STANDARDIZED MEAN DIFFERENCES FOR COMPLEX MULTILEVEL MODELS: PARAMETRIC AND NONPARAMETRIC ESTIMATION

A Dissertation

by

HOK CHIO LAI

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Oi-man Kwok
Co-Chair of Committee,	Myeongsun Yoon
Committee Members,	Dudley L. Poston Jr.
	Victor L. Willson
Head of Department,	Victor L. Willson

August 2015

Major Subject: Educational Psychology

ABSTRACT

This dissertation comprises three separate but interrelated manuscripts exploring methods for estimating the standardized mean difference effect size with several complex multilevel data structures. Multilevel modeling techniques are becoming more popular in handling data with multilevel structure in educational and behavioral research. However, unlike traditional single level research, methodological studies about multilevel effect size have been rare and those that have recently appeared had an emphasis on strictly hierarchical data structure.

In the first manuscript, I propose two methods for obtaining effect size in the two-level fully and partially cross-classified random effects models. Fully cross-classified data structure arises when individual observations are clustered by several levels that did not have a strictly hierarchical structure. For example, students may be classified by both their middle school and high school, but neither middle school is nested within high school nor vice versa. Partially cross-classified structure is a structure with an existing clustering in both the treatment and the control condition, but with the addition of an artificial clustering level only present in the treatment condition. The study will include derivation of the formulas, verification of their performances with Monte Carlo simulation, and illustration of their use with real data examples.

The second manuscript discusses two similar methods for obtaining effect size with two-level partially nested data. Partially nested data arises in randomized trials where the intervention creates artificial clustering, but no such clustering is present in the comparison group. In this manuscript I will present derivation of the formulas for the two methods, verify their performances with simulated data, illustrate their use with a real data example, and discuss the impact of failing to honor the partially nested structure on

effect size estimates.

The third manuscript explores the use of the bootstrap to estimate multilevel standardized mean difference. I will discuss various bootstrap methods, both parametric and nonparametric, to obtain effect size estimates for two-level studies. Their performances will be compared with analytical methods under conditions of excessive skewness and kurtosis in level-1 and level-2 random effects and varying design features.

ACKNOWLEDGEMENTS

I would like to thank the members of my committee, Dr. Kwok, Dr. Yoon, Dr. Willson, and Dr. Poston, for their guidance and support throughout the course of this dissertation. I especially owe a debt to Dr. Kwok, for all his time and mentorship that created the ideal environment for me to start my academic career.

I would also like to thank Dr. Deborah Simmons for kindly sharing with me the data for the Early Reading Intervention Project (supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324E060067 to Texas A&M University; the opinions expressed are those of the authors and do not represent views of the U.S. Department of Education) for the empirical example in the first manuscript.

Finally, I would like to express my gratitude and love to Grace Mak, who has accompanied me and provided me unconditional support throughout the long, long four years.

TABLE OF CONTENTS

Page

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

CHAPTER I

INTRODUCTION

In the past few decades there have been two important trends for quantitative research in the behavioral sciences. One is the increasing popularity of multilevel models (Goldstein, 2011b; Hox, 2010), which is synonymously called variance component models (Searle, Casella, & McCulloch, 2006), hierarchical linear models (Raudenbush & Bryk, 2002), and linear mixed modeling (Littell, Milliken, Stroup, & Wolfinger, 1996). Traditional analyses such as multiple regression assume that the observations are independent, which roughly means that knowing one individual's score says nothing about another individual's score. For many situations in the social sciences, however, data are collected in clusters, with examples like students in classrooms and schools, employees in organizations, clients in treatment groups, and residents in countries. Because individuals in the same cluster share the same environment, they may be more similar to each other than to someone from another cluster. Therefore, knowing an individual's score gives some information about the score of another individual in the same cluster, and the assumption of independent observation is violated.

Multilevel models are developed to address the data dependency issue (Aitkin & Longford, 1986). They provide a flexible framework for specifying level-specific regression models (Raudenbush & Bryk, 2002), and for separating the effect of a lower-level predictor into the individual-level effect and the cluster-level effect. It also allows the effect of lower-level predictors to vary across clusters by positing a distribution of the regression slopes. This is a much more efficient way of modeling the slopes than fitting separate regression models for each cluster. With continuing improvement in algorithms for estimation (e.g., Bates, 2010; Goldstein, 1986; Longford, 1987) and in

usability of computer programs (e.g., SPSS `MIXED`, SAS `PROC MIXED`, HLM, and MLwiN), multilevel modeling has already become part of the standard training for behavioral researchers.

The second important trend that has revolutionized quantitative research in the behavioral sciences is what was called the "effect size movement" (Robinson, Whittaker, Williams, & Beretvas, 2003). In the past two to three decades, many authors and editors have discussed the problems associated with significance testing (Carver, 1978; Cohen, 1994; Harlow, Mulaik, & Steiger, 1997; Kline, 2013; Schmidt, 1996). One of the major concerns was that *p*-value was frequently mistreated as an indicator of how strong or "significant" the result is. However, because the *p*-value is usually a function of the sample size (Thompson, 1996), a negligible effect can be "highly significant" with a large sample while a substantial effect may be "non-significant" just because the sample size is not large enough.

Recognizing such weakness in significance testing, some authors have proposed the use of effect size as a mean to quantify an effect of interest (Grissom & Kim, 2012; Kirk, 1996; Snyder & Lawson, 1993; Wilkinson & Task Force on Statistical Inference, 1999). Many journals have then made effect size reporting mandatory (Huberty, 2002). Similarly, several professional associations, which includes the *American Educational Research Association* (AERA, 2006), the *American Psychological Association* (APA, 2010), and the *National Center for Education Statistics* (NCES, 2012), have gradually made effect size reporting almost a necessary step in reporting the results.

The two most commonly used families of effect size are strength of association and group difference (Grissom & Kim, 2012; Kirk, 1996; Rosenthal, 1994). In a review of 32 review papers on effect size reporting practices, Peng, Chen, Chiang, and Chiang (2013) found most of them concluding that the unadjusted proportion of variance accounted for, or R^2, and the standardized mean difference (SMD), with Cohen's *d* as an example, are

the most commonly used effect size measures. The R^2 effect size measures the proportion of variance in the outcome variable explained by the predictor or set of predictors, and is more naturally associated with observational or correlational studies with continuous predictors. On the other hand, SMD expressed the difference in outcome scores in standard deviation (*SD*) units between two groups, and is more naturally associated with experimental or quasi-experimental studies where the intervention variable contains treatment arms. That being said, one should note that both R^2 and Cohen's d can be used for both observational and experimental studies, and at least for single-level studies formulas are available for converting R^2 to d or vice versa (Lipsey & Wilson, 2001).

The two trends, however, have not mixed well yet. On one hand, the analogue for R^2 in multilevel analyses had been developed for a while (Raudenbush & Bryk, 2002; Snijders & Bosker, 1994), mainly due to the nature of multilevel models being an extension of the conventional multiple linear regression. Still, as noted in Peugh (2010), "no consensus exists as to the effect sizes that are most appropriate" (p. 97, see also J. K. Roberts, Monaco, Stovall, & Foster, 2011). On the other hand, whereas cluster randomized trials that implement randomization and interventions at the group-level is quite popular in education and medical sciences, the analogue of Cohen's d in multilevel studies was not discussed until Hedges (2007). Given the complexity of multilevel models, the variability in multilevel data structures, and the importance of quantifying effects of interest, much more research efforts are needed to study multilevel effect sizes.

This dissertation comprised three manuscripts representing my research efforts on multilevel SMD. In the first manuscript, I apply the framework for deriving multilevel SMD from Hedges (2007) to the cross-classified data structure as well as its variant, the partially cross-classified data structure. As pointed out by Beretvas (2011), the two-level hierarchical structure represents only an idealized and unrealistic simplification of the real data structure. In education, for example, students are not clustered in only one way. The

data may be collected from students that are clustered by both their middle schools and high schools, or perhaps by both their schools and neighborhoods of living. In such cases there are two sources of clustering, but the two sources do not conform to a hierarchical relationship; Instead they are *crossed*. Analyses that appropriately model such structure has been developed and used more frequently in the past few years.

The partially cross-classified structure is a variation in which, for a sample of participants already clustered by one level, the intervention creates an extra level of clustering for only the treatment arm but not for the control arm. For instance, to examine the effect of emotion management group, a researcher may implement the intervention with groups of students from different classrooms (in order to reduce the probability of having close friends in a group), so students in the treatment arm is cross-classified by intervention groups and classrooms, whereas those in the control arm is only nested within classrooms. Another example would be an example where students in the treatment arm changes memberships of reading groups but those in the control arm does not, as presented first manuscript.

The main goal of the second manuscript is to derive SMD for the partially nested structure, where clustering occurs only in the treatment arm but not in the control arm. Such a data structure is probably less complicated than the cross-classified one, but it is perhaps the most common structure for cluster-randomized trials, as Bauer, Sterba, and Hallfors (2008) found that more studies used it than the two-level hierarchical structure where clustering occurs in both arms. It is not difficult to see why. In education, many interventions like reading groups and those that facilitate cooperative learning are group-based. Similarly, in psychology, there are treatment groups for addiction, family problems, and other psychological problems or developments. Typically no intervention is implemented for the control arm, resulting in data with a partially nested structure.

In the first two manuscripts, I propose two analytical methods for obtaining SMD

and its sampling variance (i.e., the squared value of the standard error, SE^2) for each design, verify their performances, and demonstrate their usage with real data.

As presented in Hedges (2007), Hedges (2011), and the first two manuscripts, the analytical formulas for obtaining SMD is long and complex. Although they are important as they outline the influence of different elements, such as sample size and cluster size, on the point and variance estimates of SMD, they may be inconvenient to use for behavioral researchers. Also, because in most situations the sample effect size is not normally distributed, one needs to invoke noncentral probability distributions to obtain confidence intervals (CIs). Furthermore, given that there are many different multilevel designs, it is impossible or at least very tedious to derive formulas for SMD and its variance for each design.

Therefore, in the third manuscript I examine whether the bootstrap (Efron, 1982), a popular resampling technique, can be a good general technique for obtaining SMD and other effect size measures with different data structures. The advantage of the bootstrap is that it requires only the specification of the point estimator for the effect size; The sampling variance is approximated by resampling. One type of the bootstrap, the nonparametric one, also has the added advantage that it handles violation of the nonnormality assumption automatically. In this manuscript, I review five methods of SMD estimation: the ANOVA method (partitioning sum of squares), the model-based method (using model estimates of variance components), the parametric bootstrap, the residual bootstrap, and the case bootstrap. The first two are analytical methods discussed in the first two manuscripts, and with them one can construct CIs using either asymptotic normal theory or the noncentral t distribution. For the bootstrap methods I consider in the manuscript the percentile CI and the bias-corrected and accelerated CI. The 10 CIs will be compared based on their empirical coverage probability and their width.

CHAPTER II

STANDARDIZED MEAN DIFFERENCES IN TWO LEVEL CROSS- CLASSIFIED RANDOM EFFECTS MODELS*

Overview

Multilevel modeling techniques are becoming more popular in handling data with multilevel structure in educational and behavioral research. Recently researchers have paid more attention to cross-classified data structure that naturally arises in educational settings. However, unlike traditional single level research, methodological studies about multilevel effect size have been rare and those that have recently appeared had an emphasis on strictly hierarchical data structure. The present article extends the work on multilevel standardized mean differences from strictly hierarchical structure to both fully and partially cross-classified structures. Analytically derived formulae for calculating effect sizes and the corresponding sampling variances (or standard errors) are presented, verified by simulation results, and illustrated with real data examples. Implications for primary research studies and meta-analyses are discussed.

Introduction

The field of educational and psychological science has witnessed a movement from the obsession of statistical significance testing to the evaluation of effect size (Ferguson, 2009). However, for studies with multilevel data, effect size is still under-reported, even though some effect size statistics have recently been developed for multilevel data with a nested structure (e.g., standardized mean difference; Hedges, 2007, 2011) and a cross-classified structure (e.g., proportion of variance accounted for; Luo &

*Reprinted with permission from "Standardized Mean Differences in Two-Level Cross-Classified Random Effects Models" by Mark H. C. Lai and Oi-man Kwok, 2014. *Journal of Educational and Behavioral Statistics*, 39, 282–302, Copyright 2014 by the American Educational Research Association.

Kwok, 2010). As cross-classified random effects models (CCREMs) have gained increasing attention in educational research (e.g., Friedel, Cortina, Turner, & Midgley, 2010; Johnson, 2011), there is a need to develop relevant effect size measures for CCREMs. The intent of the present article is to (a) analytically develop the standardized mean difference measure for two-level CCREM, (b) verify the performance of the mathematically derived formulae using simulated data, and (c) illustrate the computation of the effect size statistic with real data example.

The American Educational Research Association (AERA, 2006, p. 37), in its *Standards for Reporting on Empirical Social Science Research in AERA Publications*, recommended the use of effect size statistics with the rationale that "[i]nterpretation of statistical analyses is enhanced by reporting magnitude of relations." The American Psychological Association (APA, 2010, p. 34), in its *Publication Manual*, took a stronger stance to deem the reporting of effect size statistics as "almost always necessary." The attention given to effect size can be attributed to three important advantages of reporting such statistics. First, effect sizes, rather than statistical significance tests, directly answer research questions such as how strong two variables are associated, or how effective an intervention is (see Thompson, 2007). Second, to date effect size is the element to be synthesized in almost all meta-analytic studies (Lipsey & Wilson, 2001). Third, effect size estimation also plays a critical role in research planning, such as power analysis (Cohen, 1988). Two of the commonly reported effect size families include the standardized mean difference (i.e., group difference divided by sample standard deviation, or the d-family) and the proportion of variance accounted for (or the r-family; Grissom & Kim, 2012). Whereas the effect size statistics in single level studies are already well-developed, effect size in multilevel modeling has appeared only recently and is generally limited to strictly nested data. Therefore, more discussion on this topic will be necessary and valuable.

Effect Size in Multilevel Analyses

Although techniques for handling data with cluster structures have been developed for several decades (e.g., Goldstein, 1986; Mason, Wong, & Entwisle, 1983), in the past ten years they have gained much more attention in educational and behavioral research. This is not surprising given that in these fields much data collected have intrinsically nested structure. For example in the field of education, students are naturally nested within classrooms, and classrooms are naturally nested within schools. Because traditional data analytic techniques ignore the multilevel structure and give incorrect standard errors (Hox, 2010), new methods are proposed that provide correct standard errors and hence accurate statistical inference. One of the most popular approach is multilevel modeling (Goldstein, 2011b), which is synonymously called hierarchical linear modeling (Raudenbush & Bryk, 2002), linear mixed modeling (Littell et al., 1996), and other similar names.

Despite the rapid growth in the number of multilevel studies, rarely did researchers utilize effect size statistics in reporting multilevel results. Most of these studies used proportion of variance accounted for, or R^2 (see Luo & Kwok, 2010; Snijders & Bosker, 2012). However, for studies with a binary covariate, such as treatment-control or male-female, the standardized mean difference is a more natural choice, and is more easily understood by researchers.

In addition to the point estimates of an effect size, its sampling variance (or standard error) is also important. As commented by Cohen (1994), it is "far more informative to provide a confidence interval" (p. 1310), and the computation of (asymptotic) confidence interval (CI) requires the sampling variance of the effect size. This is particularly important for meta-analysts (Hedges, 1981; Lipsey & Wilson, 2001), because both point and variance estimates of effect size are required to get an overall

average effect size and to understand the influence of study-level covariates including publication bias in the literature. Given the importance of the point estimate and the sampling variance of the standardized mean difference effect size, as well as the lack of discussion about them in complex multilevel models, research efforts to supplement methods for their calculation are warranted.

Recently Hedges (2007, 2009, 2011) made a seminal effort in defining standardized mean difference statistics for data with two-level and three-level nested structures. Particularly he suggested that, depending on the context, there could be different choices of standard deviations in computing the effect size. Hedges (2007) illustrated the calculation of effect size in two-level studies with an example about the effect of using connected mathematics in classrooms. In that example students were nested within classrooms, and the treatment (i.e., connected mathematics) was defined at the classroom level (i.e., level-2). He showed that the overall effect size was 0.15 (95% CI $[-0.29, 0.59]$) and the within-classroom effect size was 0.17 (95% CI $[-0.34, 0.69]$). In a three level cluster-randomized design, five possible effect size statistics can be computed depending on which variance component is invoked. The formulae given by Hedges (2007, 2009, 2011) do not require researchers to have the raw data to obtain an effect size estimate; Instead, only the estimated treatment effect (i.e., grand mean difference between the treatment and the control arm), sample sizes for all levels of clustering, and the corresponding intraclass correlations are needed. In the context of meta-analysis, Ahn, Myers, and Jin (2012) have suggested methods to estimate intraclass correlations when the original research report does not include the relevant information.

Cross-Classified Random Effects Models (CCREMs)

The number of published articles adopting the CCREM method, a more complicated structure than nested multilevel models, has increased dramatically in recent

years. A simple search in the Educational Research Information Center (*ERIC*) database with the keyword "cross-classified" found only three articles during 2000 to 2005 but 32 articles during 2006 to 2012. One reason for the increasing adoption of CCREM is that multilevel data may not always have a strictly hierarchical structure. A typical example is given by Beretvas (2011), where students are nested within both primary schools (PS) and high schools (HS), but PS is not nested within HS nor vice versa. That is, not all students in one HS come from the same PS, nor do all students from one PS go to the same HS. In this case PS and HS are labeled as crossed factors. If both PS and HS are assumed to be random effects, then CCREM can be used to analyze such kind of data. Luo and Kwok (2010) have discussed the R^2 effect size for CCREMs. However, to the best of our knowledge, no discussion has taken place about standardized mean difference for CCREMs. Standardized mean difference would be suitable, for instance, in describing the effects of a school-based intervention on students' learning, where students are nested in both schools and neighborhoods. Based on the framework of previous studies (Hedges, 2007, 2011), in the present article we develop effect sizes for CCREMs through mathematical derivation, and evaluate their performances using both simulated and real data sets.

The purpose of the present article is to analytically develop the standardized mean difference measure for two-level CCREMs for both balanced and unbalanced designs, and to verify the performance of the mathematically derived formulae. Because of the complexity of the formulae, we also provide real data examples for pedagogical purposes so that applied researchers can better understand how those formulae can apply to their research. In the following sections we would (a) briefly introduce the notations for a two-level CCREM with two crossed factors; (b) discuss two estimation approaches to obtain the standardized mean difference, D, and the corresponding sampling variance, $V(D)$ (where $V(\cdot)$ denotes the variance operator), for balanced design CCREMs; (c)

empirically verify their performance through simulations; and (d) illustrate their calculations with real data having a cross-classified structure. The discussion of partially cross-classified random effect models (PCCREMs) included the same elements.

Model and Notation

In a *balanced* design with the cross-classification of two random effects A and B at level-2, let J and K be the number of clusters in effect A and in effect B respectively. In the context of education, A can be classrooms in a school and B can be neighborhoods. To make things more concrete in the following sections, we would use an hypothetical example where effect A is the classroom effect and effect B is the neighborhood effect, although the notation is equally applicable to other contexts such as therapy grouping effect by classroom effect, or in longitudinal settings with person effect by time effect. As a result there are J classrooms and K neighborhoods, and $J \times K$ combinations of classroom and neighborhood, or $J \times K$ cells. Further let $n_{jk} = n$ be the number of students in each cell with index $i = 1, \ldots, n$. In addition, assume that classrooms are randomly assigned to treatment condition or control condition. Because the word "group" can refer to either people in one of the treatment conditions or people from one of the classroom, to avoid confusion, in subsequent discussions the group receiving treatment is referred to as the *treatment arm* whereas the group in the control condition is referred to as the *control arm* (Bauer et al., 2008). For example, a researcher can randomly assign half of the classrooms to adopt a new reading instruction and the other half to use the traditional approach. Thus, classrooms are nested within treatment arms but neighborhoods and treatment arms are crossed. In this case students from the same classroom must have the same treatment status, whereas students from the same neighborhood can have different treatment statuses if they come from different classrooms.

Let $j = 1, \ldots, J^T$ and $j = 1, \ldots, J^C$ be the index of classroom for the treatment

(T) and the control (C) arms where $J^T + J^C = J$, and $k = 1, \ldots, K^T$ and $k = 1, \ldots, K^C$ be the index of neighborhood. Furthermore, the sets of neighborhoods in the treatment arm and in the control arm may be completely overlapped, partially overlapped, or completely separated. Let K_{overlap} be the number of overlapping clusters, and the three possible situations are then $K_{\text{overlap}} = K^T = K^C = K$ (complete overlapping), $K_{\text{overlap}} = (K^T + K^C) - K > 0$ (partial overlapping), or $K_{\text{overlap}} = 0$ and $K = K^T + K^C$ (complete separation). Therefore, the sample size for the treatment group is $N^T = J^T \times K^T \times n$, that for the control group is $N^C = J^C \times K^C \times n$, and the total sample size is $N = N^T + N^C$. The model can then be specified as

$$Y_{ijk} = \gamma_{00} + \gamma_{10}(\text{TREAT}_j) + \mu_{0j} + \nu_{0k} + \epsilon_{ijk}, \tag{1}$$

where Y_{ijk} refers to the score of the ith student in the jth classroom and the kth neighborhood, and TREAT_j the treatment status variable dummy coded as 0 (control) and 1 (treatment). γ_{00} is the grand mean of the control arm in the sample, γ_{10} is the mean difference between the treatment arm and the control arm, μ_{0j} is the magnitude of the effect of the jth classroom, ν_{0k} is the magnitude of the effect of the kth neighborhood, and ϵ_{ijk} is the within-cell residual (i.e., the student effect). Usually researchers do not estimate the interaction effect between random effects for simplicity (Shi, Leite, & Algina, 2010). Also, following (Hedges, 2007), it is assumed that the treatment effect does not interact with random effects A and B.

In a balanced design, the variance of Y can be partitioned into three independent components, which are denoted as σ^2_W, the within cluster variance; σ^2_A, the classroom-level variance or the variance due to the random effect A; and σ^2_B, the neighborhood-level variance or the variance due to the random effect B. There are several methods to obtain an estimate of these variance components, such as the ANOVA method,

full maximum likelihood, and restricted maximum likelihood (Searle et al., 2006). As discussed in later sections obtaining estimates of these variance components are the key to computing an effect size.

Intraclass Correlation

The intraclass correlation (ICC) quantifies the degree to which two randomly drawn observations within a cluster are correlated. In CCREMs there are different possible ICCs depending on how a cluster is defined. For instance, for observations in the same classroom (random effect A) but in different neighborhoods (random effect B), the ICC can be defined as:

$$\rho_A = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_B^2 + \sigma_W^2} = \frac{\sigma_A^2}{\sigma_T^2}, \tag{2}$$

where $\sigma_T^2 = \sigma_A^2 + \sigma_B^2 + \sigma_W^2$. Similarly, for observations in the same neighborhood but in different classrooms, the ICC can be defined as:

$$\rho_B = \frac{\sigma_B^2}{\sigma_A^2 + \sigma_B^2 + \sigma_W^2} = \frac{\sigma_B^2}{\sigma_T^2}. \tag{3}$$

Standardized Mean Differences for Fully Cross-Classified Data

In educational research, the standardized mean difference is defined as the ratio of (a) the difference between the population means of the treatment arm and of the control arm to (b) a standard deviation. Hedges (2009) defined different effect sizes associated with different levels. For example, a researcher may be interested in how an intervention is effective in group level, and can use only the between level standard deviation while ignoring the within group variations. Similarly in CCREM one can consider using σ_W, σ_A, σ_B, $\sqrt{\sigma_W^2 + \sigma_A^2}$, $\sqrt{\sigma_W^2 + \sigma_B^2}$, or σ_T. Perhaps the issue can be made simpler by reminding that in single-level studies, standardized mean difference between the treatment

and the control arms can be converted to an R^2 effect size due to the binary treatment dummy variable (e.g., 0 = control, 1 = treatment). Because in education treatment is usually a variable at the second or higher level, as in the classroom-neighborhood example where treatment is on the classroom level, generally the treatment will not explain within-level (i.e., student-level) variance (Snijders & Bosker, 1994). Therefore, in our opinions, σ^2_W is in general not justified unless one assumes that the treatment effect stays the same, whether it is individually-randomized or cluster-randomized. For meta-analysts the decision often depends on the nature of the other studies. If there are single-site studies in the list, generally choosing variance components of classroom or neighborhood levels makes the comparison in meta-analysis difficult (Hedges, 2007). Because in education often data are cross-classified (e.g., Beretvas, 2011), and researchers are usually interested in generalizing the effect to a broader population of students (or other level-1 units), σ_T is a more natural choice. Thus, in subsequent mathematical derivation we focused on using σ_T.

On the population level the effect size is defined as

$$\delta_T = \frac{\mu^T_{\bullet\bullet\bullet} - \mu^C_{\bullet\bullet\bullet}}{\sigma_T}, \tag{4}$$

where $\mu^T_{\bullet\bullet\bullet}$ and $\mu^C_{\bullet\bullet\bullet}$ are the population means of the treatment and of the control arm respectively. In a balanced design, the average of the cell means, $\bar{Y}^T_{\bullet\bullet\bullet} = \sum \bar{Y}^T_{\bullet jk}/(J^T K)$ and $\bar{Y}^C_{\bullet\bullet\bullet} = \sum \bar{Y}^C_{\bullet jk}/(J^C K)$, are unbiased and efficient estimators of $\mu^T_{\bullet\bullet\bullet}$ and $\mu^C_{\bullet\bullet\bullet}$. Thus, the difference between the two averaged cell means is an unbiased and efficient estimator of the numerator of δ_T. However, the observed total variance

$$S^2_T = \frac{\sum_{k=1}^{K^T}\sum_{j=1}^{J^T}\sum_{i=1}^{n}\left(Y_{ijk} - \bar{Y}^T_{\bullet\bullet\bullet}\right)^2 + \sum_{k=1}^{K^C}\sum_{j=1}^{J^C}\sum_{i=1}^{n}\left(Y_{ijk} - \bar{Y}^C_{\bullet\bullet\bullet}\right)^2}{N-2} \tag{5}$$

is in general a biased estimator of the total population variance σ_T^2 in random effect models. We would present two methods to obtain a consistent estimate of the population effect size δ_T, one by multiplying a correction factor to S_T and the other by utilizing the computer estimates of the variance components to obtain σ_T. The first method is based on the expected mean squares of the random effects, and the estimated effect size is denoted as D_1 in this paper. It is both efficient and consistent on balanced data structure where cells have (roughly) equal size, and will be useful for meta-analysts when the primary research studies did not present estimates of σ_T. As shown later in the simulation results it is also robust to unbalanced design. The second method is based on the estimated variance components of the random effects, and the estimated effect size is denoted as D_2. It is efficient for both balanced and unbalanced data, and is easier to compute than the first method, provided that the point and variance (or standard error) estimates of variance components are available. It will be useful for both researchers working with primary data and meta-analysts having access to the required information.

Estimation of D_1

With a balanced data structure assumed, and when the sets of clusters of random effect B in the treatment arm and in the control arm overlap completely (e.g., students receiving treatment come from the same set of neighborhoods as those in the control arm), the sample estimator of δ_T, D_1, and the corresponding sampling variance $V(D_1)$ are:

$$D_1 = \frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{S_T}\sqrt{1 - \frac{2(Kn-1)\rho_A + (Jn-2)\rho_B}{N-2}}, \tag{6}$$

and

$$V(D_1) = \frac{1}{\tilde{N}}\left[1 + (Kn-1)\rho_A\right] + D_1^2\left(\frac{Kn\check{N}_K\rho_A^2 + Jn\check{N}_J\rho_B^2 + (N-2)\bar{\rho}^2 + 2\check{N}_K\bar{\rho}\rho_A + 2\check{N}_J\bar{\rho}\rho_B}{2\left[(N-2) - 2(Kn-1)\rho_A - (Jn-2)\rho_B\right]^2}\right), \quad (7)$$

where $\tilde{N} = N^T N^C/(N^T + N^C)$, $\check{N}_K = N - 2Kn$, $\check{N}_J = N - Jn$, and $\bar{\rho} = 1 - \rho_A - \rho_B$. See Appendix A for detailed derivations. On the other hand, if the sets of clusters of random effect B in the treatment arm are different to those in the control arm (e.g., students from certain neighborhoods are all in the treatment arm, and students from some other neighborhoods are all in the control arm), the approximated sampling variance $V(D)$ is:

$$V(D_1) = \frac{1}{\tilde{N}}\left[1 + (Kn-1)\rho_A + (1-r_K)(Jn-2)\rho_B\right] + D_1^2\left(\frac{Kn\check{N}_K\rho_A^2 + Jn\check{N}_J\rho_B^2 + (N-2)\bar{\rho}^2 + 2\check{N}_K\bar{\rho}\rho_A + 2\check{N}_J\bar{\rho}\rho_B}{2\left[(N-2) - 2(Kn-1)\rho_A - (Jn-2)\rho_B\right]^2]}\right), \quad (8)$$

where $r_K = \sqrt{(K_{\text{overlap}})^2/(K^T \times K^C)}$ is the correlation of the random effect B between the treatment and the control arm, K^T and K^C are the numbers of effect B clusters specific to the treatment and the control arm, and K_{overlap} is the number of overlapping clusters. Note that K is now defined as the total number of B-clusters such that $K = K^T + K^C - K_{\text{overlap}}$. Equations (6), (7), and (8) outline the influence of cluster size, number of clusters, and intraclass correlations on the effect size estimates and its sampling variance.

