Estimating Standardized Effect Sizes for Two- and Three-Level Partially Nested Data

Mark H. C. Lai

University of Cincinnati

Oi-man Kwok

Texas A&M University

Author Note

This is an Accepted Manuscript of an article published by Taylor & Francis in Structural Equation Modeling: A Multidisciplinary Journal on 01/11/2016, available online: https://www.tandfonline.com/10.1080/00273171.2016.1231606.

Mark H. C. Lai, School of Education, University of Cincinnati; Oi-man Kwok, Department of Educational Psychology, Texas A&M University.

Correspondence concerning this article should be addressed to Mark Lai (Email: mark.lai@uc.edu), School of Education, University of Cincinnati, Cincinnati, OH 45221-0022.

This article is based on part of the first author's doctoral dissertation.

Abstract

Although previous research has discussed an effect size estimator for partially nested cluster randomized designs, the existing estimator (a) is not efficient when used with primary data, (b) can be biased when the homogeneity of variance assumption is violated, and (c) has not yet been empirically evaluated for its finite sample properties. The present paper addresses these limitations by proposing an alternative maximum likelihood estimator for obtaining standardized mean difference effect size and the corresponding sampling variance for partially nested data, as well as the variants that do not make an assumption of homogeneity of variance. The typical estimator, denoted as d (d_W with pooled *SD* and d_C with control arm *SD*), requires input of summary statistics such as observed means, variances, and the intraclass correlation, and is useful for meta-analyses and secondary data analyses; the newly proposed estimator $\hat{\delta}$ ($\hat{\delta}_W$ and $\hat{\delta}_C$) takes parameter estimates from a correctly specified multilevel model as input and is mainly of interest to researchers doing primary research. The simulation results showed that the two methods (d and $\hat{\delta}$) produced unbiased point and variance estimates for effect size. As expected, in general, $\hat{\delta}$ was more efficient than d with unequal cluster sizes, especially with large average cluster size and large intraclass correlation. Furthermore, under heterogeneous variances, $\hat{\delta}$ demonstrated a greater relative efficiency with small sample size for the unclustered control arm. Real data examples, one from a youth preventive program and one from an eating disorder intervention, were used to demonstrate the methods presented. In addition, we extend the discussion to a scenario with a three-level treatment arm and an unclustered control arm, and illustrate the procedures for effect size estimation using a hypothetical example of multiple therapy groups of clients clustered within therapists.

Keywords: Effect size, Partially nested, Partial clustering, Multilevel

Estimating Standardized Effect Sizes for Two- and Three-Level Partially Nested Data

Effect size statistic is important in educational research and is the core concept in the statistical reform in the behavioral sciences (Cumming, 2014; Kline, 2013; Wilkinson & Task Force on Statistical Inference, 1999). Nevertheless, even though effect size reporting has become more common for educational and behavioral studies (Peng, Chen, Chiang, & Chiang, 2013), to date, researchers have rarely paid attention to the complicating issues for defining and estimating standardized effect size for multilevel studies, and the discussion of effect size for complex multilevel designs has mostly been missing.

This paper discusses effect size for a special but not uncommon multilevel design—the *partially nested design*. Although Hedges and Citkowicz (2015) have proposed formulas for estimating different effect sizes with two-level partially nested data, they only discussed effect size estimation using summary statistics, and focused only on the special case where the variance components at the lowest level are assumed to be homogeneous across the treatment and the control arms. In this paper, we illustrate how to estimate sample effect size using maximum likelihood (ML) when raw data are available, and show that these ML estimators are generally more efficient than the one proposed by Hedges and Citkowicz (2015). In addition, we extend the methods for estimating effect size to partially nested data with unequal variances across the treatment and control arms, and evaluated the performances of our proposed methods through simulation studies. Moreover, we extend the discussion of effect size to the design with a three-level treatment arm and a one-level control arm.

Brief Review on Effect Size

In the past few decades effect size has gained increased attention in social science research. For substantive researchers, effect size is crucial in the design phase for sample size planning in order to achieve a desired level of precision of parameter estimates and statistical power (Kelley, 2013); in the analysis and interpretation phase it also gives a sense of the magnitude of a treatment or an intervention (Ellis, 2010; Nakagawa & Cuthill, 2007). For meta-analysts, effect size is the building block of research that summarizes and synthesizes quantitative research findings for the existing literature on a specific issue (Lipsey & Wilson, 2001). Given the importance of effect size for quantitative research, both the American Educational Research Association (AERA; 2006) and the American Psychological Association (APA; 2010) have explicitly recommended using effect size statistics to interpret quantitative research findings.

Effect size measures are well-developed in single-level studies. For experimental or quasi-experimental studies with two treatment groups, or *arms*, the term used in this paper to avoid confusion with clusters, researchers commonly used *standardized mean difference* to quantify the intervention effect in standard deviation unit. Synthesizing 32 reviews (from a total of 116 journals) about effect size reporting practices before 1999 and between 1999 and 2010, Peng et al. (2013) found that the average effect size reporting rate increased from 29.6% before 1999 to 54.7% since 1999, and increased from 42.2% to about 72% for APA/AERA journals. Peng et al. also found that standardized mean difference (in particular, Cohen's *d*) was among the two most commonly reported effect size statistic, alongside with the unadjusted R^2—variance accounted for effect size.

Multilevel Effect Size

Because of their ability to provide the strongest evidence for causal inference when properly implemented, randomized experiments have long been regarded as the gold standard for the social sciences (e.g., Campbell & Stanley, 1963). However, for the majority of research questions in the social sciences, randomization on an individual basis is not always feasible. For example, in studies of instructional intervention, most of the time it is impossible to assign students within the same classroom to receive different instructions, so randomization may often occur at the classroom, rather than individual level. Similarly, for a study of family therapies, it is not reasonable to assign family members to receive different interventions given that family is the unit for the intervention. In such studies where data have a naturally clustered structure, and ignoring the clustering generally results in underestimated standard errors of treatment effects, undercoverage in confidence intervals, and inflated Type I error rates of statistical tests. Larger degree of underestimation is associated with a larger intraclass correlation (ICC) and a larger cluster size (as a function of the design effect, see Hox, 2010). For example, in a simulation study, Wampold and Serlin (2000) showed that ignoring a nested factor of four providers each delivering treatment to 10 patients inflated the true zero treatment effect, $\omega^2 = 0$, to a medium effect size, $\hat{\omega}^2 = .067$, on average. Multilevel modeling has long been suggested as a flexible technique that accounts for the non-independence among observations and gives more accurate statistical inferences (Goldstein, 1986; Mason, Wong, & Entwisle, 1983; Raudenbush & Bryk, 2002).

Although multilevel modeling has been studied in the methodological literature for decades, only recently have researchers started to define and discuss effect size measures for cluster-randomized studies. A review of the articles published in 2015 in three journals: *American Educational Research Journal*, *Journal of Consulting and Clinical Psychology*, and *Journal of Educational Psychology* identified 17 articles involving a cluster-randomized trial. Although in 14 out of the 17 articles (82.3%) the authors reported at least a measure of standardized effect size, and in eight articles (47.1%) Cohen's *d* or a similar index was used, none of the eight articles explicitly talked about issues in multilevel effect size or cited relevant literature. Instead, all authors either simply adopted the definition of Cohen's *d* in single-level studies without any adjustment for clustering or fail to provide any information how they obtained the effect size index in a multilevel context.

As discussed by Hedges (2007) and Hedges and Citkowicz (2015), the direct application of single-level effect size formulas to multilevel data would lead to modest bias in the point estimate but would severely underestimate the standard errors (*SE*) of the effect size. Hedges (2007) presented a real data example wherein the estimated effect size for a connected mathematics curriculum was very similar whether using formulas for single-level and for multilevel studies, but where using the formulas for estimating the variance of single-level effect size would underestimate the true variability by about 5.5 times. Given that the variability of effect size is to be reported in primary studies (e.g. APA, 2010; Peng et al., 2013; Thompson, 2002) and that the variance estimate of sample effect size is commonly used in meta-analyses (Lipsey & Wilson, 2001), it is important to ensure that methods for estimating the variance of multilevel effect size are available and correctly applied.

Our literature review shows that researchers recognized the need to report some sorts of effect size measures for cluster randomized trials but were unaware of the issues of taking into account the multilevel structure in computing effect size statistics, especially for standardized mean differences. A potential reason for this finding is that formal discussion of the extension of Cohen's d-type effect size to multilevel context was not present until Hedges first formally proposed them in 2007 for two-level cluster-randomized trials, followed by Hedges (2011) for three-level trials, Lai and Kwok (2014) for cross-classified data, and Hedges and Citkowicz (2015) for partially nested data (under the homogeneity of variance assumption). Also, there are still disagreements on the most appropriate way to define multilevel effect size statistics (Peugh, 2010), particularly on the choice of standardization (Hedges, 2007), as discussed below.

Partially Nested Design

Nevertheless, the clustered structure may not be the same in different treatment arms. In some cases clustering is a product of the intervention, and the control arm is left ungrouped. For example in the study by Compas et al. (2009) on children of depressed parents, the treatment arm received family-based intervention, whereas the control arm was assigned to a self-study condition. In another randomized trial, Kirschner, Paas, Kirschner, and Janssen (2011) compared the effects of collaborative learning to the control arm of individual learning. Following the previous literature we call such data structure *partially nested* (e.g., Bauer, Sterba, & Hallfors, 2008; Lohr, Schochet, & Sanders, 2014; Moerbeek & Wong, 2008; also called *partially clustered* in Hedges & Citkowicz, 2015). Although research on methods dealing with such data can be found early in Myers, DiCecco, and Lorch (1981) and later in Wehry and Algina (2003), reviews of the existing literature shows that applied researchers seldom adopt appropriate analyses. For example, Bauer et al. (2008) found that 32% of the randomized experiments during 2003 to 2005 in four clinical research journals had a partially nested data structure, which was more common than the fully nested design; however, none of them used the appropriate analyses. Similarly, in a review of 34 articles in public health journals with an individually randomized trials where clustering was created in the treatment arm, Pals et al. (2008) found only two articles using analyses that took into account the clustering effect. Finally, Sanders (2011) noted that 13% of experiments in educational research in 2007 to 2009 used partially nested data, and only two of them used suitable analyses.

For partially nested data researchers either ignored the clustering in the treatment arm and analyzed the data with the conventional t test or single-level regression, or created artificial grouping for the control arm and analyzed the data with standard multilevel modeling. As pointed out by Bauer et al. (2008), Korendijk (2012), and Sanders (2011), ignoring the clustering resulted in underestimation of the standard errors of the treatment effect, whereas the use of artificial grouping in the control arm resulted in biased estimates of the treatment effect and the variance components when the within-cluster variance in the treatment arm is different from the control arm variance (i.e., with heterogeneous variance). Although the impact of ignoring clustering in partially nested design is smaller than in fully nested design, with the typical ICC of .22 in education (Hedges & Hedberg, 2007) and a cluster size of 20 in the treatment arm, ignoring clustering still leads to an underestimation of the variance of the sample effect size by two times.

Hedges and Citkowicz (2015) discussed one effect size estimation approach for two-level partially nested data, but their approach does not correspond to maximum likelihood estimation, which is more efficient with unbalanced cluster sizes, and can produce biased estimates when the homogeneity of variance assumption is violated. In the following sections, we introduce the notations for two-level partially nested designs and discuss two approaches to estimating effect size and obtaining confidence intervals (CIs) for partially nested data: d (d_W with pooled SD, and d_C with control arm SD), which is useful for meta-analysts, and $\hat{\delta}$ ($\hat{\delta}_W$ and $\hat{\delta}_C$), which is useful for researchers conducting primary research. We also report on two simulation studies carried out to evaluate the performances of the four estimators as well as the accuracy of the variance estimates.

In addition, demonstrations are given for computing point and interval estimates for effect size using two real data examples. Finally, we extend the discussion of effect size estimation to three-level partially nested designs.

Effect Size Estimation for Two-Level Partially Nested Design

Model and Notations

Let us consider the situation outlined in Bauer et al. (2008), where participants were randomly assigned to either the treatment or the control arms on an individual basis. Those in the treatment arm were assigned to subgroups and received the treatment, but those in the control arm formed no clustering structure. Let Y_{ij}^{T} $(i = 1, \ldots, n_j;\ j = 1, \ldots, m)$ be the scores of the outcome Y for the ith observation in the jth cluster of the treatment arm, and Y_{i}^{C} $(i = 1, \ldots, N^C)$ be the outcome for the ith observation in the control arm, so that there are m clusters in the treatment arm and n_j observations in the jth cluster. Denote the sample size of the treatment arm and of the control arm as $N^T = \sum_{j=1}^{m} n_j$ and N^C, with the total sample size $N = N^T + N^C$. In a balanced design, $n_1 = \ldots = n_j = n$ and so $N^T = mn$. The control arm is not clustered, so there is no j subscript.

