given by

$$\tilde{n} = \frac{1}{N^T} \sum_{i=1}^m n_i^2. \quad (23)$$

Note that when all of the n_i s are equal to n , Expression 23 reduces to n , and Expression 22 reduces to Expression 5.

As in the case of equal cluster sizes, the estimator d_T is approximately normally distributed, but now the variance is

$$\begin{aligned} v_T &= \frac{N \left[1 + \left(\frac{\tilde{n} N^C}{N} - 1 \right) \rho \right]}{N^T N^C} + \frac{\left[(N-2)(1-\rho)^2 + A\rho^2 + 2(N^T \tilde{n})(1-\rho)\rho \right] \delta_T^2}{2 \left[(N-2)(1-\rho) + (N^T \tilde{n})\rho \right]^2} \\ &= \frac{N \left[1 + \left(\frac{\tilde{n} N^C}{N} - 1 \right) \rho \right]}{N^T N^C} + \frac{\delta_T^2}{2h}, \end{aligned} \quad (24)$$

where h is the effective degrees of freedom of S_T^2 , given by

$$h = \frac{\left[(N-2)(1-\rho) + (N^T \tilde{n})\rho \right]^2}{(N-2)(1-\rho)^2 + A\rho^2 + 2(N^T \tilde{n})(1-\rho)\rho}, \quad (25)$$

and A is the auxiliary constant, given by

$$A = \frac{(N^T)^2 \sum_{i=1}^m n_i^2 + \left(\sum_{i=1}^m n_i^2 \right)^2 - 2N^T \sum_{i=1}^m n_i^3}{(N^T)^2 = N^T \tilde{n} + \tilde{n}^2 - 2 \sum_{i=1}^m n_i^3 / N^T}. \quad (26)$$

Note that when the n_i s are all equal to n , the constant A reduces to $(N^T - n)n$, Expression 24 reduces to Expression 6, and Expression 25 reduces to Expression 7.

Finally, the correction factor for adjusting the t statistic for the impact of clustering becomes

$$f = \frac{(N-2) - (N^C + \tilde{n} - 2)\rho}{(N-2) \left[1 + \left(\frac{\tilde{n}N^C}{N} - 1 \right) \rho \right]}, \quad (27)$$

and the t statistic adjusted for clustering becomes

$$t_A = t_{\text{Naive}} \sqrt{f}, \quad (28)$$

which has h degrees of freedom, given in Expression 25.

Estimation of δ_W from S_W

When cluster sizes are unequal, the effect size estimator d_W is not affected; that is, it remains as in Expression 13. However, the variance of the estimator when cluster sizes are unequal becomes

$$v_W = \frac{N}{N^T N^C} \left(\frac{1 + \left(\frac{\tilde{n}N^C}{N} - 1 \right) \rho}{1 - \rho} \right) + \frac{\delta_W^2}{2(N^C - 1)}, \quad (29)$$

where $\tilde{n}$ is given in Expression 23 above.

Conclusions

Experiments sometimes involve cluster sampling in one or both treatment groups. When there is clustering in only one group, such experiments are often improperly analyzed by ignoring the potential impact of clustering on significance tests or effect size calculation. 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), we demonstrate how significance levels and confidence intervals adjusted for the impact of clustering can be obtained if the intraclass correlation is known or plausible values can be imputed. One might argue that the intraclass correlations that are imputed may themselves be

incorrect. Although this is certainly true, making no adjustment is equivalent to imputing zero for the intraclass correlation, and that is likely to be farther from the truth than a careful imputation or, better still, an imputation of a range of values for a sensitivity analysis. Such procedures provide more accurate significance levels and are suitable for setting bounds on the results.

Clustering in a single treatment group raises the conceptual issue of what effect size parameter is of interest, because more than one definition of effect size is possible. Such clustering is likely to have a modest impact on the magnitude of the effect size estimates, but often will have a nonnegligible impact on the variance of the effect size estimate. The potentially large impact on the variance can have an equally large impact on the weights given to the effect size in a meta-analysis. The methods given in this article provide a way to obtain more accurate estimates of both the effect size estimate and its variance.

Author note This article is based in part on work supported by the US Institute of Education Sciences (IES) under Grant Nos. R305D11032 and R305B1000027, and by the National Science Foundation (NSF) under Grant No. 0815295. Any opinions, findings, and conclusions or recommendations are those of the authors and do not necessarily represent the views of the IES or the NSF.

