Bootstrap Confidence Intervals for Multilevel Standardized Effect Size

Mark H. C. Lai

University of Southern California

Author Note

Mark H. C. Lai, Department of Psychology, University of Southern California.

Correspondence concerning this article should be addressed to Mark Lai (Email: hokchiol@usc.edu), Department of Psychology, University of Southern California, Los Angeles, CA 90089-1061.

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

The Version of Record of this manuscript has been accepted and will appear in *Multivariate Behavioral Research*, https://doi.org/10.1080/00273171.2020.1746902.

Abstract

Although many methodologists and professional organizations have urged applied researchers to compute and report effect size measures accompanying tests of statistical significance, discussions on obtaining confidence interval (CI) for effect size with clustered/multilevel data have been scarce. In this paper, I explore the bootstrap as a viable and accessible alternative for obtaining CIs for multilevel standardized mean difference effect size for cluster-randomized trials. A simulation was carried out to compare 17 analytic and bootstrap procedures for constructing CIs for multilevel effect size, in terms of empirical coverage rates and width, for both normal and nonnormal data. Results showed that, overall, the residual bootstrap with studentized CI had the best coverage rates (94.75% on average), whereas the residual bootstrap with basic CI had better coverage in small samples. These two procedures for constructing CIs showed better coverage than using analytic methods for both normal and nonnormal data. In addition, I provide an illustrative example showing how bootstrap CIs for multilevel effect size can be easily obtained using the statistical software R and the R package `bootmlm`. I strongly encourage applied researchers to report CIs to adequately convey the uncertainty of their effect size estimates.

Keywords: effect size, multilevel, cluster-randomized trial, bootstrap, standardized mean difference, robustness, nonnormal data

Bootstrap Confidence Intervals for Multilevel Standardized Effect Size

Although many methodologists have urged researchers to compute and report effect size measures accompanying tests of statistical significance (e.g., Cohen, 1990; Cumming, 2014; Kelley & Preacher, 2012; Thompson, 2007), discussions on interval estimates of effect size with clustered and multilevel data have been scarce (e.g., Hedges, 2007; Snijders & Bosker, 1994), despite recommendations from many professional organizations (AERA, 2006; APA, 2010; Appelbaum et al., 2018) and authors (Hedges, 2008; Thompson, 2002). One plausible reason is that the computational formulas for confidence intervals (CIs) for effect size with multilevel data (Hedges, 2007) have not been implemented in standard statistical software packages for multilevel modeling, and those formulas rely on asymptotic theories that may not hold for finite samples. In addition, multilevel data comprise a collection of data structures with varying numbers of clustering levels and relations between levels (i.e., nested vs. crossed), which makes it tedious to derive complex formulas for CIs for multilevel research designs where analytic formulas for effect size and the corresponding sampling variance have not been derived. Recognizing such difficulties, in this manuscript I explore the *bootstrap* (Efron, 1982) as a general and robust alternative for obtaining CIs for multilevel effect size.

The two most common types of effect size statistics are standardized mean difference (e.g., Cohen's d) and proportion of variance accounted for (i.e., R^2) effect sizes (Grissom & Kim, 2012; Rosenthal, 1994). Given that this paper mainly focus on the former, I use the term "effect size" to mean standardized mean difference, unless otherwise specified.

In the past two decades, effect size reporting has been the central theme in the "statistical reform" in 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 researchers to report CIs so that the uncertainty associated with a sample effect size can be adequately quantified (AERA, 2006; APA, 2010; Hedges, 2008; Thompson, 2002). 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 requirement that was reiterated in its most recent *Journal Article Reporting Standard* (Appelbaum et al., 2018). A similar statement the AERA (2006) reporting standards stated 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). Methodological scholars, such as Hedges (2008) and Thompson (2002), have also made similar recommendations.

Whereas point estimates of effect size measures are regularly reported in single-level studies, the CIs for effect size estimates are still rarely attached. Peng, Chen, Chiang, and Chiang (2013) reviewed 32 review papers of effect size reporting practices in published articles from 116 journals in education and psychology, and noted that whereas the median effect size reporting rates were 58.0% after 1999, few of the studies reported CIs, contrary to the APA and AERA guidelines. In a review of articles in the *Journal of Experimental Psychology: General*, an APA journal, Fritz, Morris, and Richler (2012) found that the reporting rate for CIs for effect size was zero in 71 articles (see also Byrd, 2007, for a similar finding). This is in sharp contradiction to the existing guidelines on effect size reporting.