Estimation of D_2

The derivation of D_1 is based on assumptions that (a) the cluster size is constant and (b) the ICCs are known or estimated with a reasonable accuracy. In real research these assumptions may not hold. If either (a) or (b) or both (a) and (b) are violated, then D_1 and $V(D_1)$ calculated from equations (6) and (7) can be biased and inefficient. For unbalanced

designs, the close forms of D_1 and $V(D_1)$ are very complex and are functions of the cell sizes in addition to the components in (6) and (7). Because information about the cell sizes are rarely available from published research reports, it is difficult to obtain efficient estimates of δ_T and $V(\delta_T)$ for unbalanced data starting from expected mean squares. However, if consistent estimates of the variance components (from maximum likelihood, restricted maximum likelihood, or Bayesian estimation, etc) are available, researchers can use both the point estimates and the standard errors of the random effects to calculate the effect size. Specifically, if estimates of the treatment effect and the variance components, $\hat{\gamma}_{10}$, $\hat{\sigma}^2_W$, $\hat{\sigma}^2_A$, $\hat{\sigma}^2_B$, and their corresponding variances (the squared values of the standard errors), $V(\hat{\gamma}_{10})$, $V(\hat{\sigma}^2_W)$, $V(\hat{\sigma}^2_A)$, and $V(\hat{\sigma}^2_B)$ can be obtained, then we get

$$D_2 = \frac{\hat{\gamma}_{10}}{\sqrt{\hat{\sigma}^2_W + \hat{\sigma}^2_A + \hat{\sigma}^2_B}}, \tag{9}$$

$$V(D_2) = \frac{V(\hat{\gamma}_{10})}{\hat{\sigma}^2_W + \hat{\sigma}^2_A + \hat{\sigma}^2_B} + \frac{D_2^2\left[V(\hat{\sigma}^2_W) + V(\hat{\sigma}^2_A) + V(\hat{\sigma}^2_B)\right]}{4(\hat{\sigma}^2_W + \hat{\sigma}^2_A + \hat{\sigma}^2_B)^2}. \tag{10}$$

Derivations of (9) and (10) can be found in Appendix A.

If meta-analysts can obtain neither information about the degree of imbalance of the data nor unbiased estimates of the variance components, the best they can do is to compute the effect size assuming a balanced design and replace n in (6) and (7) with the average cell size, $N/(JK)$, to obtain D_1. If information about ICC are not obtainable, they can put in a reasonable guess of ρ_A and ρ_B by referring to research with similar designs and variables. There are also several articles summarizing what typical ICCs are for various designs and areas (Hedges & Hedberg, 2007; Murray & Blitstein, 2003).

Monte Carlo Study for Evaluating the Two Effect Size Estimation Approaches Under Unbalanced Designs

We used a $3 \times 2 \times 2 \times 2$ full factorial simulation study to empirically check the performance of D_1 and D_2 with unbalanced designs. The design factors included (a) population effect size ($\delta_T = 0.2, 0.5$, or 0.8), (b) number of clusters in random effect A (e.g., classrooms) per treatment status ($J^T = J^C = 20$ or 50), (c) average cell size ($n = 0.25$ or 1), and (d) ICC of random effect A ($\rho_A = .10$ or .25) (which are common values used in previous simulation studies). In the first simulation we generated data such that the treatment arm and the control arm shares the same set of neighborhoods. In other words, $K^T = K^C = K = K_{\text{overlap}}$. K was set to equal to J^T and J^C, and ρ_B was fixed to .1. The imbalance of data structure was similar across conditions in which 20% of the cells had an expected cell count that was 10 times larger than that of the remaining 80% of cells (see Appendix B for more details). Such a data structure is similar to that described in Beretvas (2008) where students are nested within the cross-classification of schools and neighborhoods. Across conditions Pearson's contingency coefficients ranged from .84 to .92, showing that the degree of imbalance was quite strong. R 3.0.1 (R Core Team, 2013) was used to generate 500 data sets for each conditions, with $\mu^C_{\bullet\bullet\bullet} = 0$ and $\sigma^2_W = 1.0$. All random effects were normally distributed.

The estimation of D_1 and $V(D_1)$ was performed in R with ρ_A and ρ_B being fixed to the population value. On the other hand, D_2 and $V(D_2)$ were obtained using `MODEL CONSTRAINT` in the `TYPE=CROSSCLASSIFIED` procedure in Mplus 7.0 (L. K. Muthén & Muthén, 1998–2012) with the default non-informative prior Bayesian estimation.

As shown in Table 1, both effect size estimators had relative bias less than 5%. The relative *SE* bias was stronger, but the impact was small as the percentage coverage of the 95% symmetric CI[1] was close to nominal value and ranged from 93.6% to 95.6%.

This is within the expected interval [92.7%, 96.6%] when the true coverage is 95% with 500 replications, and thus we conclude that the performance of both estimators was satisfactory. As expected D_1 was less efficient under unbalanced structure, but the loss of efficiency was little. Specifically, the relative efficiency, $RE = V(D_2)/V(D_1)$, of D_1 was lowest when both J^T and n were small and when ρ_A was large, but it was still acceptable as $RE = 86\%$. In summary, both approaches to obtain δ_T and $V(\delta_T)$ performed well in unbalanced designs.

Next we generated data with $K_{\text{overlap}} > 0$. We kept $K = J^T = J^C$, but set $K^T = K^C = 3J^T/5$ and so $K_{\text{overlap}} = K/5$. In other words, 20% of the clusters in random effect B were overlapped between the treatment and the control arms, with a correlation approximately equals to $r_k = \sqrt{(1/5)^2/[(3/5) \times (3/5)]} \approx 0.33$. The results of the simulation are shown in Table 2. For all conditions the coverage of 95% CI of both D_1 and D_2 were acceptable and ranged from 91.8% to 97.4%, which is not too far away from the expected CI when the true coverage is 95%. However, compared to occasions with complete overlapping, here the RE of D_1 relative to D_2 dropped to 70.8% to 87.9%, indicating that D_1 is substantially less efficient than D_2. Contrary to the pattern in occasions with sets of clusters completely overlapped, here RE was lowest when n was large and ρ_A was small.

Real Data Example

We would use part of the National Educational Longitudinal Study data set (NELS:88; Ingles, Abraham, Karr, Spencer, & Frankel, 1990) to illustrate the calculation of D_1 and D_2. This longitudinal study followed a nationally representative sample of students starting from their eighth grade, and recorded students' experiences in a variety of areas such as home and working. A hypothetical research question is whether availability of a mathematics club in middle school predicted students' mathematics

Table 1
Simulation Results for Unbalanced CCREMs (With Clusters in Random Effect B Completely Overlapped)

				Coverage		Relative Bias		Relative *SE* Bias		
δ_T	J^T	N^T	ρ_A	D_1	D_2	D_1	D_2	D_1	D_2	$RE(D_1, D_2)$
0.2	20	100	.10	94.8	96.0	4.3	1.9	−5.8	3.5	89.2
			.25	94.4	94.8	4.4	−4.1	−9.1	0.3	87.2
		400	.10	96.4	94.2	−2.9	−0.7	2.2	2.2	95.0
			.25	94.8	93.8	−3.2	1.3	0.1	−6.1	93.6
	50	625	.10	94.0	93.0	−0.9	−3.4	−7.4	−8.2	93.0
			.25	93.2	94.0	−1.9	−3.5	−10.1	−3.4	91.2
		2500	.10	94.6	92.4	0.4	0.3	−0.1	−14.2	97.1
			.25	94.0	91.4	0.2	2.2	−1.3	−20.2	96.4
0.5	20	100	.10	95.0	95.4	2.0	−1.2	−7.3	2.3	89.5
			.25	94.2	94.2	2.1	−3.7	−10.0	−0.1	87.4
		400	.10	96.0	95.0	−1.1	−1.7	2.8	4.0	95.1
			.25	94.8	93.4	−1.1	−1.1	0.1	−5.3	93.6
	50	625	.10	94.4	93.4	−0.3	−1.7	−5.9	−6.1	93.1
			.25	93.4	94.0	−0.6	−1.8	−8.9	−1.8	91.4
		2500	.10	94.8	92.2	0.3	−0.0	−0.7	−14.1	97.1
			.25	94.2	91.8	0.3	0.6	−1.4	−19.3	96.1
0.8	20	100	.10	94.6	94.6	1.4	−2.0	−9.0	1.0	90.0
			.25	93.8	93.8	1.5	−3.6	−11.0	−0.6	87.8
		400	.10	96.2	95.0	−0.6	−1.9	3.2	5.5	95.5
			.25	95.2	93.4	−0.6	−1.7	−0.1	−4.3	93.7
	50	625	.10	94.4	93.6	−0.1	−1.3	−4.4	−3.8	93.1
			.25	94.2	94.2	−0.3	−1.4	−7.7	0.7	90.9
		2500	.10	94.8	92.4	0.2	−0.2	−1.4	−13.4	97.1
			.25	94.0	92.6	0.3	0.1	−1.6	−18.7	96.5

Note. Based on 500 replications for each condition. Pearson's contingency coefficients for all conditions ranged from .84 to .92, indicating strong associations between the clustering of random effects A and B. δ_T = population effect size. $J^T = J^C$ = number of clusters of random effect A in the treatment (control) arm; Number of clusters of random effect $B = J^T$. $N^T = N^C$ = total sample size of the treatment (control) arm. ρ_A = intraclass correlation of effect A; $\rho_B = 0.1$ for all conditions. Coverage refers to the percentage of replications in which the 95% confidence interval includes δ_T. For 500 replications, the Monte Carlo coverage percentage have a confidence interval of [92.7%, 96.6%] if the true coverage percentage is 95 %. RE(D_1, D_2) = relative efficiency of estimator D_1 to the Mplus (version 7.0) estimation of D_2 using `TYPE=CROSSCLASSIFIED` and `ESTIMATOR=BAYES`, which was computed by dividing the sampling variance of the later by that of the former.

Table 2
Simulation Results for Unbalanced CCREMs (With Clusters in Random Effect B Partially Overlapped)

				Coverage		Relative Bias		Relative *SE* Bias		
δ_T	J^T	N^T	ρ_A	D_1	D_2	D_1	D_2	D_1	D_2	$RE(D_1, D_2)$
0.2	20	100	.10	92.6	94.4	3.2	−5.4	−9.4	−0.7	81.0
			.25	93.8	92.6	5.0	−1.4	−9.6	−11.5	81.6
		400	.10	95.8	97.4	1.9	−5.6	9.9	18.8	72.5
			.25	96.0	95.0	3.7	−9.2	9.5	6.4	77.5
	50	625	.10	94.4	94.2	0.6	−10.2	1.4	11.7	83.7
			.25	93.4	93.0	2.5	−10.2	−7.2	3.0	87.2
		2500	.10	94.2	93.4	0.9	−8.4	1.0	−3.1	70.8
			.25	93.8	92.8	−0.1	−15.7	−5.4	−17.0	80.4
0.5	20	100	.10	92.8	94.4	1.5	−5.8	−8.6	0.5	80.8
			.25	94.0	93.4	2.4	−4.1	−8.7	−9.6	81.0
		400	.10	96.0	97.2	0.9	−3.9	10.0	19.6	72.8
			.25	95.8	95.0	1.7	−5.8	9.4	7.0	77.6
	50	625	.10	94.6	94.0	0.3	−4.9	2.3	13.2	84.3
			.25	93.2	93.0	1.1	−4.7	−7.3	3.2	87.7
		2500	.10	94.6	93.2	0.5	−3.7	0.8	−2.7	71.0
			.25	94.0	91.8	0.1	−6.8	−5.5	−16.4	80.7
0.8	20	100	.10	93.0	94.0	1.0	−6.0	−7.7	1.9	81.0
			.25	93.6	94.0	1.8	−4.8	−7.8	−7.5	80.8
		400	.10	95.6	96.8	0.7	−3.5	9.9	20.4	73.5
			.25	95.6	94.8	1.1	−5.0	9.2	7.3	78.3
	50	625	.10	95.2	94.0	0.3	−3.6	3.1	14.3	85.1
			.25	93.0	92.6	0.8	−3.3	−7.4	3.8	87.9
		2500	.10	94.6	93.0	0.4	−2.5	0.5	−2.5	71.4
			.25	94.0	91.8	0.2	−4.7	−5.6	−15.7	81.0

Note. Based on 500 replications for each condition. Pearson's contingency coefficients for all conditions ranged from .84 to .92, indicating strong associations between the clustering of random effects A and B. δ_T = population effect size. $J^T = J^C$ = number of clusters of random effect A in the treatment (control) arm; Number of clusters of random effect $B = J^T$. $N^T = N^C$ = total sample size of the treatment (control) arm. ρ_A = intraclass correlation of effect A; $\rho_B = 0.1$ for all conditions. Coverage refers to the percentage of replications in which the 95% confidence interval includes δ_T. For 500 replications, the Monte Carlo coverage percentage have a confidence interval of [92.7%, 96.6%] if the true coverage percentage is 95 %. RE(D_1, D_2) = relative efficiency of estimator D_1 to the Mplus (version 7.0) estimation of D_2 using `TYPE=CROSSCLASSIFIED` and `ESTIMATOR=BAYES`, which was computed by dividing the sampling variance of the later by that of the former.

achievement at 10th grade. Using only cases with complete data on all the variables related to the analysis, the data consisted of 15,611 students cross-classified by 986 middle schools (MS, 383 with math club) and 1,418 high schools (HS, $K^T = 625$, $K^C = 940$, so $K_{\text{overlap}} = 147$). The average cell size was thus $15,611/(986 \times 1,418) = 0.0112$. Only 147 HS had both students from MS's with math club and those from MS's without math club, so correlation r_K of the HS effect was $\sqrt{147^2/(625 \times 940)} = .192$ for the two treatment arms. Using the SPSS `mixed` procedure, we estimated the grouping effect of availability of mathematics club, the variance components for within cluster, MS (random effect A), and HS (random effect B), as well as their standard errors. The grouping effect was estimated as 0.731 ($SE = 0.362$), and the variance components were estimated as $\hat{\sigma}^2_A = 18.784$ ($SE = 1.781$), $\hat{\sigma}^2_B = 7.293$ ($SE = 1.445$), and $\hat{\sigma}^2_W = 76.296$ ($SE = 0.902$). Using equations (9) and (10), the effect size D_2 for the grouping effect is 0.0722 ($SE = 0.0358$, 95% CI $[0.002, 0.142]$), indicating a small effect size. The estimated value of D_1 (with the sample estimated $\rho_A = 0.183$ and $\rho_B = 0.0712$) is 0.116 ($SE = 0.0338$, 95% CI $[0.050, 0.182]$). Under such an extreme unbalanced data structure D_2 is expected to be more accurate than D_1, although the difference is not truly substantial when the 95% CI is also taken into account.

Standardized Mean Differences for Partially Cross-Classified Data

Thus far we have considered cross-classified data, where observations in both the treatment and the control arms are cross-classified by effects A and B. However, there are designs where only observations in the treatment arm are cross-classified, but the observations in the control arm are nested only in effect A but not in effect B. In this article we denote such a data structure as partially cross-classified. This is similar to the partially nested design discussed in Bauer et al. (2008) where the observations in the treatment arm are nested within random effect A whereas those in the control arm are not.

The difference is that in partially cross-classified data there is one more level of nested structure, random effect B, present in both the treatment and the control arms.

Consider a hypothetical example, where students from J^T classrooms are randomly assigned to K emotion management groups (i.e., the treatment), and those from J^C other classrooms do not receive any treatment. Further, assume that each emotion management group includes students from different classrooms to avoid situations where group members are very familiar with each other. Suppose that a researcher is interested in the effectiveness of the emotion management group on students' life satisfaction (Y). Such a design can be represented by the model equation

$$Y_{ijk} = \gamma_{00} + \gamma_{10}(\text{TREAT}_j) + \mu_{0j} + \nu_{0k}(\text{TREAT}_j) + \epsilon_{ijk}, \tag{11}$$

where TREAT_j is the treatment status dummy coded as 0 (control) and 1 (treatment), μ_{0j} is the classroom effect, ν_{0k} is the emotion management grouping effect that is only present in the treatment arm, and ϵ_{ijk} is the student effect. γ_{00} is the grand mean of Y of the control arm, and γ_{10} is the treatment effect. Further assume that both treatment conditions share the same total variance σ^2_T and the same variance of effect A σ^2_A. The variance of effect B in the treatment arm is σ^2_B, the within-cell variance of the treatment arm is $\sigma^2_{W|\text{TREAT}}$, and that of the control arm is $\sigma^2_{W|\text{CON}}$. Let n^T_{jk} be the size of the cells in the treatment arm and n^C_j be the cluster size in the control arm. We denote such a model as a partially cross-classified random effect model (PCCREM). Finally, we assume that A and B have no interaction, and ρ_A and ρ_B are defined the same way as in (2) and (3).

Estimation of D_1

As shown in Appendix A the sample estimator D_1 and $V(D_1)$ of the effect size δ_T are given as

$$D_1 = \frac{\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}}{\bar{S}^2_T}\sqrt{\frac{W^T\beta^T + W^C\beta^C}{W^T + W^C}}, \tag{12}$$

$$V(D_1) = \frac{1 + (Kn^T - 1)\rho_A + (J^T n^T - 1)\rho_B}{N^T} + \frac{1 + (n^C - 1)\rho_A}{N^C} + \frac{D^2}{2(W^T\beta^T + W^C\beta^C)}, \tag{13}$$

where $\bar{S}^2_T = (W^T S^2_{T|\text{TREAT}} + W^C S^2_{T|\text{CON}})/(W^T + W^C)$ is the weighted average of the total variances of the treatment arm, $S^2_{T|\text{TREAT}}$, and of the control arm, $S^2_{T|\text{CON}}$, with weights

$$W^T = \frac{(N^T - 1)^2}{Kn^T\check{N}^T_K\rho^2_A + J^T n^T\check{N}^T_J\rho^2_B + (N^T - 1)\bar{\rho}^2 + 2\check{N}^T_K\bar{\rho}\rho_A + 2\check{N}^T_J\bar{\rho}\rho_B},$$

$$W^C = \frac{(N^C - 1)^2}{(N^C - 1) - 2(n^C - 1)\rho_A + (n^C - 1)\left[N^C - (n^C - 1)\right]\rho^2_A},$$

where $\check{N}^T_K = N^T - Kn^T$, $\check{N}^T_J = N^T - J^T n^T$, and $\bar{\rho} = 1 - \rho_A - \rho_B$, and

$$\beta^T = 1 - \frac{(Kn^T - 1)\rho_A + (J^T n^T - 1)\rho_B}{N^T - 1},$$

$$\beta^C = 1 - \frac{(n^C - 1)\rho_A}{N^C - 1}.$$

Estimation of D_2

When maximum likelihood or other unbiased estimates of the fixed effect, the variance components, and their sampling variances are available, one can calculate the

standardized mean difference and its sampling variance as

$$D_2 = \frac{\hat{\gamma}_{10}}{\sqrt{\hat{\sigma}^2_T}}, \tag{14}$$

$$V(D_2) = \frac{V(\hat{\gamma}_{10})}{\hat{\sigma}^2_T} + \frac{D_2^2\left[V(\hat{\sigma}^2_T)\right]}{4(\hat{\sigma}^2_T)^2}, \tag{15}$$

where $\hat{\sigma}^2_T$ is the weighted average of the total estimated variances of the two treatment arms by their respective sampling variances (i.e., $\hat{\sigma}^2_{T|\text{TREAT}}$ and $\hat{\sigma}^2_{T|\text{CON}}$, with $\hat{\sigma}^2_{T|\text{TREAT}} = \hat{\sigma}^2_{W|\text{TREAT}} + \hat{\sigma}^2_A + \hat{\sigma}^2_B$ for the treatment arm and $\hat{\sigma}^2_{T|\text{CON}} = \hat{\sigma}^2_{W|\text{CON}} + \hat{\sigma}^2_A$ for the control arm), and $V(\hat{\sigma}^2_T)$ is given by

$$\frac{1}{\left[V(\hat{\sigma}^2_{W|\text{TREAT}}) + V(\hat{\sigma}^2_A) + V(\hat{\sigma}^2_B)\right]^{-1} + \left[V(\hat{\sigma}^2_{W|\text{CON}}) + V(\hat{\sigma}^2_A)\right]^{-1}}. \tag{16}$$

If the weighted average of the two variance components is difficult to obtain, researchers can replace $\hat{\sigma}^2_T$ by $\hat{\sigma}^2_{W|\text{CON}} + \hat{\sigma}^2_A$, which results in some loss of efficiency, but the loss is in general minor unless the sample size of the treatment arm is much larger than that of the control arm. Derivations of (14) and (15) can be found in Appendix A.

Monte Carlo Study for Evaluating the Two Effect Size Estimation Approaches Under Unbalanced Designs

Similar to what we did for fully cross-classified designs, we used simulations to empirically check the performance of D_1 and D_2 for PCCREMs. The simulation conditions were the same as those used for the previous CCREM simulation study except that the random effect B was not present in the control arm, and the associated variances were added to $\rho_{W|\text{CON}}$. For each condition 500 data sets were generated in R, where D_1 for each condition was also computed. For the calculation of D_2 the `PROC MIXED` (Littell et al., 1996) procedure in SAS 9.3 was used.

As shown in Table 3, both D_1 and D_2 have relative bias less than 5% and coverage percentage of the 95% CI ranging from 92.8% to 95.2%, so their performances are satisfactory. In general D_1 was less efficient under unbalanced structure when ρ_A was large and the average cell size was small, but the loss of efficiency was negligible (minimum relative efficiency being 96.4%). In summary, both approaches to estimate the effect size δ_T and $V(\delta_T)$ perform similarly well in the chosen unbalanced designs.

Real Data Example

For illustrative purpose the data used in Coyne et al. (2013) was analyzed. The data consisted of 103 kids receiving Early Reading Intervention (ERI). The treatment arm consisted of 70 kids who have changed group membership based on their performance over the course of the study, which created a cross-classified data structure given that the initial group membership of each kid was different from the final group membership. In other words, kids were cross-classified by the initial and final group memberships in the treatment arm. On the other hand, the control arm consisted of 33 kids who were randomly assigned to groups at the beginning of the study and were kept in the same group over the course of study, which created a strictly hierarchical structure for this group of kids. The dependent variable is the score on a word identification test at the final stage. There were 19 groups among those receiving intervention (i.e., $J^T = 19$) and 10 groups among those receiving regular reading instruction (i.e., $J^C = 10$) at the the initial stage. At the final stage, those in regular reading stayed in original their groups, but some of those receiving intervention had moved to different groups. In summary, students receiving regular reading instruction were nested within initial groupings, whereas those receiving intervention were cross-classified by the initial and final groupings ($K = 19$). The design was not balanced, and the average cell size was 0.187.

Using SPSS `mixed` with an approach analogous to Bauer et al. (2008), the

Table 3
Simulation Results for Unbalanced PCCREMs

				Coverage		Relative Bias		Relative *SE* Bias		
δ_T	J^T	N^T	ρ_A	D_1	D_2	D_1	D_2	D_1	D_2	$RE(D_1, D_2)$
0.2	20	100	0.10	94.2	94.6	1.9	2.3	−6.4	0.4	98.8
			0.25	94.0	95.0	2.1	3.0	−8.4	0.2	96.0
		400	0.10	95.0	93.0	−3.5	−3.3	−1.2	−0.4	100.5
			0.25	94.6	93.6	−3.9	−3.2	−1.6	−1.8	99.9
	50	625	0.10	93.2	92.8	−0.5	0.3	−13.1	−9.3	97.9
			0.25	93.6	93.2	−1.5	−0.1	−14.5	−10.0	96.1
		2500	0.10	94.8	94.0	1.0	0.9	−4.3	−3.7	98.7
			0.25	94.8	93.2	0.9	0.7	−3.1	−2.2	98.3
0.5	20	100	0.10	94.6	95.2	0.9	1.6	−7.2	−0.4	98.8
			0.25	94.8	95.6	1.0	1.7	−9.1	−0.9	96.2
		400	0.10	94.8	93.6	−1.3	−1.1	−0.9	−0.1	100.2
			0.25	94.4	93.8	−1.3	−1.0	−1.7	−2.5	99.9
	50	625	0.10	93.4	93.2	−0.1	0.3	−11.6	−7.9	97.7
			0.25	92.8	93.4	−0.5	0.1	−13.2	−9.0	96.0
		2500	0.10	94.4	94.4	0.5	0.5	−4.6	−4.1	98.5
			0.25	94.6	94.0	0.5	0.4	−2.9	−2.5	98.2
0.8	20	100	0.10	95.0	95.0	0.7	1.4	−7.9	−1.1	99.0
			0.25	94.8	95.2	0.8	1.4	−9.6	−2.1	96.5
		400	0.10	94.6	93.6	−0.8	−0.6	−0.4	0.1	99.9
			0.25	93.8	93.8	−0.7	−0.5	−1.7	−3.4	100.0
	50	625	0.10	93.4	93.4	0.0	0.3	−9.9	−6.5	97.5
			0.25	93.4	93.6	−0.2	0.2	−11.6	−8.1	96.0
		2500	0.10	95.0	94.0	0.3	0.3	−4.8	−4.5	98.2
			0.25	95.0	94.6	0.4	0.3	−2.5	−2.9	98.2

Note. Based on 500 replications for each condition. δ_T = population effect size. Pearson's contingency coefficients for all conditions ranged from .84 to .92, indicating strong associations between the clustering of random effects A and B for the treatment arm. J^T = number of clusters of random effect A in the treatment arm; Number of clusters of random effect B for the treatment arm = J^T. $N^T = N^C$ = total sample size of the treatment (control) arm. ρ_A = intraclass correlation of effect A; $\rho_B = 0.1$ for the treatment group for all conditions. Coverage refers to the percentage of replications in which the 95% confidence interval includes δ_T. For 500 replications, the Monte Carlo coverage percentage have a confidence interval of [92.7%, 96.6%] if the true coverage percentage is 95 %. RE(D_1, D_2) = relative efficiency of the estimator D_2 to the estimator D_2 computed from SAS 9.3 with `DDFM=SATTERTHWAITE` and `METHOD=REML`, which was computed by dividing the sampling variance of the later by that of the former.

grouping effect was estimated as 9.202 (SE = 3.492), and the variance components were estimated as $\hat{\sigma}^2_A$ = 3.401 (SE = 15.217), $\hat{\sigma}^2_B$ = 100.140 (SE = 42.621), $\hat{\sigma}^2_{W|\text{TREAT}}$ = 97.723 (SE = 20.549), and $\hat{\sigma}^2_{W|\text{CON}}$ = 176.504 (SE = 50.411). Assuming that the total variances of both treatment conditions are comparable, ρ_A and ρ_B were estimated as .018 and .550 respectively. For the estimation of D_1, the additional inputs were $\bar{Y}^T_{\bullet\bullet\bullet}$ = 105.06, $\bar{Y}^C_{\bullet\bullet}$ = 95.73, $S^2_{T|\text{TREAT}}$ = 181.83, and $S^2_{T|\text{CON}}$ = 199.19. The computed D_1 = 0.675 (SE = 0.264, 95% CI [0.157, 1.192]), which can be interpreted such that on average students receiving intervention scored .67 SD higher on word identification than those receiving regular instructions. Using equations (14) and (15), the effect size D_2 for the grouping effect is 0.682 (SE = 0.261, 95% CI [0.171, 1.193]). Both approaches gave similar point and interval estimates, and both indicated a moderate to large intervention effect with the 95% CI not including zero.

Conclusion

Unlike single-level research studies in which effect sizes are regularly reported, effect size statistics for multilevel studies, in particular standardized mean difference, are still not fully investigated. Effect size is extremely important because it directly quantifies the effect of interest (e.g., the effect of the treatment, gender difference), regardless of whether the study consists of single-level or multilevel data.

Our article has included analytically derived formulae of the standardized mean difference for fully and partially cross-classified treatment-control arm designs, as well as methods for obtaining the effect size when reliable and consistent estimates of variance components are available. Although the analytical formulae for D_1 are tedious to use and can lose efficiency when the design is unbalanced or when the sample size is small, they are nevertheless important. In secondary analyses and meta-analyses where the clustering is not taken into account in the original analyses or when information about the variance

components is not available, D_1 can still be computed when the following information are available: number of clusters, cluster size, and intraclass correlations are available. For occasions where intraclass correlations are not available, Hedges (2007) provided an example of substituting values reported from other studies or with an educated guess, and Ahn et al. (2012) suggested quantitative procedures to estimate the ICCs. In addition, one can perform sensitivity analyses to examine whether different choices of ICCs result in substantial differences in the estimated effect size and the corresponding standard errors (Hedges, 2007, 2011).