Appendix

The derivations use a theorem given in the Appendix of Hedges (2007b). The theorem states that if $Y \sim N(\mu, a\sigma^2)$ and S^2 is a quadratic form in normal variates that is independent of Y , so that the $E\{S^2\} = b\sigma^2$ and $V\{S^2\} = 2c\sigma^4$, where a , b , and c are known constants, then

$$T = \sqrt{\frac{b}{a}} \left(\frac{Y}{S} \right) \quad (30)$$

has approximately the noncentral t distribution, with

$$\nu = b^2 / c \quad (31)$$

degrees of freedom and the noncentrality parameter

$$\theta = \sqrt{\frac{b}{a}} \left(\frac{\mu}{\sigma} \right).$$

Consequently,

$$D = \frac{Y\sqrt{b}}{S} \quad (32)$$

is a consistent estimate of the effect size $\delta = \mu/\sigma$ with approximate variance

$$a + \frac{c\delta^2}{2b^2}. \quad (33)$$

An approximately unbiased estimate of δ is given by $DJ(b^2/c)$, where the function $J(x)$ is given by Expression 9.

To derive the effect size estimate in this article, we apply the theorem with $Y = \bar{Y}_{..}^T - \bar{Y}_{..}^C$, $S = S_T$ and $\delta = (\mu_{..}^T - \mu_{..}^C)/\sigma$. To use this theorem, we must derive the constants a , b , and c .

Equal cluster sizes

Under the model, the sampling distribution of the mean difference $\bar{Y}_{..}^T - \bar{Y}_{..}^C$ is normal with variance

$$\begin{aligned} \sigma_W^2 \left(\frac{1}{N^T} + \frac{1}{N^C} \right) + \frac{\sigma_B^2}{m} &= \frac{N\sigma_T^2}{N^T N^C} \left[1 + \left(\frac{nN^C}{N} - 1 \right) \rho \right] \\ &= \sigma_T^2 \left[\frac{N(1-\rho) + nN^C\rho}{N^T N^C} \right]. \end{aligned}$$

Therefore, the constant a is given by the last term in square brackets above when the effect size δ_T is being estimated using S_T , or that term divided by $(1 - \rho)$ when the effect size δ_W is being estimated using S_W [because $\sigma_T^2 = \sigma_W^2/(1 - \rho)$].

The model also implies that the observations in the treatment group are nested within clusters so that the sum of squares within the treatment group SST can be decomposed into between-cluster and within-cluster sums of squares ($SSBT$ and $SSWT$, respectively), so that

$$SST = SSWT + SSBT,$$

where

$$SSWT = \sum_{i=1}^m \sum_{j=1}^n \left(Y_{ij}^T - \bar{Y}_{i.}^T \right)^2$$

and

$$SSBT = n \sum_{i=1}^m \left(\bar{Y}_{i.}^T - \bar{Y}_{..}^T \right)^2,$$

and where $\bar{Y}_{i.}^T$ is the mean of the observations in the i th cluster and $\bar{Y}_{..}^T$ is the mean of all of the observations in the treatment group. Let SSC be the sum of squares about the mean in the control group, so that

$$SSC = \sum_{i=1}^{N^C} \left(Y_i^C - \bar{Y}_{..}^C \right)^2,$$

where $\bar{Y}_{..}^C$ is the mean of all the observations in the control group. Thus, the total pooled within-group variance S_T^2 can be written as

$$S_T^2 = \frac{SST + SSC}{N^T + N^C - 2} = \frac{SSWT + SSBT + SSC}{N^T + N^C - 2},$$

where each of $SSWT$, $SSBT$, and SSC is statistically independent and each is distributed as a constant times a chi-square.

Using standard results for expected mean squares, the expected values of $SSWT$, $SSBC$, and SSC are

$$\begin{aligned} E\{SSWT\} &= m(n-1)\sigma_W^2, \\ E\{SSBT\} &= (m-1)(\sigma_W^2 + n\sigma_B^2) = (m-1)\sigma_W^2 + (m-1)n\sigma_B^2, \\ E\{SSC\} &= (N^C-1)\sigma_W^2, \end{aligned}$$

so the expected value of S_T^2 is

$$E\{S_T^2\} = \sigma_W^2 + \left(\frac{N^T - n}{N - 2} \right) \sigma_B^2 = \sigma_T^2 \left[\frac{(N^C + n - 2)\rho}{N - 2} \right].$$

Therefore, the constant b is given by the last term in square brackets above when the effect size δ_T is being estimated using S_T or $b = 1$ when the effect size δ_W is being estimated using S_W (since the expected value of S_W^2 is σ_W^2).