Myers et al. (1981) proposed a pseudogroup and a quasi-F approach to analyzing such data. With the development of multilevel modeling, a model predicting the response variable Y_{ij} can then be conceptualized by the level-1 model (Bauer et al., 2008)

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\mathrm{TREAT}_{ij}) + \varepsilon_{ij}, \tag{1}$$

and the level-2 model

$$\beta_{0j} = \gamma_{00}, \tag{2}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}, \tag{3}$$

where TREAT is a dummy variable coded as 1 = treatment arm and 0 = control arm. Thus, for the treatment arm there are the level-2 random effect, u_{1j}, and the level-1 error term, ε_{ij}, whereas for the control arm there is only the error term ε_{ij}. Here β_{0j} is the within-cluster regression intercept for cluster j, which is assumed to remain constant across clusters and equals γ_{00}, the population mean of the control arm. The random slope, β_{1j}, denotes the difference between the mean of the jth cluster in the treatment arm and the control arm mean, μ_C; under a balanced design its mean across all js is γ_{10}, which can be unbiasedly estimated by the difference between the sample means, $\bar{Y}^T_{..} - \bar{Y}^C$. The cluster-specific random effect is captured by u_{1j} with $V(u_{1j}) = \sigma^2_B$. The level-1 error term is, ε_{ij}, and without assuming homogeneity of variance we have $V(\varepsilon_{ij}|\text{TREAT} = 0) = \sigma^2_C$ for the control arm and $V(\varepsilon_{ij}|\text{TREAT} = 1) = \sigma^2_{W|T}$ for the treatment arm. When the clustering involves random assignment and the treatment effect does not change the within-cluster variability, it is reasonable to assume constant variance across both the treatment and the control arms, so that $\sigma^2_{W|T} = \sigma^2_C = \sigma^2_W$ (Bauer et al., 2008). The error terms u_{1j} and ε_{ij} both have an expected value of zero and are independent, and in general we assume that they both follow a normal distribution. Note that the sum of the variance components within the treatment arm is $\sigma^2_{W|T} + \sigma^2_B$, whereas that within the control arm has only one component, σ^2_C. Thus, the variances of two arms differ, unless $\sigma^2_B = 0$. Let us define the ICC for the treatment arm as ρ, where

$$\rho = \frac{\sigma^2_B}{\sigma^2_{W|T} + \sigma^2_B}. \tag{4}$$

Such a model can easily be analyzed using common statistical packages for multilevel modeling; indeed, Bauer et al. (2008) provided codes for fitting such a model in SPSS and SAS. In a simulation study, Sanders (2011) showed that Bauer et al.'s use of a fixed intercept and a random treatment effect is better than treating the intercept as random with a fixed treatment effect, an approach found to have inflated Type I error rate (Wehry & Algina, 2003). In another simulation study, Baldwin, Bauer, Stice, and Rohde (2011) further showed that Bauer et al.'s partially clustered multilevel models produced unbiased and efficient parameter estimates while keeping the Type I error rate at nominal level (see also Talley, 2013). Finally, Sterba et al. (2014) showed that the above model can be reparameterized and analyzed using structural equation modeling (SEM) software.

Effect Size Estimations

In treatment-control arm studies, the most common effect size statistic is the standardized mean difference (Cohen, 1988; Hedges, 1981),

$$\delta = \frac{\mu^T - \mu^C}{\sigma}, \tag{5}$$

where μ^T and μ^C are the population means of the treatment and of the control arm respectively and σ is the pooled within-arm standard deviation.

Choice of standard deviation. Hedges (2007) and Hedges and Citkowicz (2015) noted that with multilevel data, the concept of effect size is vague. That is because σ can be defined as σ_W (with homoscedasticity assumed, i.e., $\sigma^2_{W|T} = \sigma^2_C$), σ_B, or $\sqrt{\sigma^2_W + \sigma^2_B}$, each with a different target of generalization. If homoscedasticity is not assumed, we have two additional choices of σ_C and $\sqrt{\sigma^2_{W|T} + \sigma^2_B}$. The choice generally depends on which *SD* is more natural in the population. For example, if the population is naturally unclustered and the treatment is considered artificially created, σ_W is to be preferred in defining the population effect size; Hedges and Citkowicz (2015) provided expressions for the point and variance estimates of effect size using that definition (d_W in their paper, which we subsequently denoted as $d_{W\mathrm{HC}}$ to distinguish it from d_W that we propose in this paper). Research studies with a partially nested design usually fall into this category (see also Heo, Litwin, Blackstock, Kim, & Arnsten, 2014; Lohr et al., 2014). On the other hand, if the clustered condition is an established norm (e.g., students in classrooms) and the researcher is interested in an individually (i.e., unclustered) implemented intervention (e.g., homeschooling), $\sqrt{\sigma^2_W + \sigma^2_B}$ is preferable in defining the population effect size. In this case, it may be more intuitive to think of the arm with the clustered intervention as the control and the one with the individually administered intervention as the treatment. For partially nested data such as the example given in Bauer et al. (2008), because the clustering is artificially created in the treatment arm and does not naturally occur in the general population, σ_W would be a better denominator for effect size. Therefore, the present study only discusses standardized effect size using σ_W as the denominator, which is analogous to Cohen's *d* with homogeneous variance (and is analogous to Glass's [1976] effect size when using only the control arm *SD*).

Assuming Homogeneity of Variance

Let $\bar{Y}^T_{..}$ and $\bar{Y}^C$ be the grand means of the treatment arm and of the control arm, respectively, and $\bar{Y}^T_{.j}$ the mean of the jth cluster in the treatment arm. First, we define the within-cluster variance of the treatment arm and the variance of the control arm as

$$S^2_{W|T} = \frac{\sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(Y^T_{ij} - \bar{Y}^T_{.j}\right)^2}{N^T - m}, \tag{6}$$

$$S^2_C = \frac{\sum_{i=1}^{N^C} \left(Y^C_i - \bar{Y}^C\right)^2}{N^C - 1}. \tag{7}$$

When the within-cluster variance in the treatment arm equals the variance of the control arm at the population (i.e., $\sigma^2_{W|T} = \sigma^2_C$), a situation discussed in Bauer et al. (2008), we can let S^2_W be the pooled within-cluster level variance such that

$$S^2_W = \frac{(N^T - m)S^2_{W|T} + (N^C - 1)S^2_C}{N - m - 1}, \tag{8}$$

and let $S^2_{B|T}$ be the between-cluster mean squares in the treatment arm, where

$$S^2_{B|T} = \frac{\sum_{j=1}^{m} n_j \left(\bar{Y}^T_{.j} - \bar{Y}^T_{..}\right)^2}{m - 1}. \tag{9}$$

We first consider situations where equal within-treatment arm variance holds, and then consider the case when heteroscedasticity is present (see also Moerbeek & Wong, 2008).

We first consider estimating the population effect size under homogeneous variance with

$\sigma^2_{W|T} = \sigma^2_C = \sigma^2_W$, where the population effect size is defined as $\delta_W = \gamma_{10}/\sigma_W$, a situation studied in Hedges and Citkowicz (2015) (although they define effect size using only the control arm *SD*). We then consider effect size estimation under heterogeneous variance with $\sigma^2_{W|T} \neq \sigma^2_C$. Next we present two approaches, namely, d_W and d_C (with the use of summary statistics) and $\hat{\delta}_W$ and $\hat{\delta}_C$ (with the use of maximum likelihood estimates), to estimate population effect size from a sample of partially nested data.[1]

Using summary statistics. The materials in this section are similar to Hedges and Citkowicz (2015) with minor differences in notations. Using summary statistics, the sample effect size is

$$d_W = \frac{\bar{Y}^T_{\cdot\cdot} - \bar{Y}^C}{S_W}, \tag{10}$$

and

$$V(d_W) = \frac{1 + (\tilde{n} - 1)\rho}{N^T(1 - \rho)} + \frac{1}{N^C} + \frac{d^2_W}{2(N - m - 1)}, \tag{11}$$

where S_W is as defined in (8) and $\tilde{n} = \sum_{j=1}^m n^2_j/N^T$, which reduces to n under a balanced design. The formulas for d and $V(d_W)$ assume that the population ICC, ρ, is known, but we can plug in a sensible estimate based on theory or replace it by a moment estimator $(S^2_{B|T} - S^2_W)/[S^2_{B|T} + (n_U - 1)S^2_W]$ where $n_U = (N^T - \tilde{n})/(m - 1)$. The derivation of (10) and (11) may be found in the Appendix. Note that for unbalanced designs, the grand mean is no longer an efficient estimator of the mean of the control arm, so d_W is not the most efficient estimator for δ_W (i.e., variance of d_W is larger than the second method described below).

Using maximum likelihood estimation. If consistent estimates of γ_{10} (fixed effect) and σ_W (random effect) and their associated variance estimates (or standard error estimates) are accessible, we can use the following equations based on the estimated variance components (see the Appendix for derivation)

$$\hat{\delta}_W = \frac{\hat{\gamma}_{10}}{\hat{\sigma}_W}, \tag{12}$$

$$V(\hat{\delta}_W) = \frac{V(\hat{\gamma}_{10})}{\hat{\sigma}^2_W} + \frac{\hat{\delta}^2_W V(\hat{\sigma}^2_W)}{4\hat{\sigma}^4_W}. \tag{13}$$

If the maximum likelihood or restricted maximum likelihood estimates of $\hat{\gamma}_{10}$ and $\hat{\sigma}_W$ are available, which are asymptotically unbiased, consistent (i.e., converged to the population value), and efficient (i.e., with minimum variance) under general conditions, then by the invariance property of the maximum likelihood (Casella & Berger, 2002) $\hat{\delta}_W$ is also the maximum likelihood estimate of δ_W and is asymptotically unbiased, consistent, and efficient, even for conditions with unbalanced data. Thus, when relevant information is available, $\hat{\delta}_W$ is a better estimator than d_W.

Using Only the *SD* of the Control Arm

For single-level studies, Glass (1976) suggested computing the effect size using only the standard deviation of the control arm if there is evidence or reason to believe that the treatment changes the variance of the score distribution. Similarly, in partially nested design, the within-cluster variance, σ_W, may be affected by the treatment. In this case, we agree with Glass (1976) that the control arm *SD* would be a more natural choice for standardization as it represents the variability of the general population without intervention. Therefore, we define the population effect size δ_C as

$$\delta_C = \frac{\mu^T - \mu^C}{\sigma^C}, \tag{14}$$

where σ^C is the population *SD* of the control arm[2]. Bauer et al. (2008) also discussed a model with $\sigma^2_C \neq \sigma^2_{W|T}$ and provided SPSS and SAS code for fitting partially nested models with heterogeneous variance.

When σ^2_C is not equal to $\sigma^2_{W|T}$, $\delta_C \neq \delta_W$, as they have different denominators. Consider a hypothetical example with $N^T = N^C = 100$ and $m = 20$ with a equal cluster sizes in the treatment arm. Assume a medium effect size at the population level such that $\delta_C = 0.5$. If the level-1 variance of the treatment arm is only half of the control arm variance such that $\sigma^2_{W|T} = 0.5\sigma^2_C$, using equation (8) the pooled variance is expected to be $[(100 - 20)0.5\sigma^2_C + (100 - 1)\sigma^2_C]/(200 - 20 - 1) = 0.78\sigma^2_C$. Thus, the pooled *SD* is $\sqrt{0.78} = 0.88$ times the control arm *SD*. As a result, when we estimate $\delta_C = 0.5$ using the pooled *SD* with this example, we expect to obtain a value that is $1/0.88 = 1.13$ times the true δ_C, or .57. It is obvious from (8) that the pooled *SD* is smaller than the control arm *SD* (i.e., overestimated effect size) when $\sigma^2_{W|T} < \sigma^2_C$, and the pooled *SD* is larger than the control arm *SD* (i.e., underestimated effect size) when $\sigma^2_{W|T} > \sigma^2_C$. It should also be clear that using the pooled *SD* gives a more biased estimate of δ_C when $N^T - m$ is large relative to N^C.

Using summary statistics. A sample estimator of δ_C, d_C can be obtained as

$$d_C = \frac{\bar{Y}^T_{\cdot\cdot} - \bar{Y}^C}{S_C}, \tag{15}$$

$$V(d_C) = \upsilon\frac{1 + (\tilde{n} - 1)\rho}{N^T(1 - \rho)} + \frac{1}{N^C} + \frac{(d_C)^2}{2(N^C - 1)}, \tag{16}$$

where S_C has been defined in equation (7) and $\upsilon = \sigma^2_{W|T}/\sigma^2_C$ is the variance ratio between the treatment and the control arms, which can be estimated as

$$\hat{\upsilon} = \frac{N^C - 1}{N^C - 3}\frac{S^2_{W|T}}{S^2_C}.$$

The sample estimator $\hat{\upsilon}$ is unbiased when the cluster sizes are balanced as the factor $(N^C - 1)/(N^C - 3)$ corrects for the bias in the F-ratio $S^2_{W|T}/S^2_C$. Note that $V(d_C) > V(d_W)$ when $\upsilon = 1$, so d_W is preferred when variance can be assumed equal. Also note that even though d_C shares some similarity to $d_{W\text{HC}}$ proposed by Hedges and Citkowicz (2015, p. 1299; also denoted as d_W in their paper), our current estimator does not require the homogeneity of variance assumption whereas theirs assumes that the homogeneity of variance assumption holds.