We have also suggested a method to estimate effect size D_2 using maximum likelihood or Bayesian estimates of variance components. It is easier to implement and we thus recommend its use when raw data are available. To facilitate future replication and research synthesis we also recommend researchers analyzing primary data to report the effect size and its sampling variance, or at least the estimated values and the sampling variances (or the standard errors) of the variance components.

Given the complexity associated with the effect size estimation equation for CCREMs, a logical question would be when a researcher can ignore one level of clustering (i.e., random effect B) but still get a good estimate of the effect size and the sampling variance. We have reanalyzed the two real data examples by ignoring one level of clustering, and it appears that when the two crossed random effects are highly correlated, omitting one random effect does not lead to substantial differences in point and interval estimates of effect size. This makes sense because when the two effects share a lot of common information, and most of the information is still preserved when one effect was omitted (see Luo & Kwok, 2009). On the other hand, if the crossed random effects were only weakly correlated or uncorrelated (such as when the design is balanced), in general the bias on the estimated sampling variance increases when number of clusters K and the intraclass correlation ρ_B of the omitted random factor is large, based on

equations (7) and (8). For example, assuming a balanced design, when $\delta_T = 0.5$, $J = K = 30$, $n = 1$, $\rho_A = 0.25$, $\rho_B = 0.1$, ignoring the clustering of B results in an underestimation of $V(D_1)$ by 19.1% (from 0.046 to 0.037); when $n = 10$ and other things being unchanged but $\rho_B = 0.2$, then $V(D_1)$ is underestimated by 32.2% (from 0.055 to 0.037); and when $n = 1$ but K is doubled to 60, $V(D_1)$ is underestimated by 32.8% (from 0.028 to 0.019); This would result in CIs of the effect size that are too narrow and not valid (see Hedges, 2011). Nevertheless, because CCREMs are complex models, further studies are needed to fully understand the impact of ignoring one or more levels of clustering on effect size estimation in real research.

The present article is limited to only two-level CCREM with two crossed random effects, which is an extension of two-level multilevel models. However, the framework can be extended to CCREMs with three or more levels and with three or more random effects, or to CCREMs where treatment is defined as a level-1 variable (i.e., the treatment is individually-randomized). Future research can investigate perhaps effect size estimations in more complicated designs, as well as in other models in the multilevel family such as the multiple membership models. Simulation studies comparing different variance component estimation methods (e.g., Bayesian vs. REML) in the process of computing effect size are also highly encouraged. Also, in this paper we assumed that the effect size of interest has σ_T as the denominator for standardization. There are occasions where researchers may be interested in effect size with σ_B or σ_W or other alternatives as the denominator, but they are left for discussions in future studies. Finally, procedures to convert standardized mean difference effect sizes with multilevel structure into proportion of variance accounted for effect sizes (see Luo & Kwok, 2010) will be highly valuable for research synthesis methodology.

Notes

[1]Relative bias = $(\sum_{i=1}^{500} D^{(i)}/500 - \delta_T)/\delta_T$, where $D^{(i)}$ is the computed effect size D_1 or D_2 for the ith replication. Relative SE bias = $[\sum_{i=1}^{500} \hat{SE}(D^{(i)})/500 - SD(D)]/SD(D)$, where $\hat{SE}(D^{(i)})$ is the estimated standard error of D for the ith replication, and $SD(D)$ is the standard deviation of D across 500 replications. Ninety-five percent confidence interval is computed as $[D^{(i)} - 1.96 \times \hat{SE}(D^{(i)}), D^{(i)} + 1.96 \times \hat{SE}(D^{(i)})]$.

CHAPTER III

STANDARDIZED MEAN DIFFERENCES IN TWO-LEVEL PARTIALLY NESTED MODELS

Overview

The present paper discussed two methods to obtain standardized mean difference effect size and the corresponding sampling variance for partially nested cluster randomized designs. The first method requires input of summary statistics such as observed means, variances, and intraclass correlation, and would be useful for meta-analyses and secondary data analyses. The second method takes estimated variance components as input and would be of interest for primary researchers. The simulation results showed that the two methods were unbiased and had adequate confidence interval coverage, although the first method underestimated the variability of D when cluster sizes were small, intraclass correlation was high, and the distribution of the cluster sizes was extremely unbalanced. Real data from a youth preventive program are used to demonstrate the method. Furthermore, I also discuss biases on D under incorrect modeling of partially nested data, and show that the bias increases with larger intraclass correlation and cluster size.

Introduction

Effect size statistic is important in educational research. Indeed, it is the core concept in statistics reform in the behavioral sciences (Cumming, 2014; Kline, 2013; Wilkinson & Task Force on Statistical Inference, 1999). For primary researcher, it is crucial in the designing phase for sample size planning in order to achieve a desired level of statistical power or precision in parameter estimation (Kelley, 2013); In the analysis and interpretation phase it also gives a sense about the magnitude of a treatment or an

intervention (Ellis, 2010; Nakagawa & Cuthill, 2007). For meta-analysts, it is the building block of their research that summarizes and synthesizes a bunch of mixed findings (Lipsey & Wilson, 2001). The American Educational Research Association (2006) explicitly recommended using effect size statistics to interpret research findings. However, whereas effect size reporting has become more common for single-level studies (Peng et al., 2013), it is still rare for multilevel studies. This manuscript aims to provide methods to obtain effect size estimates with a special but not uncommon multilevel design—partially nested design.

Brief Review on Single-Level Effect Size

As discussed in Nakagawa and Cuthill (2007) and Peng and Chen (2014), there are multiple definitions of effect size, some with reference to a null hypothesis (Grissom & Kim, 2012; Kramer & Rosenthal, 1999; Thompson, 2002), some as a population parameter (Hedges, 1981), and some as a sample estimator (Nakagawa & Cuthill, 2007). In a recent paper, Kelley and Preacher (2012) defines effect size as broad as "a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" (p. 137), which may include many index that are not generally regarded as an effect size[1]. Regardless of the definition, however, most of them include the essential characteristic that effect size should be able to quantify the magnitude of an effect, where an effect can be in the context of intervention, prediction, or causation. This will form the working definition of effect size for this manuscript.

Effect size measures are well-developed in single-level studies. For experimental or quasi-experimental studies with two groups, or *arms* as used in this paper to avoid confusion with clusters, a straight forward effect size measure is the mean difference between the treatment arm and the control arm in the original metric of the outcome measure. However, the metric in psychological measurement usually does not have

intrinsic meaning (Blanton & Jaccard, 2006; Sechrest, McKnight, & McKnight, 1996), and the mean difference in the original metric may not be interpretable. Also, different studies may use different measure for the same outcome construct, and so raw mean differences may not be comparable across studies. As a compromise, the mean difference is convert to standard deviation unit, or *standardized*, in order to establish a common ground for cross-study comparison. Peng et al. (2013) synthesized 16 articles reviewing effect size reporting practices after 1999, and found that *standardized mean difference* (in particular Cohen's d) is among the two most commonly reported effect size statistic, alongside with the unadjusted R^2, or variance accounted for effect size.

Contrary to the trend in single-level studies, for multilevel data R^2-type of effect size was more dominant, and multilevel R^2 was discussed much earlier in the literature (e.g., Snijders & Bosker, 1994). Methods to estimate standardized mean differences, denoted as D in this manuscript, were first formally proposed by Hedges only in 2007 for two-level cluster-randomized trials. In this paper I aim to extend the work of Hedges to partially nested designs, where clustering is limited to one but not the other arm (see Bauer et al., 2008; Lee & Thompson, 2005; Moerbeek & Wong, 2008; C. Roberts & Roberts, 2005). Specifically I propose two approaches, one useful for meta-analysts and the other useful for primary researchers, for estimating D and its sampling variance (or standard errors). An example is given for the use of the formulae in real data, and the consequence on the estimated effect size of ignoring the clustered structure in the data would be discussed.

Effect Size With Partially Nested Design

Because of their ability to provide the strongest evidence for causal inference when properly implemented, randomized experiments has 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 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. As another example, for a study of family therapies, it is not reasonable to assign family members to receive different interventions given that the intervention itself has family as its unit. In such studies where data have a naturally clustered structure, multilevel modeling has long been suggested as a flexible technique which accounts for the non-independence among observations (Goldstein, 1986; Mason et al., 1983; Raudenbush & Bryk, 2002).

Nevertheless, the clustered structure may not be the same in different treatment arms. In some cases the 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 were assigned to self-study condition. In another randomized trial Kirschner, Paas, Kirschner, and Janssen (2011) compared the effects of collaborative learning and individual learning. Following previous literature I call such kind of data structure *partially nested* (e.g., Bauer et al., 2008; Moerbeek & Wong, 2008). Bauer et al. (2008) found in their literature review 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. Later Sanders (2011) found that 13% of experiments in educational research in 2007 to 2009 with 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., Korendijk (2012, chapter 4), and Sanders, the

first approach resulted in the underestimation of the standard errors of the treatment effect, whereas the second approach resulted in biased estimates of the treatment effect when the within-cluster variance in the treatment arm is different than that within the control arm (also called *heteroscedasticity*), and also biased estimates of variance components.

Although multilevel modeling techniques has been studied in the methodology literature for decades, only recently did researchers define and discuss effect size measures for clustered randomized studies (Hedges, 2007, 2009, 2011). In the following sections I would introduce the notations, suggest methods to obtain D and $V(D)$ (where $V(\cdot)$ denotes the variance operator), as well as confidence interval (CI) for D; illustrate the methods with real data; and discuss the impact of ignoring the clustering for both primary studies and meta-analyses.

Model and Notations

Consider the situation outlined in Bauer et al. (2008), where participants were randomly assigned to 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 and Y_j^C be the scores of the outcome Y for the ith observation in the jth cluster of the treatment arm and for the jth observation in the control arm respectively. Note that with this setting I treat each observation in the control arm as a pseudo cluster (Sanders, 2011). Denote the sample size of the treatment arm and of the control arm as N^T and N^C, with the total sample size $N = N^T + N^C$. In the treatment arm, let J be the number of clusters with index $j = 1, \ldots, J$, and let $i = 1, \ldots, n_j$ be the index of the observation within the jth cluster in the treatment arm. In a balanced design we have $n_1 = \ldots = n_j = n$, and thus $N^T = Jn$. In the control arm, $j = J + 1, \ldots, J + N^C$, and the i subscript is dropped. Let $\bar{Y}_{\bullet\bullet}^T$ and $\bar{Y}_{\bullet}^C$ be the grand means of the treatment arm and of the control arm respectively, and $\bar{Y}_{\bullet j}^T$ be the mean of the jth

cluster in the treatment arm. When the pooled within-cluster variance in the treatment arm equals the variance of the control arm, a situation described in Bauer et al. (2008), I can let S_W^2 be the pooled within-cluster level variance, where

$$S_W^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}\left(Y_{ij}^T - \bar{Y}_{\bullet j}^T\right)^2 + \sum_{j=J+1}^{J+N^C}\left(Y_j^C - \bar{Y}_{\bullet}^C\right)^2}{N - J - 1}, \tag{17}$$

and let $S_{B|\text{TREAT}}^2$ be the between-cluster mean squares in the treatment arm, where

$$S_{B|\text{TREAT}}^2 = \frac{\sum_{j=1}^{J} n_j\left(\bar{Y}_{\bullet j}^T - \bar{Y}_{\bullet\bullet}^T\right)^2}{J - 1}. \tag{18}$$

In this manuscript I mainly consider situations where equal within-level variance hold. See Moerbeek and Wong (2008) for discussion when heteroscedasticity is present.

A model predicting the response variable Y_{ij} can then be conceptualized by the level-1 model (Bauer et al., 2008)

$$Y_{ij} = \beta_{0j} + \beta_{1j}(\text{TREAT}_{ij}) + \varepsilon_{ij}, \tag{19}$$

and the level-2 model

$$\beta_{0j} = \gamma_{00}, \tag{20}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}. \tag{21}$$

Here β_{0j} is the within-cluster regression intercept for cluster j, which is assumed fixed across clusters and equals γ_{00}. In a balanced design γ_{00} equals $\bar{Y}_{\bullet}^C$. β_{1j} can be regarded as the difference between $\bar{Y}_{\bullet j}^T$ and $\bar{Y}_{\bullet}^C$, and under a balanced design its mean across all js is

$\gamma_{10} = \bar{Y}^T_{\bullet\bullet} - \bar{Y}^C_{\bullet}$. The cluster-specific random effect is captured by u_{1j} with $V(u_{1j}) = \sigma^2_B$. ε_{ij} is the level-1 residual, and its variance, $V(\varepsilon_{ij}) = \sigma^2_W$, is assumed constant across both the treatment and the control arms, which is reasonable when the clustering involves random assignment and the treatment effect does not change the within-cluster variability. Note that the sum of the variance components within the treatment arm is $\sigma^2_W + \sigma^2_B$, whereas that within the control arm is only σ^2_W. Thus, the within treatment arm variances differ unless $\sigma^2_B = 0$. Define the intraclass correlation (ICC) for the treatment arm as ρ, where

$$\rho = \frac{\sigma^2_B}{\sigma^2_W + \sigma^2_B}. \tag{22}$$

Such a model can be easily analyzed in common statistical packages for multilevel modeling (Baldwin, Bauer, Stice, & Rohde, 2011; Bauer et al., 2008), or can be reparameterized and analyzed with structural equation modeling (SEM) software (Sterba et al., 2014).

Effect Size Estimation Using Summary Statistics

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

$$\delta = \frac{\Delta\mu}{\sigma}, \tag{23}$$

where $\Delta\mu$ is the population treatment effect (i.e., the mean difference between the two arms) and σ is the pooled within treatment standard deviation. Hedges (2007) commented that with multilevel data, the concept of effect size is vague. That happens because σ can refer to σ_W (with homoscedasticity assumed), σ_B, or $\sqrt{\sigma^2_W + \sigma^2_B}$, each with a different target of generalization. For example, choosing σ_W implies looking at the average treatment effect within cluster, and choosing $\sqrt{\sigma^2_W + \sigma^2_B}$ implies looking at the effect size

in a population that is naturally clustered. For partially nested data such as the example given in Bauer et al. (2008), because the clustering is part of the treatment and does not naturally occur in the general population, σ_W would be a better choice to define δ.

Using summary statistics and assuming an approximately balanced design, the effect size is

$$D_1 = \frac{\Delta\bar{Y}}{S_W}, \tag{24}$$

and

$$V(D_1) = \frac{1 + (n-1)\rho}{N^T(1-\rho)} + \frac{1}{N^C} + \frac{D_1^2}{2(N-J-1)}, \tag{25}$$

where $\Delta\bar{Y} = \gamma_{10} = \bar{Y}^T_{\bullet\bullet} - \bar{Y}^C_{\bullet}$, S_W as defined in (17), and using the equality

$$\mathrm{E}(S^2_{B|\mathrm{TREAT}}) = \sigma^2_W + n\sigma^2_B, \tag{26}$$

ρ can be estimated by $(S^2_{B|\mathrm{TREAT}} - S^2_W)/[S^2_{B|\mathrm{TREAT}} + (n-1)S^2_W]$. The derivation of (24) and (25) can be found in Appendix C. Note that for unbalanced designs, one can replace n with the mean of n_j, that is, $\bar{n} = N^T/J$; However, the grand mean is no longer an efficient estimator of the mean of the control arm, and so D_1 is not efficient (i.e., variance of D_1 is larger than the second method described below).

Effect Size Estimation Using Estimated Variance Components

If consistent estimates of γ_{10} (fixed effect), σ_W (random effect), and their associated estimated variances (or standard errors) are accessible, one can use the following equations based on the estimated variance components (see Appendix C for

derivation)

$$D_2 = \frac{\hat{\gamma}_{10}}{\hat{\sigma}_W}, \tag{27}$$

$$V(D_2) = \frac{V(\hat{\gamma}_{10})}{\hat{\sigma}^2_W} + \frac{D_2^2 V(\hat{\sigma}^2_W)}{4\hat{\sigma}^4_W}. \tag{28}$$

If maximum likelihood estimates of γ_{10} and σ_W, which are, under general conditions, asymptotically unbiased, consistent (i.e., converged to the population value), and efficient (i.e., with minimum variance), then D_2 is also asymptotically unbiased, consistent, and efficient, even for conditions with unbalanced data. Thus, when relevant information is available, D_2 is a better estimator than D_1.

Constructing Approximate Confidence Interval for *D*

Like any other point estimates such as the sample mean, the sample D_1 and D_2 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 American Psychological Association (2010) strongly encouraged the reporting of CI alongside with an effect size measures.

Based on the Central Limit Theorem, both D_1 and D_2 will be normally distributed with a large sample size. Therefore, an approximate $(1 - \alpha) \times 100\%$ CI for D_1 and D_2 would be

$$[\hat{D} + z_{1-\alpha/2} SE(\hat{D}), \hat{D} + z_{1-\alpha/2} SE(\hat{D})],$$

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 $z_{.975} \approx 1.96$.

As noted in Hedges (2007, pp. 371–379), under the model described in equations (19) to (21) with normally distributed residuals, both D_1 and D_2, when

multiplied by a constant, follow an approximate noncentral t distribution. Because the noncentral t distribution is skewed, Cumming and Finch (2001) warned that the common rule of thumb of using approximation with a normal distribution when degrees of freedom is larger than 30 may not hold. Nevertheless, in our simulation conditions (see Appendix D), with $df > 75$ the asymptotic intervals closely matched the noncentral t interval in terms of coverage probability and width. As clustered data commonly has a large size, the simple asymptotic method will be sufficient for many occasions. Nevertheless, for small sample the noncentral method can be used, as described in Appendix D.

A Simulation Study Comparing the Performance of D_1 and D_2

A simulation study was used to check the performance of D_1 and D_2, and their analytically derived variances. For each condition 500 data sets were generated in R (R Core Team, 2014) using the above model defined in equations (19) to (21) with $\delta = 0.5$. Because the methods were derived analytically, the purpose of the simulation was mainly to check how robust D_1 and D_2 are under extreme conditions, including small cluster size, few number of clusters, and extreme unbalanced cluster sizes.

Design Factors

The simulation employed a $2 \times 2 \times 2 \times 2 \times 2$ design with five design factors, as described below.

Intraclass correlation, ICC. The two conditions of ICC were .1 and .5. The former represents the normal ICC level in education (Hedges & Hedberg, 2007) and the latter represents an extreme case.

Number of clusters in the treatment arm, J. According to Kreft and De Leeuw (1998), 30 is the recommended minimum sample size for using multilevel modeling. In this simulation I included conditions with either 15 or 30 clusters to represent extreme

and minimum J values.

Average cluster size, $\bar{n}$. The $\bar{n}$ values were either 5 or 25. The former represents an insufficient level by most standards (e.g. Hox, 2010), but is nevertheless typical for longitudinal or family data. The latter is chosen to represent typical classroom size in the US.

Sample size ratio between the two arms, $N^T : N^C$. The clustering in the treatment arm reduces information contained in the sample. That is, even when the level-1 sample sizes for both the treatment and the control arms are equal, the treatment arm with clustering has a smaller *effective sample size* (see Hox, 2010; Moerbeek & Wong, 2008). In this simulation I used the conditions where $N^T : N^C = 1$ or 5.

Distribution of cluster sizes in the treatment arm. Although for simplicity, in the previous derivation of D_1 I assumed equal cluster sizes, such an assumption seldom holds in real research. Even when equal cluster size was emphasized in research planning, nonresponses due to various reasons render the final sample unbalanced. Unequal cluster sizes further reduces the effective sample size, particularly when the variability of the cluster sizes are large relative to the mean (Candel & Breukelen, 2009). Therefore, I would also investigate the impact of unequal cluster sizes on the performances of D_1 and D_2. Because the cluster size can be considered count data with strictly positive values, a suitable distribution to model cluster sizes would be the zero-truncated negative binomial[2](see James, 1953, for an example in modeling size of pedestrian groups). The larger the variance of the negative binomial, the more unbalanced the cluster sizes. For this simulation I generate the group sizes from a zero-truncated Poisson (a special case where the variance roughly equals the mean) or from an extreme zero-truncated negative binomial with variance roughly 10 times the mean.

Data Generation and Analyses

For each of the 32 simulation conditions, I set σ_W^2 to 1.0, and the variance component σ_B^2 was computed based on the ICC. Then I use R (R Core Team, 2014) to generate 500 sets of between-level cluster sizes and cluster-specific effect u_1 (with mean 0 and variance σ_W^2, assuming normality) for the treatment arm. Then for both the treatment and the control arms, the individual-specific effect was generated (with mean 0 and variance σ_W^2, and the outcome scores were generated according to equations (19) to (21).

For each data set, D_1 and $V(D_1)$ were obtained as described in equations (24) and (25). To obtain D_2 and $V(D_2)$, I first analyzed each data set with the partially nested model using `lme4` (Bates, Maechler, Bolker, & Walker, 2014). Because `lme4` does not compute the estimated *SE* of the variance components (which is done intentionally; see Bates, 2011, for detail), I used the parametric bootstrap *SE* with 200 resamples instead. Then D_2 and $V(D_2)$ were calculated based on equations (27) and (28). The R code implementing the whole simulation were shown in Appendix F.

Evaluation Criteria

Relative *SE* bias. For both D_1 and D_2, the percentage relative *SE* bias was computed as

$$\frac{\sum_{i=1}^{500} \hat{SE}(D^{(i)})/500 - SD(\mathbf{D})}{SD(\mathbf{D})} \times 100\%,$$

where $\hat{SE}(D^{(i)})$ is the squared root value of the estimated variance of D for the ith replication, and $SD(\mathbf{D})$ is the *SD* of the 500 estimated values of D, and is also denoted as the *empirical SE*. Following Hoogland and Boomsma (1998), a relative *SE* bias with an absolute value larger than 10% would be considered *unacceptable*.

Coverage of 95% CI. Ideally a 95% CI should have have a .95 confidence coefficient, which in the frequentist sense means that in the long run, 95% of the CIs

constructed at the sample level should include the population parameter. However, due to sampling variations, different degrees of approximations, and violations of assumptions, the results from the simulation will show deviations. In this simulation, the empirical coverage percentage is computed as

$$\frac{\text{Number of replications with CI covering } \delta}{500}$$

Following L. K. Muthén and Muthén (2002), I consider coverage between 91% and 98% as acceptable.

Root Mean Squared Squares (*RMSE*). Even though both D_1 and D_2 are roughly unbiased estimators of δ, under unbalanced cluster sizes D_1 is expected to be inefficient, meaning that it will have a larger sampling variance. Therefore, the *RMSE* is also computed, which is defined as

$$\sqrt{\frac{\sum_{i=1}^{500}(D^{(i)} - \delta)^2}{500}}.$$

When both of the estimators are approximately unbiased, *RMSE* mainly reflects their sampling variability. An estimator with lower *RMSE* is more *efficient*, and thus is preferred.

Simulation Results

As expected, both D_1 and D_2 were approximately unbiased (with relative bias $< 5\%$). Table 4 showed the simulation results. In terms of relative *SE* bias, both D_1 and D_2 were in acceptable range when cluster sizes followed a Poisson distribution. Under extreme unbalanced conditions, however, D_1 showed substantial *SE* bias when ICC = .5 (*SE* was underestimated by 11% to 26%). The coverage of CI was substantially smaller and ranged from 84% to 91% for D_1 under those conditions. On the other hand, the *SE*

bias of D_2 remain less than 10%, and the CI for D_2 showed adequate coverage. Under conditions where D_1 showed substantial *SE* bias, D_1 had a larger *RMSE* and was thus less efficient. Otherwise the efficiency was similar between the two estimators.

Real Data Illustration

The summary of the multilevel analysis provided in Model 1 of Bauer et al. (2008, p. 231) would be used to demonstrate the usage equations (27) and (28). The data concerned the effectiveness of the Reconnecting Youth (RY) preventive intervention program, which involved grouping 325 adolescents into 41 classes. There were two other comparison arms called *control* (n = 675) and *typical* (n = 598) that did not receive treatment and were not clustered into higher level units. The outcome variable is deviant peer bonding. The fixed effects included dummy variables representing the memberships of the treatment arm and of the *typical* arms, as well as those representing the schools they attended. The two random effects were the person level residuals (which was assumed constant across arms) and the class level residuals.

Here I only focused on the treatment effect of RY compared to *control*, which had a coefficient $\hat{\gamma}_{10} = 0.19$. Using equation (27) it is straight forward to see that the effect size of RY = $0.19/\sqrt{0.789} = 0.214$. Bauer et al. (2008) did not report the sampling variance nor the standard error for the effect of RY. However, they did report that the t value equaled 2.63, and thus *SE* of the effect of RY could be estimated as $0.19/2.63 = 0.0722$. Similarly, for the level-1 residual variance, its standard error could be obtained as $0.789/26.73 = 0.0295$. Substituting $V(\hat{\gamma}_{10}) = 0.0722^2$, $D_2 = 0.214$, $\hat{\sigma}^2_W = 0.789$, and $V(\hat{\sigma}^2_W) = 0.0295^2$ into the formula for $V(D_2)$, that is, equation (28), I got $0.0722^2/0.789 + (0.214^2)(0.0295^2)/(4 \times 0.789^2) = 0.0066$ (or *SE* = 0.0814). Then the approximate 95% symmetric confidence interval could be obtained as $0.214 \pm z_{.025}(0.0814)$ (where $z_{.025}$ is the .25 quantile for the standard normal

Table 4
Percentage Relative Standard Error Bias and Mean Squared Errors of D_1 and D_2 Across Different Conditions

Distribution					RBias($SE(\hat{D})$)		95% CI Coverage		$RMSE(D)$	
of n_j	J	ρ	$N^T : N^C$	average n	D_1	D_2	D_1	D_2	D_1	D_2
Poisson	15	.10	1	5	1.6	3.5	95.6	96.2	0.18	0.18
				25	−9.3	−8.0	92.6	92.4	0.12	0.12
			5	5	1.4	−0.2	94.4	94.0	0.29	0.30
				25	0.3	0.6	95.4	95.6	0.15	0.15
		.50	1	5	−8.1	−0.1	92.0	93.6	0.33	0.31
				25	−8.8	−6.5	91.4	91.6	0.29	0.28
			5	5	−2.7	−2.1	94.8	94.8	0.39	0.40
				25	0.0	1.6	94.8	94.4	0.28	0.28
	30	.10	1	5	2.2	4.2	96.2	97.2	0.13	0.13
				25	0.0	0.4	94.8	95.0	0.08	0.08
			5	5	4.8	4.1	95.6	95.4	0.20	0.20
				25	1.7	2.6	95.2	95.0	0.11	0.11
		.50	1	5	−1.4	4.6	94.0	94.2	0.22	0.21
				25	−0.8	0.1	93.6	93.6	0.19	0.19
			5	5	1.8	4.7	95.8	96.2	0.27	0.26
				25	0.4	2.5	94.0	94.6	0.20	0.20
NB	15	.10	1	5	−5.3	3.1	93.8	96.2	0.19	0.19
				25	−17.7	−6.9	89.2	92.2	0.13	0.12
			5	5	−3.2	−1.6	94.0	94.2	0.31	0.31
				25	−4.3	1.1	94.2	96.0	0.16	0.15
		.50	1	5	−26.0	−0.3	84.2	93.8	0.40	0.32
				25	−21.8	−5.8	87.0	92.0	0.33	0.28
			5	5	−20.4	−3.6	88.4	96.2	0.48	0.41
				25	−12.1	2.2	91.8	94.4	0.32	0.28
	30	.10	1	5	−8.6	−0.2	93.0	95.0	0.14	0.14
				25	−9.4	−0.1	92.4	95.4	0.09	0.08
			5	5	0.1	4.0	96.0	96.4	0.21	0.21
				25	−3.2	2.1	94.6	95.8	0.11	0.11
		.50	1	5	−23.8	0.9	89.2	95.8	0.28	0.23
				25	−14.7	−0.2	91.0	94.4	0.22	0.19
			5	5	−15.6	4.3	91.2	95.2	0.32	0.27
				25	−11.1	1.8	91.6	94.0	0.23	0.20

Note. The population effect size is $\delta = 0.5$. ρ = intraclass correlation of the treatment arm. n = average cluster size. RBias($SE(\hat{\theta})$) = percentage relative standard error bias, which is calculated as $[\sum SE(\hat{\theta}_j)/R - SD(\hat{\boldsymbol{\theta}})]/SD(\hat{\boldsymbol{\theta}}) \times 100$, where $SE(\hat{\theta}_j)$ is the estimated standard error for the jth replication, R is the number of replications, $SD(\hat{\boldsymbol{\theta}})$ is the standard deviation of the R estimates of θ. $RMSE$ = mean squared error [$RMSE(D)$ = bias$(D) + V(D)$]. NB = negative binomial distribution with mean equals to the average n and variance approximately equals to 3 × average n.

distribution), which equals [0.054, 0.374].