The variances of $SSWT$, $SSBC$, and SSC are

$$\begin{aligned} V\{SSWT\} &= \frac{2}{(N-2)^2} \left[\frac{m(n-1)\sigma_W^4}{(N^C-1)} + \frac{2(m-1)(\sigma_W^2 + n\sigma_B^2)^2}{(N-2)^2} \right], \\ V\{SSBT\} &= \frac{2(m-1)(\sigma_W^2 + n\sigma_B^2)^2}{(N-2)^2} = \frac{2(m-1)\sigma_W^4}{(N^C-1)} + \frac{2(m-1)n^2\sigma_B^4}{\sigma_W^4} + \frac{4(m-1)n\sigma_W^2\sigma_B^2}{\sigma_W^4}, \\ V\{SSC\} &= \frac{2}{(N-2)^2} \left[\frac{m(n-1)\sigma_W^4}{(N^C-1)} + \frac{2(m-1)(\sigma_W^2 + n\sigma_B^2)^2}{(N-2)^2} \right], \end{aligned}$$

so that the variance of S_T^2 is

$$\begin{aligned} V S_T^2 &= \frac{2(N-2)\sigma_W^4 + 2n^2(m-1)\sigma_B^4 + 4n(m-1)\sigma_B^2\sigma_W^2}{(N-2)^2} \\ &= 2\sigma_T^4 \left[\frac{(N-2)(1-\rho)^2 + n^2(m-1)\rho^2 + 2n(m-1)\rho(1-\rho)}{(N-2)^2} \right]. \end{aligned}$$

Therefore, the constant c is given by the last term in square brackets above when the effect size δ_T is being estimated using S_T , or $c = 1/(N^C - 1)$ when the effect size δ_W is being estimated using S_W (since the variance of S_W^2 is $2\sigma_W^4/(N^C - 1)$).

Inserting the expressions above for a , b , and c into Eqs. 31, 32, and 33 and simplifying gives Expressions 7, 5, and 6, respectively. Also, a direct argument using Eq. 30 leads to Expression 20.

Unequal cluster sizes

When the number of observations per cluster varies, the derivation follows the same lines, but the algebra is slightly more complex. Under the model, the sampling distribution of the mean difference $\bar{Y}_{..}^T - \bar{Y}_{..}^C$ is normal with variance

$$\frac{N\sigma_T^2}{N^T N^C} \left[1 + \left(\frac{\tilde{n} N^C}{N} - 1 \right) \rho \right] = \sigma_T^2 \left[\frac{N(1-\rho) + \tilde{n} N^C \rho}{N^T N^C} \right].$$

Therefore, the constant a is given by the last term in square brackets above when the effect size δ_T is being estimated using S_T or that term divided $(1 - \rho)$ when the effect size δ_W is being estimated using S_W [because $\sigma_T^2 = \sigma_W^2/(1 - \rho)$].

When the cluster sizes are unequal, $SSWT$ and $SSBT$ are defined as

$$SSWT = \sum_{i=1}^m \sum_{j=1}^{n_i} \left(Y_{ij}^T - \bar{Y}_{i.}^T \right)^2$$

and

$$SSBT = \sum_{i=1}^m n_i \left(\bar{Y}_{i.}^T - \bar{Y}_{..}^T \right)^2,$$

and SSC is the same as is defined in the [Equal cluster sizes](#) section. They are still independent of one another, but $SSBT$ no longer has a chi-square distribution. As in the equal-cluster-sizes case,

$$E\{SSWT\} = m(n-1)\sigma_W^2,$$

but now

$$E\{SSBT\} = (m-1) \left(\sigma_W^2 + \tilde{n} \sigma_B^2 \right),$$

due to the varying cluster size n_i (see, e.g., Eqs. 77 and 78 on page 70 of Searle, Casella, & McCulloch, 1992). Therefore,

$$E\{S_T^2\} = \sigma_W^2 + \left(\frac{N^T - \tilde{n}}{N-2} \right) \sigma_B^2 = \sigma_T^2 \left[1 - \frac{(N^C + \tilde{n} - 2)\rho}{N-2} \right],$$

and the constant b is given by the last term in square brackets above when the effect size δ_T is being estimated using S_T , or $b = 1$ when the effect size δ_W is being estimated using S_W (since the expected value of S_W^2 is σ_W^2).