Using maximum likelihood estimation. If reasonable point and variance estimates for γ_{10} and σ^2_C can be obtained, δ_C and its sampling variance can be estimated by plugging in the maximum likelihood estimates:

$$\hat{\delta}_C = \frac{\hat{\gamma}_{10}}{\hat{\sigma}_C}, \tag{17}$$

$$V(\hat{\delta}_C) = \frac{V(\hat{\gamma}_{10})}{\hat{\sigma}_C^2} + \frac{(\hat{\delta}_C)^2 V(\hat{\sigma}_C^2)}{4\hat{\sigma}_C^4}. \tag{18}$$

Choosing between the control arm *SD* and the pooled *SD*. Whereas the use of the pooled *SD* can produce inconsistent effect size estimate, the use of the control arm *SD* is consistent without assuming homogeneity of variance, and should be a more defensible option when strong evidence of the equality between $\sigma^2_{W|T}$ and σ^2_C is not present. On the other hand, by comparing expressions (11) and (16), it is obvious that, when the homogeneity assumption holds such that $\delta_W = \delta_C$, d_W (and $\hat{\delta}_W$) is also a consistent but more efficient estimator of the population effect size than d_C (and $\hat{\delta}_C$), especially when N^T and N^C are small and the effect size is large. Therefore, when the total sample size is small, if the pooled sample *SD*, S_W, has a similar value as the control arm *SD*, S_C, using the pooled *SD* provides a more precise effect size estimate. Although we can test the homogeneity of variance hypothesis (H_0: $\sigma^2_{W|T} = \sigma^2_C$) by comparing nested multilevel models with a likelihood ratio test, the power to detect violation of such homogeneity is not known. As the homogeneity of variance assumption is more restrictive, we recommend using d_C rather than d_W (and $\hat{\delta}_C$ over $\hat{\delta}_W$) unless there are strong evidence or a strong rationale for imposing the homogeneity of variance assumption.

Constructing Approximate Confidence Intervals for Sample Effect Size Estimates

Like any other point estimates such as the sample mean, sample effect size estimates provide absolutely no information about the uncertainty in the estimated effect size. Numerous authors have commented on the importance of reporting CI for effect size (e.g., Cumming, 2014; Grissom & Kim, 2012; Hedges, 2008; Peng et al., 2013; Thompson, 2002), and both the AERA (2006) and the APA (2010) strongly encourage reporting CI along with an effect size estimate.

Let $\tilde{\delta}$ be one of the sample effect size estimator discussed in this paper (i.e., $\tilde{\delta} = d_W$, d_C, $\hat{\delta}_W$, or $\hat{\delta}_C$). Based on the central limit theorem (cf. Casella & Berger, 2002), $\tilde{\delta}$ will be consistent and normally distributed with a large sample size. Therefore, an approximate $(1 - \alpha) \times 100\%$ CI for $\tilde{\delta}$ may be obtained as

$$[\tilde{\delta} - z_{1-\alpha/2} SE(\tilde{\delta}), \tilde{\delta} + z_{1-\alpha/2} SE(\tilde{\delta})], \tag{19}$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile in the standard normal distribution. For example, for the commonly reported 95% CI, one uses $z_{.975} \approx 1.96$. Also, in practice $SE(\tilde{\delta}) = \sqrt{V(\tilde{\delta})}$ has to be estimated from the sample using equation (11) for d_W, (13) for $\hat{\delta}_W$, (16) for d_C, and (17) for $\hat{\delta}_C$.

Simulation to Evaluate the Performance of d_W, d_C, $\hat{\delta}_W$, and $\hat{\delta}_C$

A simulation study was used to evaluate the performance of d_W, $\hat{\delta}_W$, d_C, and $\hat{\delta}_C$, and their analytically derived variances for two-level partially nested design. The design factors of the simulation were as follows: population effect size ($\delta = .2, .5, .8$), ICC ($\rho = .1, .25, .5$), number of clusters in the treatment arm ($m = 10, 30, 100$), average cluster size ($\bar{n} = 5, 10, 25, 50$), degree of unbalanced cluster sizes ($\tilde{n}{:}\bar{n} = 1, 1.64$), sample size ratio between the two arms ($N^T{:}N^C = 1, 5$), and the variance ratio of the level-1 error terms between the two arms ($\upsilon = \sigma^2_{W|T}/\sigma^2_C = .5, 1, 2$). For each of the 1,296 simulation conditions, 2,000 data sets were generated in R (R Core Team, 2015) using the model defined in equations (1) to (3), with σ_C fixed to one (and larger ICC corresponding to larger σ^2_B. The effect size estimates d_W and d_C and their variances were easily obtained in R using equations (10), (11), (15), and (16), whereas for $\hat{\delta}_W$ and $\hat{\delta}_C$ we obtained the estimated variance components using the R packages `lme4` (Bates, Mächler, Bolker, & Walker, 2015) and `nlme` (Pinheiro, Bates, DebRoy, Sarkar, & R Core Team, 2015), respectively.

Because the methods were derived analytically, the purpose of the simulation was mainly to determine how robust the analytical results were under the extreme conditions of small cluster size, unbalanced cluster sizes, and small number of clusters. Therefore, we do not present every detail of the simulation study here (the tables of the full simulation results may be obtained from the first author). For each condition, we considered the point estimate to have unacceptable bias when the absolute value of the standardized bias was larger than .40 (Collins, Schafer, & Kam, 2001). In turn, we considered the variance (or standard error) estimate to have unacceptable bias when the absolute value of the relative *SE* bias was larger than .10 (Hoogland & Boomsma, 1998). Finally, with unbalanced cluster sizes, d_W and d_C were expected to be inefficient, meaning that they had a larger sampling variance. Therefore, we also computed the relative efficiency of $\hat{\delta}_W$ relative to d_W, $RE(\hat{\delta}_W, d_W) = MSE(d_W)/MSE(\hat{\delta}_W)$ (and similarly for $RE[\hat{\delta}_C, d_C]$), as the ratio of their mean squared errors (*MSE*; i.e., squared bias plus sampling variance). If $RE(\hat{\delta}_W, d_W) > 1$, $\hat{\delta}_W$ is more efficient than d_W.

Simulation Results. Figure 1 shows the standardized biases of d_W and $\hat{\delta}_W$ when the homogeneity assumption is met (i.e., $\upsilon = 1$) and those of d_C and $\hat{\delta}_C$ for all conditions. As illustrated, all four estimators showed little bias in the point estimates except for the combination of $m = 10$ and $\bar{n} = 5$, but even under such conditions with small sample sizes the standardized biases were mostly under 20%, well within the acceptable range suggested by Collins et al. (2001). Figure 2 shows the relative *SE* biases with respect to the four effect size estimators. As illustrated, the relative *SE* biases were generally within 10% in absolute values except for $\hat{\delta}_W$ and $\hat{\delta}_C$ for a few conditions with $m = 10$, $\bar{n} = 5$, and unbalanced cluster sizes. In summary, the standardized biases and the relative *SE* biases for all four estimators were generally small, confirming the adequacy of the analytical results even for extreme conditions. Note that Baldwin et al. (2011) found that both the fixed effects and the variance components were unbiasedly recovered using the correctly specified partially nested model, so our results were consistent with theirs.

The relative efficiency of the maximum likelihood estimator to using summary statistics was found to depend mainly on the imbalance of the cluster sizes, ICC, and average cluster size. As shown in Figure 3, when the design was balanced (i.e., $\tilde{n} = \bar{n}$), the relative efficiency was very close to 1.00, as $\hat{\delta}$ was generally identical to d_W and $\hat{\delta}_C$ was generally identical to d_C except when the estimated ICC was at the boundary. On the other hand, when cluster sizes were not all the same, $\hat{\delta}_W$ became more efficient with a larger ICC and a larger $\bar{n}$, and when $\bar{n} = 50$ and $\rho = .5$, $\hat{\delta}_W$ were 1.4 to 1.6 times more efficient than d_W, meaning that the sampling variance of d_W was 1.4 to 1.6 times bigger than that of $\hat{\delta}_W$. The results were generally similar for $\hat{\delta}_C$ and d_C.

Real Data Illustration

Example 1. The summary of the multilevel analysis provided in Model 1 of Bauer et al. (2008, p. 231) was used to demonstrate the usage of equations (12) and (13) for effect size estimation. The data concerned the effectiveness of the Reconnecting Youth (RY) preventive intervention program, which involved participants in 9th- to 11th-grade from five schools in the Southwest and four schools on the Pacific coast in the United States (see Cho, Hallfors, & Sánchez, 2005, for more information of the sample and the study). A total of 1,370 students at high risk of dropout, drug use, and emotional distress were randomly assigned to eitherthe intervention or the control arm. In the intervention arm, or the RY arm, only $N^T = 370$ out of 695 students were retained. These students were grouped into 41 classes to receive a one-semester curriculum with the intention of developing skills and a supporting group environment. On the other hand, students in the control arm ($N^C = 675$) were not specifically grouped. There was also a comparison group of low-risk students called the *typical* arm, but for this illustration we only focus on the comparison of the RY and the control arm.

Results showed that grouping high-risk students resulted in some negative outcomes in the posttest after six months—the outcome variable of interest here was deviant peer bonding. The fixed effects included dummy variables representing the memberships of the RY arm and of the control arm, as well as those representing the schools the students attended. The two random components were the person-level residual variance (which was assumed to be constant across arms) and the class-level residual variance.

The treatment effect of RY compared to control had a coefficient $\hat{\gamma}_{10} = 0.19$, $t(68.3) = 2.63$ on deviant peer bonding, with $\hat{\sigma}_W^2 = 0.789$, $z = 26.73$, and $\hat{\sigma}_B^2 = 0.053$. Using equation (12), it is clear that the effect size of RY was $\hat{\delta}_W = 0.19/\sqrt{0.789} = 0.214$. To estimate $SE(\hat{\delta}_W)$, we need *SE*s of $\hat{\gamma}_{10}$ and of $\hat{\sigma}_W^2$, which Bauer et al. (2008) did not directly report. However, using the values of the test statistics, we could estimate the *SE* of the effect of RY as $0.19/2.63 = 0.0722$ and that of σ_W^2 as $0.789/26.73 = 0.0295$. Substituting $V(\hat{\gamma}_{10}) = 0.0722^2$, $\hat{\delta} = 0.214$, $\hat{\sigma}_W^2 = 0.789$, and $V(\hat{\sigma}_W^2) = 0.0295^2$ into equation (13)for $V(\hat{\delta}_W)$, we got

$0.0722^2/0.789 + (0.214^2)(0.0295^2)/(4 \times 0.789^2) = 0.0066$ (or $SE = 0.0814$). The approximate 95% symmetric CI for the sample effect size could be obtained as $0.214 \pm z_{.025}(0.0814)$, which equals $[0.054, 0.374]$. Therefore, we found that RY had a small but nonzero effect of increasing deviant peer bonding for high-risk students.

If the parameter estimates for the correctly specified multilevel model are not available, we can still estimate the effect size d by using some estimates or representative values of the ICC, ρ. If we substituted $\hat{\rho} = \hat{\sigma}^2_B/(\hat{\sigma}^2_B + \hat{\sigma}^2_W) = 0.063$ to equation (12) and (13) and assume S_W to be the same as $\hat{\sigma}_W$ and $\bar{Y}^T_{\cdots} - \bar{Y}^C = \hat{\gamma}_{10}$, we would get the same estimate $d_W = \hat{\delta}_W = 0.214$ and a smaller estimated SE of 0.076, assuming an equal cluster size of 9.02. As Bauer et al. (2008) noted that there were 5 to 15 students in each intervention group, a more conservative estimate of $SE(d_W)$ would be obtained by using $\tilde{n} = 1.5\bar{n} = 13.53$, resulting in an estimated SE of 0.0816, much closer to the estimate obtained by maximum likelihood. If we were to use the variance formulas in Hedges and Citkowicz (2015) for d_W, we would get $SE = 0.0817$, which is slightly larger than $SE(d_W)$ found using equation (13) in this paper, as Hedges and Citkowicz only used the SD for the control arm while assuming homoscedasticity. Finally, if the formulas for single-level effect size were used, we would get $SE_{\text{naive}}(d) = \sqrt{1/370 + 1/675 + 0.214^2/[2(370 + 675 - 2)]} = 0.065$, an underestimation of the SE taking into account the clustering by slightly more than 20%.

Example 2. We illustrate the estimation of δ_C using the summary from Baldwin et al. (2011), a re-analysis of the data from Stice, Shaw, Burton, and Wade (2006). The data are from a dissonance-based eating disorder prevention intervention for 480 female adolescents in four intervention arms: dissonance intervention, healthy-weight management, expressive writing, and assessment-only control. The dissonance intervention arm showed the biggest decrease in thin-ideal internalization (TII) compared to the assessment-only control arm, so in this illustration we only use results pertaining to these two arms. In the intervention arm, $N^T = 114$ adolescents in 17 clusters ($\bar{n} = 6.7$), whereas in the baseline arm $N^C = 126$ unclustered participants. Using a likelihood ratio test Baldwin et al. found that the heterogeneous variance model fitted the data better than a homogeneous variance model, so we use the estimates from the heterogeneous variance model to estimate the effect size for the dissonance intervention on TII.