Effect Size Using Only the *SD* of the Control Arm

For single level study, Glass (1976) suggested to compute 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, could be affected by the treatment. The effect size δ^C would be defined as

$$\delta^C = \frac{\Delta\mu}{\sigma^C}, \tag{29}$$

where σ^C is the standard deviation of the control group. First I define the within-cluster variance of the treatment group and the variance of the control group as

$$S_W^2 = \frac{\sum_{j=1}^{J}\sum_{i=1}^{n}(Y_{ij}^T - \bar{Y}_{\bullet\bullet}^T)^2}{N^T - J},$$
$$S_C^2 = \frac{\sum_{j=J+1}^{J+N^C}(Y_j^C - \bar{Y}_{\bullet}^C)^2}{N^C - 1}.$$

A sample estimator D_1^C can be obtained as

$$D_1^C = \frac{\Delta\bar{Y}}{S_C} \tag{30}$$

$$V(D_1^C) = \kappa\frac{1 + (n-1)\rho}{N^T(1-\rho)} + \frac{1}{N^C} + \frac{(D_1^C)^2}{2(N^C - 1)}, \tag{31}$$

where $\kappa = S_W^2/S_C^2$ is the estimated variance ratio between the treatment and the control arms. Note that $V(D_1^C) > V(D_1)$ when $\kappa = 1$, so D_1 is preferred when variance can be assumed equal.

If reasonable point and variance estimates of the variance components for σ_W^2 and

σ_C^2 can be obtained, then D_2^C and its sampling variance can be estimated as

$$D_2^C = \frac{\hat{\gamma}_{10}}{\hat{\sigma}_C}, \tag{32}$$

$$V(D_2^C) = \frac{V(\hat{\gamma}_{10})}{\hat{\sigma}_C^2} + \frac{D_2^2 V(\hat{\sigma}_C^2)}{4\hat{\sigma}_C^4}. \tag{33}$$

Effects of Ignoring the Clustering Structure on *D*

When the clustering in the treatment arm is not modeled, the impacts on D and $V(D)$ are functions of the intraclass correlation ρ, the average cluster size n, and the total sample size ratio of the treatment arm and the control arm. In general it leads to underestimation of both D and its sampling variance. Table 5 showed the expected percentage relative bias of the estimated D and its estimated variance for some combinations of ρ and n when the sample sizes for both the treatment and the control arms equaled to 200. Even for a small intraclass correlation of .10, $V(D)$ would be underestimated by 9 to 15% for n between 2 to 8. For $\rho = .50$ and $n = 8$, the effect size is expected to be underestimated by 18% whereas its sampling variance is expected to be underestimated by 60%. Increases in both ρ and n, which contribute to the increase in *design effect* $= 1 + (n - 1)\rho$ (see B. O. Muthén & Satorra, 1995), lead to more severe underestimated $V(D)$, whereas only increases in ρ leads to more severe underestimation of the effect size point estimate.

Because one of the most popular way in combining multiple effect sizes in a meta-analysis is to use inverse variance weights (Lipsey & Wilson, 2001), an underestimated $V(D)$ can lead to biased results. Because cluster-randomized trials usually have a medium to large level-1 sample size, ignoring the clustering in those studies may incorrectly lead to results that are largely only driven by a few studies.

Table 5
Percentage Relative Bias of Effect Size and Its Variance When Clustering of the Treatment Group is Ignored

ρ	average n	RBias(D)	RBias[$V(D)$]
.10	2	−2.66	−8.90
	4	−2.65	−10.68
	8	−2.64	−14.49
.30	2	−9.23	−25.04
	4	−9.21	−30.23
	8	−9.17	−38.95
.50	2	−18.32	−43.19
	4	−18.28	−50.00
	8	−18.21	−59.77

Note. ρ = intraclass correlation of the treatment arm. n = average cluster size. RBias(θ) = percentage relative bias, which is calculated as $(\sum \hat{\theta}_j / R - \theta_{\text{true}})/\theta_{\text{true}} \times 100$.

Conclusion

The present paper proposed two methods to estimate effect size D for partially nested design. This helps primary researchers working with such designs to appreciate the practical significance of their results, and is a tool for meta-analysts synthesizing effects of group interventions. I also showed that when the clustering of one arm is not accounted for, the estimated D and $V(D)$ showed negative bias, and the degree of bias was magnified with larger design effect. Educators working with similar designs should incorporate effect size presented here in addition to statistical significance to evaluate treatment efficacy, and for studies with large sample size the point and interval estimates of D are much more informative.

There are a few limitations of this paper. First, the calculations of D_1 and D_2, and particularly their variances, can be tedious. Researchers in substantive areas may prefer

more automated procedures. In Appendix E I presented sample codes for estimating D_2 using the SEM approach in Mplus. Future study may investigate other methods such as bootstrapping. Second, the simulation results in this study only apply to the simple situation with two arms and no covariates. Impact of additional complexity on effect size estimation can be further addressed in the future.

Notes

[1]For example, under such conceptualization a p-value may also be called an effect size, if the "question of interest" is something like the likelihood that the treatment has an effect. This may be somehow counterintuitive.

[2]In the simulation, I generate zero-truncated negative binomial numbers as follow: (a) Get $f(0)$, the cumulative density at 0, in the given negative binomial distribution; (b) Generate a uniform random value, u in the range $[f(0), 1]$; (c) Get the u quantile of the given negative binomial distribution. Appendix F shows a R functional `ZeroTruncate` that convert a standard distribution to the zero-truncated version.

CHAPTER IV

BOOTSTRAP CONFIDENCE INTERVAL FOR MULTILEVEL EFFECT SIZE

Overview

Although many methodologists have urged the use of effect size measures accompanying tests of statistical significance, discussions on obtaining confidence intervals (CIs) multilevel effect sizes has been rare. In this manuscript I explore the bootstrap as a viable and accessible alternative for obtaining CIs for multilevel standardized mean differences. A simulation is carried out to compare 10 procedures for constructing CIs in terms of empirical coverage probability. Results showed that, across all simulation conditions, the semiparametric bootstrap with the bias-corrected and accelerated CI and the model-based analytical methods with asymptotic symmetric CI performed the best, with the former being more robust to violation of the normality assumption.

Introduction

Although many methodologists have urged the use of effect size measures accompanying tests of statistical significance (e.g. Cohen, 1990; Cumming, 2014; Kelley & Preacher, 2012; Thompson, 2007), discussions on effect size estimation for multilevel data has been rare (e.g. Hedges, 2007; Snijders & Bosker, 1994). Much rarer is the discussion on obtaining interval estimates of multilevel effect size. One reason is that the computational formulas for confidence intervals (CIs) for effect size with multilevel data can be extremely complex (Hedges, 2007), even with the use of asymptotic theory that may not hold for finite samples. In addition, whereas traditional single-level data can be regarded as one type of data structure (with the assumption of simple random sampling), multilevel data comprise a collection of data structures with varying numbers of clustering

levels and relations between levels (i.e., nested vs. crossed). This makes it tedious to derive complex formulas for CIs for each type of multilevel data structures. Recognizing such difficulties, in this manuscript I explore the *bootstrap* (Efron, 1982), a type of resampling techniques, as a viable alternative for obtaining CIs for multilevel effect size.

In the past two decades, effect size reporting has been the central theme in the "statistical reform" in the behavioral sciences (e.g. Kline, 2013; Thompson, 2002). Many professional organizations, including the American Educational Association (AERA, 2006), the American Psychological Association (APA, 2010), the International Committee of Medical Journal Editors (Schulz, Altman, Moher, & CONSORT Group, 2010), and the National Center for Education Statistics (NCES, 2012), have guidelines for reporting effect size.

In addition to reporting point estimates of effect size measures, many sources have also encouraged the use of CI to quantify the uncertainty associated with a sample effect size. For example, the APA publication manual (APA, 2010) stated that "[w]henever possible, [researchers should] provide a confidence interval for each effect size reported to indicate the precision of estimation of the effect size" (p. 34). A similar statement is found in the AERA (2006) reporting standards that "there should be included . . . [a]n indication of the uncertainty of that index of effect (such as a standard error or a confidence interval" (p. 37). Hedges (2008) and Thompson (2002) have also made similar recommendations.

Whereas point estimates of effect size measures, such as standardized mean difference (SMD) and proportion of variance accounted for (the d-family and the r-family; Grissom & Kim, 2012; Rosenthal, 1994), are regularly reported in single-level studies, the CIs are still rarely attached. Peng et al. (2013) reviewed 32 review papers of effect size reporting practices in published articles from 116 journals in education and psychology, and found that whereas the median effect size reporting rates were 58.0% after 1999, some of the review papers reported that the reporting rates for CIs for effect

size were essentially zero (Byrd, 2007; Fritz, Morris, & Richler, 2012). This is in sharp contradiction to the existing guidelines on effect size reporting.

Although different researchers have proposed methods for obtaining CIs for effect size, substantive researchers may not be familiar with them. Even for single-level studies, the analytical formulas for the sampling variance of effect size is not simple, and as would be discussed later some method for obtaining CIs invokes noncentral distributions, which is seldom part of the quantitative training for behavioral researchers. For the simplest multilevel structure with two nested levels, the variance of SMD already fills two lines of space (Hedges, 2007); For the more complicated cross-classified structure the variance of SMD takes three full lines. On the other hand, computer intensive methods such as the bootstrap (Efron & Tibshirani, 1993) requires computation of only the point but not the variance estimates, which greatly simplifies the analytical load on substantive researchers. It also has the added advantage of handling automatically some violations of assumptions such as the normality of random effects, which makes it an ideal method for obtaining CIs for effect size. Although the present study concerns mainly the use of the bootstrap for SMD with two-level data, the method can easily apply to other types of effect size measures and to more complicated data structures.

The Bootstrap

Efron (1982) has popularized the bootstrap method for obtaining standard errors and variances of some statistical estimators when closed form solutions are difficult or impossible to obtain. Probably one of the applications of the bootstrap that are most familiar to behavioral researchers is for mediation analyses (e.g., Preacher & Hayes, 2004). For mediation analyses the sampling distribution of indirect effects is in general skewed even when multivariate normality holds, making the standard procedures of significance testing and CI construction biased. Indeed, MacKinnon, Lockwood, and

Williams (2004) showed that the bias-corrected bootstrap outperformed other methods in the study for constructing CI for the indirect effect. For single-level studies, Kelley (2005) and Chen and Peng (2014) recommended the bootstrap method as the approach for estimating SMD, especially when the normality assumption is violated.

Although there are different types of bootstrap methods with different implementations, they generally follow the same general framework:

1. Get an approximated population distribution (formally called the cumulative distribution function, or CDF), denoted as $\hat{F}$, from the sample data $\mathbf{x}$;
2. Simulate a large number, R, of independent samples (i.e., sampling with replacement) from $\hat{F}$, each with the same size as the original sample $\mathbf{x}$, and denote the ith sample as $\mathbf{x}_i^*$;
3. Compute the target estimator $T(\mathbf{x}_i^*)$, such as the mean or the effect size, for each sample;
4. Obtain the empirical sampling distribution of T as the distribution of the R values of $T(\mathbf{x}_i^*)$.

Note that $\mathbf{X}$ can include more than one observed variables. After the empirical sampling distribution is obtained, various methods can be applied to obtain *SE*s and CIs for T. Usually R needs to be large (say 1000 or more) when CI is of interest (Davison & Hinkley, 1997).

Types of Bootstrap

Three variations of the bootstrap that has been discussed most often in the literature is the parametric bootstrap, the semiparametric bootstrap, and the nonparametric bootstrap (Davison & Hinkley, 1997; Efron & Tibshirani, 1993). The main differences among the three lie in how $\hat{F}$ is defined in step 1.

Parametric bootstrap. In the parametric bootstrap, the family of distribution is specified for $\hat{F}$, but the parameters of $\hat{F}$ are estimated from the sample data. For example, if a researcher is interested in the mean of a variable X and think that the distribution of X in the population is at least approximately normal, the researcher can specify $\hat{F}$ as the CDF of a normal random variable, denoted as $N(\bar{x}, s_x^2)$, where $\bar{x}$ and s_x^2 are the sample mean and variance of X. This method relies on the assumption that the family of distribution of $\hat{F}$ is specified correctly.

Semiparametric bootstrap. When there are at least two variables in the data, the semiparametric bootstrap can be used. For example, if one is interested in the relation between two variables, X and Y, from a sample of size N, one can specify the regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

and use least squares methods to estimate β_0, β_1, and the residual ε_i. If the distribution of ε_i is specified, then the model is fully parametric. However, in the semiparametric bootstrap, one resamples from the *empirical distribution* of ε_i rather than the joint distribution of X and Y. The empirical distribution of ε assigns a probability of $1/N$ to each of the N ε values, and thus is a discrete distribution. Each bootstrap sample $\mathbf{x}_i^*$ then includes the same original values of X and the new Y values are computed as

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i^*,$$

where ε_i^* is a resampled value of the residual from the empirical distribution. The method makes the assumption that the functional form between X and Y (e.g., linear or quadratic) is specified correctly; However, it makes no assumption on the residuals, thus the name *semiparametric* bootstrap is also called the semiparametric bootstrap (Davison &

Hinkley, 1997).

Nonparametric bootstrap. In the nonparametric bootstrap, each observation is assigned a probability of $1/N$. Bootstrap, samples each of size N, are then drawn with replacement from the original sample. Note that when a case is selected into a bootstrap sample, the observed values on all variables are kept. Thus, one can think of $\hat{F}$ as a discrete distribution with each element being the vector of observed values for an observation. Because no distributional assumptions are made on $\hat{F}$, it is called the *nonparametric* bootstrap.

As noted in Davison and Hinkley (1997), because the nonparametric bootstrap does not rely on distributional assumptions, it is expected to outperform the other two methods when the model is misspecified. On the other hand, When the model is specified correctly, the parametric bootstrap and the semiparametric bootstrap can produce more efficient results (i.e., with smaller sampling variance), and are more stable when N is small.

Models and Notations

As an initial effort to compare methods for constructing CIs with multilevel data, this study focuses on SMD for the simplest but most commonly used multilevel structure—the two-level hierarchical structure. There are plenty of examples of this structure in educational and psychological research, including students nested within classrooms and then schools, citizens nested within regions and countries, and employees nested within organizations. In many educational studies, researchers have no choice but to implement randomization and interventions at the classroom or the school level rather than at the individual level.

Methods for inference of treatment effect in cluster-randomized trials were developed a long time ago (e.g. Goldstein, 1986; Mason et al., 1983). In the multilevel

modeling framework, the level-1 model is

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}, \tag{34}$$

and the level-2 model is

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{TREATMENT}_j + u_{0j}. \tag{35}$$

In the model, y_{ij} is the outcome values of the ith individual in the jth cluster, TREATMENT is the dummy variable with 1 being the treatment arm and 0 being the control arm, and β_{0j} is the cluster mean of the jth cluster. γ_{00} is the grand mean of the control arm, and γ_{01} is the treatment effect that represents the grand mean difference between the treatment and the control arms. The level-1 and level-2 residuals are ε_{ij} and u_{0j}, which are assumed independent. Typically, one assumes $\varepsilon_{ij} \sim N(0, \sigma_W^2)$ and $u_{0j} \sim N(0, \sigma_B^2)$, where the variances are constant across clusters and treatment arms. The parameters can be estimated using standard statistical packages such as SPSS and SAS, as well as specialized programs such as HLM and MLwiN.

Obtaining CI for SMD With Two-Level Data

Effect size measures of these studies, on the other hand, were not discussed until recently (Hedges, 2007). The effect size of SMD is defined as the ratio between the treatment effect and the standard deviation (*SD*) of the outcome. Whereas such a definition causes no confusion for single-level studies, it is ambiguous in multilevel studies several different *SD*s can be used, including the within-cluster *SD*, the between-cluster *SD*, and the total *SD*. Hedges (2007) viewed the issue in the meta-analysis framework, and suggested that the choice should depend on the nature of other studies in the synthesis. For example, if in most other studies data are collected

from a single site, the within-cluster *SD* may be a better choice.

It is not a purpose of this study to argue which *SD* should be used. Indeed, any effect size can be estimated with the bootstrap as long as the estimator can obtained from the original sample one. I chose the total *SD* in this study simply because it uses more information in the data and theoretically can be converted to a variance accounted for effect size Snijders and Bosker (1994).

Analysis of Variance Method

Using the total *SD* of the outcome, the population SMD for a two-level treatment-control arm design is defined as

$$\delta_T = \frac{\gamma_{01}}{\sqrt{\sigma_W^2 + \sigma_B^2}}. \tag{36}$$

When each cluster contains the same number of observations n, in other words, when cluster sizes are constant, a sample estimator of δ_T can be defined as (Hedges, 2007, p. 349)

$$d_T = \left(\frac{\bar{Y}_{\bullet\bullet}^T - \bar{Y}_{\bullet\bullet}^C}{S_T}\right)\sqrt{1 - \frac{2(n-1)\rho}{N-2}}, \tag{37}$$

where $\bar{Y}_{\bullet\bullet}^T$ and $\bar{Y}_{\bullet\bullet}^C$ are the grand means of the outcome for the treatment and the control arms, ρ is the intraclass correlation (ICC), and S_T is the pooled total sample standard deviation such that

$$S_T^2 = \frac{\sum_{j=1}^{J^T}\sum_{i=1}^{n}(Y_{ij}^T - \bar{Y}_{\bullet\bullet}^T)^2 + \sum_{j=1}^{J^C}\sum_{i=1}^{n}(Y_{ij}^C - \bar{Y}_{\bullet\bullet}^C)^2}{N-2}$$

with J^T and J^C being the number of clusters in the treatment and the control arm respectively. The variance of d_T is

$$V(d_T) = \left(\frac{N^T + N^C}{N^T N^C}\right)(1 + (n-1)\rho) + \delta_T^2 \left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}{2[(N-2) - 2(n-1)\rho]^2}\right). \quad (38)$$

In reality δ is not known and so one has to replace it with d_T. Because the method is derived from decomposing the sum of squares, some authors denoted it as the analysis of variance (ANOVA) method (Searle et al., 2006).

When the cluster sizes are not constant, Hedges (2007) also derived a formula for $V(d_T)$. However, the formula is quite complex and requires the information about the size of each cluster, which usually happens only when researchers have the raw data. In that case the method described later would be much simpler to use, and thus should be preferred.

With the estimates d_T and $V(d_T)$, there are two methods to construct CI for d_T. First, based on the central limit theorem (see Casella & Berger, 2002), $d_T/\sqrt{V(d_T)}$ has an asymptotic standard normal distribution. Therefore, a symmetric $(1-\alpha) \times 100\%$ CI can be obtained as

$$[d_T - z_{1-\alpha/2}\sqrt{V(d_T)}, d_T + z_{1-\alpha/2}\sqrt{V(d_T)}], \quad (39)$$

where $z_{1-\alpha/2}$ denotes the $1-\alpha/2$ quantile of the standard normal distribution. As noted by Hedges (2007), the distribution of d_T is better approximated by a scaled noncentral t distribution. Cumming and Finch (2001) pointed out that the degrees of freedom of the noncentral t needs to be larger than 60 for the normal approximation to work properly. For small samples, it would be better to use the noncentral t distribution to form

asymmetric CI. In particular, the random variable

$$d_T \left(\frac{N^T N^C [1 + (n-1)\rho]}{N^T + N^C} \right)$$

has an approximate noncentral t distribution with noncentrality parameter $ncp = \delta_T N^T N^C [1 + (n-1)\rho]/(N^T + N^C)$ and degrees of freedom

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

Using the method discussed in Kelley (2005) and Steiger and Fouladi (1997), and replacing δ_T by d_T in the ncp, we can obtain an asymmetric $(1-\alpha) \times 100\%$ CI as

$$[t_{\alpha/2,\nu,ncp}/ncp, t_{1-\alpha/2,\nu,ncp}/ncp], \tag{40}$$

where $t_{\alpha/2,\nu,ncp}$ is the $\alpha/2$ quantile of the noncentral t distribution with degrees of freedom df.

Model-Based Approach

When raw data is available, researchers can fit the model as described in (34) and obtain $\hat{\gamma}_{01}$, $\hat{\sigma}^2_W$, and $\hat{\sigma}^2_B$ as the parameter estimates, as well as their variances (i.e., squared value of the *SE*s). Then the sample effect size can be estimated as (see Hedges, 2009)

$$\hat{\delta}_T = \frac{\hat{\gamma}_{01}}{\sqrt{\hat{\sigma}^2_W + \hat{\sigma}^2_B}} \tag{41}$$

with variance

$$V(\hat{\delta}_T) = \frac{V(\hat{\gamma}_{01})}{\hat{\sigma}^2_W + \hat{\sigma}^2_B} + \frac{\delta^2_T [V(\hat{\sigma}^2_W) + V(\hat{\sigma}^2_B)]}{4(\hat{\sigma}^2_W + \hat{\sigma}^2_B)^2}. \tag{42}$$

Note that some programs, such as the `lme4` package in R (R Core Team, 2014), does not report the asymptotic variance of the variance components. One can instead obtain the variance using parametric bootstrap with 200 resamples, which in my experience is quite close to the value based on asymptotic theory. Ames (2013) has shown that the model-based approach outperformed the ANOVA approach in terms of bias in the estimated variance, especially when the sample size was small.

The asymptotic symmetric CI can then obtained by replacing d_T and $V(d_T)$ by $\hat{\delta}_T$ and $V(\hat{\delta}_T)$ respectively in equation (39). The noncentral CI can similarly be obtained by plugging in

$$ncp = \sqrt{\frac{\hat{\sigma}^2_W + \hat{\sigma}^2_B}{V(\hat{\gamma}_{01})}}$$

$$df = \frac{2(\hat{\sigma}^2_W + \hat{\sigma}^2_B)^2}{V(\hat{\sigma}^2_W) + V(\hat{\sigma}^2_B)}$$

into expression (40).

Parametric Bootstrap

The parametric bootstrap for multilevel data (Goldstein, 2011a) looks like the semiparametric bootstrap discussed previously where residuals are sampled and new response variable is computed based on the model. What makes it parametric is that the distributions of the residuals are specified as normal. Specifically, ε^* is drawn from $N(0, \hat{\sigma}^2_W)$ and u^* is drawn from $N(0, \hat{\sigma}^2_B)$, and new response y^*_{ij} is computed based on (34) and (35). The multilevel model is then refitted to the new bootstrap data, and $\hat{\delta}^*_T$ is computed with equation (41).

There are numerous ways to construct CIs with the bootstrap (Davison & Hinkley, 1997; Efron & Tibshirani, 1993; Van der Leeden, Meijer, & Busing, 2008; Wu, 1986). For this study I only focus on two of them: *percentile* CI and the *bias-corrected and*

accelerated (BCa) CI. In the percentile CI (Efron & Tibshirani, 1993) with confidence level equal $(1 - \alpha)$, one simply obtain the $\alpha/2$ and the $1 - \alpha/2$ quantiles from the distribution of the bootstrapped values $\hat{\delta}^*_T$. The percentile CI has the advantage of being easy to compute and understand. It also does not make any distributional assumption on the estimator, so the two confidence limits need not be symmetric. On the other hand, as noted by Davison and Hinkley (1997) and Van der Leeden et al. (2008), the percentile CI tends to produce biased confidence limits, especially with the nonparametric bootstrap. Specifically it is not very accurate when the estimator is biased or when the original sample size is small (Efron & Tibshirani, 1993).

One can improve the performance of the percentile method with the (BCa) method. Like the percentile method, the BCa method took the two quantile values in the distribution of the bootstrapped values of the estimator. However, instead of using the $\alpha/2$ and the $1 - \alpha/2$ quantiles, one uses α_L and α_U defined as (Efron & Tibshirani, 1993, p. 185)

$$\alpha_L = \Phi\left(w + \frac{w + z_{\alpha/2}}{1 - a(w + z_{\alpha/2})}\right)$$
$$\alpha_U = \Phi\left(w + \frac{w + z_{1-\alpha/2}}{1 - a(w + z_{1-\alpha/2})},\right)$$

where $\Phi(\cdot)$ is the CDF for the standard normal distribution (and can be obtained with the `pnorm` function in R and the `NORM.S.DIST` with `CUMULATIVE=TRUE` in Microsoft Excel). The two correcting factors are w and a, with w correcting for median bias and a

correcting for the skewness of the distribution of the estimator, and can be estimated as

$$w = \Phi^{-1}\left(\frac{\#\{\hat{\delta}_T^* < \hat{\delta}_T\}}{R+1}\right)$$
$$a = \frac{\sum_{j=1}^{J} l_j^3}{6(\sum_{j=1}^{J} l_j^2)^{3/2}},$$

with $\Phi^{-1}(.)$ being the quantile function (or the inverse CDF) of the standard normal distribution and l_j being the influence values (Davison & Hinkley, 1997) of the estimator and can be estimated using the grouped jackknife (Van der Leeden et al., 2008)[1]. $\#\{\hat{\delta}_T^* < \hat{\delta}_T\}$ is the number of bootstrapped samples with $\hat{\delta}_T^* < \hat{\delta}_T$. Note that when the estimator is unbiased, one has $w = 0$; When the influence function is symmetric, one has $a = 0$. With both conditions $\alpha_L = \alpha/2$ and $\alpha_U = 1 - \alpha/2$, and the BCa CI is equivalent to the percentile CI.

Semiparametric Bootstrap

For multilevel models, one can obtain two types of level-2 residuals: one being ordinary least squares (OLS) estimates and the other being the shrinkage estimates (see Hox, 2010). Whereas the shrunken residuals are biased, they have a smaller mean squared error (*MSE*) (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). There is no clear answer regarding which type of residuals should be used. Van der Leeden, Busing, and Meijer (1997) found that using the shrunken residuals produced a smaller *MSE* than using the OLS residuals. However, Carpenter, Goldstein, and Rasbash (2003) noted that the variance of the shrunken residuals was too small and did not match the estimated values of the variance components. They proposed a transformation method to make the variance-covariance matrix of the shrunken residuals matches that of the variance components; They also showed that their procedures outperformed the parametric bootstrap in terms of CI coverage when the data was generated from a chi-squared

distribution with one degree of freedom. Goldstein (2011a) further modified the transformation method to preserve the dependencies across levels. Vallejo Seco, Ato García, Fernández García, and Livacic Rojas (2013) showed that the semiparametric bootstrap produced smaller root mean squared error *RMSE*. In this study the method by Goldstein (2011a) is included for comparison. The same procedures apply to the semiparametric bootstrap for the percentile and the BCa CIs.

Nonparametric Bootstrap

The nonparametric bootstrap assumes that each observation is independent (Davison & Hinkley, 1997), which is clearly violated with multilevel data (Goldstein, 2011a; Hox, 2010; J. K. Roberts & Fan, 2004; Van der Leeden et al., 2008). While there is multiple modifications proposed (J. K. Roberts & Fan, 2004; Van der Leeden et al., 2008), the major ones are (a) to resample only clusters, and keep the level-1 units in a cluster intact, and (b) resample first the clusters, and within each clusters resample the level-1 units. Both methods resulted in bootstrap samples with level-1 sample size different from the original size. For simplicity, only method (a) is included in this study for comparison.

It should be noted that because the nonparametric bootstrap makes much less assumptions than the parametric and the semiparametric bootstraps, it requires more information from the data. Therefore, its performance would be poor compared to the other two methods, even when the assumptions for the latter two methods are violated (Efron & Tibshirani, 1993; Van der Leeden et al., 2008). Thai, Mentré, Holford, Veyrat-Follet, and Comets (2013) found that in longitudinal linear-mixed models, the semiparametric bootstrap and the nonparametric bootstrap performed similarly when there were at least 100 individuals.

Simulation Study

I conducted a simulation study comparing the performances of ten methods of conducting CIs: symmetric and noncentral intervals for d_T and $\hat{\delta}_T$, and percentile and BCa intervals for the parametric, residual, and nonparametric bootstraps. Six design factors were used as described below.

Design Factors

Intraclass correlation (ICC). The three ICC levels were chosen to reflect the common values in educational research based on a review of articles in 2011–2013 in *American Educational Research Journal* and *Child Development* (see also Hedges & Hedberg, 2007): .05, .10, and .20, to see how the methods for constructing CIs perform. As found in Vallejo Seco et al. (2013), the performance of the semiparametric bootstrap got worse with increasing ICC.

Distribution of u_0. Because one of the goal of the present study is to recommend methods to construct CIs when the normality assumption is violated, the level-2 residuals, u_0, followed either a normal distribution or a scaled chi-squared distribution defined as $\sigma_B^2(\chi_1^2 - 1)/2$. For all conditions I set σ_W^2 as 1.0, and so $\sigma_B^2 = \text{ICC}/(1 - \text{ICC})$. The χ_1^2 distribution is positively skewed with skewness ≈ 2.82 and kurtosis ≈ 12, and has been often used in previous literature to examine the impact of nonnormality (e.g., Carpenter et al., 2003; Maas & Hox, 2004).

Distribution of ε. The levels for the distribution of level-1 residuals, ε, were the same as those for the distribution of u_0, except that the variance of the distribution is σ_W^2 instead of σ_B^2. This allows examination of the relative impact of nonnormality in level-1 and level-2 residuals.