Finally, because the between-cluster variance component estimate in the treatment group is

$$\hat{\sigma}_B^2 = \frac{MSBT - MSWT}{\tilde{n}(m-1)},$$

and because it follows that S_T^2 can be written as a function of the between- and within-cluster variance components,

$$S_T^2 = \frac{SSWC + 2SSWT + \tilde{n}(m-1)\hat{\sigma}_B^2}{N-2}.$$

Therefore, the variance of S_T^2 is given by

$$\begin{aligned} (N-2)^2 V\{S_T^2\} &= V\{SSWC\} + 4V\{SSWT\} \\ &\quad + \tilde{n}^2 (m-1)^2 V\{\hat{\sigma}_B^2\} \\ &\quad + 8Cov(SSWT, \hat{\sigma}_B^2). \end{aligned}$$

Using Expressions 95 and 102 from pages 74 and 75 of Searle et al. (1992) for the variances of the sums of squares, and the variance component estimates and Searle et al.'s Expression 96 for the covariance term, and then simplifying yields

$$V\{S_T^2\} = \frac{2\sigma_T^4 \left[(N-2)(1-\rho)^2 + A\rho^2 + 2(N^T - \tilde{n})(1-\rho)\rho \right]}{(N-2)^2},$$

where A is the auxiliary constant in Expression 26. Therefore, constant c is given by the term in square brackets above when the effect size δ_T is being estimated using S_T , or $c = 1/(N^C - 1)$ when the effect size δ_W is being estimated using S_W [since the variance of S_W^2 is $2\sigma_W^4/(N^C - 1)$].

Inserting the expressions above for a , b , and c into Eqs. 31, 32, and 33 and simplifying gives Expressions 25, 22, and 24, respectively, and a direct argument using Eq. 30 leads to Expression 28.

References

- Baldwin, S. A., Murray, D. M., Shadish, W. R., Pals, S. L., Holland, J. M., Abramowitz, J. S., & Watson, J. (2011). Intraclass correlation associated with therapists: Estimates and applications in planning psychotherapy research. *Cognitive Behaviour Therapy*, 40, 15–33. doi:10.1080/16506073.2010.520731
- Bauer, D. J., Sterba, S. K., & Hallfors, D. D. (2008). Evaluating group-based interventions when control participants are ungrouped. *Multivariate Behavioral Research*, 43, 210–236. doi:10.1080/00273170802034810
- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, 29, 30–59. doi:10.3102/0162373707299550
- Chaplin, D., & Capizzano, J. (2006). *Impacts of a summer learning program: A random assignment study of Building Educated Leaders for Life (BELL)*. Washington, DC: Urban Institute.
- Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research*. London, UK: Arnold.
- 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. (2007a). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics*, 32, 151–179. doi:10.3102/1076998606298040
- Hedges, L. V. (2007b). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 341–370. doi:10.3102/1076998606298043
- Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, 36, 346–380. doi:10.3102/1076998610376617
- Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, 29, 60–87. doi:10.3102/0162373707299706
- Hedges, L. V., & Hedberg, E. C. (2013). Intraclass correlations and covariate outcome correlations for planning 2 and 3 level cluster randomized experiments in education. *Evaluation Review*, 37, 13–57.
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York, NY: Academic Press.
- Hoover, D. R. (2002). Clinical trials of behavioral interventions with heterogeneous teaching subgroup effects. *Statistics in Medicine*, 21, 1351–1364. doi:10.1002/sim.1139
- Kubany, E. S., Hill, E. E., Owens, J. A., Iannce-Spencer, C., McCaig, M. A., Tremayne, K. J., & Williams, P. L. (2004). Cognitive trauma therapy for battered women with PTSD (CTT-BW). *Journal of Consulting and Clinical Psychology*, 72, 3–18. doi:10.1037/0022-006X.72.1.3
- Lee, K. J., & Thompson, S. G. (2005). The use of random effects models to allow for clustering in individually randomized trials. *Clinical Trials*, 2, 163–173. doi:10.1191/1740774505cn.0820a
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications, Inc.
- Lohr, S., Schochet, P. Z., & Sanders, E. (2014). Partially nested randomized controlled trials in education research: A guide to design and analysis. (NCER 2014-2000) Washington, DC: US Department of Education, National Center for Education Research, Institute of Education Sciences. Retrieved from <http://ies.ed.gov/>
- Pals, S. L., Murray, D. M., Alfano, C. M., Shadish, W. R., Hannan, P. J., & Baker, W. L. (2008). Individually randomized group treatment trials: A critical appraisal of frequently used design and analytical approaches. *American Journal of Public Health*, 98, 1418–1424.
- Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, 2, 152–162. doi:10.1191/1740774505cn0760a
- Schnurr, P. P., Friedman, M. J., Engel, C. C., Foa, E. B., Shea, M. T., Chow, B. K., & Bernardy, N. (2007). Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *Journal of the American Medical Association*, 297, 820–830.
- Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: Wiley.
- Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance. *Psychological Methods*, 5, 425–433. doi:10.1037/1082-989X.5.4.425