The treatment effect had a coefficient of −0.44, $t(23.2) = -5.51$ on deviant peer bonding, with $\sigma^2_{W|T} = 0.34$, $z = 6.84$, $\sigma^2_C = 0.20$, $z = 7.87$, and $\sigma^2_B = 0.03$. Using equation (17), it is clear that the effect size of dissonance intervention was $\hat{\delta}_C = -0.44/\sqrt{0.20} = -0.98$. To estimate $SE(\hat{\delta}_C)$, we need *SE*s of $\hat{\gamma}_{10}$ and of $\hat{\sigma}^2_{W|T}$, which Baldwin et al. (2011) did not directly report. Again, using the values of the test statistics, we could estimate the *SE* of the effect of RY as $-0.44/-5.51 = 0.080$ and that of $\hat{\sigma}^2_{W|T}$ as $0.20/7.87 = 0.0254$. Substituting $V(\hat{\gamma}_{10}) = 0.080^2$, $\hat{\delta}_C = -0.98$, $\hat{\sigma}^2_W = 0.20$, and $V(\hat{\sigma}^2_W) = 0.0254^2$ into the formula for $V(\hat{\delta}_C)$, that is, equation (18), we got $0.080^2/0.20 + (-0.98)^2(0.0254^2)/(4 \times 0.20^2) = 0.0358$ (or $SE = 0.189$). Then the approximate 95% symmetric CI could be obtained as $-0.98 \pm z_{.025}(0.189)$, which equals $[-1.355, -0.613]$. Therefore, we found that the dissonance intervention had a large effect on reducing TII.

If instead we assumed homogeneity of variance with an estimated treatment effect of −0.44, $t(28.6) = -5.25$ and $\hat{\sigma}^2_W = 0.27$, $z = 14.88$, we would get an estimated effect size $\hat{\delta}_W = -0.85$, with $SE = 0.164$ and 95% symmetric CI of $[-1.168, -0.528]$, and thus an underestimated effect size as compared to $\hat{\delta}_C$.

Efficiency Lost of d_C and $\hat{\delta}_C$ When Homogeneity of Variance Holds

Although d_C and $\hat{\delta}_C$ are preferable to d_W and $\hat{\delta}_W$ in the sense that the former two do not make the assumption of homogeneity of variance, they are less efficient estimators when homogeneity of variance holds. Comparing equations (11) and (16), we see that the loss of efficiency in d_C may be attributed to the replacement of $N - m - 1$ by $N^C - 1$ in the denominator of the last term and the noise introduced in computing the sample variance ratio υ. To pinpoint the efficiency lost in d_C and $\hat{\delta}_C$, we computed the relative efficiencies $RE(d_C, d_W)$ and $RE(\hat{\delta}_C, \hat{\delta}_W)$ using the data in Simulation 1. ANOVA results showed that there were essentially no difference between $RE(d_C, d_W)$ and $RE(\hat{\delta}_C, \hat{\delta}_W)$ (with $\eta^2 = .0008$ for all its main and interaction effects). The major factors influencing relative efficiency was the main effect of N^T:N^C ($\eta^2 = .31$), followed by the main effects of average cluster size $\bar{n}$ ($\eta^2 = .14$), number of clusters J ($\eta^2 = .11$), and population effect size δ ($\eta^2 = .09$). As shown in Table 1, d_C was almost as efficient ($RE = 0.95$ to 1.0) as d_W when N^T:$N^C = 1$. When the control arm sample size N^C is small relative to N^T, d_C is less efficient, especially when J and $\bar{n}$ are small and δ was large. For example, when $m = 10$, $\bar{n} = 5$, and $\delta = 0.8$, the relative efficiency was only 0.66; if δ was changed to 0.2, the relative efficiency would increase to 0.79; if, in addition, $\bar{n}$ was changed to 25, the relative efficiency would increase to 0.95; and when $m \geq 30$, $\bar{n} = 25$, and $\delta = 0.2$, the relative efficiency was 0.99, in which case, d_C was almost as efficient as d_W.

Three-Level Experimental Arm

The previous discussion represents a relatively simple two-level partially nested scenario. In real data, the research design can be extended in multiple ways, such as having more than two clustering levels or including multiple treatment conditions. In this paper, we extend the discussion of effect size to a scenario where the treatment arm follows a three-level hierarchical structure and the control arm is unclustered, which we denote as a 3T1C design. We chose to discuss this design as it has been mentioned in several previous studies (Heo et al., 2014; Lohr et al., 2014; Sterba, 2015), and is a natural extension of the two-level partially nested design. An example was given in Sterba (2015) where, in the treatment arm, there were multiple therapists each managing several therapy groups of clients, whereas the control arm consisted of wait-list clients who were unclustered. Therefore, the treatment arm had a three-level structure but the control arm had only one level; the research question was to evaluate the effectiveness of therapy groups.

Model and Notations

Consider a design with a three-level treatment arm and a one-level control arm. In the treatment arm, there are m therapists indexed by $k = 1, \ldots, m$, where the kth therapist is in charge of p_k therapy groups, indexed by $j = 1, \ldots, p_k$. The jth therapy group for the kth therapist consists of n_{jk} clients, indexed by $i = 1, \ldots, n_{jk}$. Let N^T be the number of clients such that

$N^T = \sum_{k=1}^{m} \sum_{j=1}^{p} n_{jk}$. The control arm, on the other hand, consists of N^C unclustered wait-list clients. A model for this design may be formulated as

$$Y_{ijk} = \gamma_{000} + \gamma_{100}(\text{TREAT}_{ijk}) + r_{1jk}(\text{TREAT}_{ijk}) + u_{10k}(\text{TREAT}_{ijk}) + \varepsilon_{ijk}, \tag{20}$$

where TREAT is a dummy variable coded as 1 = treatment arm and 0 = control arm, γ_{000} is the grand mean for the control arm, and γ_{100} denotes the average treatment effect. The random effect term r_{1jk} is the clustering effect of therapy group j of therapist k, u_{10k} is the level-3 effect of the kth therapist, and ε_{ijk} is the level-1 error term for individual i in therapy group j of therapist k. For a balanced design, $i = 1, \ldots n$, $j = 1, \ldots, p$, and $k = 1, \ldots, m$ for the treatment arm and $i = 1, \ldots, N^C$ for the control arm. It is assumed that, for the treatment arm, $\varepsilon_{ijk} \sim N(0, \sigma^2_{W|T})$, $r_{1jk} \sim N(0, \sigma^2_2)$, and $u_{10k} \sim N(0, \sigma^2_3)$, and, for the control arm, $\varepsilon_{ijk} \sim N(0, \sigma^2_C)$.

For the above model, we can impose the equality constraint $\sigma^2_{W|T} = \sigma^2_C = \sigma^2_W$. Under this homogeneity assumption, the population effect size may be defined as $\delta_W = \gamma_{100}/\sigma^2_W$. Let us define the pooled level-1 sample variance as

$$S^2_W = \frac{\sum_{k=1}^{m} \sum_{j=1}^{p_k} \sum_{i=1}^{n_{jk}} (Y_{ijk} - \bar{Y}^T_{\ldots}) + \sum_{i=1}^{N^C} (Y_i - \bar{Y}^C)}{N - P - 1},$$

where $N = N^T + N^C$ is the overall sample size and $P = \sum_{k=1}^{m} p_k$ is the total number of level-2 units. Under a balanced design $P = mp$. An unbiased estimator for the population effect size is

$$d_W = \frac{\bar{Y}^T_{\ldots} - \bar{Y}^C}{S_W} \tag{21}$$

and, under a balanced design, its approximate sampling variance is

$$V(d_W) = \frac{1 + (n-1)\rho_2 + (n_2 - 1)\rho_3}{N^T(1 - \rho_2 - \rho_3)} + \frac{1}{N^C} + \frac{d^2}{2(N - P - 1)}, \tag{22}$$

where $\rho_2 = \sigma^2_2/(\sigma^2_W + \sigma^2_2 + \sigma^2_3)$ and $\rho_3 = \sigma^2_3/(\sigma^2_W + \sigma^2_2 + \sigma^2_3)$ are the ICCs for level 2 and level 3, respectively, for the treatment arm, and $n_2 = pn$ is the number of level-1 units in each level-3 cluster.

If the homogeneity of variance assumption is not imposed, it is more natural to estimate the population effect size $\delta_C = \gamma_{100}/\sigma_C$. The sample effect size estimator and its sampling variance are

$$d_C = \frac{\bar{Y}^T_{\cdots} - \bar{Y}^C}{S_C}, \tag{23}$$

and

$$V(d_C) = \upsilon \frac{1 + (n-1)\rho_2 + (n_2 - 1)\rho_3}{N^T(1 - \rho_2 - \rho_3)} + \frac{1}{N^C} + \frac{(d_C)^2}{2(N^C - 1)}. \tag{24}$$

As suggested in Hedges and Citkowicz (2015), we can plug in the ICCs, ρ_2 and ρ_3, with some reasonable estimates based on the literature or on substantive knowledge. Alternatively, we can obtain the method of moment estimates using the equations given in Searle, Casella, and McCulloch (2006, p. 429).

Unequal cluster sizes. With unequal cluster sizes in the treatment arm at level 1 (i.e., $n_{jk} \neq n_{j'k'}$ for some $j \neq j'$ and $k \neq k'$) and/or level 2 (i.e., $p_k \neq p_{k'}$ for some $k \neq k'$), we can replace n_2 in equations (21) to (24) by

$$\tilde{n}_2 = \frac{1}{N^T} \sum_{k=1}^{m} \left(\sum_{j=1}^{p_k} n_{jk} \right)^2$$

and replace n by

$$\tilde{n} = \frac{1}{N^T} \sum_{k=1}^{m} \sum_{j=1}^{p_k} n_{jk}^2.$$

On the other hand, if maximum likelihood estimates and the corresponding sampling variances of the treatment effect $\hat{\gamma}_{100}$ and of the pooled within-cluster *SD*, $\hat{\sigma}_W$, are available, we can plug in the MLEs similar to equations (12) and (13) to obtain the MLE of the population effect size

$$\hat{\delta}_W = \frac{\hat{\gamma}_{100}}{\hat{\sigma}_W},$$

$$V(\hat{\delta}_W) = \frac{V(\hat{\gamma}_{100})}{\hat{\sigma}_W^2} + \frac{\hat{\delta}_W^2 V(\hat{\sigma}_W^2)}{4\hat{\sigma}_W^4}.$$

If homogeneity of variance is not assumed, we can instead plug in the MLE for the control arm *SD*, $\hat{\sigma}_C$, in the place of $\hat{\sigma}_W$, and estimate effect size using the above equations.

Simulation to Evaluate the Performance of the Estimators in the 3T1C Design

We also evaluated the performance of the four effect size estimators for the 3T1C design using simulation. However, as shown in the results for the two-level partially nested design, the effect size estimators generally performed as expected except for conditions with high ICC, small number of clusters, small and unbalanced cluster sizes. Therefore, it is sufficient to evaluate the performance of the four estimators using some extreme conditions. For the 3T1C design, we used a smaller scale of simulation conditions with $m = 10$, $\bar{p}$ (average number of level-2 units in a level-3 cluster) $= 10$, $\bar{n} = 5$ or 50, $\rho_3 = .1, .25, .5$, $\rho_2 = .1, .2$, $\delta = .2, .5, .8$, $N^T{:}N^C = 1, 5$, and $\upsilon = .5, 1, 2$. The clusters at level 2 are unbalanced such that half of the clusters have size $9\bar{p}/5 = 18$ and the other half have size $\bar{p}/5 = 2$. Similarly, half of the clusters at level 1 have size $9\bar{n}/5$ and the other half have size $\bar{n}/5$. For each simulation condition, we evaluated the standardized bias and relative *SE* bias for each effect size estimator using 2,000 replications.

The simulation results for the 3T1C design are shown in Figure 4 (for standardized biases) and Figure 5 (for relative *SE* biases). We evaluated d_W and $\hat{\delta}_W$ only for the conditions with $\upsilon = 1$. For those conditions, the standardized biases were all smaller than .10 in absolute values, and the relative *SE* biases were all smaller than .05 in absolute values. Therefore, when the homogeneity assumption was met, the performance of d_W and $\hat{\delta}_W$ and their variance estimators was very good even with small sample sizes, unbalanced cluster sizes, and high ICC. We evaluated d_C and $\hat{\delta}_C$ with respect to all three values of υ. For all conditions, the standardized biases were smaller than the cut-off of .40, but for four conditions (with $\bar{n} = 5$, $\rho_3 = .1$, $\delta_C = 0.8$, $N^T{:}N^C = 5$, and $\upsilon = .5$ or 1) both d_C and $\hat{\delta}_C$ were slightly more biased, with standardized biases between .20 to and .21. Similarly, the *SE* estimates for d_C and $\hat{\delta}_C$ were unbiased for all conditions with relative *SE* bias $< .10$ in absolute values, except for six conditions with $\bar{n} = 5$, $\rho_3 = .5$, $N^T{:}N^C = 5$, $\upsilon = 2$ where the relative *SE* biases for d_C were between 12.3% to 13.2%.