Number of clusters, J. It is generally agreed that multilevel models require at least 30 clusters (e.g., Hox, 2010). Indeed, Flynn and Peters (2004) found that the 95% CI

coverage with the semiparametric bootstrap was only about 91% when there was 24 cluster and the ICC was 0.1. In this study, J was set to either 20, 30, or 70 based on the literature review, and there was equal number of clusters in the treatment and the control arms (i.e., $J^T = J^C$). The level with 20 clusters matches the most extreme condition in the simulation by Carpenter et al. (2003).

Average cluster size, $\bar{n}$. There were two levels for $\bar{n}$: 5 for small and 25 for medium. The small value is typical for longitudinal or family data (and matches the conditions in Thai et al., 2013) , whereas the medium value is chosen to represent typical classroom size in the US.

Distribution of cluster sizes, $P(n)$. Among the five methods for estimating δ_T, only the ANOVA method assumes equal cluster sizes. However, Ames (2013) showed that the model-based estimates of effect size can also be biased with unequal cluster sizes. Unequal cluster sizes also reduces the effective sample size (Candel & Breukelen, 2009), which may make both the analytical methods and the bootstrap methods less stable. One possible family of distributions to model cluster sizes, which are strictly positive numbers, would be the zero-truncated negative binomial (James, 1953). By varying the dispersion or the variance of the negative binomial one can control the degree of imbalance of the cluster sizes. For this study I set the ratio between the variance and the mean of the group sizes to be either 1 (i.e., the zero-truncated Poisson) or 10.

Data Generation and Analyses

In this study there is a total of 3 (ICC) $\times$ 2 (distribution of u_0) $\times$ 2 (distribution of ε) $\times$ 3 (J) $\times$ 2 ($\bar{n}$) $\times$ 2 (distribution of n) = 144 conditions. A thousand sets of u_0 and ε will be first generated independently in R (R Core Team, 2014) according to the predefined distribution (i.e., normal or chi-squared) for each condition, and the outcome values will be computed based on the model described in equation (34). The population

effect size δ_T was fixed to 0.5 to represent a medium effect.

For each data set, the 95% CIs using the 10 methods of interest will be obtained in R. With the exception of the two CIs for d_T, all CIs calculation required fitting mixed models, which is performed using `lme4` (Bates et al., 2014). The `lme4` package includes a method `bootMer` to do parametric bootstrap and semiparametric bootstrap. However, the procedure for "reflating" the shrunken residuals is not yet implemented, and the nonparametric bootstrap procedure is not available. Therefore, I have written my own implementation of the three bootstrap methods in R as the `SimpleBoot` function, as shown in Appendix G (together with the full R code of the simulation).

Evaluation Criteria

Coverage of 95% CI. If the methods for constructing CI are accurate, then I would expect in 950 out of the 1,000 replications the CIs constructed would include the population value $\delta_T = 0.5$, with a standard error of $\sqrt{.95 \times .05/1,000} \approx 0.7\%$. The empirical coverage percentage will be calculated as

$$\frac{\text{Number of replications with CI covering } \delta_T}{1000}$$

The closer this percentage is to 95% the better the CI performs.

Results

Convergence Rate

Across conditions and methods and procedures to obtain CI, the convergence percentage is at least 98%. The conditions with lowest convergence rate were mainly from semiparametric bootstrap with low ICC (i.e., .05) and small average cluster size (i.e., 5), and the main cause for nonconvergence is that with small clustering effect and small cluster size sometimes the random effect covariance matrix is singular, and so the step of

"reflating" the predicted random effects $\tilde{u}_0$ and $\tilde{\varepsilon}$ in semiparametric bootstrap failed.

Type of CI

Given that the current study includes a large number of simulation conditions, I first reduced the conditions by choosing, for each of the five methods to obtain CI, the better procedure in terms of coverage (i.e., normal vs. noncentral *t* for the analytical methods and percentile vs. BCa for the bootstrap methods). As ideally the coverage should be exactly 95% and for many conditions the empirical coverage was less than 95%, I computed the *root mean squared error* (*RMSE*) of the coverage for each condition as

$$RMSE = \sqrt{\frac{\sum(\text{Empirical percentage coverage} - 95\%)^2}{1000}}.$$

Surprisingly, simpler procedures to compute CIs appear to work better for methods relying on the normality assumption. For the ANOVA method, $RMSE = 0.039$ for symmetric CIs and 0.040 for noncentral *t* CIs; for the model-based method, $RMSE = 0.016$ for symmetric CI and 0.017 for noncentral *t* CI; for parametric bootstrap, $RMSE = 0.019$ for percentile CI and 0.020 for BCa CI. The BCa CIs had better coverage for semiparametric bootstrap ($RMSE = 0.017$) and nonparametric bootstrap ($RMSE = 0.021$) than the percentile CIs ($RMSE = 0.020$ and 0.023 respectively).

For subsequent analyses, the results only pertain to symmetric CI for ANOVA and model-based methods, percentile CI for parametric bootstrap, and BCa CI for semiparametric and nonparametric bootstrap.

Results of Logistic Regression

Because of the large number of conditions (144×5 methods to obtain CI), a logistic regression is conducted to determine the effect of all main effects and all two-way interactions on the coverage. Given the large sample size, and that a factorial design is

employed, I computed the McFadden's pseudo R^2 for each effect as

$$R^2 = 1 - \frac{\text{Deviance}_{\text{Model}}}{\text{Deviance}_{\text{Null}}},$$

where Deviance = $-2 \times$ log-likelihood and the null model refers to the model with only the intercept. The term with the biggest R^2 is the type of methods ($R^2 = 21.8\%$), followed by the type of methods $\times$ cluster size distribution ($P(n)$) interaction ($R^2 = 13.5\%$), the main effect of $P(n)$ (7.3%), the main effects of number of clusters (J) (6.5%), distribution of ε (4.5%), and distribution of u_0 (3.7%). Other terms with $R^2 > 1.0$ can be found in Table 6. Mean empirical coverage by conditions were shown in Table 7, 8, and 9.

Table 6
Summary of Logistic Regression Results With Empirical Confidence Interval Coverage as the Dependent Variable

Effect	*df*	Deviance	Pseudo R^2
CI Method	4	539.86	21.76
Number of Clusters, J	2	161.01	6.49
Average Cluster Size, $\bar{n}$	1	46.56	1.88
Cluster Size Distribution, $P(n)$	1	181.79	7.33
Distribution of ε, $f(\varepsilon)$	1	110.53	4.46
Distribution of u_0, $f(u_0)$	1	90.80	3.66
CI Type $\times$ ICC	8	69.00	2.78
CI Type $\times$ J	8	36.15	1.46
CI Type $\times$ $P(n)$	4	334.09	13.47
CI Type $\times$ $f(\varepsilon)$	4	46.82	1.89
CI Type $\times$ $f(u_0)$	4	42.69	1.72
ICC $\times$ $\bar{n}$	2	68.12	2.75
ICC $\times$ $P(n)$	2	31.04	1.25
$J \times \bar{n}$	2	36.66	1.48
$\bar{n} \times f(\varepsilon)$	1	30.01	1.21

Note. Only effects with pseudo $R^2 > .01$ are shown. *df* = degrees of freedom. CI = confidence interval. ICC = intraclass correlation.

Table 7
Mean and Median Confidence Interval (CI) Coverage for the Two Analytical Methods

CI Method	ICC	J	$P(n)$	Mean Coverage	Median Coverage
ANOVA	Poisson	20	.05	.929	.931
			.1	.933	.933
			.2	.928	.927
		30	.05	.936	.937
			.1	.934	.935
			.2	.937	.937
		70	.05	.938	.940
			.1	.936	.938
			.2	.937	.931
	NB	20	.05	.913	.913
			.1	.905	.905
			.2	.885	.890
		30	.05	.917	.919
			.1	.900	.898
			.2	.887	.886
		70	.05	.920	.921
			.1	.907	.905
			.2	.884	.886
Model-Based	Poisson	20	.05	.933	.933
			.1	.935	.933
			.2	.934	.933
		30	.05	.937	.938
			.1	.938	.940
			.2	.941	.941
		70	.05	.941	.944
			.1	.942	.942
			.2	.945	.944
	NB	20	.05	.934	.933
			.1	.933	.934
			.2	.931	.931
		30	.05	.935	.937
			.1	.931	.929
			.2	.938	.936
		70	.05	.940	.941
			.1	.942	.940
			.2	.940	.944

Note. NB = Negative Binomial. ICC = intraclass correlation. J = number of clusters. $P(n)$ = distribution of cluster sizes.

Table 8
Mean and Median Confidence Interval (CI) Coverage for the Parametric and the Semiparametric Bootstrap Methods

CI Method	ICC	J	$P(n)$	Mean Coverage	Median Coverage
Parametric	Poisson	20	0.05	0.930	0.931
Bootstrap			0.1	0.931	0.931
			0.2	0.931	0.931
		30	0.05	0.934	0.935
			0.1	0.934	0.935
			0.2	0.937	0.935
		70	0.05	0.937	0.939
			0.1	0.940	0.940
			0.2	0.943	0.942
	NB	20	0.05	0.930	0.932
			0.1	0.933	0.933
			0.2	0.929	0.929
		30	0.05	0.933	0.934
			0.1	0.928	0.929
			0.2	0.933	0.933
		70	0.05	0.939	0.942
			0.1	0.938	0.938
			0.2	0.937	0.940
Semiparametric	Poisson	20	0.05	0.933	0.938
Bootstrap			0.1	0.937	0.937
			0.2	0.932	0.930
		30	0.05	0.940	0.940
			0.1	0.937	0.938
			0.2	0.940	0.939
		70	0.05	0.944	0.945
			0.1	0.943	0.945
			0.2	0.945	0.944
	NB	20	0.05	0.937	0.940
			0.1	0.935	0.936
			0.2	0.927	0.928
		30	0.05	0.937	0.937
			0.1	0.932	0.935
			0.2	0.934	0.929
		70	0.05	0.938	0.939
			0.1	0.939	0.938
			0.2	0.938	0.940

Note. NB = Negative Binomial. ICC = intraclass correlation. J = number of clusters. $P(n)$ = distribution of cluster sizes.

Table 9
Mean and Median Confidence Interval (CI) Coverage for the Nonparametric Bootstrap Methods

CI Method	ICC	J	$P(n)$	Mean Coverage	Median Coverage
Nonparametric Bootstrap	Poisson	20	0.05	0.917	0.915
			0.1	0.923	0.923
			0.2	0.922	0.922
		30	0.05	0.933	0.933
			0.1	0.930	0.929
			0.2	0.934	0.935
		70	0.05	0.941	0.940
			0.1	0.941	0.939
			0.2	0.942	0.942
	NB	20	0.05	0.927	0.925
			0.1	0.927	0.926
			0.2	0.924	0.923
		30	0.05	0.931	0.935
			0.1	0.930	0.933
			0.2	0.931	0.931
		70	0.05	0.940	0.940
			0.1	0.940	0.938
			0.2	0.939	0.940

Note. NB = Negative Binomial. ICC = intraclass correlation. J = number of clusters. $P(n)$ = distribution of cluster sizes.

CI methods. The main effect of CI methods had the largest R^2, which is mainly attributed to the relatively low coverage of the ANOVA methods (91.8%) compared to other methods (mean empirical coverage > 93.2%). Overall, the model-based method and the semiparametric bootstrap performed the best with mean coverage of 93.7%, followed by parametric bootstrap with 93.4% and nonparametric bootstrap with 93.2%.

***P*(*n*).** As expected when the cluster sizes became more unbalanced, empirical coverage dropped. Specifically, with $P(n)$ having a zero-truncated Poisson distribution with variance approximately equaled to $\bar{n}$, the mean empirical coverage was 93.6%; with

$P(n)$ having a zero-truncated negative binomial distribution with variance approximately equaled to $10 \times \bar{n}$, the mean empirical coverage dropped to 92.7%.

CI methods × $P(n)$. It should be noted that the difference in coverage with different $P(n)$ was more salient with the ANOVA method, where the mean coverage was 93.4% for Poisson and 90.2% for negative binomial (see Figure 1). This was not surprising given that the formulae for the ANOVA method was derived for balanced design. For all the remaining four methods, although the coverage was better in general for the Poisson conditions, the differences were all less than 0.4%, showing that these 4 methods were robust to unbalanced cluster sizes.

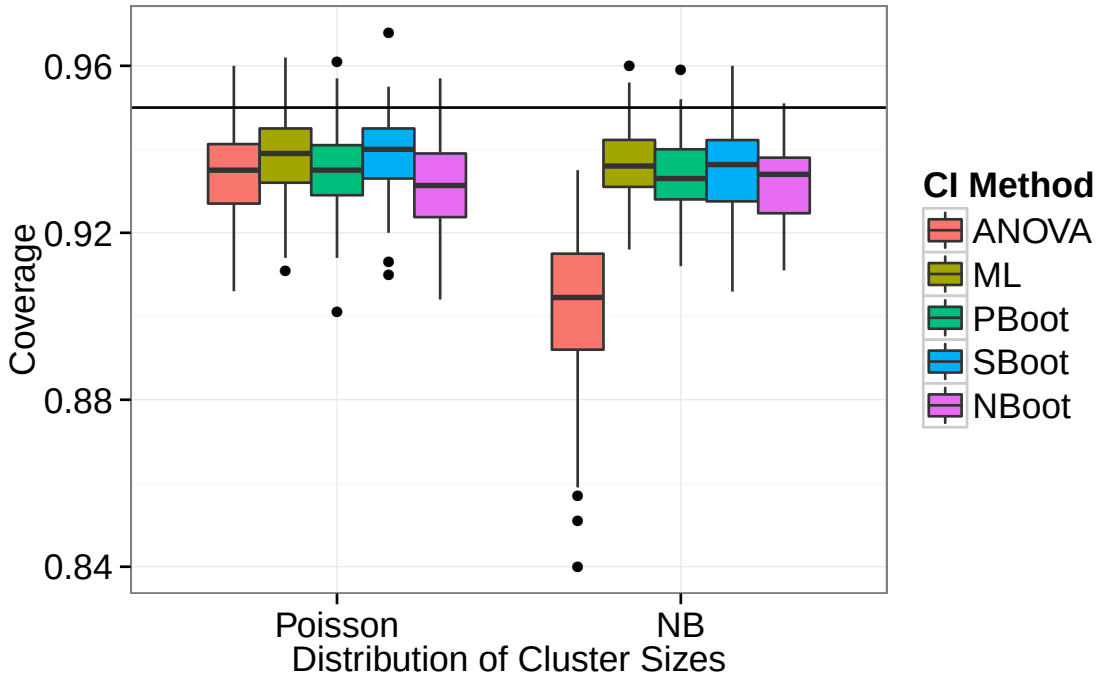

Figure 1: Boxplots showing the empirical coverage of CI methods by distribution of cluster sizes across conditions.

ICC. Whereas the main effect for ICC was small (R^2 = .0067), as shown in Figure 2 it was found that for the ANOVA method coverage dropped with increasing ICC (92.6% for ICC = 0.05, 91.9% for ICC = 0.10, 91.0% for ICC = 0.20). For the other four methods the differences were at most 0.2%.

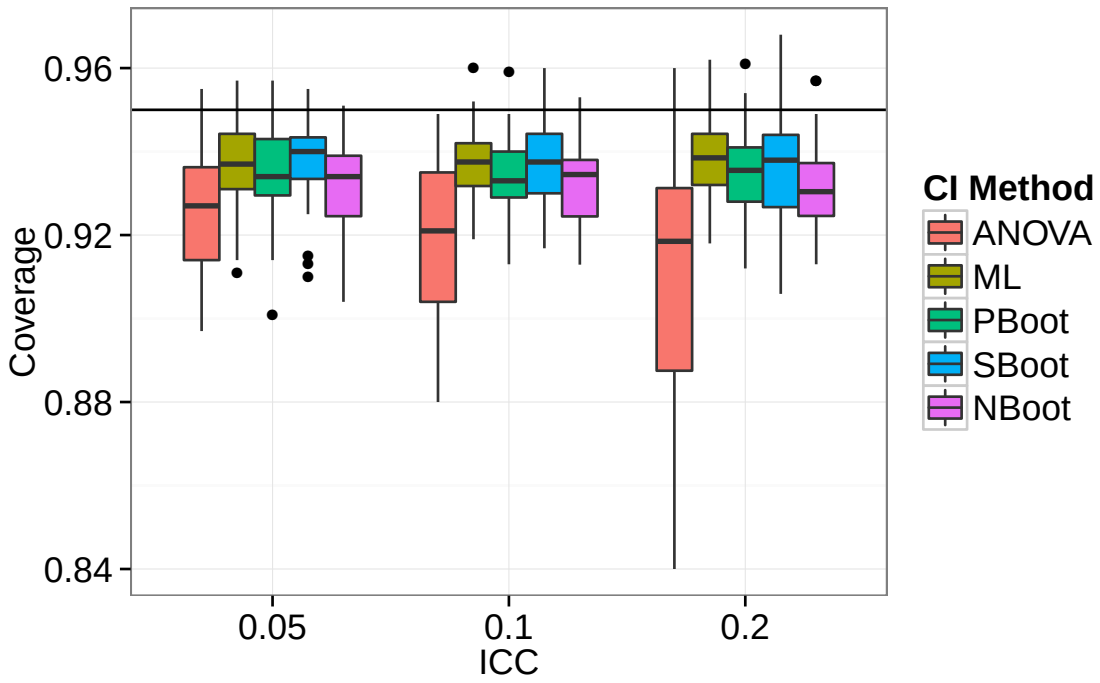

Figure 2: Boxplots showing the empirical coverage of CI methods by intraclass correlation across conditions.

***J*.** As expected, more clusters helped to achieve a better CI coverage. The difference was most prominent for nonparametric bootstrap: As shown in Figure 3, with J = 20, the mean coverage for the nonparametric bootstrap CI was only 92.4%, compared to 93.3% for the model-based method, 93.1% for parametric bootstrap, and 93.4% for semiparametric bootstrap. However, when J = 70, nonparametric bootstrap (94.0%) had

comparable performance with the parametric bootstrap (93.9%), model-based method (94.2%), and the semiparametric bootstrap (94.1%).

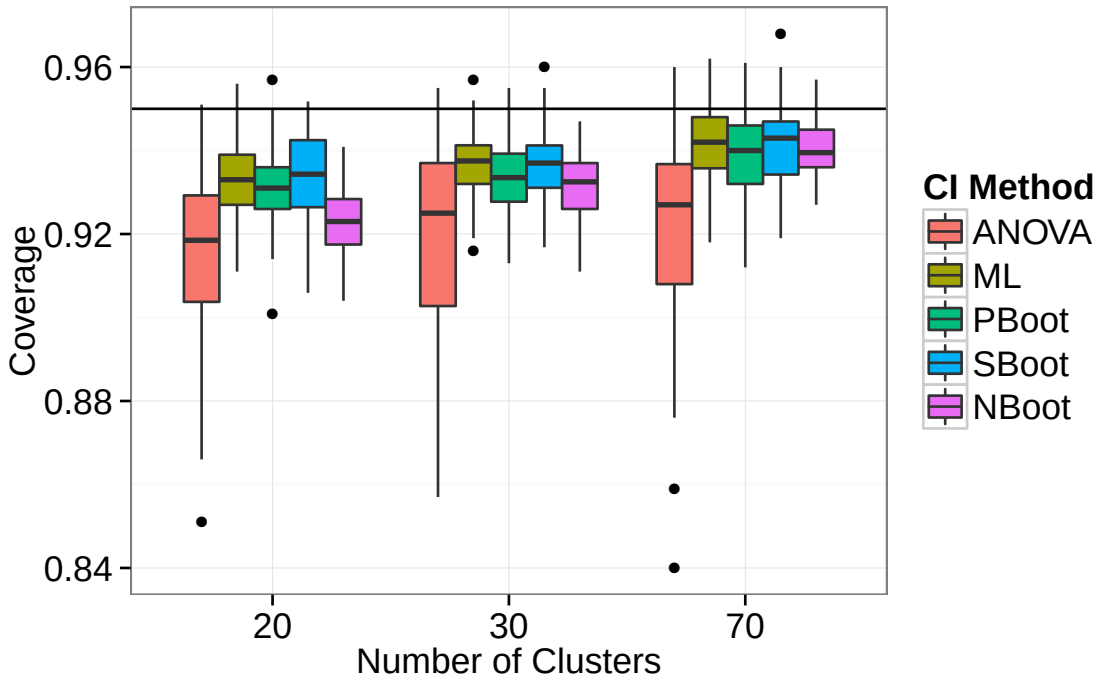

Figure 3: Boxplots showing the empirical coverage of CI methods by number of clusters across conditions.

n̄. Compared to the effect of J, the effect of average cluster sizes was smaller (see Figure 4). In general larger $\bar{n}$ resulted in better coverage, but the difference between $\bar{n} = 5$ and $\bar{n} = 25$ was 0.8% for ANOVA method, 0.5% for semiparametric bootstrap, and around 0.3% or less for the other three methods. It was found that average cluster size was more important with a larger J, as the difference in empirical coverage between $\bar{n} = 5$ and $\bar{n} = 25$ was only 0.02% for $J = 20$ but 0.8% for $J = 70$.

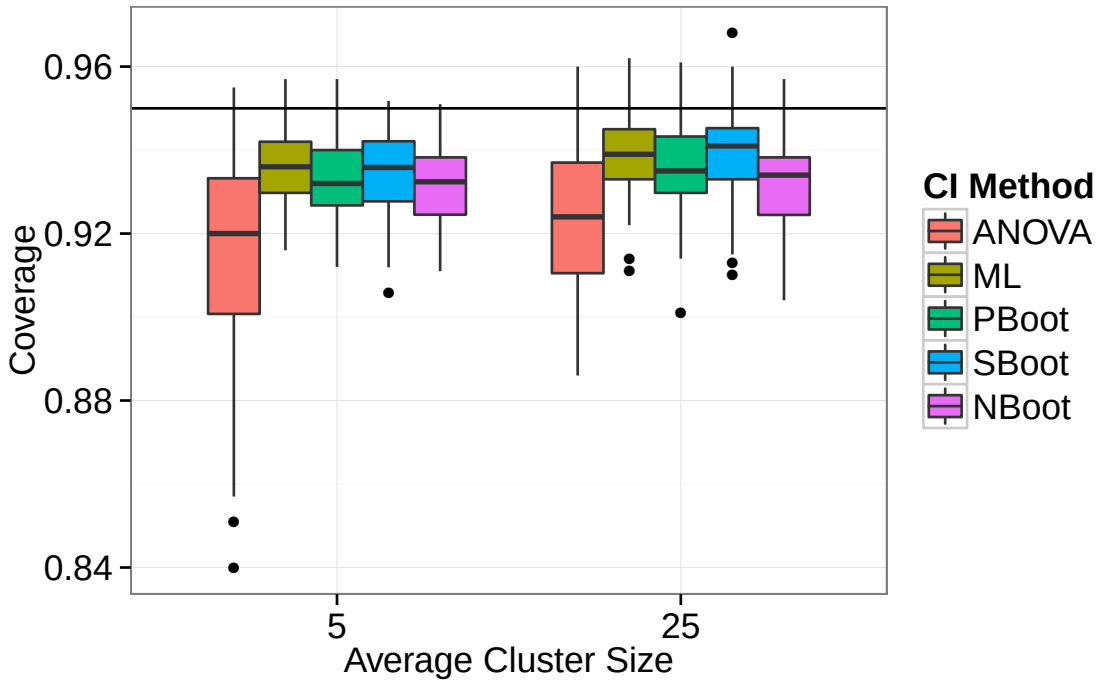

Figure 4: Boxplots showing the empirical coverage of CI methods by average cluster size across conditions.

***f*(*u*$_0$).** Surprisingly, when the level-2 random effects followed a χ^2_1 distribution, the coverage was better than when the random effects followed a normal distribution. The mean coverage being 92.9% when $f(u_0)$ is normal and 93.5% when $f(u_0)$ is chi-squared. As shown in Figure 5, with the exception of the nonparametric bootstrap, all other four methods showed better coverage when $f(u_0)$ is chi-squared, with the difference ranging between 0.4% to 1.2%.

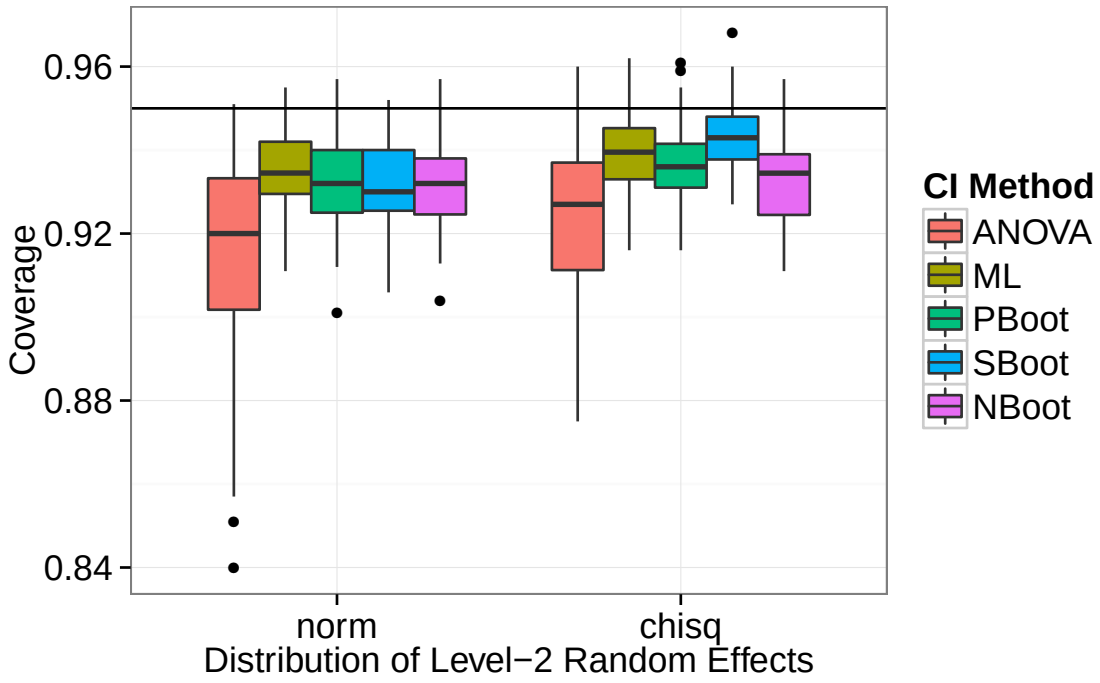

Figure 5: Boxplots showing the empirical coverage of CI methods by distribution of level-2 random effects across conditions.

***f*(ε).** As expected, and different from the results with the level-2 random effects, when normality of level-1 random effects did not hold, the coverage was suboptimal, with mean coverage being 93.5% when $f(\varepsilon)$ is normal and 92.9% when $f(u_0)$ has a chi-square distribution with one degree of freedom. As shown in Figure 6, for the three methods (i.e., ANOVA, model-based, and parametric bootstrap) that assume normality, the difference in coverage was about 1%. On the other hand, for the semiparametric bootstrap, the coverage under different $f(\varepsilon)$ were both 93.7%; whereas for the nonparametric bootstrap, the two coverage rates were 93.3% and 93.0%.

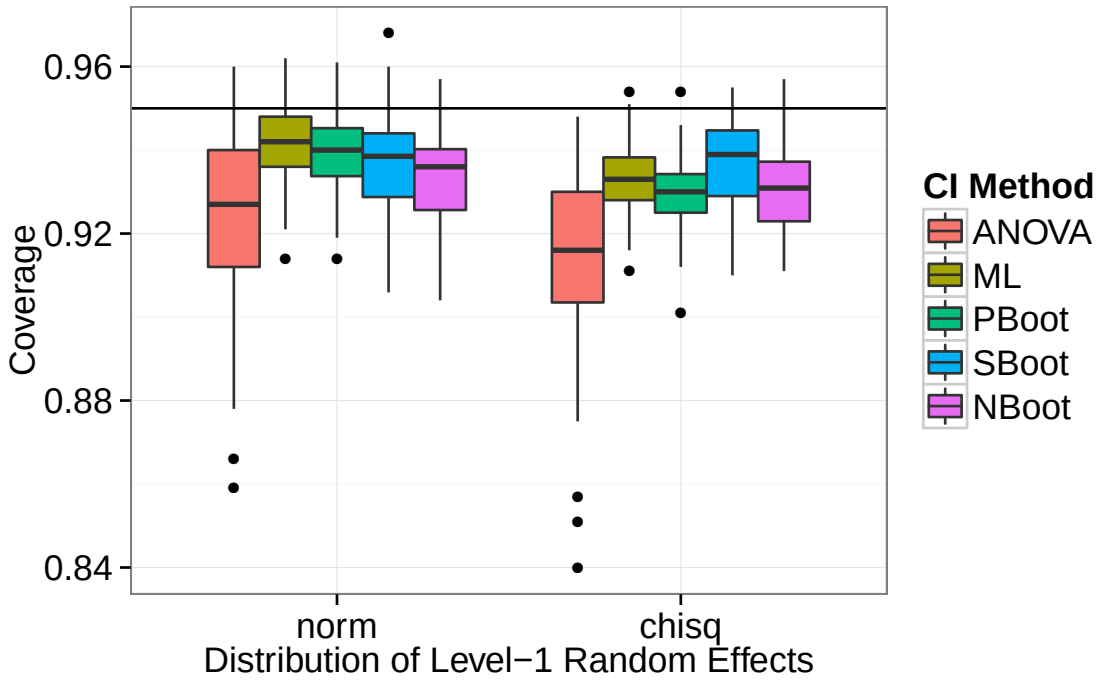

Figure 6: Boxplots showing the empirical coverage of CI methods by distribution of level-1 random effects across conditions.

Discussion

Despite recommendations in the field (e.g. AERA, 2006; APA, 2010; Hedges, 2008), very rarely did researchers report CIs for effect size in single level studies, and little attention has been given to assist substantive researchers in computing CIs for the newly developed multilevel effect size. The present study compared the performance of the bootstrap methods to the analytical methods in obtaining CIs, and results supported both the parametric bootstrap and the semiparametric bootstrap as viable alternatives when analytical methods are not available. Below I summarize the results of the study.

Performances of the Five Methods to Construct CI

Although I found differences in performance among the five methods to compute CI for multilevel SMDs, the difference tended to be small as for most of the conditions the empirical coverage was above 91%. Indeed, if all the clusters in a real data set have equal size, all five methods would give CIs with similar performance. If, however, the cluster sizes were not equal, then the ANOVA method should not be used.

Overall, the model-based method and the semiparametric bootstrap produced CIs with highest coverage. Therefore, when either one of these two methods were available, they should be used. However, each of these two methods have its limitations. The model-based method relied on the normality assumption for the random effect in each level. Although the present study showed that chi-squared distribution for level-2 random effects did not lead to lower coverage for the model-based method, it was still a question whether it can be generalized to distributions that were multimodal or deviate from normal in a different way than the chi-squared distribution with one degree of freedom did. Future research effort is needed to understand the robustness of the model-based method to nonnormality of higher-level random effects. Also, the present result clearly showed that nonnormality at level-1 resulted in a suboptimal performance for the model-based method. Therefore, methods that did not rely on the normality assumption such as the semiparametric and the nonparametric bootstrap will produce CIs with better coverage with nonnormal data. Another limitation of the model-based method is that the variance estimator is specific to a given multilevel structure. Although formulae for computing the variance of multilevel SMD have been derived for three-level data (Hedges, 2011) and for cross-classified data (Lai & Kwok, 2014), there are many more different multilevel structures such as the multiple membership structure (Beretvas, 2011), structure with more than two crossed factors, and the partially nested structure

(Bauer et al., 2008). It would be tedious if not impossible to derive the formulae for the variance estimator every time researchers have data with a new structure.

On the other hand, the semiparametric bootstrap did not make the normality assumption and performed better than the model-based approach with extremely nonnormal data. However, it had a slightly worse coverage than the model-based approach when normality holds. Also, the need to "reflate" the residuals for each level introduces the risk that the Cholesky decomposition or its inverse may be difficult to compute, especially in models with more random effects in multiple levels. Also, because the semiparametric bootstrap is not a built-in feature for most software packages that do multilevel analyses, researchers may need to program the procedure on their own or wait for other methodologists to make the procedure accessible in different software packages.

The percentile method with parametric bootstrap showed slightly lower but very similar coverage than the model-based method, and showed lower but acceptable coverage than the semiparametric bootstrap when normality did not hold. As the parametric bootstrap is easy to implement and is available either as a built-in feature in `lme4` in R and as user-written scripts in SAS and in SPSS. Therefore it can be a good alternative for multilevel data when the model-based method is not yet accessible. The nonparametric bootstrap, on the other hand, had lower coverage than the parametric bootstrap when normality holds and showed no advantage when normality is violated, and so was not recommended.

Suggested Practice in Reporting CIs for Multilevel Effect Size

It is inconsistent for researchers to regularly report measures of uncertainty for sample statistics such as mean, regression coefficients, but not reporting *SE*s or CIs for effect size. After all, effect size aims to quantify the effect of interest in an interpretable way. A significant barrier for substantive researchers to adhere to the reporting guidelines

is the complexity associated with CI computations for effect size, especially with multilevel data. The goal of the present study is to demonstrate to and familiarize researchers with some of the tools they can use to obtain CIs for multilevel effect size. Based on the results of the present study, I have a few suggestions for reporting CIs for two-group multilevel studies:

1. When normality is not severely violated, use model-based methods for data with two-level of clustering (Hedges, 2007), three-level of clustering (Hedges, 2011), cross-classified structure (Lai & Kwok, 2014), and two-level partially nested structure (Lai, 2014, Chapter III of this dissertation), as the formulae were already developed;
2. When normality is in doubt or when analytical formulae are not available, use the semiparametric bootstrap if available and accessible;
3. If neither the model-based method nor the semiparametric bootstrap CI are available, obtain the CIs by the parametric bootstrap.

Limitations

There are several limitations of the present study that should be addressed in future studies. First, I only considered the chi-squared distribution with $df = 1$ as the condition for nonnormal random effects. It is possible that the results may be different if the random effects follow a different nonnormal distribution, although given the results in this study the difference is not likely to be large. Second, it is not obvious why the coverage for all methods were higher under level-2 nonnormality. Inspection of the data showed that in those conditions, the point estimate of effect size is closer to the population value $\delta_T = 0.5$. As the effect size estimator is know to be positively biased especially when sample size is small (Hedges, 1981, 2007), it is possible that the bias got canceled when the distribution deviated from normal, thus resulting in better coverage. Future theoretical

work is needed to understand the impact of nonnormality of point and interval estimates of effect size. Third, the bootstrap procedures studied in the present study only represents three procedures that are better known to researchers (Van der Leeden et al., 2008). Other bootstrap procedures (Field, Pang, & Welsh, 2010; Owen & Eckles, 2012, e.g.) have been developed and may perform better for multilevel effect size. Fourth, the bootstrap procedures described in the present study are not yet available in some multilevel software packages, and future effort is needed to make them more accessible to substantive researchers. Finally, simulations done in the present study only pertains to the basic two-level strictly hierarchical structure. Future research should utilize more complex structures to verify whether the bootstrap procedures still perform well in those designs.

Notes

[1]In the conventional jackknife estimate, $l_i = T(\mathbf{x}) - T_{(i)}(\mathbf{x})$ represents the changes in $T(\mathbf{x})$ when the ith observation is deleted. As noted in Van der Leeden et al. (2008), as the level-1 observations are not independent for multilevel data, one can perform jackknife only on the highest level, so l_j is the changes in $\hat{\delta}_T$ when the jth cluster is deleted. In this paper I used a simplified version than the one used in Van der Leeden et al.

CHAPTER V

CONCLUSIONS

This dissertation discusses parametric and nonparametric estimations for multilevel effect size. As effect size has been regularly reported in single level studies, there is no reason to not adhere to the same standard for multilevel studies. Because the definition and estimation of multilevel standardized mean difference (SMD) was developed relatively recently (Hedges, 2007), much more work is needed. One direction is to extend the analytical formulae to more complex multilevel data structure, and Hedges (2011) did exactly the same for three-level strictly nested structure. The other direction is to utilize estimation methods that are more flexible and can easily handle structures with different complexities; Otherwise it would be tedious to develop analytical methods for each novel multilevel structure.

The first and the second manuscript of this dissertation extended analytical methods to commonly utilized non-hierarchical structures. In the first manuscript, I developed ANOVA and model-based methods to obtain SMD and the corresponding sampling variance for *cross-classified* and *partially cross-classified* data structure. It can be argued that in reality, multilevel data are often not strictly hierarchical, as people share more than one environment. For example students are clustered by both schools and neighborhood. As Luo and Kwok (2009) have shown that ignoring a crossed factor can lead to incorrect standard error estimates of the coefficients, it is important to develop appropriate method to compute effect size and the corresponding sampling variance for cross-classified data. Simulation results from the first manuscript showed that both the ANOVA method and the model-based method have acceptable performance, with the later being more robust to unequal cell sizes. The methods were also demonstrated with

NELS:88 and the ERI data sets.

The second manuscript dealt with the *partially nested* structure. In such structure, data in the treatment arm is clustered because of the intervention, but data in the control arm is not. Although such data are not commonly seen in multilevel observational studies, it nevertheless is popular in clinical trial and experimental studies (Bauer et al., 2008; Sanders, 2011), where SMD is a more natural effect size choices. Again I developed ANOVA and model-based methods to obtain SMD and the corresponding sampling variance for such design. Simulation results showed that under slight imbalance of the cluster sizes, both methods performed very well in terms of CI coverage (all above 92.4%), except under the condition with ICC = .5 and average n = 25 (i.e., high *design effect*) where the CI coverage was still above 91%. Both methods had similar performances as they had similar root mean squared errors. On the other hand, under extreme imbalance of cluster sizes, only the model-based method maintained the same performance, whereas the CI coverage with the ANOVA method suffered the most with high design effect and dropped below 90%. The two methods were demonstrated with the Reconnecting Youth preventive intervention program (Bauer et al., 2008). Finally, a modified estimator using only the standard deviation of the control arm was discussed, and the effect of ignoring the clustering on the treatment arm on the effect size and variance estimates were reported.

In the third manuscript, the focus is no longer only on the analytical procedures to compute multilevel effect size, but more on the potential usefulness of the bootstrap as a more flexible alternative. Whereas analytical formulae to estimate point estimate for SMD is straight forward and can easily be generalized to more complex data structure not yet discussed, this is not true for the sampling variance or standard error of multilevel SMD. Even when such formulae can be developed, they are likely to be complex and not user-friendly in their look. On the other hand, the bootstrap is originally developed to deal

with the problem of obtaining sampling variance and CIs for complex estimators (Efron & Tibshirani, 1993). By adapting the bootstrap to multilevel data (Van der Leeden et al., 2008) it can potentially be applied to a lot of multilevel problems, including the problem of obtaining CIs for multilevel SMD in the third manuscript. The simulation results showed that whereas the semiparametric bootstrap performed the best out of the three bootstrap methods studied, the model-based method was quite robust to nonnormality of the random effects. After all, the difference between the model-based method and the three bootstrap methods in terms of empirical CI coverage was relatively small, and in practice the choice is likely to be determined by the ease of use and availability of each methods. Thus, the results supported the potential usefulness of the bootstrap when reporting multilevel effect size.

The findings of the three manuscripts in this dissertation can be summarized in the following recommendations.

1. Always report some kind of effect size for cluster randomized trials or multilevel studies with binary predictors, either the mean difference in the original metric of the outcome or the SMD;
2. Attach with the effect size measure the corresponding *SE* and/or CI;
3. When computing CI for multilevel SMD, use either the model-based method or the semiparametric bootstrap when available;
4. If neither the model-based method nor the semiparametric bootstrap CI are available, obtain the CIs by the parametric bootstrap.

REFERENCES

Ahn, S., Myers, N. D., & Jin, Y. (2012). Use of the estimated intraclass correlation for correcting differences in effect size by level. *Behavior Research Methods*, *44*, 490–502. doi:10.3758/s13428-011-0153-1

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society. Series A (General)*, *149*, 1–43. doi:10.2307/2981882

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

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

Ames, A. J. (2013). Accuracy and precision of an effect size and its variance from a multilevel model for cluster randomized trials: A simulation study. *Multivariate Behavioral Research*, *48*, 592–618. doi:10.1080/00273171.2013.802978

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

Bates, D. (2010). *lme4: Mixed-effects modeling with R.* Retrieved from `http://lme4.r-forge.r-project.org/lMMwR/lrgprt.pdf`

Bates, D. (2011). *Mixed models in R using the lme4 package Part 3: Inference based on profiled deviance.* Retrieved from `http://lme4.r-forge.r-project.org/slides/2011-01-11-Madison/3ProfilingH.pdf`

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using S4 classes (R Package Version 1.1-7) [Computer program]. Retrieved

from http://CRAN.R-project.org/package=lme4

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

Beretvas, S. N. (2008). Cross-classified random effects models. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 161–197). Charlotte, NC: Information Age.

Beretvas, S. N. (2011). Cross-classified and multiple membership models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 313–334). New York, NY: Routledge.

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *American Psychologist*, *61*, 27-41.

Byrd, J. K. (2007). A call for statistical reform in EAQ. *Educational Administration Quarterly*, *43*, 381–391. doi:10.1177/0013161X06297137

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

Candel, M. J. J. M., & Breukelen, G. J. P. V. (2009). Varying cluster sizes in trials with clusters in one treatment arm: Sample size adjustments when testing treatment effects with linear mixed models. *Statistics in Medicine*, *28*, 2307–2324. doi:10.1002/sim.3620

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *52*, 431–443. doi:10.1111/1467-9876.00415

Carver, R. P. (1978). The case against statistical significant testing. *Harvard Educational Review*, *48*, 378–399. doi:10.1080/00220973.1993.10806591

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

Chen, L.-T., & Peng, C.-Y. J. (2014). The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes. *Behavior Research Methods*. Advance online publication. doi:10.3758/s13428-014-0461-3.

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

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. doi:10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:10.1037/0003-066X.49.12.997

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. doi:10.1037/a0016930

Coyne, M. D., Simmons, D. C., Hagan-Burke, S., Simmons, L. E., Kwok, O.-m., Kim, M., . . . Rawlinson, D. M. (2013). Adjusting beginning reading intervention based on student performance: An experimental evaluation. *Exceptional Children*, *80*, 25–44.

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

Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, *61*, 532–574. doi:10.1177/0013164401614002

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application*.

Cambridge, UK: Cambridge University.

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia, PA: Society for industrial and applied mathematics.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. New York, NY: Chapman and Hall. doi:10.1111/1467-9639.00050

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

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538. doi:10.1037/a0015808

Field, C. A., Pang, Z., & Welsh, A. H. (2010). Bootstrapping robust estimates for clustered data. *Journal of the American Statistical Association*, *105*, 1606–1616. doi:10.1198/jasa.2010.tm09541

Flynn, T. N., & Peters, T. J. (2004). Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *BMC Health Services Research*, *4*, article 33. doi:10.1186/1472-6963-4-33

Friedel, J. M., Cortina, K. S., Turner, J. C., & Midgley, C. (2010). Changes in efficacy beliefs in mathematics across the transition to middle school: Examining the effects of perceived teacher and parent goal emphases. *Journal of Educational Psychology*, *102*, 102–114. doi:10.1037/a0017590

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*, 2–18. doi:10.1037/a0024338

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized

least squares. *Biometrika*, 43–56. doi:10.2307/2336270

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

Goldstein, H. (2011b). *Multilevel statistical models* (4th ed.). Hoboken, NJ: Wiley.

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

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (Eds.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.

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. doi:10.3102/10769986006002107

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

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

Hedges, L. V. (2009). Effect sizes in nested designs. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 337–355). New York, NY: Russell Sage Foundation.

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

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure

modeling: An overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329–367. doi:10.1177/0049124198026003003

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

Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological Measurement*, *62*, 227–240. doi:10.1177/0013164402062002002

Ingles, S. J., Abraham, S. Y., Karr, R., Spencer, B. D., & Frankel, M. R. (1990). *National Educational Longitudinal Study of 1988: Base year student component data file user's manual.* Washington, DC: U. S. Department of Education.

James, J. (1953). The distribution of free-forming small group size. *American Sociological Review*, *18*. doi:10.2307/2087444

Johnson, I. Y. (2011). Contingent instructors and student outcomes: An artifact or a fact? *Research in Higher Education*, *52*, 761–785. doi:10.1007/s11162-011-9219-2

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*, 51–69. doi:10.1177/0013164404264850

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.

Kelley, K., & Preacher, K. J. (2012). On effect size. *Psychological Methods*, *17*, 137–152. doi:10.1037/a0028086

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759. doi:10.1177/0013164496056005002

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. doi:10.1016/j.learninstruc.2011.01.001

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`

Kramer, S. H., & Rosenthal, R. (1999). Effect sizes and significance levels in small-sample research. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 59–79). Thousand Oaks, CA: Sage.

Kreft, I., & De Leeuw, J. (1998). *Introducing multilevel modeling*. London, UK: Sage.

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. doi:10.3102/1076998614532950

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:10.1191/1740774505cn082oa

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

Littell, R. C., Milliken, G. A., Stroup, W. W., & Wolfinger, R. D. (1996). *SAS system for mixed models*. Cary, NC: SAS.

Longford, N. T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, *74*, 817–827. doi:10.2307/2336476

Luo, W., & Kwok, O.-m. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, *44*, 182–212.

doi:10.1080/00273170902794214

Luo, W., & Kwok, O.-m. (2010). Proportional reduction of prediction error in cross-classified random effects models. *Sociological Methods & Research*, *39*, 188–205. doi:10.1177/0049124110384062

Maas, C. J. M., & Hox, J. J. (2004). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics & Data Analysis*, *46*, 427–440. doi:10.1016/j.csda.2003.08.006

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128. doi:10.1207/s15327906mbr3901

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through multilevel linear model. *Sociological Methodology*, *14*, 72–103. doi: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. doi:10.1002/sim.3115

Murray, D. M., & Blitstein, J. L. (2003). Methods to reduce the impact of intraclass correlation in group-randomized trials. *Evaluation Review*, *27*, 79–103. doi:10.3102/0162373707299706

Muthén, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267–316. doi:10.2307/271070

Muthén, L. K., & Muthén, B. O. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*, 599–620. doi:10.1207/S15328007SEM0904_8

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

National Center for Educational Statistics. (2012). *NCES statistical standards* (rev. ed.). Washington, DC: U. S. Department of Education.

Owen, A. B., & Eckles, D. (2012). Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, *6*, 895–927. doi:10.1214/12-AOAS547

Peng, C.-Y. J., & Chen, L.-T. (2014). Beyond Cohen's *d*: Alternative effect size measures for between-subject designs. *The Journal of Experimental Education*, *82*, 22–50. doi:10.1080/00220973.2012.745471

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. doi:10.1007/s10648-013-9218-2

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

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers*, *36*, 717–731. doi:10.3758/BF03206553

R Core Team. (2013). R: A language and environment for statistical computing (Version 3.0.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team. (2014). R: A language and environment for statistical computing (Version 3.1.1) [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.

Roberts, C., & Roberts, S. A. (2005). Design and analysis of clinical trials with clustering effects due to treatment. *Clinical Trials*, *2*, 152–162. doi:10.1191/1740774505cn076oa

Roberts, J. K., & Fan, X. (2004). Bootstrapping within the multilevel/hierarchical linear modeling framework: A primer for use with SAS and SPLUS. *Multiple Linear Regression Viewpoints*, *30*, 23–34.

Roberts, J. K., Monaco, J. P., Stovall, H., & Foster, V. (2011). Explained variance in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 219–230). New York, NY: Routledge.

Robinson, D. H., Whittaker, T. A., Williams, N. J., & Beretvas, S. N. (2003). It's not effect sizes so much as comments about their magnitude that mislead readers. *The Journal of Experimental Education*, *72*, 51–64. doi:10.1080/00220970309600879

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York, NY: Russel Sage Foundation.

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).

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, *1*, 115–129. doi:10.1037/1082-989X.1.2.115

Schulz, K. F., Altman, D. G., Moher, D., & CONSORT Group. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *PLoS Medicine*, *7*, e1000251. doi:10.1371/journal.pmed.1000251

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

Sechrest, L., McKnight, P., & McKnight, K. (1996). Calibration of measures for psychotherapy outcome studies. *American Psychologist*, *51*, 1065–1071. doi:10.1037/0003-066X.51.10.1065

Shi, Y., Leite, W., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*, 1–15. doi:10.1348/000711008X398968

Snijders, T. A. B., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*, 342–363. doi:10.1177/0049124194022003004

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Snyder, P., & Lawson, S. (1993). Evaluating results using corrected and uncorrected effect size estimates. *The Journal of Experimental Education*, *4*, 334–349. doi:10.1080/00220973.1993.10806594

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 221–257). Mahwah, NJ: Erlbaum.

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. doi:10.1080/00273171.2014.882253

Thai, H.-T., Mentré, F., Holford, N. H. G., Veyrat-Follet, C., & Comets, E. (2013). A comparison of bootstrap approaches for estimating uncertainty of parameters in linear mixed-effects models. *Pharmaceutical Statistics*, *12*, 129–140.

doi:10.1002/pst.1561

Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher*, *25*, 26–30. doi:10.3102/0013189X025002026

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

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*, 423–432. doi:10.1002/pits.20234

Vallejo Seco, G., Ato García, M., Fernández García, M. P., & Livacic Rojas, P. E. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema*, *25*, 520–528. doi:10.7334/psicothema2013.58

Van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (1997). *Bootstrap methods for two-level models* (Technical Report PRM 97-04). Leiden, Netherlands: Leiden University.

Van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401–433). New York, NY: Springer.

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

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *The Annals of Statistics*, *14*, 1261–1295. doi:10.1214/aos/1176350142

APPENDIX A
DERIVATION OF SAMPLING DISTRIBUTIONS OF SMD WITH CROSS-CLASSIFIED DATA

The following steps to find an unbiased estimator were developed based on the work of Hedges (2007, pp. 360-362). Here we extended the framework to CCREMs with some differences in notation. Note that the theorem section below is general in the sense that it can be used to derive effect size statistics for designs other than those presented in the current paper. For example, Hedges (2011) used it for data with three-level nested structure. In future studies researchers may apply the same theorem to more complicated designs such as those with more than two crossed random effects or those with crossed random effect in lower level.

Theorem

Suppose that $\Delta\bar{Y} \sim N(\Delta\mu, a\sigma^2/\tilde{N})$ is a random variable that represents the grand mean difference of the outcome variable Y in the sample, and a is a variance inflation factor due to clustering, which equals one in data with simple random sampling but is larger than one in clustered data; $\tilde{N}$ is a function of the sample sizes that relates the variance of Y, σ^2, to the variance of $\bar{Y}$.

Let S^2 be a sample estimator of σ^2, the population variance component of Y, where σ is the denominator of the population effect size. Assuming Y and all random effects are normally distributed, S^2 is a quadratic form of normally distributed variables that is independent of $\Delta\bar{Y}$. For example, when the variance component of interest is σ_T^2 (i.e., the sum of all variance components of Y, see the section "Intraclass Correlation"), the sample estimator S^2 can be chosen as the pooled observed total variance within treatment status of Y, or it can be the maximum likelihood estimate $\hat{\sigma}_T^2$. Let $E(S^2) = b\sigma^2$

and $V(S^2) = 2c\sigma^4$, where a, b, c, and $\tilde{N}$ are known constants that are fixed by the study design (as shown later). As Searle et al. (2006) suggested, S^2 has an approximate sampling distribution of the product of a chi-squared with h degrees of freedom (df) and a constant k, which implies that S^2/k would follow a chi-squared distribution with h df. Because a chi-squared has an expected value equals df and a variance equals to $2df$, it follows that $\mathrm{E}(S^2) = kh = b\sigma^2$ and $V(S^2) = 2k^2h = 2c\sigma^4$. Solving the two equations we get $k = c\sigma^2/b$ and $h = b^2/c$.

Then define

$$T = \frac{\frac{\Delta\bar{Y}}{\sqrt{a\sigma^2/\tilde{N}}}}{\sqrt{\frac{S^2/(c\sigma^2/b)}{b^2/c}}} = \sqrt{\frac{\tilde{N}b}{a}}\left(\frac{\Delta\bar{Y}}{S}\right). \tag{A1}$$

The middle of the equality shows that the numerator of T is a normal random variable with variance equals 1, and the denominator of T is the square root of the ratio between a chi-squared and its degrees of freedom. Therefore, T follows a noncentral t distribution that has $df = b^2/c$ and a noncentral parameter θ equals the expectation of the numerator, that is

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

where $\delta = \Delta\mu/\sigma$. In the present study we are interested in the effect size $\delta_T = \Delta\mu/\sigma_T$. As suggested by Hedges (1981), by substituting $S = S_T$, where S_T is defined in equation (5), we get

$$D = \sqrt{b}\frac{\Delta\bar{Y}}{S_T} = T\sqrt{\frac{a}{\tilde{N}}} \tag{A2}$$

as a consistent estimator of δ_T (i.e., $D \to \delta_T$ in large sample) with approximate variance

$$\frac{a}{\tilde{N}} + \frac{c\delta_T^2}{2b^2}. \tag{A3}$$

From (A2), the effect size of interest, D, can be obtained by dividing the grand

mean difference of Y in the sample by the sample mean squares (which equals sum of squares divided by the degrees of freedom), and then multiplied by a correction factor $\sqrt{b}$, the square root of the ratio of the expected mean square to the variance component. In single-level data, the expected mean square equals the variance component, and thus $b = 1$ and no correction is needed. In data with cluster structure, however, $b \neq 1$ in general and so correction is needed.

An approximately unbiased estimator of δ_T would be $g = DJ(b^2/c)$ (Hedges, 2007, p. 361), where $J(x) \approx 1 - 3/(4x - 1)$. The difference between D and g is negligible for large sample size (Hedges, 1981). In multilevel studies, the bias of D is a function of b^2/c that increases when either or both of the average cluster size and the number of clusters increase. Roughly speaking, the bias is small as the total sample size is large. For example, with $K = K^T = K^C = 10$, $J^T = J^C = 10$, $\rho_A = \rho_B = .25$, $n = 0.5$ (i.e., on average each cell has 0.5 student), and thus $N^T = N^C = 50$, the expected values of D and g differ only by about 3%. So throughout the paper it is assumed that D is the effect size estimator of interest. Because D and $V(D)$ are defined solely in terms of a, b, c, and $\tilde{N}$, given the grand mean difference of Y of the treatment and of the control arms and their pooled unadjusted observed variance S_T^2, the next tasks are to express the constants a, b, c, and $\tilde{N}$ in terms of summary statistics in CCREMs and PCCREMs, and substituted these constants into (A2) and (A3) so that D and $V(D)$ can be defined in terms of summary statistics of the data.

D_1 for Fully Cross-Classified Data

Assuming a balanced design, the variance of the unweighted average of the cell means in the treatment arm will be

$$\begin{aligned} V(\bar{Y}^T_{\bullet\bullet\bullet}) &= \frac{\sigma^2_W}{J^T K n} + \frac{\sigma^2_A}{J^T} + \frac{\sigma^2_B}{K} \\ &= \frac{\sigma^2_T}{N^T}\left[1 + (Kn-1)\rho_A + (J^T n - 1)\rho_B\right]. \end{aligned}$$

Similar expression will apply to $V(\bar{Y}^C_{\bullet\bullet\bullet})$. It is assumed that the treatment arm and the control arm share the same intraclass correlations, and have the same cluster sizes. We thus get

$$V(\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet\bullet}) = \left(\frac{1}{N^T} + \frac{1}{N^C}\right)\sigma^2_T[1 + (Kn-1)\rho_A + (1-r_K)(Jn-2)\rho_B],$$

where r_K is the correlation due to the fact that observations in the treatment and the control arm may share some B-clusters, and in a balanced design it equals $\sqrt{(K_{\text{overlap}})^2/(K^T \times K^C)}$. For the special case of $K_{\text{overlap}} = 0, K^T = K^C = K$, we get $r_K = 1$ and thus the last term in the bracket $(1 - r_K)(Jn - 2)\rho_B$ equals zero, which explains the difference between equations (7) and (8).

Substituting $\sigma = \sigma_T$ into the above theorem, we get $\tilde{N}$ and a such that

$$\tilde{N} = \frac{1}{1/N^T + 1/N^C} = \frac{N^T N^C}{N^T + N^C}, \tag{A4}$$

$$a = 1 + (Kn-1)\rho_A + (1-r_K)(Jn-2)\rho_B. \tag{A5}$$

To find b and c, we need to consider the sampling distribution of the observed variance S^2_T. First define the three mean squares (MS) in a given data with cross-classified

structure:

$$SS_A = Kn \sum_{j=1}^{J^T} (\bar{Y}^T_{\bullet j \bullet} - \bar{Y}^T_{\bullet\bullet\bullet})^2 + Kn \sum_{j=1}^{J^C} (\bar{Y}^C_{\bullet j \bullet} - \bar{Y}^C_{\bullet\bullet\bullet})^2,$$

$$SS_B = J^T n \sum_{k=1}^{K} (\bar{Y}^T_{\bullet\bullet k} - \bar{Y}^T_{\bullet\bullet\bullet})^2 + J^C n \sum_{k=1}^{K} (\bar{Y}^C_{\bullet\bullet k} - \bar{Y}^C_{\bullet\bullet\bullet})^2,$$

$$SS_W = \sum_{j=1}^{J^T} \sum_{k=1}^{K} \sum_{i=1}^{n} (\bar{Y}^T_{ijk} - \bar{Y}^T_{\bullet\bullet\bullet})^2 + \sum_{j=1}^{J^C} \sum_{k=1}^{K} \sum_{i=1}^{n} (\bar{Y}^C_{ijk} - \bar{Y}^C_{\bullet\bullet\bullet})^2 - SS_A - SS_B.$$

Note that the interaction between A and B is assumed zero. The corresponding degrees of freedom for SS_A, SS_B, and SS_W are $J - 2$, $K - 1$, and $N - J - K + 1$. The sample MS can then be obtained as SS/df.