Illustration Using Simulated Data

Sterba (2015) provided a simulated data example to demonstrate the estimation of various partially nested design. Based on the results of Table II (p. 2) of that paper, we simulated a data set with normally distributed random effects and level-1 error term for a three-level treatment arm and a one-level control arm design. Our hypothetical research question was to estimate the treatment effect of therapy groups in reducing depression. In the treatment arm, nine therapists each lead five therapy groups, each consisting of five patients. Therefore, there were $9 \times 5 \times 5 = 225$ patients in the treatment arm. The control arm consisted of 45 patients on the wait-list, who were unclustered. Using the `lme4` package in R, we estimated that the unstandardized treatment effect was $\hat{\gamma}_{100} = -1.788$, $SE = 0.252$. If the homogeneity assumption was assumed, the estimated pooled level-1 variance component was $\hat{\sigma}^2_W = 1.466$ ($SE = 0.139$), $\hat{\sigma}^2_2 = 0.188$, $\hat{\sigma}^2_3 = 0.148$. Using maximum likelihood, we can obtain $\hat{\delta}_W = -1.477$, $SE = 0.219$, and 95% CI $[-1.91, -1.05]$. Therefore, we can be 95% confident that the group-based treatment reduced the depression outcome by 1.05 to 1.91 standard deviations compared to the wait-list patients.

If only summary statistics are available with a mean difference between the treatment arm and the control arm of -1.788, the pooled level-1 *SD* of 1.211, the estimated ICCs of .105 at level 3 (i.e., therapists) and .084 at level 2 (i.e., therapy groups), we can use equations (21) and (22) to obtain $d_W = -1.477$, $SE = 0.220$. Because of the equal cluster sizes at both level 1 and level 2, the two estimators d_W and $\hat{\delta}_W$ are theoretically identical except for some samples where the estimated ICCs using summary statistics are negative but the maximum likelihood estimates of the ICCs are constrained to 0. If the formulas for single-level effect size were used, we would get $SE_{\text{naive}}(d) = \sqrt{1/225 + 1/45 + (-1.477)^2/[2(225 + 45 - 2)]} = 0.175$, an underestimation of the *SE* by slightly more than 25%.

On the other hand, if the homogeneity assumption was relaxed, the unstandardized treatment effect was $\hat{\gamma}_{100} = -1.788$, $SE = 0.254$. The estimated variance components of the treatment arm were: $\hat{\sigma}^2_{W|T} = 1.455$, $\hat{\sigma}^2_2 = 0.188$, $\hat{\sigma}^2_3 = 0.150$. For the control arm, $\hat{\sigma}^2_C = 1.513$ ($SE = 0.262$). Using maximum likelihood, we obtained $\hat{\delta}_C = -1.454$, $SE = 0.258$, and 95% CI

$[-1.96, -0.95]$. Therefore, we can be 95% confident that the group-based treatment reduced the depression outcome by 0.95 to 1.96 standard deviations compared to the wait-list patients. Very similar results for d_C may be obtained using summary statistics and equations (23) and (24). Although the magnitudes of $\hat{\delta}_W$ and $\hat{\delta}_C$ are similar, the variability of $\hat{\delta}_C$ ($SE = 0.258$) is larger than $\hat{\delta}_W$ ($SE = 0.219$) because $\hat{\delta}_C$ only uses the control arm *SD* and, in our example, there were only 45 observations in the control arm.

Conclusion

Despite a movement in the field of social and behavioral research towards effect size reporting in the past two decades, changes have been relatively slow for multilevel studies. It is also extremely rare to see researchers doing primary research reporting *SE* or confidence intervals for effect size estimates in multi- or even single-level studies (Peng et al., 2013). A major reason for the scarcity of adequate effect size reporting involves the great variations in multilevel designs; that is, the data may be clustered in different ways with different numbers of levels across different groups. Despite recent efforts to supplement the literature with methods for estimating effect size for the most commonly used designs, more attention needs to be paid to other variations in multilevel data and to evaluating the performances of existing methods.

The contributions of the present paper are fourfold. First, the study addressed the limitations of Hedges and Citkowicz's (2015) work on effect size estimation for two-level partially nested designs by relaxing the homogeneity of variance assumption. As it is a strong assumption that the treatment would not alter the within-cluster variability, it is important that applied researchers receive guidance on computing effect size using only the control arm *SD*. Second, the present paper discussed an estimation approach using maximum likelihood that is easier and more accurate with primary data, which, in turn, reduces the burden for applied researchers to report effect size estimates and the corresponding confidence intervals. This study helps researchers working with such a design to quantify and understand the practical significance of their results in addition to relying only on statistical significance tests, as studies with large sample size the point and interval estimates of d are much more informative than simply stating the p-value. Our discussion also provides tools for meta-analysts synthesizing effects of group interventions. Third, using simulation results, the present paper is the first to show that all four effect size estimators discussed—d_W, $\hat{\delta}_W$, d_C, and $\hat{\delta}_C$—performed well in most of the simulation conditions, and $\hat{\delta}_W$ and $\hat{\delta}_C$ produced more precise results for researchers doing primary research. Thus, our results added to the previous simulation results regarding model parameter estimates (e.g., Baldwin et al., 2011) and analytical results on effect size (Hedges & Citkowicz, 2015) when partially nested data are available. Finally, to the best of our knowledge the present study is the first to consider effect size for a three-level treatment one-level control design.

Based on our simulation results, when the required input is available, the newly proposed maximum likelihood estimators $\hat{\delta}_W$ and $\hat{\delta}_C$ is to be preferred over d_W and d_C as $\hat{\delta}_W$ and $\hat{\delta}_C$ yielded more precise effect size estimates with a smaller standard error, especially under conditions with unbalanced cluster sizes and large *design effect* (i.e., a large average cluster size together with a large ICC; Kish, 1965). On the other hand, when the maximum likelihood estimators cannot be obtained, a situation that is common when conducting meta-analyses, d_W and d_C are still viable alternatives as they are efficient with balanced cluster sizes. That is, although they are less efficient than the maximum likelihood estimators under conditions with unbalanced cluster sizes and large design effect, they still provide appropriate sampling variance estimates for the estimated effect size. Moreover, it is important to avoid applying the formulas for single-level effect size to multilevel data as that can result in severely underestimated sampling variance, thus incorrectly inflating the contribution of the effect size estimates from the multilevel studies towards the synthesized effect size.

Our simulation results also suggested that the efficiency loss as a result of using only the control arm *SD* instead of the pooled *SD* was relatively small except for conditions with high treatment-to-control sample size ratio, small level-1 and level-2 sample sizes, and a large effect size. When there were at least 30 clusters and an average of 25 participants for each cluster, the efficiency loss was negligible. The benefit of using only the control arm *SD* to estimate effect size is that it does not make the assumption of equal variances across treatment arms. Therefore, we recommend the use of d_C and $\hat{\delta}_C$ with $m \geq 30$ and $\bar{n} \geq 25$ unless one has a theoretical justification and is very confident that the variances are equal across treatment arms.

Despite the contributions of this discussion, a few limitations must be noted. First, to calculate d and $\hat{\delta}$, and particularly their variances, can be tedious. As a result, substantive researchers may prefer more automated procedures. Thus, it is recommended that future study investigate other methods such as bootstrapping (e.g., Goldstein, 2011). Second, we assumed that the control arm standard deviation is the preferred metric for standardization, which is consistent with how effect sizes are developed in single-level studies (e.g., Grissom & Kim, 2012). However, as noted in Hedges and Citkowicz (2015), for multilevel data and in certain situations the clustered treatment arm may be considered more natural. If the total *SD* for the clustered arm is to be used to define effect size, one can still easily obtain sample effect size estimates by changing the denominator, but the expression for the sampling variance is more complex and needs further investigation.

Third, we only considered designs with an unclustered control arm. As discussed in Heo et al. (2014), Lohr et al. (2014), and Sterba (2015), a partially nested design with a three-level treatment arm and a two-level control arm is also common, and the fact that the control arm is clustered brings additional complexity to defining standardized effect size and deriving the sampling variance of effect size estimator. It is recommended a follow up study be devoted to discussing ways to estimate effect size with this design. Fourth, the simulation results in this study apply only to simple situations with two arms and no covariates. Therefore, it is recommended that impact of additional complexity on effect size estimation be further addressed in the future. Finally, although from our simulation results the performance of $\hat{\delta}_W$ and $\hat{\delta}_C$ was acceptable even for small numbers of clusters and cluster sizes with an unbalanced design, under such conditions maximum likelihood can produce standard errors for fixed and the random effects that are too small (e.g., Maas & Hox, 2004), so researchers may consider using some version of corrected standard errors for small sample size (e.g., Kenward & Roger, 1997) when computing effect size estimates. Future research is needed to demonstrate whether the use of corrected standard errors leads to substantial improvement in the point and variance estimates of effect size for partially nested designs.

Footnotes

[1] Note that the estimator $d_{W\mathrm{HC}}$ in Hedges and Citkowicz (2015) is different from the d_W in our discussion here. Their $d_{W\mathrm{HC}}$ is defined using only the *SD* of the control arm with the assumption of homogeneity of variance. This is conceptually similar to our d_C with the difference that our d_C is still consistent with heterogeneous variance. In summary, $d_{W\mathrm{HC}}$ is less efficient than our d_W when the homogeneity of variance assumption holds, and $d_{W\mathrm{HC}}$ is inconsistent when the homogeneity of variance assumption is violated.

[2] For single-level studies it is common to use d for the effect size estimator using the pooled standard deviation, and use Δ for the estimator using the control arm standard deviation (Grissom & Kim, 2012). In this paper, however, we use the superscript C for estimators using the control arm standard deviation as it can apply both to estimation using summary statistics and that using maximum likelihood

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40. http://dx.doi.org/10.3102/0013189X035006033

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Baldwin, S. A., Bauer, D. J., Stice, E., & Rohde, P. (2011). Evaluating models for partially clustered designs. *Psychological Methods*, *16*, 149–165. http://dx.doi.org/10.1037/a0023464

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. http://dx.doi.org/10.18637/jss.v067.i01

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. http://dx.doi.org/10.1080/00273170802034810

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Casella, G., & Berger, R. L. (2002). *Statistical inference* (2nd ed.). Pacific Grove, CA: Thomson Learning.

Cho, H., Hallfors, D. D., & Sánchez, V. (2005). Evaluation of a high school peer group intervention for at-risk youth. *Journal of Abnormal Child Psychology*, *33*, 363–374. http://dx.doi.org/10.1007/s10802-005-3574-4

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Collins, L. M., Schafer, J. L., & Kam, C.-M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, *6*, 330–351. http://dx.doi.org/1082-989X.6.4.330

Compas, B. E., Forehand, R., Keller, G., Champion, J. E., Rakow, A., Reeslund, K. L., . . . Cole,

D. A. (2009). Randomized controlled trial of a family cognitive-behavioral preventive intervention for children of depressed parents. *Journal of Consulting and Clinical Psychology*, *77*, 1007–1020. http://dx.doi.org/10.1037/a0016930

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*, 7–29. http://dx.doi.org/10.1177/0956797613504966

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8. http://dx.doi.org/10.3102/0013189X005010003

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika*, 43–56. http://dx.doi.org/10.2307/2336270

Goldstein, H. (2011). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 163–171). New York, NY: Routledge.

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications* (2nd ed.). New York, NY: Routledge.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*, 107–128. http://dx.doi.org/10.3102/10769986006002107

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341–370. http://dx.doi.org/10.3102/1076998606298043

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*, 167–171. http://dx.doi.org/10.1111/j.1750-8606.2008.00060.x

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*, 346–380. http://dx.doi.org/10.3102/1076998610376617

Hedges, L. V., & Citkowicz, M. (2015). Estimating effect size when there is clustering in one treatment group. *Behavior Research Methods*, *47*, 1295–1308.

http://dx.doi.org/10.3758/s13428-014-0538-z

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. http://dx.doi.org/10.3102/0162373707299706

Heo, M., Litwin, A. H., Blackstock, O., Kim, N., & Arnsten, J. H. (2014). Sample size determinations for group-based randomized clinical trials with different levels of data hierarchy between experimental and control arms. *Statistical Methods in Medical Research*, *72*, 181–204. http://dx.doi.org/10.1177/0962280214547381

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367. http://dx.doi.org/10.1177/0049124198026003003

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York, NY: Routledge.