We then start with the expected mean squares, $\mathrm{E}(MS)$, for a balanced design with two crossed effects A and B:

$$\mathrm{E}(MS_A) = Kn\sigma^2_A + \sigma^2_W = \sigma^2_T[1 + (Kn - 1)\rho_A - \rho_B],$$

$$\mathrm{E}(MS_B) = Jn\sigma^2_B + \sigma^2_W = \sigma^2_T[1 - \rho_A + (Jn - 1)\rho_B],$$

$$\mathrm{E}(MS_W) = \sigma^2_W = \sigma^2_T(1 - \rho_A - \rho_B).$$

Note that for a given data MS_f for an effect f equals sum of squares SS_f divided by its degrees of freedom m_f. Assuming normality of the measured criterion variable, a mean squares multiplied by its $df = m_f$ and then divided by its expectation is distributed as a chi-squared with m_f degrees of freedom (Searle et al., 2006). Thus,

$$\mathrm{E}(SS_f) = m_f \, \mathrm{E}(MS_f),$$

$$V(SS_f) = 2m_f \left[\mathrm{E}(MS_f)\right]^2.$$

Apply the above results and the degrees of freedom for two groups design to different

random effects, we get the expectations of the variance components:

$$\begin{aligned}
\mathrm{E}(SS_A) &= (J-2)\sigma_T^2[1+(Kn-1)\rho_A-\rho_B],\\
\mathrm{E}(SS_B) &= (K-1)\sigma_T^2[1-\rho_A+(Jn-1)\rho_B],\\
\mathrm{E}(SS_W) &= (N-J-K+1)\sigma_T^2(1-\rho_A-\rho_B),
\end{aligned}$$

and their variances:

$$\begin{aligned}
V(SS_A) &= 2(J-2)\sigma_T^4\left[1+(Kn-1)\rho_A-\rho_B\right]^2,\\
V(SS_B) &= 2(K-1)\sigma_T^4\left[1-\rho_A+(Jn-1)\rho_B\right]^2,\\
V(SS_W) &= 2(N-J-K+1)\sigma_T^4\left(1-\rho_A-\rho_B\right)^2.
\end{aligned}$$

Because $S_T^2 = (SS_A + SS_B + SS_W)/(N-2)$, the expected value and the variance of S_T^2 are

$$\mathrm{E}(S_T^2) = \sigma_T^2\left[1-\frac{2(Kn-1)\rho_A+(Jn-2)\rho_B}{N-2}\right],$$

and

$$V(S_T^2) = \frac{2\sigma_T^2}{(N-2)^2}\left[Kn\check{N}_K\rho_A^2 + Jn\check{N}_J\rho_B^2 + (N-2)\bar{\rho}^2 + 2\check{N}_K\bar{\rho}\rho_A + 2\check{N}_J\bar{\rho}\rho_B\right],$$

where $\check{N}_K = N-2Kn$, $\check{N}_J = N-Jn$, and $\bar{\rho} = 1-\rho_A-\rho_B$. Again, substituting $\sigma = \sigma_T$ and $S = S_T$ into the above theorem, we get

$$b = 1-\frac{2(Kn-1)\rho_A+(Jn-2)\rho_B}{N-2}, \tag{A6}$$

$$c = \frac{1}{(N-2)^2}\left[Kn\check{N}_K\rho_A^2 + Jn\check{N}_J\rho_B^2 + (N-2)\bar{\rho}^2 + 2\check{N}_K\bar{\rho}\rho_A + 2\check{N}_J\bar{\rho}\rho_B\right]. \tag{A7}$$

Substituting a in equation (A5), b in (A6), and c in (A7) into equations (A2) and (A3), we

can obtain equations (6) and (8).

D_1 for Partially Cross-Classified Data

With a balanced design, the variance of the unweighted average of the cell means in the treatment arm and the unweighted average of the group means in the control arm can be shown respectively as

$$V(\bar{Y}^T_{\bullet\bullet\bullet}) = \frac{V^T \sigma^2_T}{N^T}, \tag{A8}$$

$$V(\bar{Y}^C_{\bullet\bullet}) = \frac{V^C \sigma^2_T}{N^C}, \tag{A9}$$

where $V^T = 1 + (Kn^T - 1)\rho_A + (J^T n^T - 1)\rho_B$ and $V^C = 1 + (n^C - 1)\rho_A$. Because the two treatment conditions are assumed to be independent,

$$V(\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet}) = \sigma^2_T \left(\frac{V^T}{N^T} + \frac{V^C}{N^C} \right). \tag{A10}$$

Note that $V(\bar{Y}^T_{\bullet\bullet\bullet} - \bar{Y}^C_{\bullet\bullet})$ can no longer be separated as a and $\tilde{N}$. Instead, we substitute $a^* = a/\tilde{N}$ into (A3), and thus the variance of the estimator D of the effect size δ_T becomes

$$a^* + \frac{c\delta^2_T}{2b^2}, \tag{A11}$$

where

$$a^* = \frac{1 + (Kn^T - 1)\rho_A + (J^T n^T - 1)\rho_B}{N^T} + \frac{1 + (n^C - 1)\rho_A}{N^C}. \tag{A12}$$

The expectations of the observed total variances in the treatment arm and in the

control arm are respectively

$$\mathrm{E}(S^2_{T|\mathrm{TREAT}}) = \sigma^2_T \left[1 - \frac{(Kn^T - 1)\rho_A + (J^T n^T - 1)\rho_B}{N^T - 1}\right] = \beta^T \sigma^2_T,$$
$$\mathrm{E}(S^2_{T|\mathrm{CON}}) = \sigma^2_T \left[1 - \frac{(n^C - 1)\rho_A}{N^C - 1}\right] = \beta^C \sigma^2_T,$$

where β^T and β^C are the correction factors for clustering for the within treatment status total variances. Their variances are

$$\begin{aligned} V(S^2_{T|\mathrm{TREAT}}) &= \frac{2\sigma^4_T \left[Kn^T \check{N}^T_K \rho^2_A + J^T n^T \check{N}^T_J \rho^2_B + (N^T - 1)\bar{\rho}^2 + 2\check{N}^T_K \bar{\rho}\rho_A + 2\check{N}^T_J \bar{\rho}\rho_B\right]}{(N^T - 1)^2} \\ &= 2(W^T)^{-1}\sigma^4_T, \\ V(S^2_{T|\mathrm{CON}}) &= \frac{2\sigma^4_T \left\{(N^C - 1) - 2(n^C - 1)\rho_A + (n^C - 1)\left[N^C - (n^C - 1)\right]\rho^2_A\right\}}{(N^C - 1)^2} \\ &= 2(W^C)^{-1}\sigma^4_T, \end{aligned}$$

where $\check{N}^T_K = N^T - Kn^T$ and $\check{N}^T_J = N^T - J^T n^T$. The weighted average of S^2_T will then be

$$S^2_T = \frac{W^T S^2_{T|\mathrm{TREAT}} + W^C S^2_{T|\mathrm{CON}}}{W^T + W^C}, \tag{A13}$$

and its variance will be

$$V(S^2_T) = \frac{2\sigma^4_T}{W^T + W^C}. \tag{A14}$$

We then get

$$b = \frac{W^T \beta^T + W^C \beta^C}{W^T + W^C},$$
$$c = \frac{1}{W^T + W^C}.$$

Then we can obtain the desired effect size estimates and the sampling variance by

substituting a^*, b, and c into (A2) and (A11).

Derivation of Effect Size Estimator D_2

When unbiased point estimates (or those with negligible bias) and the sampling variances (or standard errors) of both the fixed effect of the treatment and the variance components (i.e., the random effects variance) are available, the calculations of D and $V(D)$ are greatly simplified. First, we define the population effect size to be $\delta_T = \Delta\mu/\sqrt{\sigma_T^2}$. The sample estimator of $\Delta\mu$ is then the estimated fixed effect of the treatment, $\hat{\gamma}_{01}$, and that of σ_T^2 have been defined in the sections of estimation of D_2 for CCREM and for PCCREM, which is a function of $\hat{\sigma}_A^2$, $\hat{\sigma}_B^2$, and $\hat{\sigma}_W^2$. The estimated sampling variance of $\hat{\gamma}_{01}$ is $V(\hat{\gamma}_{01})$, and that of $V(\hat{\sigma}_T^2)$ is defined again in the section pertaining to D_2 and is a function of $V(\hat{\sigma}_A^2)$, $V(\hat{\sigma}_B^2)$, and $V(\hat{\sigma}_W^2)$. All of the quantities $\hat{\gamma}_{01}$, $V(\hat{\gamma}_{01})$, $\hat{\sigma}_A^2$, $V(\hat{\sigma}_A^2)$, $\hat{\sigma}_B^2$, $V(\hat{\sigma}_B^2)$, $\hat{\sigma}_W^2$, and $V(\hat{\sigma}_W^2)$ can be obtained from multilevel modeling software such as SAS or SPSS.

Specifically, from the theorem we have $\Delta\bar{Y} \sim N(\Delta\mu, a\sigma^2/\tilde{N})$. Replacing $\Delta\bar{Y}$ by $\hat{\gamma}_{10}$ and σ^2 by $\hat{\sigma}_T^2$, we get $\hat{\gamma}_{10} \sim N(\Delta\mu, a\hat{\sigma}_T^2/\tilde{N})$. Based on the variance of $\hat{\gamma}_{10}$, we get

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

Also from the theorem, we have $E(S^2) = b\sigma^2$. Now we use $\hat{\sigma}_T^2$ as the sample estimator, so $S^2 = \hat{\sigma}_T^2$. Because $\mathrm{E}(\hat{\sigma}_T^2) = \sigma_T^2$, we have $b = 1$, which merely reflects the fact that $\hat{\sigma}_T^2$ is an unbiased estimator of σ_T^2. The theorem also defines c such that $V(S^2) = 2c\sigma^4$. When replacing S^2 by $\hat{\sigma}_T^2$ and σ by $\hat{\sigma}_T$, we get

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

Then by substituting the above results of $a/\tilde{N}$, b, and c into (A2) and (A3), we obtain

equations (9), (10), (14), and (15).

APPENDIX B

GENERATING UNBALANCED CCREM DATA

Take, for example, $J^T = J^C = K = 20$, $n = 1$, and $N^T = N^C = 400$. The sizes of the $J^T = 20$ clusters were first generated by sampling on a multinomial distribution with total counts of 400 and equal probabilities. The resulting cluster sizes were for example 21, 10, 24, 18, 22, . . . , 25, 14, 22. Then, for each cluster, the cell sizes were generated by sampling on a multinomial distribution with a vector of $K = 20$ predefined probabilities with four being .179 (high) and sixteen being .0179 (low). The configuration of cells with high and low probabilities were shown in Table B1. The same procedure was used to generate unbalanced data in other conditions.

Table B1
Cell Probabilities for Generating Cell Counts for Data With Two Crossed Random Effects

		Random Effect B																			
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	1	X	X	X	X																
	2	X	X	X	X																
	3	X	X	X	X																
	4	X	X	X	X																
	5					X	X	X	X												
	6					X	X	X	X												
	7					X	X	X	X												
	8					X	X	X	X												
	9									X	X	X	X								
Random	10									X	X	X	X								
Effect A	11									X	X	X	X								
	12									X	X	X	X								
	13												X	X	X	X					
	14												X	X	X	X					
	15												X	X	X	X					
	16												X	X	X	X					
	17																X	X	X	X	
	18																X	X	X	X	
	19																X	X	X	X	
	20																X	X	X	X	

Note. Cells with an "X" have a probability that is 10 times of that for cells with no "X". For example, if rows represent classrooms and columns represent neighborhoods, then students from schools 1 to 4 are 10 times more likely to come from neighborhoods 1 to 4 than to come from neighborhoods 5 to 20.

APPENDIX C

DERIVATION OF SMD FOR PARTIALLY NESTED DESIGNS

The following steps to find an unbiased estimator were based on the work of Hedges (2007, pp. 360-362). The theorem part is also given in Appendix A on page 98 of this dissertation, and is not repeated in this appendix. As shown in equation (A2) and (A11), our task is to express a^*, b, and c in terms of known quantities, and substitute them into the two equations.

Derivation of D_1 for Partially Nested Designs

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. Denote $\bar{Y}^T_{\bullet\bullet}$ and $\bar{Y}^C_{\bullet}$ as the grand means for the treatment arm and the control arm, with corresponding sampling variance

$$V(\bar{Y}^T_{\bullet\bullet}) = \frac{\sigma^2_W + n\sigma^2_B}{N^T} = \sigma^2_W \frac{1 + n(1-\rho)}{N^T(1-\rho)}, \tag{C1}$$

$$V(\bar{Y}^C_{\bullet}) = \frac{\sigma^2_W}{N^C}. \tag{C2}$$

The last equality for $V(\bar{Y}^T_{\bullet\bullet})$ 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

$$\Delta Y = \bar{Y}^T_{\bullet\bullet} - \bar{Y}^C_{\bullet}, \tag{C3}$$

with sampling variance

$$V(\bar{Y}^T_{\bullet\bullet} - \bar{Y}^C_{\bullet}) = \sigma^2_W \left[\frac{1 + (n-1)\rho}{N^T(1-\rho)} + \frac{1}{N^C} \right]. \tag{C4}$$

The expectation and variance of the variance components would then be

$$\begin{aligned}
\mathrm{E}(SS_{W|\mathrm{TREAT}}) &= (N^T - J)\sigma_W^2, \\
\mathrm{E}(SS_{W|\mathrm{CON}}) &= (N^C - 1)\sigma_W^2, \\
V(SS_{W|\mathrm{TREAT}}) &= 2(N^T - J)\sigma_W^4, \\
V(SS_{W|\mathrm{CON}}) &= 2(N^C - 1)\sigma_W^4.
\end{aligned}$$

Because $S_W^2 = (SS_{W|\mathrm{TREAT}} + SS_{W|\mathrm{CON}})/(N^T - J + N^C - 1)$,

$$\mathrm{E}(S_W^2) = \frac{(N^T - J)\sigma_W^2 + (N^C - 1)\sigma_W^2}{N^T - J + N^C - 1} = \sigma_W^2$$

and

$$V(S_W^2) = \frac{2(N^T - J)\sigma_W^4 + 2(N^C - 1)\sigma_W^4}{(N^T - J + N^C - 1)^2} = \frac{2\sigma_W^4}{N - J - 1}.$$

Hence

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

We can now substitute a^*, b, and c into (A2) and (A11) to get the expressions in the main text.

If only the standard deviation of the control arm is used, and the homoscedasticity assumption is not made, then

$$V(\bar{Y}_{\bullet\bullet}^T) = \kappa\sigma_C^2 \frac{1 + n(1-\rho)}{N^T(1-\rho)}$$

and $c = 1/(N^C - 1)$, resulting in the formula for $V(D_1^C)$ in equation (31).

Derivation of D_2 for Partially Nested Designs

The procedure to obtain D_2 for CCREMs (see page 106 in Appendix A) can also be used for partially nested data. In order to calculate D_2, 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.

By definition we have $\Delta\bar{Y} \sim N(\Delta\mu, a^*\sigma^2)$. Replacing $\Delta\bar{Y}$ by $\hat{\gamma}_{10}$ and σ^2 by $\hat{\sigma}^2_W$, we have $V(\hat{\gamma}^2_{10}) = a^*\sigma^2_W$, and so

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

From the theorem, if we use $S^2 = \hat{\sigma}^2_W$ as an estimator for σ^2_W, we get $E(\hat{\sigma}^2_T) = b\sigma^2_W$. Assuming that $\hat{\sigma}^2_T$ is unbiased, that is, $\mathrm{E}(\hat{\sigma}^2_T) = \sigma^2_T$, 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 $\hat{\sigma}^2_W$ and σ by $\hat{\sigma}_W$, we get

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

Then by substituting a^*, b, and c into (A2) and (A11), one get the expressions for D_2 and $V(D_2)$.

APPENDIX D

CONSTRUCTING NONCENTRAL CONFIDENCE INTERVAL FOR EFFECT SIZE WITH PARTIALLY NESTED DATA

As shown in Appendix A, a sample estimator D, when multiplied by a constant λ, follow an approximate noncentral t distribution with noncentrality parameter equal $\lambda\delta$ with degrees of freedom (df) equal $N - J - 1$ and $2(\hat{\sigma}_W^2)^2/V(\hat{\sigma}_W^2)$ respectively. Here δ is the population effect size, and unless df is very small (say less than 15; Cumming & Finch, 2001) the expected values of D_1 and D_2 are very close to δ. Because the distribution of λD can be considered approximately independent of δ, it can be used as a pivotal quantity (Casella & Berger, 2002) for constructing CI. Denote $t_{p,\nu,\theta}$ as the p quantile of the noncentral t distribution with $df = \nu$ and noncentrality parameter = θ, then an approximate $(1 - \alpha) \times 100\%$ CI for D_1 and D_2 is obtained as (cf. Steiger & Fouladi, 1997)

$$[t_{\alpha/2,\nu,\lambda D}/\lambda, t_{1-\alpha/2,\nu,\lambda D}/\lambda],$$

where for D_1, $\nu_1 = N - J - 1$ and

$$\lambda_1 = \sqrt{\left[\frac{1 + (n-1)\rho}{N^T(1-\rho)} + \frac{1}{N^C}\right]^{-1}},$$

and for D_2, $\nu_2 = 2(\hat{\sigma}_W^2)^2/V(\hat{\sigma}_W^2)$ and $\lambda_2 = \sqrt{\hat{\sigma}_W^2/V(\hat{\gamma}_{10}^2)}$, as followed from Appendix C.

APPENDIX E

ESTIMATING EFFECT SIZE FOR PARTIALLY NESTED DATA WITH MPLUS

In structural equation modeling (SEM) software that allow multilevel data and nonlinear constraints, such as Mplus (L. K. Muthén & Muthén, 1998–2012), D_2, D_2^C, and their standard errors can be estimated simultaneously with other model parameters. The Mplus code for a simple hypothetical example is shown below.

Estimation of D_2

```
TITLE:  Effect size D2 for partially nested data;
DATA:   File = pnested.dat;
VARIABLE:
        Names = id y clus treat;
        Usevar = y treat;
        Cluster = clus;
        Within = treat;
ANALYSIS:
        Type = twolevel random;
MODEL:  %Within%
        s | y on treat;
        y (sigma2w);
        %Between%
        s; y@0;
        [s] (gamma10);
MODEL CONSTRAINT:
        new(d2);
        d2 = gamma10 / sqrt(sigma2w);
OUTPUT: Cinterval;
```

Estimation of D_2^C

```
TITLE:  Effect size D2c for partially nested data;
DATA:   File = pnested.dat;
VARIABLES:
        Names = id y clus treat;
        Usevar = y;
        Cluster = clus;
        Group = treat (0 = control 1 = treat);
ANALYSIS:
        Type = twolevel;
```

```
MODEL:  %Within%
        y;
        %between%
        y;
MODEL CONTROL:
        %Within%
        y (sigma2c);
        %Between%
        y@0;
        [y] (yc);
MODEL TREAT:
        %Within%
        y;
        %between%
        y;
        [y] (yt);
MODEL CONSTRAINT:
        new(d2c);
        d2c = (yt - yc) / sqrt(sigma2c);
OUTPUT: Cinterval;
```

APPENDIX F

R CODE FOR SIMULATION (PARTIALLY NESTED)

```
# ------------------------------------------------------------------------ #
# 2014 Mark Lai
#
# Script to run simulation to check the performance of standardized mean
# difference with partially nested data, as described in dissertation
# manuscript 2.
# ------------------------------------------------------------------------ #


# House keeping: remove all objects in workspace. -------------------------
rm(list = ls())

setwd("/mnt/Dropbox/Research/pnested_ES/sim_checking/")
source("mc_hack.R")

# Load required packages. -------------------------------------------------
library(lme4)
library(parallel)


# Define helper functions. ------------------------------------------------
ZeroTruncate <- function(dist) {
  # A function factory for generating random numbers from a zero-truncated
  # version of the given distribution. It works by first sampling in a
  # uniform distribution with range (F0, 1), where F0 is the cdf at 0.
  # The random numbers are then inverted to the corresponding quantiles.
  #
  # Args:
  #   dist: A character string of the kernel of the distribution.
  #         E.g., "norm" for normal, "pois" for Poisson, and
  #         "nbinom" for negative binomial.
  #
  # Returns:
  #   A function for generating random numbers. The first argument `n` is
  #   the number of observations, and the other arguments are the same
  #   as those in the non-zero-truncated counterparts.
  pdist <- get(paste0("p", dist))
  qdist <- get(paste0("q", dist))
  function(n, ...) {
    qdist(runif(n, pdist(0, ...), 1), ...)
  }
}
```

```
GenClusID <- function(nclus, ave_csize, var_inflat = NULL) {
  # Convert cluster sizes of length = nclus to a vector of cluster ids
  # of length = N
  #
  # Args:
  #   nclus: Number of clusters.
  #   ave_csize: Average cluster size.
  #   var_inflat: Ratio of variance to the mean of the distribution of
  #               cluster sizes. When it is null (default), the Poisson is
  #               used. Otherwise the negative binomial is used to get the
  #               desired variance.
  #
  # Returns:
  #   A sorted vector of length exactly equals N.
  if (is.null(var_inflat)) {
    csizes_unscaled <- ZeroTruncate("pois")(nclus, ave_csize)
  } else {
    csizes_unscaled <-
      ZeroTruncate("nbinom")(nclus, mu = ave_csize,
                             size = ave_csize / (var_inflat - 1))
  }

  N <- nclus * ave_csize
  csizes <- round(prop.table(csizes_unscaled) * N, 0)
  nclus <- length(csizes_unscaled)
  N_org <- sum(csizes)
  extra_pos <- sample(which(csizes > 2), abs(N_org - N),
                      prob = csizes[csizes > 2])
  csizes[extra_pos] <- csizes[extra_pos] - sign(N_org - N)
  clus_id <- rep(seq_len(nclus), times = csizes)
  return(clus_id)
}


CalcD1 <- function(mean_T, mean_C, s2_w, s2_b, N_T, N_C, J, icc, n) {
  # Compute effect size estimates D1 for partially nested data.
  # The input can be numeric values or vectors.
  #
  # Args:
  #   mean_T: Grand mean(s) of the outcome for the treatment arm.
  #   mean_C: Grand mean(s) of the outcome for the control arm.
  #   s2_w: Pooled within-level variance(s).
  #   s2_b: Between-level ariance(s) for the treatment arm.
  #   N_T: Total (Level-1) sample size(s) for the treatment arm.
  #   N_C: (Level-1) sample size(s) for the control arm.
  #   J: Number(s) of clusters in the treatment arm.
  #   icc: Intraclass correlation(s).
  #
  # Returns:
  #   A list including a vector of estimated D1 values and another vector of
```

```
  #   the corresponding variances.
  n <- N_T / J
  if (missing(icc)) { # compute ICC if necessary
    icc <- (s2_b - s2_w) / (s2_b + (n - 1) * s2_w)
  }

  est_D1 <- (mean_T - mean_C) / sqrt(s2_w)
  est_VarD1 <- (1 + (n - 1) * icc) / N_T / (1 - icc) + 1 / N_C +
    est_D1^2 / 2 / (N_T + N_C - J - 1)
  return(list(D1 = est_D1, VarD1 = est_VarD1))
}

CalcD2 <- function(gam10, var_gam10, sigma2, var_sigma2) {
  # Compute effect size estimates D2 for partially nested data.
  # The input can be numeric values or vectors.
  #
  # Args:
  #   gam10: Estimated gamma_10(s) (i.e., treatment effect).
  #   var_gam10: Estimated variance(s) of gamma_10.
  #   sigma2: Estimated pooled within-level variance component(s).
  #   var_sigma2: Estimated variance(s) of sigma2.
  #
  # Returns:
  #   A list including a vector of estimated D2 values and another vector of
  #   the corresponding variances.
  est_D2 <- gam10 / sqrt(sigma2)
  est_VarD2 <- var_gam10 / sigma2 + est_D2^2 * var_sigma2 / 4 / (sigma2)^2
  return(list(D2 = est_D2, VarD2 = est_VarD2))
}


# Wrapper function for computing D1 and D2 from raw data.
GetD1s <- function(data_T, data_C, clus_id, N_T, N_C, n_clus) {
  # A wrapper for computing D1 from matrices of datasets.
  #
  # Args:
  #   data_T: A matrix where each column is a data vector for the treatment.
  #   data_C: A matrix where each column is a data vector for the control.
  #   clus_id: A matrix where each column is a vector of cluster ID for
  #            the treatment.
  #   N_T: Within-level sample size for the treatment group.
  #   N_C: Within-level sample size for the control group.
  #   n_clus: Number of clusters in the treatment group.
  #
  # Returns:
  #   A list with four sublists: (a) D1, (b) variances of D1,
  #   (c) degrees of freedom for noncentral t approximation,
  #   (d) scaling factor for the noncentrality parameter.
  means_T <- colMeans(data_T)
  means_C <- colMeans(data_C)
  between_data_T <- vapply(  # Replace all data points in data_T by the group means
```

```
    seq_along(means_T), function(i) ave(data_T[ , i], clus_id[ , i]),
    FUN.VALUE = vector("numeric", N_T)
  )
  within_data <- rbind(data_T - between_data_T,  # group-mean centered data
                       data_C - mean(data_C))
  ComputeMS <- function(x, df, m = 0) {
    # Quicker function to compute mean squares (MS).
    # x = data (can be vector or matrix), df = degrees of freedom,
    # m = centering matrix; default to 0, meaning that x already centered.
    x_centered <- x - m
    diag(crossprod(x_centered)) / df
  }
  s2_b <- ComputeMS(between_data_T, n_clus - 1,
                    rep(1, N_T) %*% t(means_T)) # create centering matrix
  s2_w <- ComputeMS(within_data, N_T + N_C - n_clus - 1)
  n <- N_T / n_clus
  icc <- (s2_b - s2_w) / (s2_b + (n - 1) * s2_w)
  c(CalcD1(means_T, means_C, s2_w, s2_b, N_T, N_C, n_clus, icc),
    df1 = list(rep(N_T + N_C - n_clus - 1, length(means_T))),
    lambda1 = list(((1 + (n - 1) * icc) / N_T / (1 - icc) + 1 / N_C)^(-0.5)))
}

GetD2s <- function(data_all, clus_id, treat) {
  # A wrapper for computing D1 from matrices of datasets.
  #
  # Args:
  #   data_all: A matrix where each column is a combined data vector.
  #   clus_id: A matrix where each column is a vector of cluster ID.
  #            Each unit in the control group is treated as a cluster with
  #            unique ID.
  #   clus_id: A matrix where each column is a vector of cluster ID for
  #            the treatment.
  #   treat: A vector of 0 (treatment) and 1 (control).
  #
  # Returns:
  #   A list with four sublists: (a) D2, (b) variances of D2,
  #   (c) degrees of freedom for noncentral t approximation,
  #   (d) scaling factor for the noncentrality parameter.
  GetS2Boot <- function(ss, x) {
    # Compute sigma2w for bootstrap sample.
    foo <- try(sigma(refit(x, ss))^2, silent = TRUE)
    if (inherits(foo, "try-error")) NA
    else foo
  }
  ExtractParam <- function(data, clus_id, treat) {
    # Run mixed model and extract required parameters for CalcD2
    df <- data.frame(y = data, cid = clus_id)
    m1 <- lmer(y ~ treat + (treat - 1 | cid), data = df)
    # parametric bootstrap to obtain sampling variance
    ss <- simulate(m1 , nsim = 200)
    boot_var_sigma2 <- var(vapply(ss, GetS2Boot, FUN.VALUE = 1.0, x = m1),
```

```
                          na.rm = TRUE)
    return(c(est_gam01 = fixef(m1)["treat"],
             est_var_gam01 = vcov(m1)["treat", "treat"],
             est_sigma2 = sigma(m1)^2,
             est_var_sigma2 = boot_var_sigma2))
  }

  param_D2 <-  # extracted parameters
    vapply(seq_len(ncol(data_all)),
           function(i) ExtractParam(data_all[ , i], clus_id[ , i], treat),
           numeric(4))
  c(CalcD2(param_D2[1, ], param_D2[2, ], param_D2[3, ], param_D2[4, ]),
    df2 = list(2 * param_D2[3, ]^2 / param_D2[4, ]),
    lambda2 = list(sqrt(param_D2[3, ] / param_D2[2, ])))
}


RunSim <- function(nrep, n_clus, clus_size, N_ratio, icc, pop_ES,
                   csize_dist = c('pois', 'nbinom', 'balance'),
                   sigma2 = 1.0, seed = 50, simID = NULL,
                   save_each = FALSE) {
  # A function that simulate data, compute D1, D2, and their corresponding
  # variances, and evaluate the bias, efficiency, and the Mean-squared
  # error (MSE).
  #
  # Args:
  #   nrep: Number of replications for the simulation.
  #   n_clus: Number of clusters for the treatment arm.
  #   clus_size: (Average) Cluster size for the treatment arm.
  #   N_ratio: Ratio of within-level sample size between treatment and
  #            control arm.
  #   icc: Intraclass correlation.
  #   pop_ES: Population effect size.
  #   csize_dist: Distribution of cluster sizes. If `balance`, all clusters
  #               have the same size; if `pois`, small imbalance occurs;
  #               if `nbinom`, severe imbalance happens.
  #   sigma2: Within-level variance component; default to 1.0.
  #   seed: Random seed for data generation; default to 50.
  #   simID: Optional Condition ID; default to NULL.
  #   save_each: If TRUE, the function returns NULL, and the output for each
  #              condition is saved to the current working directory as
  #              "simresult-i.rds", where i = simID. If FALSE, the output
  #              will be print to console.
  #
  # Returns:
  #   If save_each = FALSE (default), returns a list composed of 8 sublists,
  #   in the order of: D1, VarD1, df1, lambda1, D2, Var2, df2, lambda2.
  #   If save_each = TRUE, then returns NULL.
  if (is.null(simID) & save_each == TRUE) {
    cat("No input on simID. Cannot save output to disk.
        Print to console instead")
```

```
    save_each = FALSE
  }

  set.seed(seed)  # set the seed
  # initialize values of sample sizes, between-level variance,
  #   dummy variable for treatment.
  N_T <- clus_size * n_clus
  N_C <- floor(N_T / N_ratio)
  N <- N_T + N_C
  tau_00 <- sigma2 * icc / (1 - icc)  # between-level variance component
  treat <- rep(c(1, 0), c(N_T, N_C))
  # Generate data (data_T, data_C) for all replications
  error_w <- matrix(rnorm(nrep * N, sd = sqrt(sigma2)),
                    ncol = nrep)
  error_b <- matrix(rnorm(nrep * n_clus, sd = sqrt(tau_00)),
                    ncol = nrep)
  if (csize_dist == 'balance') {  # same size for all clusters
    clus_id <- matrix(rep(1:n_clus, each = clus_size), nrow = N_T, ncol = nrep)
  } else {  # If `nbinom`, the variance is 10 times the mean
    var_inflat <- if (csize_dist == 'pois') NULL else 10
    clus_id <- replicate(nrep, GenClusID(n_clus, clus_size, var_inflat))
  }

  data_T <- vapply(seq_len(nrep), function(i) error_b[clus_id[ , i], i],
                   FUN.VALUE = vector("numeric", N_T)) +
    error_w[1:N_T, ] + pop_ES * sqrt(sigma2)
  data_C <- error_w[(N_T + 1):N, ]
  data_all <- rbind(data_T, data_C)
  D1_out <- GetD1s(data_T, data_C,  # Compute D1 (with partial vectorization)
                   clus_id, N_T, N_C, n_clus)
  # Compute D2 (with lapply, then with vectorization)
  clus_id <- rbind(clus_id, matrix((1:N_C) + n_clus, nrow = N_C, ncol = nrep))
  D2_out <- GetD2s(data_all, clus_id, treat)

  cat("Finish simulation Condition", simID, "\n")
  flush.console()
  if (save_each) saveRDS(c(D1_out, D2_out), paste0("simresult-", simID, ".rds"))
  else c(D1_out, D2_out)
}

# Test simulation function ------------------------------------------
# result <- RunSim(20, 30, 5, 1, .1, .5, "balance")
# cat("Mean estimated value of D1 =", mean(result$D1),
#     "\nMean estimated variance of D1 =", mean(result$VarD1),
#     "\nEmpirical variance of D1 =", var(result$D1),
#     "\n% SE Bias =", (mean(sqrt(result$VarD1)) / sd(result$D1) - 1) * 100, "%",
#     "\nMean estimated value of D2 =", mean(result$D2),
#     "\nMean estimated variance of D2 =", mean(result$VarD2),
#     "\nEmpirical variance of D2 =", var(result$D2),
#     "\n% SE Bias =", (mean(sqrt(result$VarD2)) / sd(result$D2) - 1) * 100,
#     "%\n")
```

```
# Run Simulation. ---------------------------------------------------------
# Define design factors and constants.
DESIGNFACTOR <- expand.grid(n_clus = c(15, 30),
                            clus_size = c(5, 25),
                            N_ratio = c(5, 1),
                            icc = c(.1, .5),
                            csize_dist = c('pois', 'nbinom'))
POP_ES <- .5
NREP <- 500
COND_TO_RUN <- 1:32  # define which conditions to run
# Run and time the simulation.
system.time(
  simresult <-
    mcMap(RunSim, nrep = nrep, n_clus = DESIGNFACTOR[COND_TO_RUN, 1],
          clus_size = DESIGNFACTOR[COND_TO_RUN, 2],
          N_ratio = DESIGNFACTOR[COND_TO_RUN, 3],
          icc = DESIGNFACTOR[COND_TO_RUN, 4],
          csize_dist = DESIGNFACTOR[COND_TO_RUN, 5],
          pop_ES = POP_ES, simID = COND_TO_RUN, save_each = TRUE,
          mc.cores = 8L)  # Use 4 cores in Linux; not applicable with Windows.
)
```