Kelley, K. (2013). Effect size and sample size planning. In T. D. Little (Ed.), *The Oxford handbook of quantitative methods: Vol. 1. foundations* (pp. 206–222). New York, NY: Oxford University.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*, 983–997. http://dx.doi.org/10.2307/2533558

Kirschner, F., Paas, F., Kirschner, P. A., & Janssen, J. (2011). Differential effects of problem-solving demands on individual and collaborative learning outcomes. *Learning and Instruction*, *21*, 587–599. http://dx.doi.org/10.1016/j.learninstruc.2011.01.001

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

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Korendijk, E. J. H. (2012). *Robustness and optimal design issues for cluster randomized trials* (Doctoral dissertation, Utrecht University, Utrecht, Netherland). Retrieved from `http://dspace.library.uu.nl/handle/1874/240965`

Lai, M. H. C., & Kwok, O.-m. (2014). Standardized mean differences in two-level cross-classified random effects models. *Journal of Educational and Behavioral Statistics*, *39*, 282–302. http://dx.doi.org/10.3102/1076998614532950

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

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: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from `http://ies.ed.gov/`

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137. http://dx.doi.org/10.1046/j.0039-0402.2003.00252.x

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through multilevel linear model. *Sociological Methodology*, *14*, 72–103. http://dx.doi.org/10.2307/270903

Moerbeek, M., & Wong, W. K. (2008). Sample size formulae for trials comparing group and individual treatments in a multilevel model. *Statistics in Medicine*, *27*, 2850–2864. http://dx.doi.org/10.1002/sim.3115

Myers, J. L., DiCecco, J. V., & Lorch, R. F., Jr. (1981). Group dynamics and individual performances: Pseudogroup and quasi-F analyses. *Journal of Personality and Social Psychology*, *40*, 86–98. http://dx.doi.org/10.1037//0022-3514.40.1.86

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*, 591–605. http://dx.doi.org/10.1111/j.1469-185X.2007.00027.x

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 analytic approaches. *American Journal of Public Health*, *98*(8), 1418–1424. http://dx.doi.org/10.2105/AJPH.2007.127027

Peng, C.-Y. J., Chen, L.-T., Chiang, H.-M., & Chiang, Y.-C. (2013). The impact of APA and AERA guidelines on effect size reporting. *Educational Psychology Review*, *25*, 157–209.

http://dx.doi.org/10.1007/s10648-013-9218-2

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology*, *48*, 85–112. http://dx.doi.org/10.1016/j.jsp.2009.09.002

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2015). nlme: Linear and nonlinear mixed effects models (R package version 3.1-122) [Computer software manual]. Retrieved from http://CRAN.R-project.org/package=nlme

R Core Team. (2015). R: A language and environment for statistical computing (Version 3.2.2) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Sanders, E. A. (2011). *Multilevel analysis methods for partially nested cluster randomized trials* (Doctoral dissertation). Available from ProQuest dissertations and theses dababase. (UMI No. 3452760).

Searle, S. R., Casella, G., & McCulloch, C. E. (2006). *Variance components* (2nd ed.). Hoboken, NJ: Wiley.

Sterba, S. K. (2015). Partially nested designs in psychotherapy trials: A review of modeling developments. *Psychotherapy Research*. Advance online publication. http://dx.doi.org/10.1080/10503307.2015.1114688

Sterba, S. K., Preacher, K. J., Forehand, R., Hardcastle, E. J., Cole, D. A., & Compas, B. E. (2014). Structural equation modeling approaches for analyzing partially nested data. *Multivariate Behavioral Research*, *49*, 93–118. http://dx.doi.org/10.1080/00273171.2014.882253

Stice, E., Shaw, H., Burton, E., & Wade, E. (2006). Dissonance and healthy weight eating disorder prevention programs: A randomized efficacy trial. *Journal of Consulting and Clinical Psychology*, *74*, 263–275. http://dx.doi.org/10.1037/0022-006X.74.2.263

Talley, A. E. (2013). *The impact of nonnormal and heteroscedastic level one residuals in partially clustered data* (Master's thesis). Retrieved from https://tdl-ir.tdl.org/tdl-ir/handle/2152/22630

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *3*, 25–32. http://dx.doi.org/10.3102/0013189X031003025

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*(4), 425–433. http://dx.doi.org/10.1037/1082-989X.5.4.425

Wehry, S., & Algina, J. (2003). Type I error rates of four methods for analyzing data collected in a groups vs individuals design. *Journal of Modern Applied Statistical Methods*, 400–413. Retrieved from http://digitalcommons.wayne.edu/jmasm/vol2/iss2/13/

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604. http://dx.doi.org/10.1037/0003-066X.54.8.594

Table 1

Mean Relative Efficiency of Effect Size Estimation Using Control Arm SD Relative to Using Pooled SD Under Homogeneity of Variance

N^T:N^C	m	$\bar{n}$	δ	$RE(d_C, d)$
1	10	5	0.2	0.98
			0.5	0.97
			0.8	0.95
		25	0.2	1.00
			0.5	0.99
			0.8	0.98
	30	5	0.2	0.99
			0.5	0.98
			0.8	0.96
		25	0.2	1.00
			0.5	0.99
			0.8	0.98
	100	5	0.2	1.00
			0.5	0.99
			0.8	0.96
		25	0.2	1.00
			0.5	1.00
			0.8	0.99
5	10	5	0.2	0.79
			0.5	0.74
			0.8	0.66
		25	0.2	0.95
			0.5	0.93
			0.8	0.89
	30	5	0.2	0.94
			0.5	0.88
			0.8	0.81
		25	0.2	0.99
			0.5	0.96
			0.8	0.91
	100	5	0.2	0.98
			0.5	0.93
			0.8	0.85
		25	0.2	0.99
			0.5	0.96
			0.8	0.93

Note. N^T = sample size of the treatment arm. N^C = sample size of the control arm. m = number of clusters in the treatment arm. $\bar{n}$ = average cluster size. δ = population effect size. $RE(d_C, d)$ = efficiency of effect size estimator using only control arm *SD* relative to that using the pooled *SD*.

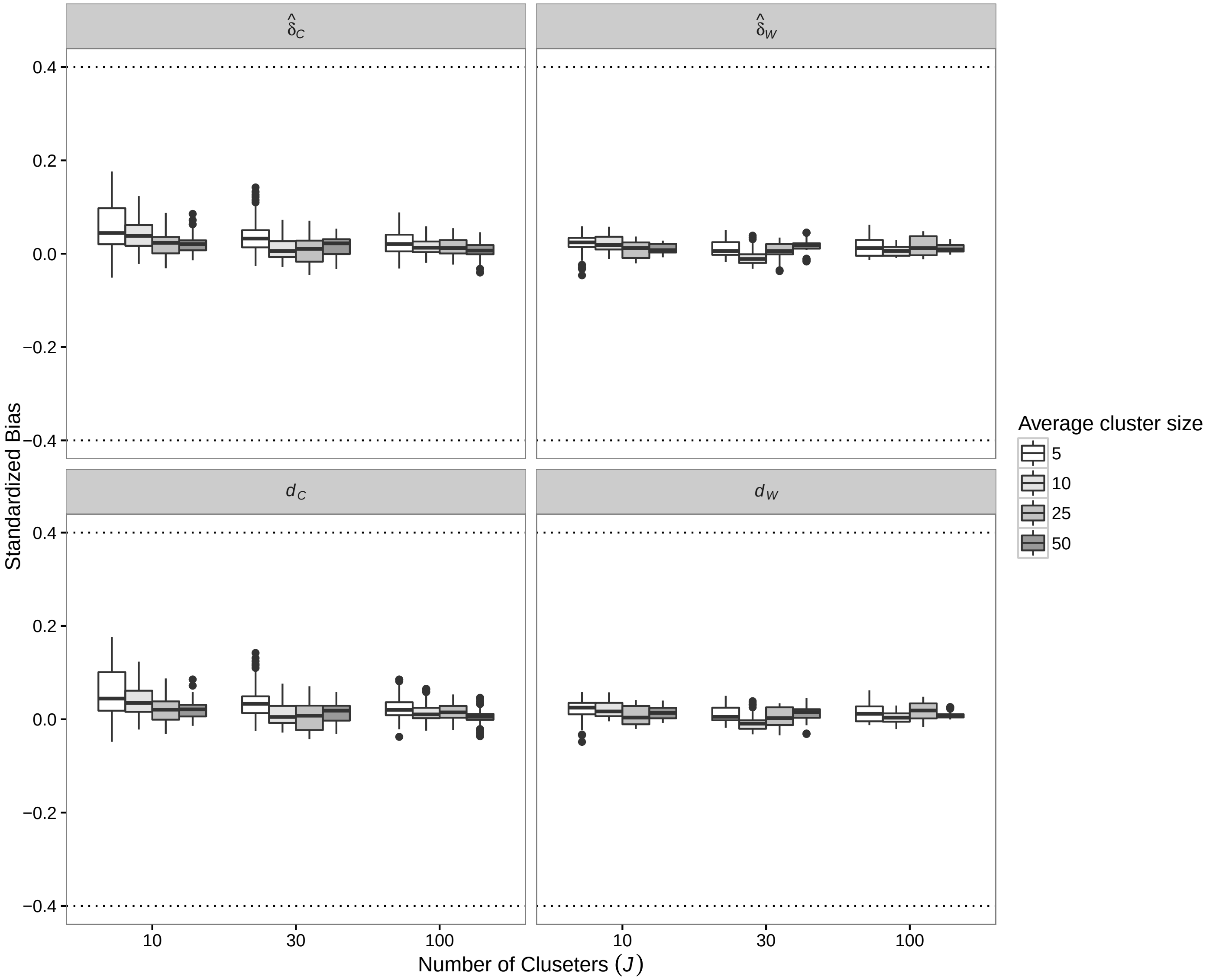

Figure 1. Distribution of standardized biases for two-level partially nested design. $\hat{\delta}_W$ = Effect size estimator using maximum likelihood with the pooled *SD*; $\hat{\delta}_C$ = Effect size estimator using maximum likelihood with the control arm *SD*; d_W = Effect size estimator using summary statistics with the pooled *SD*; d_C = Effect size estimator using summary statistics with the control arm *SD*. Results for $\hat{\delta}_W$ and d_W are only for conditions with homogeneous level-1 variances across the treatment and the control arms; those for $\hat{\delta}_C$ and d_C are for both homogeneous and heterogeneous variances.

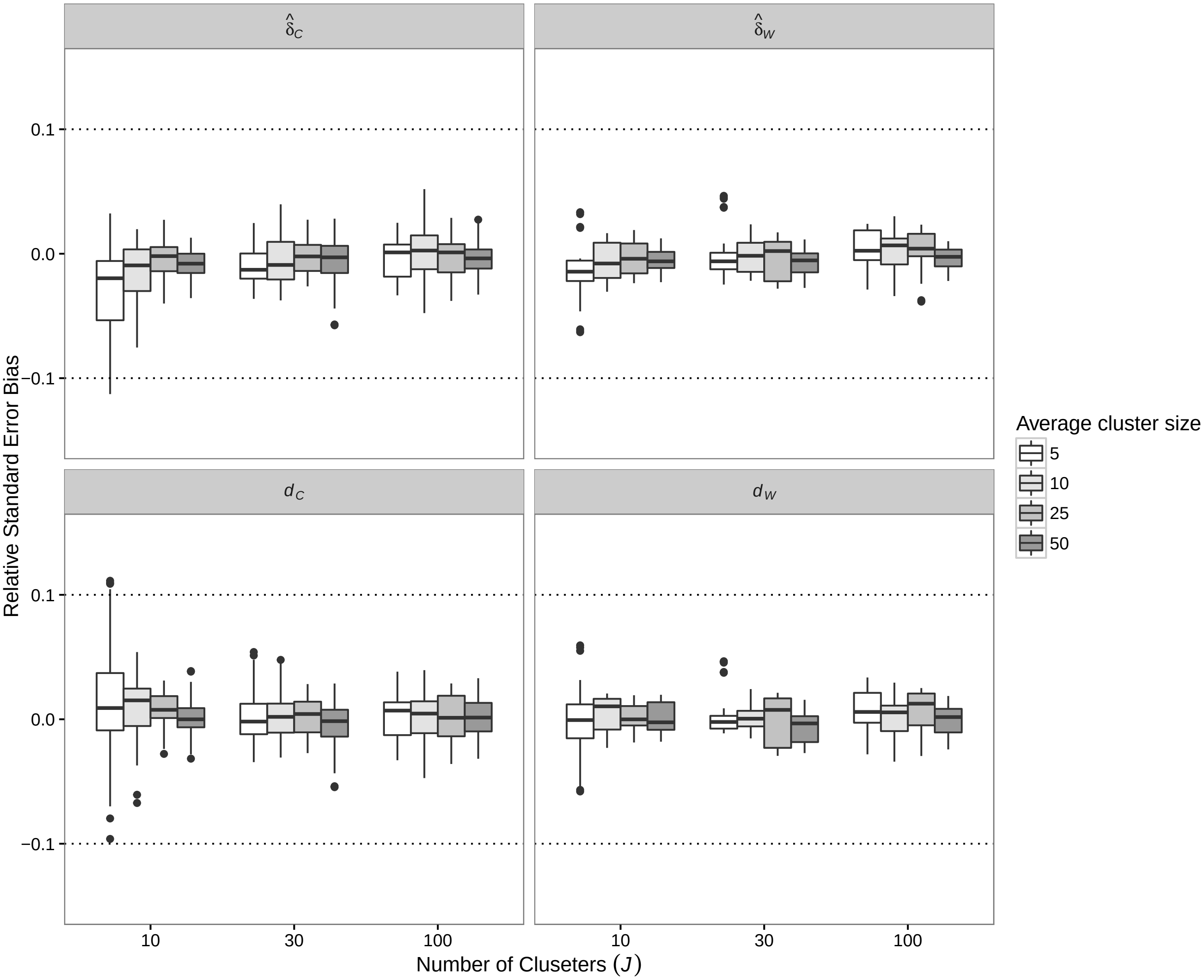

Figure 2. Distribution of relative *SE* biases for two-level partially nested design. $\hat{\delta}_W$ = Effect size estimator using maximum likelihood with the pooled *SD*; $\hat{\delta}_C$ = Effect size estimator using maximum likelihood with the control arm *SD*; d_W = Effect size estimator using summary statistics with the pooled *SD*; d_C = Effect size estimator using summary statistics with the control arm *SD*. Results for $\hat{\delta}_W$ and d_W are only for conditions with homogeneous level-1 variances across the treatment and the control arms; those for $\hat{\delta}_C$ and d_C are for both homogeneous and heterogeneous variances.

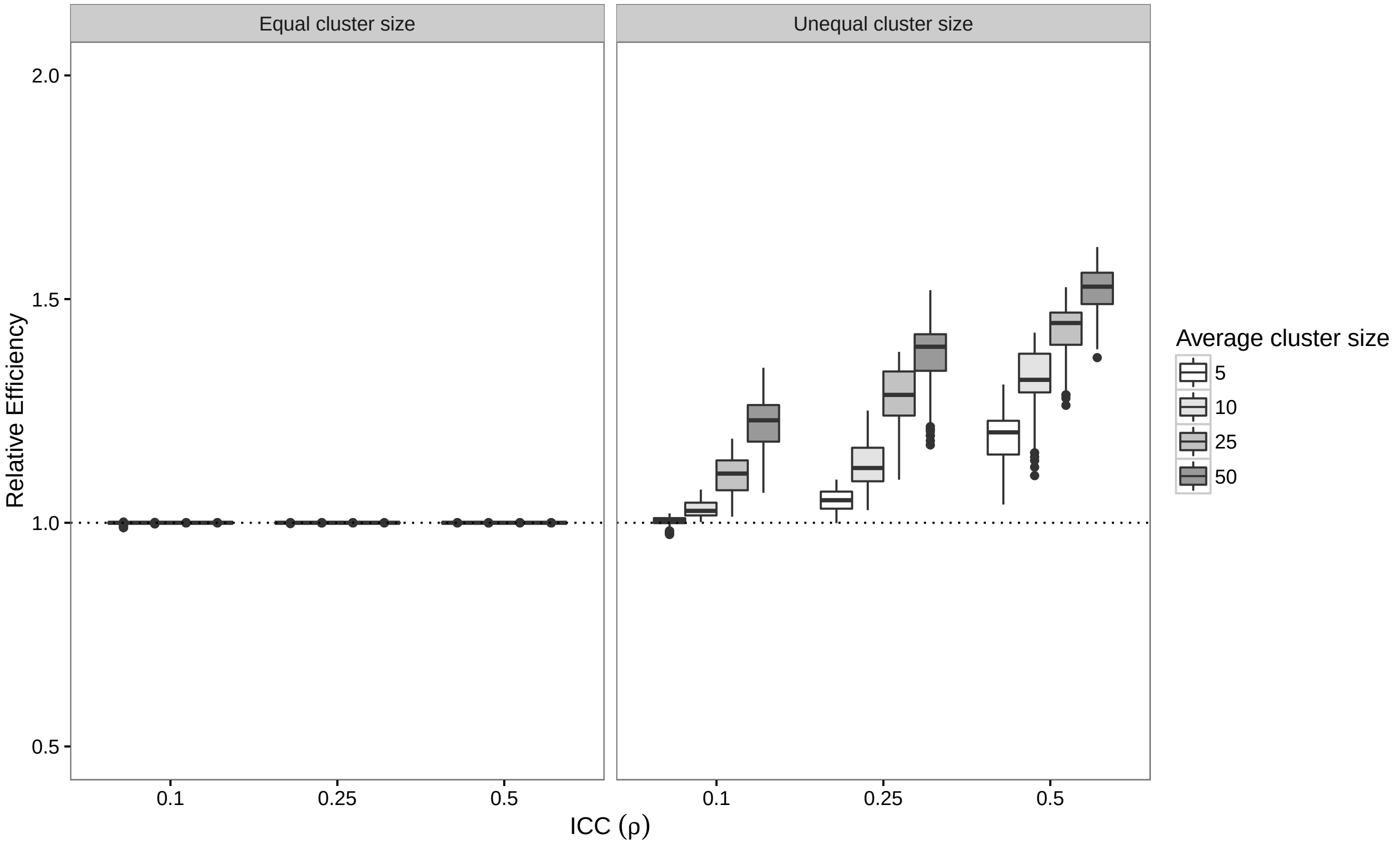

Figure 3. Efficiency of $\hat{\hat{\delta}}_W$ (and $\hat{\hat{\delta}}_C$) relative to d_W (and d_C). Relative efficiency was computed as the ratio between the mean squared errors for $\hat{\hat{\delta}}_W$ (and $\hat{\hat{\delta}}_C$) and for d_W (and d_C).

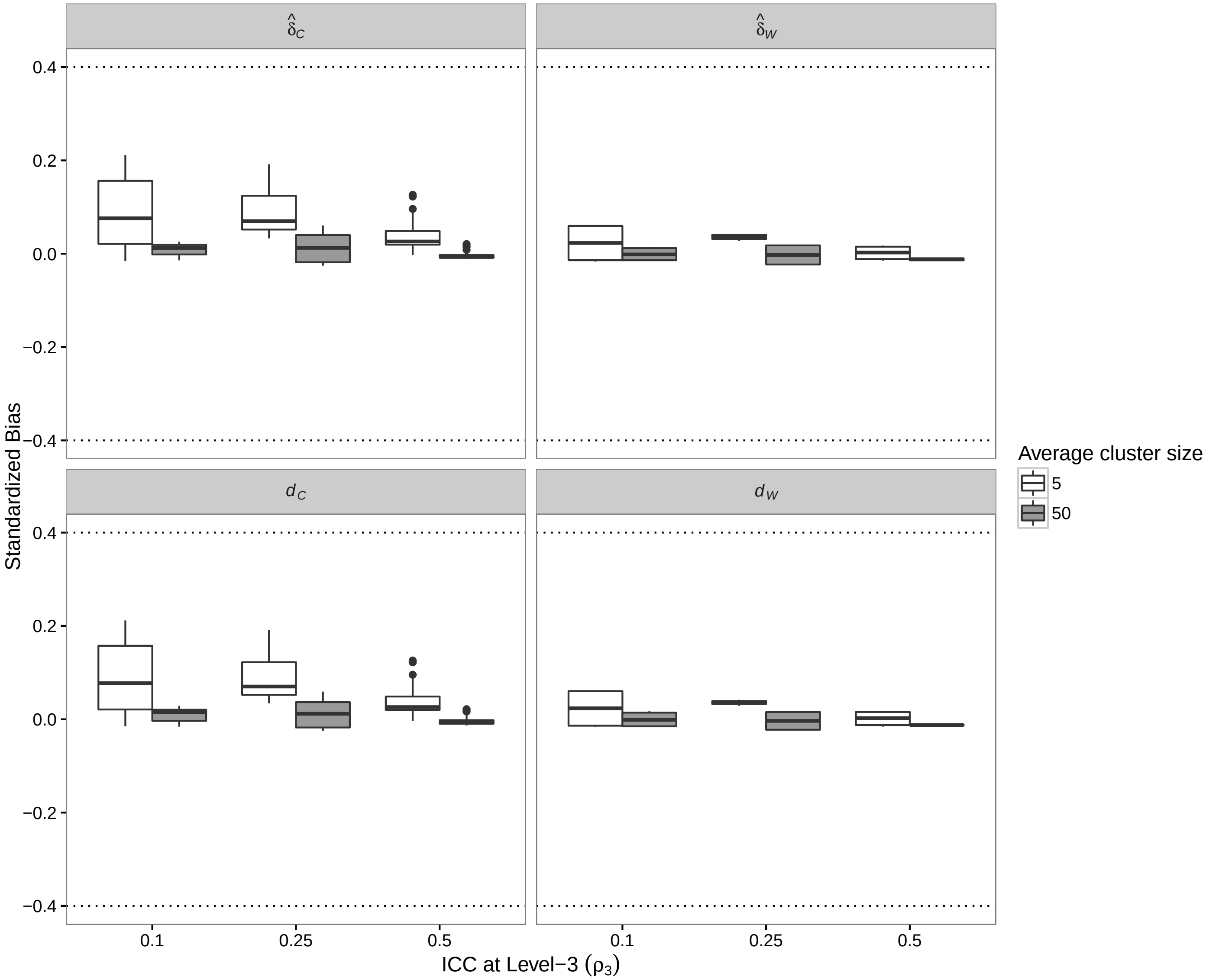

Figure 4. Distribution of standardized biases for three-level partially nested design. $\hat{\delta}_W$ = Effect size estimator using maximum likelihood with the pooled *SD*; $\hat{\delta}_C$ = Effect size estimator using maximum likelihood with the control arm *SD*; d_W = Effect size estimator using summary statistics with the pooled *SD*; d_C = Effect size estimator using summary statistics with the control arm *SD*. Results for $\hat{\delta}_W$ and d_W are only for conditions with homogeneous level-1 variances across the treatment and the control arms; those for $\hat{\delta}_C$ and d_C are for both homogeneous and heterogeneous variances.

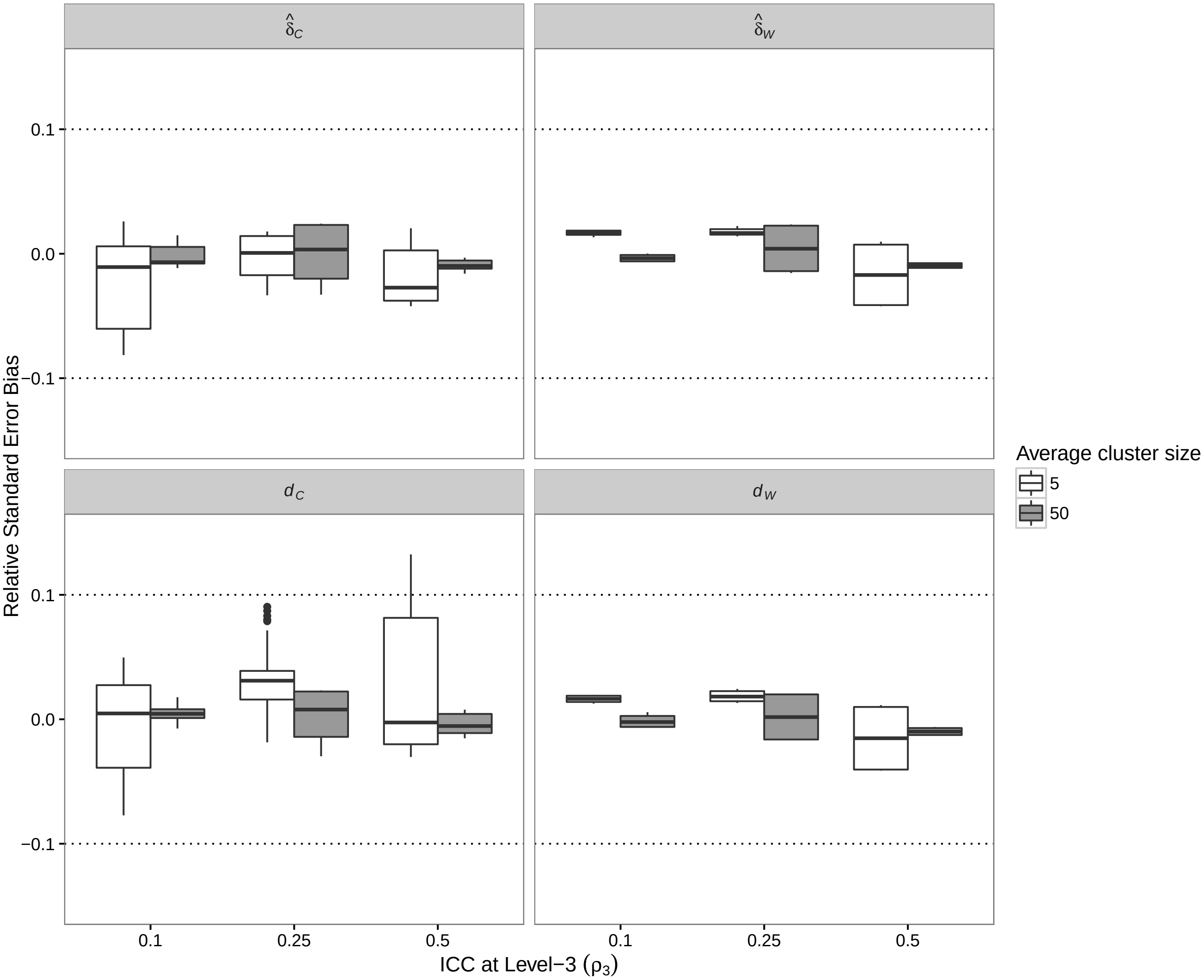

Figure 5. Distribution of relative *SE* biases for three-level partially nested design. $\hat{\delta}_W$ = Effect size estimator using maximum likelihood with the pooled *SD*; $\hat{\delta}_C$ = Effect size estimator using maximum likelihood with the control arm *SD*; d_W = Effect size estimator using summary statistics with the pooled *SD*; d_C = Effect size estimator using summary statistics with the control arm *SD*. Results for $\hat{\delta}_W$ and d_W are only for conditions with homogeneous level-1 variances across the treatment and the control arms; those for $\hat{\delta}_C$ and d_C are for both homogeneous and heterogeneous variances.