APPENDIX G

R CODE FOR SIMULATION (BOOTSTRAP EFFECT SIZE)

```
# ------------------------------------------------------------------------ #
# 2015 March 5, Mark Lai
#
# Script to run simulation to check the performance of bootstrap
# standardized mean difference with partially nested data, as described
# in dissertation manuscript 3.
# ------------------------------------------------------------------------ #


# House keeping: remove all objects in workspace. -------------------------
rm(list = ls())
dir.create("result", showWarnings = FALSE)

# Load required packages. -------------------------------------------------
library(lme4, quietly = TRUE)

# Define helper functions. ------------------------------------------------
ZeroTruncate <- function(dist) {
  # A function factory for generating random numbers from a zero-truncated
  # version of the given distribution. It works by first sampling in a
  # uniform distribution with range (F0, 1), where F0 is the cdf at 0.
  # The random numbers are then inverted to the corresponding quantiles.
  #
  # Args:
  #   dist: A character string of the kernel of the distribution.
  #         E.g., "norm" for normal, "pois" for Poisson, and
  #         "nbinom" for negative binomial.
  #
  # Returns:
  #   A function for generating random numbers. The first argument `n` is
  #   the number of observations, and the other arguments are the same
  #   as those in the non-zero-truncated counterparts.
  pdist <- get(paste0("p", dist))
  qdist <- get(paste0("q", dist))
  function(n, ...) {
    qdist(runif(n, pdist(0, ...), 1), ...)
  }
}


GenClusID <- function(nclus, ave_csize, var_inflat = NULL) {
  # Convert cluster sizes of length = nclus to a vector of cluster ids
  # of length = N
```

```
  #
  # Args:
  #   nclus: Number of clusters.
  #   ave_csize: Average cluster size.
  #   var_inflat: Ratio of variance to the mean of the distribution of
  #               cluster sizes. When it is null (default), the Poisson is
  #               used. Otherwise the negative binomial is used to get the
  #               desired variance.
  #
  # Returns:
  #   A sorted vector of length exactly equals N.
  if (is.null(var_inflat)) {
    csizes_unscaled <- ZeroTruncate("pois")(nclus, ave_csize)
  } else {
    csizes_unscaled <-
      ZeroTruncate("nbinom")(nclus, mu = ave_csize,
                             size = ave_csize / (var_inflat - 1))
  }

  N <- nclus * ave_csize
  csizes <- round(prop.table(csizes_unscaled) * N, 0)
  csizes[csizes == 0] <- 1
  nclus <- length(csizes_unscaled)
  N_org <- sum(csizes)
  extra_pos <- sample(which(csizes >= 2), abs(N_org - N),
                      prob = csizes[csizes >= 2])
  csizes[extra_pos] <- csizes[extra_pos] - sign(N_org - N)
  clus_id <- rep(seq_len(nclus), times = csizes)
  return(clus_id)
}


GetConfint <- function(d, type = c("asymptotic", "noncentral"),
                       ase, df, lambda) {
  # Get central (asymptotic) or noncentral confidence interval for multilevel
  # effect size.
  #
  # Args:
  #   d: Estimated effect size(s).
  #   type: Type of confidence interval.
  #   ase: Asymptotic standard error(s); used only for type = "asymptotic".
  #   df: Degrees of freedom(s) of noncentral t distribution; used only for
  #       type = "noncentral".
  #   lambda: Scaling factor(s) such that the noncentrality parameter of
  #           the noncentral t distribution is lambda * d; used only for
  #           type = "noncentral".
  #
  # Returns:
  #   A list of two vectors of upper and lower confidence limits.
  type = match.arg(type)
  if (type == "asymptotic") {
```

```
    return(list(as_ul = d + qnorm(.975) * ase,
                as_ll = d - qnorm(.975) * ase))
  } else if (type == "noncentral") {
    return(list(nc_ul = qt(.975, df, lambda * d) / lambda,
                nc_ll = qt(.025, df, lambda * d) / lambda))
  }
}


SimpleBoot <- function(x, FUN, nsim,
                       type = c("parametric", "semiparametric", "case"),
                       parallel_boot = FALSE, mc.cores = 1L + parallel_boot) {
  # Generate bootstrap samples of an estimator for fitted model object of lmer.
  # Random slope is not yet supported.
  #
  # Args:
  #   x: Fitted model object of class `lmerMer`.
  #   FUN: Function to be applied to each bootstrap samples.
  #   nsim: Number of bootstrap samples
  #   type: Type of bootstrapping. Parametric bootstrap generates both level-1
  #         and level-2 residuals from normal distributions, with variance
  #         equal to the estimated variance components. Nonparametric
  #         bootstrap samples "reflated" level-1 and level-2 residuals with
  #         with replacement from the original model. See Goldstein (2011)
  #         for detail.
  #   parallel_boot: Whether to use parallel computing with `mclapply`. Default
  #                  is FALSE.
  #   mc.cores: Number of cores to be used. Default is 2 for
  #             parallel_boot = TRUE.
  #
  # Returns:
  #   A vector of length = nsim of bootstrap statistics.
  type <- match.arg(type)
  if (type == "parametric") {
    ss <- simulate(x, nsim = nsim, use.u = FALSE)
  } else {
    group <- as.numeric(getME(x, "flist")[[1]])
    N <- getME(x, "N")
    if (type == "semiparametric") {
      n_clus <- max(group)
      vcs <- c(getME(x, "theta"), 1) * sigma(x)
      Qt <- cbind(as.matrix(ranef(x)[[1]][group, ]), residuals(x, "response"))
      S <- cov(Qt) * (N - 1) / N
      if (any(vcs == 0)) {  # handling when tau00 is zero
        A <- vcs / sqrt(diag(S))
        A[!is.finite(A)] <- 0
        A <- diag(A)
      } else {
        U_R <- diag(vcs)
        U_S <- try(base::chol(S), silent = TRUE)
        if (inherits(U_S, "try-error")) return(rep(NA, nsim))
```

```
        A <- try(solve(U_S, U_R), silent = TRUE)
        if (inherits(A, "try-error")) return(rep(NA, nsim))
      }
      Qt_star <- Qt %*% A
      fixed <- model.matrix(x) %*% fixef(x)
      U_star <- Qt_star[!duplicated(group) , -ncol(Qt_star), drop = FALSE]
      e_star <- Qt_star[ , ncol(Qt_star), drop = FALSE]
      Z <- getME(x, "Z")
      ss <- replicate(nsim, fixed +
                        as.matrix(Z %*% U_star[sample(n_clus, replace = TRUE), ,
                                               drop = FALSE]) +
                        e_star[sample(N, replace = TRUE), , drop = FALSE],
                      simplify = FALSE)
    } else {
      BootCase <- function(x, group, N) {
        new_index2 <- c(sample(unique(group), replace = TRUE))
        new_index1 <- lapply(new_index2, function(i) seq_len(N)[group == i])
        group_length <- vapply(new_index1, length, FUN.VALUE = integer(1))
        new_group <- rep(seq_along(new_index2), group_length)
        org_data <- x@frame
        fname <- names(getME(x, "flist"))
        new_data <- org_data[unlist(new_index1), , drop = FALSE]
        new_data[fname] <- new_group
        new_data
      }
      ss <- replicate(nsim, BootCase(x, group, N), simplify = FALSE)
    }}
  ffun <- local({
    type
    FUN
    refit
    x
    function(newsample) {
      if (type != "case") foo <- try(FUN(refit(x, newsample)), silent = TRUE)
      else {
        use_REML <- as.logical(getME(x, "REML"))
        foo <- try(FUN(lmer(formula(x), data = newsample, REML = use_REML)),
                   silent = TRUE)
      }
      if (inherits(foo, "try-error")) NA
      else foo
    }
  })

  if (!parallel_boot) vapply(ss, ffun, FUN.VALUE = 1.0, USE.NAMES = FALSE)
  else as.numeric(mclapply(ss, ffun, mc.cores = mc.cores))
}


CalcDT <- function(mean_T, mean_C, s2_t, icc, n, N_T, N_C) {
  # Compute effect size estimates D1 for nested data.
```

```
  # The input can be numeric values or vectors.
  #
  # Args:
  #   mean_T: Grand mean(s) of the outcome for the treatment arm.
  #   mean_C: Grand mean(s) of the outcome for the control arm.
  #   s2_t: Pooled sum of  variance component(s).
  #   icc: Intraclass correlation(s).
  #   n: (Average) cluster size.
  #   N_T: Total (Level-1) sample size(s) for the treatment arm.
  #   N_C: (Level-1) sample size(s) for the control arm.
  #
  # Returns:
  #   A list including a vector of estimated D1 values and two vectors of
  #   the upper and lower confidence limits.
  N <- N_T + N_C
  const_a <- 1 + (n - 1) * icc
  const_b <- 1 - (2 * (n - 1) * icc) / (N - 2)
  const_c <- ((N - 2) * (1 - icc)^2 + n * (N - 2 * n) * icc^2 +
                2 * (N - 2 * n) * icc * (1 - icc)) / (N - 2)^2
  N_tilde <- N_T * N_C / (N_T + N_C)
  est_DT <- (mean_T - mean_C) * sqrt(const_b / s2_t)
  est_VarDT <- const_a / N_tilde + const_c * est_DT^2 / 2 / const_b^2
  as_ci_DT <- GetConfint(est_DT, "asymptotic", ase = sqrt(est_VarDT))
  nc_ci_DT <- GetConfint(est_DT, "noncentral", df = const_b^2 / const_c,
                         lambda = sqrt(N_tilde / const_a))
  c(list(DT = est_DT), as_ci_DT, nc_ci_DT)
}

CalcDTM <- function(gam10, var_gam10, sigma2_T, var_sigma2_T) {
  # Compute effect size estimates D2 for nested data.
  # The input can be numeric values or vectors.
  #
  # Args:
  #   gam10: Estimated gamma_10(s) (i.e., treatment effect).
  #   var_gam10: Estimated variance(s) of gamma_10.
  #   sigma2_T: Estimated pooled sum of variance component(s).
  #   var_sigma2_T: Estimated variance(s) of sigma2_T.
  #
  # Returns:
  #   A list including a vector of estimated D2 values and two vectors of
  #   the upper and lower confidence limits.
  est_DTM <- gam10 / sqrt(sigma2_T)
  const_a_star <- var_gam10 / sigma2_T
  const_c <- var_sigma2_T / 2 / sigma2_T^2
  est_VarDTM <- const_a_star + est_DTM^2 * const_c / 2
  as_ci_DTM <- GetConfint(est_DTM, "asymptotic", ase = sqrt(est_VarDTM))
  nc_ci_DTM <- GetConfint(est_DTM, "noncentral", df = 1 / const_c,
                          lambda = 1 / sqrt(const_a_star))
  c(list(DTM = est_DTM), as_ci_DTM, nc_ci_DTM)
}
```

```
# Wrapper function for computing D1, D2, and bootDT from raw data or fitted
# data.
GetDTs <- function(data, treat, clus_id, N_T, N_C, n_clus) {
  # A wrapper for computing D1 from matrices of datasets.
  #
  # Args:
  #   data: A vector or matrix where each column is the raw response.
  #   treat: A vector or matrix where each column is the treatment dummy
  #          variable.
  #   clus_id: A matrix where each column is a vector of cluster ID for
  #            the treatment.
  #   N_T: Within-level sample size for the treatment group.
  #   N_C: Within-level sample size for the control group.
  #   n_clus: Number of clusters in the treatment group.
  #
  # Returns:
  #   A list with five sublists: (i) D1,
  #   (ii & iii) upper and lower asymptotic confidence limits of D1,
  #   (iv & v) upper and lower noncentral confidence limits of D1.
  N <- N_T + N_C
  means_T <- colMeans(data[treat == 1, ])
  means_C <- colMeans(data[treat == 0, ])
  between_data <- vapply(  # Replace all data points in data_T by the group means
    seq_along(means_T), function(i) ave(data[ , i], clus_id[ , i]),
    FUN.VALUE = vector("numeric", N)
  )

  ComputeMS <- function(x, df, m = 0) {
    # Quicker function to compute mean squares (MS).
    # x = data (can be vector or matrix), df = degrees of freedom,
    # m = centering matrix; default to 0, meaning that x already centered.
    x_centered <- x - m
    diag(crossprod(x_centered)) / df
  }

  grand_means <- rbind(rep(1, N_T) %*% t(means_T), rep(1, N_C) %*% t(means_C))
  s2_b <- ComputeMS(between_data, n_clus - 2, grand_means)
  s2_w <- ComputeMS(data, N - n_clus, between_data)
  n <- N / n_clus
  icc <- (s2_b - s2_w) / (s2_b + (n - 1) * s2_w)
  icc[icc < 0] <- 0
  s2_t <- ComputeMS(data, N - 2, grand_means)
  CalcDT(means_T, means_C, s2_t, icc, n, N_T, N_C)
}

GetDTMs <- function(model_all) {
  # A wrapper for computing D2 from fitted model objects.
  #
  # Args:
  #   model_all: A list with one or more fitted model objects.
```

```
  #
  # Returns:
  #   A list with five sublists: (i) D2,
  #   (ii & iii) upper and lower asymptotic confidence limits of D2,
  #   (iv & v) upper and lower noncentral confidence limits of D2.
  ExtractParam <- function(model) {
    # Run mixed model and extract required parameters for CalcD2
    # parametric bootstrap to obtain sampling variance
    GetSigma2T <- function(x) unname((1 + getME(x, "theta")^2) * sigma(x)^2)
    nboot_var <- 200
    boot_var_sigma2 <- var(SimpleBoot(model, GetSigma2T, nsim = nboot_var),
                           na.rm = TRUE) * (nboot_var - 1) / nboot_var
    c(est_gam01 = fixef(model)["treat"],
      est_var_gam01 = vcov(model)["treat", "treat"],
      est_sigma2_T = GetSigma2T(model),
      est_var_sigma2_T = boot_var_sigma2)
  }

  param_D2 <-  # extracted parameters
    vapply(model_all, ExtractParam, numeric(4))
  CalcDTM(param_D2[1, ], param_D2[2, ], param_D2[3, ], param_D2[4, ])
}

GetDTboots <- function(model_all, nsim, type, ...) {
  # A wrapper for computing bootstrap DT from fitted model objects.
  #
  # Args:
  #   model_all: A list with one or more fitted model objects.
  #   nsim: Number of bootstrap samples
  #   type: Type of bootstrapping. Parametric bootstrap generates both level-1
  #         and level-2 residuals from normal distributions, with variance
  #         equal to the estimated variance components. Nonparametric
  #         bootstrap samples "reflated" level-1 and level-2 residuals with
  #         with replacement from the original model. See Goldstein (2011)
  #         for detail.
  #   ... : Additional argument passed to SimpleBoot.
  #
  # Returns:
  #   A list with five sublists: (i) DTboot,
  #   (ii & iii) upper and lower percentile confidence limits of DTboot,
  #   (iv & v) upper and lower BCa confidence limits of DTboot.
  DTMer <- function(x) {
    # Compute DT from fitted object.
    unname(fixef(x)["treat"] / sqrt(1 + getME(x, "theta")^2) / sigma(x))
  }

  InfluenceJackm <- function(model) {
    # Jackknife estimate of influence on DT for each data point.
    group <- as.numeric(getME(model, "flist")[[1]])
    vapply(unique(group),
           function(i) DTMer(lmer(formula(model),
```

```
                                data = model@frame[group != i, ])),
          FUN.VALUE = numeric(1)) - DTMer(model)
  }

  BootDT <- function(x, nsim, type) {
    # Get percentile and BCa bootstrap.
    t <- DTMer(x)
    l_j <- InfluenceJackm(x)
    resample <- SimpleBoot(x, DTMer, nsim, type, ...)
    w <- qnorm(sum(resample <= t) / (nsim + 1))
    a <- sum(l_j^3) / sum(l_j^2)^1.5 / 6
    prob <- c(.025, .975)
    zalpha <- qnorm(prob)
    bca_p <- pnorm(w + (w + zalpha) / (1 - a * (w + zalpha)))
    # pnorm(2 * w + qnorm(prob))  # bias-corrected with no accelaration
    c(mean(resample), quantile(resample, c(prob, bca_p), na.rm = TRUE))
  }

  out <- vapply(model_all, BootDT, FUN.VALUE = numeric(5),
                nsim = nsim, type = type)
  list(DTboot = out[1, ], perc_ul = out[3, ], perc_ll = out[2, ],
       bca_ul = out[5, ], bca_ll = out[4, ])
}


RunSim <- function(nrep, n_clus, clus_size, icc, pop_ES,
                   csize_dist = c('balance', 'pois', 'nbinom'),
                   sigma2 = 1.0, lv1_dist = c("norm", "chisq"),
                   lv2_dist = c("norm", "chisq"), nboot = 10,
                   parallel_boot = FALSE, mc.cores = 1L + parallel_boot,
                   seed = 548, simID = NULL, save_each = FALSE) {
  # A function that simulate data, compute D1, D2, and their corresponding
  # variances, and evaluate the bias, efficiency, and the Mean-squared
  # error (MSE).
  #
  # Args:
  #   nrep: Number of replications for the simulation.
  #   n_clus: Number of clusters for the (treatment and control arm combined).
  #   clus_size: (Average) Cluster size for the treatment arm.
  #   icc: Intraclass correlation.
  #   pop_ES: Population effect size.
  #   csize_dist: Distribution of cluster sizes. If `balance`, all clusters
  #               have the same size; if `pois`, small imbalance occurs;
  #               if `nbinom`, severe imbalance happens.
  #   sigma2: Within-level variance component; default to 1.0.
  #   lv1_dist: Distribution for level-1 random effects. "norm" is normal;
  #             "chisq" is chi-squared.
  #   lv2_dist: Distribution for level-2 random effects. "norm" is normal;
  #             "chisq" is chi-squared.
  #   nboot: Number of bootstrap samples for each bootstrap method.
  #   parallel_boot: Whether to use parallel computing with `mclapply`. Default
```

```
#                  is FALSE.
#   mc.cores: Number of cores to be used. Default is 2 for
#             parallel_boot = TRUE.
#   seed: Random seed for data generation; default to 50.
#   simID: Optional Condition ID; default to NULL.
#   save_each: If TRUE, the function returns NULL, and the output for each
#              condition is saved to the current working directory as
#              "simresult-i.rds", where i = simID. If FALSE, the output
#              will be print to console.
#
# Returns:
#   If save_each = FALSE (default), returns a list composed of 20 sublists,
#   with 4 groups of effect size: D1, D2, DTboot (parametric),
#   DTboot (semiparametric). Each group with five lists: point estimate,
#   and 2 sets of confidence limits.
  if (save_each == TRUE) {
    if (is.null(simID)) {
      cat("No input on simID. Cannot save output to disk.
        Print to console instead\n\n")
      save_each = FALSE
    } else cat("Note: All results will be saved to the working directory.\n\n")
  }

csize_dist <- match.arg(csize_dist)
lv1_dist <- match.arg(lv1_dist)
lv2_dist <- match.arg(lv2_dist)

set.seed(seed)  # set the seed
# initialize values of sample sizes, between-level variance,
#   dummy variable for treatment.
N_T <- N_C <- clus_size * n_clus / 2
N <- N_T + N_C
tau_00 <- sigma2 * icc / (1 - icc)  # between-level variance component
treat <- rep(c(1, 0), c(N_T, N_C))
# Generate level-1 and level-2 random effects.
# chisq_df <- 1
rranef <- function(n, dist, var = 1) {
  # Generate random effect values.
  if (dist == "norm") return(rnorm(n, sd = sqrt(var)))
  else if (dist == "chisq") {
    x <- rchisq(n, df = 1)
    (x - 1) * sqrt(var / 2)
  }
}

error_w <- matrix(rranef(nrep * N, lv1_dist, var = sigma2), ncol = nrep)
error_b <- matrix(rranef(nrep * n_clus, lv2_dist, var = tau_00), ncol = nrep)
if (csize_dist == 'balance') {  # same size for all clusters
  clus_id <- matrix(rep(1:n_clus, each = clus_size), nrow = N, ncol = nrep)
} else {  # If `nbinom`, the variance is 10 times the mean
  var_inflat <- if (csize_dist == 'pois') NULL else 10
```

```
    clus_id <- replicate(nrep,
                         c(GenClusID(n_clus / 2, clus_size, var_inflat),
                           GenClusID(n_clus / 2, clus_size, var_inflat) +
                             n_clus / 2))
  }

  data <- vapply(seq_len(nrep), function(i) error_b[clus_id[ , i], i],
                 FUN.VALUE = vector("numeric", N)) + error_w +
    pop_ES * sqrt(sigma2 + tau_00) * treat
  # D1
    DT_out <- GetDTs(data, treat,  # Compute D1 (with partial vectorization)
                     clus_id, N_T, N_C, n_clus)
  data_df <- lapply(seq_len(nrep),
                    function(i) data.frame(y = data[ , i], cid = clus_id[ , i]))
  m1_all <- lapply(data_df, lmer, formula = y ~ treat + (1 | cid))
  # D2
  DTM_out <- GetDTMs(m1_all)
  # Parametric bootstrap DT
    DTboot_par_out <- GetDTboots(m1_all, nboot, "parametric",
                                 parallel_boot = parallel_boot,
                                 mc.cores = mc.cores)
  #   # Semiparametric bootstrap DT
    DTboot_spar_out <- GetDTboots(m1_all, nboot, "semiparametric",
                                  parallel_boot = parallel_boot,
                                  mc.cores = mc.cores)
  #   # Case bootstrap DT
    DTboot_npar_out <- GetDTboots(m1_all, nboot, "case",
                                  parallel_boot = parallel_boot,
                                  mc.cores = mc.cores)
    cat("Finish simulation Condition", simID, "\n")
    flush.console()
    output <- c(DT_out, DTM_out,
                DTboot_par_out,
                DTboot_spar_out,
                DTboot_npar_out)
    if (save_each) saveRDS(output, paste0("result/simresult-", simID, ".rds"))
    else output
  return(DTM_out)
}

# Test simulation function ------------------------------------------
# result <- RunSim(2, 20, 4, .4, .8, "nbinom", parallel_boot = FALSE)
# cat("Mean estimated value of D1 =", mean(result$D1),
#     "\nMean estimated variance of D1 =", mean(result$VarD1),
#     "\nEmpirical variance of D1 =", var(result$D1),
#     "\n% SE Bias =", (mean(sqrt(result$VarD1)) / sd(result$D1) - 1) * 100, "%",
#     "\nMean estimated value of D2 =", mean(result$D2),
#     "\nMean estimated variance of D2 =", mean(result$VarD2),
#     "\nEmpirical variance of D2 =", var(result$D2),
#     "\n% SE Bias =", (mean(sqrt(result$VarD2)) / sd(result$D2) - 1) * 100,
#     "%\n")
```

```
# Run Simulation. ---------------------------------------------------------
# Define design factors and constants.
DESIGNFACTOR <- expand.grid(n_clus = c(20, 30, 70),
                            clus_size = c(5, 25),
                            icc = c(.05, .1, .2),
                            pop_ES = c(.5),
                            csize_dist = c('pois', 'nbinom'),
                            lv1_dist = c('norm', 'chisq'),
                            lv2_dist = c('norm', 'chisq'),
                            stringsAsFactors = FALSE)
NREP <- 1000
COND_TO_RUN <- seq_len(nrow(DESIGNFACTOR))  # define which conditions to run
# Run and time the simulation.
time_proc <- system.time({
  jid <- seq_along(COND_TO_RUN)
  simresult <-
    mclapply(jid, function(i) {
      RunSim(nrep = NREP,
             n_clus = DESIGNFACTOR[COND_TO_RUN[i], 1],
             clus_size = DESIGNFACTOR[COND_TO_RUN[i], 2],
             icc = DESIGNFACTOR[COND_TO_RUN[i], 3],
             pop_ES = DESIGNFACTOR[COND_TO_RUN[i], 4],
             csize_dist = DESIGNFACTOR[COND_TO_RUN[i], 5],
             lv1_dist = DESIGNFACTOR[COND_TO_RUN[i], 6],
             lv2_dist = DESIGNFACTOR[COND_TO_RUN[i], 7],
             nboot = 999, simID = COND_TO_RUN[i], save_each = TRUE)
    }, mc.cores = 2L)
})

print(time_proc)
```