Appendix

Derivation of Effect Sizes for Partially Nested Designs

Theorem

The following steps to finding an unbiased estimator are based on the work of Hedges (2007, pp. 360–362). Consider a sample estimator of $\delta = \Delta\mu/\sigma$ in the form d = correction factor × $(\Delta\bar{Y}/S)$, where S is an estimator of σ with $\mathrm{E}(S^2) = b\sigma^2$ and $V(S^2) = 2c\sigma^4$. Further, let $\Delta\bar{Y} \sim N(\Delta\mu, a\sigma^2)$. It can then be shown that, by choosing $\sqrt{b}$ as the correction factor,

$$d = \sqrt{b}\frac{\Delta\bar{Y}}{S} = T\sqrt{a} \tag{A1}$$

would be approximately unbiased and consistent for δ. When S^2 has a scaled χ^2 distribution (which holds exactly with equal cluster size and approximately with unequal cluster size; see Searle et al., 2006), the random variable $T = d/\sqrt{a}$ has a noncentral t distribution with b^2/c degrees of freedom and a noncentral parameter of $\delta/\sqrt{a}$. From Hedges (1981), when $b^2/c \rightarrow \infty$, d would be normally distributed with mean δ and the asymptotic variance

$$V(d) = a + \frac{c\delta^2}{2b^2}. \tag{A2}$$

It is clear that as $b^2/c \rightarrow \infty$, $V(d)$ will be dominated by the first term a and be independent of δ. This is consistent with other large-sample results where the estimator converges to a normal distribution where the mean and the asymptotic variance are independent.

Our task is to express a, b, and c in terms of known quantities, and substitute them into equations (A1) and (A2), when δ is defined as $\Delta\mu/\sigma_W^2$.

Derivation of *d* for the Two-Level Partially Nested Design

In a balanced design where the cluster sizes in the treatment arm are equal, the sample grand mean is an unbiased and efficient estimator of the population mean in both the treatment arm and the control arm. First assume that $\sigma^2_{W|T} = \sigma^2_C = \sigma^2_W$, that is, $\upsilon = 1$. Denote $\bar{Y}^T_{\cdot\cdot}$ and $\bar{Y}^C$ as the grand means for the treatment arm and the control arm, with corresponding sampling variance

$$V(\bar{Y}^T_{..}) = \frac{\sigma^2_W + n\sigma^2_B}{N^T} = \sigma^2_W \frac{1 + n(1-\rho)}{N^T(1-\rho)}, \tag{A3}$$

$$V(\bar{Y}^C) = \frac{\sigma^2_W}{N^C}. \tag{A4}$$

The expression for $V(\bar{Y}^T_{..})$ follows from the definition of ICC such that $\rho = \sigma^2_B/(\sigma^2_B + \sigma^2_W)$. The treatment effect could be then estimated as

$$\Delta\bar{Y} = \bar{Y}^T_{..} - \bar{Y}^C, \tag{A5}$$

with sampling variance

$$V(\bar{Y}^T_{..} - \bar{Y}^C) = \sigma^2_W \left[\frac{1 + (n-1)\rho}{N^T(1-\rho)} + \frac{1}{N^C}\right]. \tag{A6}$$

The expectation and variance of the variance components would then be

$$\begin{aligned} \mathrm{E}(SS_{W|T}) &= (N^T - m)\sigma^2_W, \\ \mathrm{E}(SS_C) &= (N^C - 1)\sigma^2_W, \\ V(SS_{W|T}) &= 2(N^T - m)\sigma^4_W, \\ V(SS_C) &= 2(N^C - 1)\sigma^4_W. \end{aligned}$$

Because $S^2_W = (SS_{W|T} + SS_C)/(N^T - m + N^C - 1)$,

$$\mathrm{E}(S^2_W) = \frac{(N^T - m)\sigma^2_W + (N^C - 1)\sigma^2_W}{N^T - m + N^C - 1} = \sigma^2_W$$

and

$$V(S^2_W) = \frac{2(N^T - m)\sigma^4_W + 2(N^C - 1)\sigma^4_W}{(N^T - m + N^C - 1)^2} = \frac{2\sigma^4_W}{N - m - 1}.$$

Hence

$$\begin{aligned} a &= \frac{1+(n-1)\rho}{N^T(1-\rho)} + \frac{1}{N^C}, \\ b &= 1, \\ c &= \frac{1}{N-m-1}. \end{aligned}$$

We can now substitute a, b, and c into (A1) and (A2) to get the expressions in the main text.

Unbalanced cluster sizes. When the cluster sizes $n_1, \ldots, n_m$ are not equal, $V(\bar{Y}^T_{\cdot\cdot})$ is not just a function of $\tilde{n}$. First note that $\bar{Y}^T_{\cdot\cdot} = \sum_{j=1}^m \sum_{i=1}^{n_j} Y^T_{ij}/N^T$. Under our model and for a specific cluster j, $V(Y^T_{ij}) = \sigma^2_{W|T} + \sigma^2_B$ and $\text{Cov}(Y^T_{ij}, Y^T_{ij'}) = \sigma^2_B$ for $j \neq j'$, so $V(\sum_{i=1}^{n_j} Y^T_{ij}) = n_j\sigma^2_{W|T} + n_j^2\sigma^2_B$. As the clusters are assumed to be independent of each other, $V(\bar{Y}^T_{\cdot\cdot}) = (\sum_{j=1}^m n_j\sigma^2_{W|T} + \sum_{j=1}^m n_j^2\sigma^2_B)/(N^T)^2 = \sigma^2_{W|T}/N^T + \tilde{n}\sigma^2_B$. Therefore, the formulas for d and $V(d)$ would work by replacing n with $\tilde{n}$.

Using the *SD* of the control arm to compute d_C. If only the standard deviation of the control arm is used, and the homoscedasticity assumption is not made,

$$V(\bar{Y}^T_{\cdot\cdot}) = \upsilon\sigma^2_C \frac{1+n(1-\rho)}{N^T(1-\rho)}$$

and $c = 1/(N^C - 1)$, resulting in the formula for $V(d_C)$ in equation (16), and for unbalanced cluster sizes one can replace n by $\tilde{n}$.

Derivation of *d* for the Three-Level/One-Level Partially Nested Design

In a balanced design with equal cluster sizes at level 1 and at level 2 in the treatment arm, and assuming that $\sigma^2_{W|T} = \sigma^2_C = \sigma^2_W$ (i.e., $\upsilon = 1$), it can be shown that the sampling variance for the grand means are

$$V(\bar{Y}^T_{\ldots}) = \frac{\sigma^2_W + n\sigma^2_2 + n_2\sigma^2_3}{N^T} = \sigma^2_W \left[\frac{1 + n\rho_2 + n_2\rho_3}{N^T(1 - \rho_2 - \rho_3)}\right], \tag{A7}$$

$$V(\bar{Y}^C) = \frac{\sigma^2_W}{N^C}. \tag{A8}$$

It follows then

$$V(\bar{Y}^T_{\ldots} - \bar{Y}^C) = \sigma^2_W \left[\frac{1 + n\rho_2 + n_2\rho_3}{N^T(1 - \rho_2 - \rho_3)} + \frac{1}{N^C}\right], \tag{A9}$$

so

$$a = \frac{1 + n\rho_2 + n_2\rho_3}{N^T(1 - \rho_2 - \rho_3)} + \frac{1}{N^C}. \tag{A10}$$

It is also clear that the pooled level-1 variance is an unbiased estimator of σ^2_W and the degrees of freedom is $N - P - 1$ ($df = N^T - P - 1$ for the treatment arm; $df = N^C - 1$ for the control arm). Referring to the definition of the constants b and c in the theorem in the beginning of this Appendix, for the three-level/one-level design one obtains

$$b = 1,$$
$$c = \frac{1}{N - P - 1}.$$

We can now substitute a, b, and c into (A1) and (A2) to get the expressions in the main text.

Unbalanced cluster sizes. Without the assumption of a balanced design and under the model in equation (20), for the treatment arm one has

$$\mathrm{Cov}(y_{ijk}, y_{i'j'k'}) = \begin{cases} \mathrm{Var}(y_{ijk}) = \sigma^2_{W|T} + \sigma^2_2 + \sigma^2_3, & i = i', j = j', k = k' \\ \sigma^2_2 + \sigma^2_3, & i \neq i', j = j', k = k' \\ \sigma^2_3, & i \neq i', j \neq j', k = k' \\ 0, & i \neq i', j \neq j', k \neq k' \end{cases}.$$

Therefore, the unweighted sum of all the y values in the treatment arm, $Y^T_{...} = \sum_{k=1}^{m} \sum_{j=1}^{p_k} \sum_{i=1}^{n_{jk}} Y^T_{ijk}$, has a sampling variance

$$V(Y^T_{...}) = N^T \sigma^2_{W|T} + \sum_{k=1}^{m} \sum_{j=1}^{p_k} n^2_{jk}\sigma^2_2 + \sum_{k=1}^{m} \left(\sum_{j=1}^{p_k} n_{jk} \right)^2 \sigma^2_3. \tag{A11}$$

Dividing both sides by $(N^T)^2$, one gets $V(\bar{Y}^T_{...}) = \sigma^2_{W|T}/N^T + \tilde{n}\sigma^2_2 + \tilde{n}_2\sigma^2_3$. With b and c unchanged, the formulas for d and $V(d)$ for unbalanced designs can be obtained by replacing n with $\tilde{n}$ and n_2 with $\tilde{n}_2$.

Using the *SD* of the control arm to compute d_C. If only the standard deviation of the control arm is used, and the homoscedasticity assumption is not made, then

$$V(\bar{Y}^T_{..}) = \upsilon \frac{1 + (n-1)\rho_2 + (n_2 - 1)\rho_3}{N^T(1 - \rho_2 - \rho_3)}$$

and $c = 1/(N^C - 1)$, resulting in the formula for $V(d_C)$ in equation (24). For unbalanced cluster sizes one can replace n by $\tilde{n}$ and n_2 by $\tilde{n}_2$.

Derivation of $\hat{\delta}$ for Partially Nested Designs

In order to calculate $\hat{\delta}$ for partially nested data, we assume that estimates of the fixed effect $\hat{\gamma}_{01}$, of the within-level variance component $\hat{\sigma}^2_W$, as well as of their corresponding variance $V(\hat{\gamma}_{01})$ and $V(\hat{\sigma}^2_W)$ are available. Using the same framework for deriving the distribution of d, we have $\hat{\gamma}_{10} \sim N(\Delta\mu, a\sigma^2_W)$ and thus $V(\hat{\gamma}_{10}) = a\sigma^2_W$, so

$$a = \frac{V(\hat{\gamma}^2_{10})}{\hat{\sigma}^2_W}.$$

Based on the theorem in the beginning of this Appendix, if we use $S^2 = \hat{\sigma}^2_W$ as an estimator for σ^2_W, we get $E(\hat{\sigma}^2_W) = b\sigma^2_W$. Assuming that $\hat{\sigma}^2_W$ is approximately unbiased, that is, $\mathrm{E}(\hat{\sigma}^2_W) = \sigma^2_W$, we have

$$b = 1.$$

The theorem also states that $V(S^2) = 2c\sigma^4$, and in our case, if we replace S^2 by $\hat{\sigma}_W^2$ and σ by $\hat{\sigma}_W$, we get

$$c = \frac{V(\hat{\sigma}_W^2)}{2(\hat{\sigma}_W^2)^2}.$$

Then by substituting a, b, and c into (A1) and (A2), one get the expressions for $\hat{\delta}$ and $V(\hat{\delta})$. For the estimation of $\hat{\delta}_C$, one can simply replace the point and variance estimates of $\hat{\sigma}_W^2$ by those of $\hat{\sigma}_C^2$