Reporting CIs for effect size is important for two major reasons. First, the sample effect size is only a point estimate, just like a sample mean, correlation, or regression coefficient. It can be a noisy estimate of the population effect size, especially when the sample size is small, and the point estimate itself does not convey the amount of uncertainty associated with it. An effect size of .80 with a 95% CI [.75, .85] is perceived very differently from an effect size of .80 but with a 95% CI [.10, 1.50]. Second, CIs for effect size plays an increasingly important role in sample size planning (or power analysis). Traditionally, effect size estimates in a pilot study or previous research are commonly used for sample size calculations for subsequent studies. However, due to publication bias, many of the published research tend to have an overestimated effect size (van

Assen, van Aert, & Wicherts, 2015). Using an overestimated effect size for sample size planning perpetuates the problems of underpowered studies and noisy estimates (Maxwell, 2004). To counter that, Perugini, Gallucci, and Costantini (2014) suggested using the safeguard power analysis, which requires the use of the lower limit of an 80% or 95% CI of effect size in a previous study for sample size planning (see also Anderson & Maxwell, 2017, for alternative methods to adjust the sample size estimate when planning for a replication study).

Although different researchers have proposed methods for obtaining CIs for effect size, applied researchers may not be familiar with them or may find them difficult to use by hand. Even for single-level studies, the analytic formulas for the sampling variance of effect size are not simple, and some of these methods for obtaining CIs invoke noncentral distributions (Smithson, 2001), which is seldom part of the quantitative training for behavioral researchers. For the simplest multilevel structure with two nested levels, the analytic expression of the sampling variance of effect size already fills two lines of space (Hedges, 2007); for the more complicated cross-classified structure the variance of effect size takes three full lines (Lai & Kwok, 2014). Furthermore, these formulas or analytic methods tend to be restrictive, as they rely on large sample approximations, assume normality of the error terms and the random effects, and cannot handle extensions such as covariate adjustment and effect heterogeneity (i.e., random slopes).

On the other hand, computer intensive methods, such as the bootstrap (Efron & Tibshirani, 1993), require analytic formula of only the point but not the variance estimates of a parameter, which greatly alleviates the technical burden on applied researchers. Some forms of the bootstrap, as discussed later, also have the added advantage of automatically handling some violations of assumptions, such as the assumption of normal errors and random effects, which makes them attractive alternatives for obtaining CIs for effect size and other types of effect size statistics.

With the ultimate goal of encouraging researchers to report CIs for multilevel effect size in mind, in the current study I aim to (a) introduce and review various analytic and bootstrap methods to construct CIs for multilevel effect size; (b) compare the coverage, symmetry, and width of these CIs using simulations for both normal and nonnormal data across a variety of design features, and provide evidence-based recommendations; and (c) illustrate how one can obtain bootstrap CIs for multilevel effect size with the provided R code and the R package `bootmlm` (Lai, 2019). Although the present study concerns mainly the use of the bootstrap for effect size with two-level data, the method can easily be applied to other types of effect size measures and to more complicated data structures. Also, the scope of the present study is limited to the use of multilevel boostrapping to cross-sectional data, as longitudinal data usually requires different ways of defining effect size and require more complex multilevel models (e.g., autocorrelations, random slopes) that can be addressed in a separate paper.

The Bootstrap

Efron (1982) has popularized the bootstrap method for obtaining standard errors (*SE*s) and variances (i.e., SE^2) of statistical estimators when closed form solutions are difficult or impossible to obtain. One of the most familiar applications of the bootstrap in behavioral research is mediation analysis (e.g., Preacher & Hayes, 2004). For mediation, the sampling distribution of an indirect effect, which is a product of two path coefficient estimators, is generally skewed even when multivariate normality holds, making the standard procedures of significance testing and CI construction biased. MacKinnon, Lockwood, and Williams (2004) showed that the bias-corrected bootstrap outperformed other methods in terms of power and confidence limit accuracy for constructing CI for the indirect effect, but Hayes and Scharkow (2013) showed that it suffered from inflated Type I error rates and suboptimal coverage especially with a small sample size.

Similarly, Kelley (2005) and Chen and Peng (2015) recommended the bootstrap for estimating CIs for effect size with single-level data. Like indirect effects, even with normal data, an effect size δ is obtained as a mean difference, which is normally distributed, divided by a standard deviation, which is generally the square root of a scaled chi-squared variable, so the sampling distribution of δ follows a noncentral t distribution, which is generally skewed (Hedges, 1981). The noncentral t distribution becomes only an approximation at best when the data are not normal and when cluster sizes are not equal with multilevel data. With the bootstrap, researchers do not need to rely on asymptotic normality for multilevel effect size, which generally does not hold for finite sample size, nor do they need to derive complex expressions for the corresponding standard errors to construct CIs. Below I review the definition of effect size for a two-level cluster-randomized trial below, and then introduce various methods for constructing CIs.

Multilevel Standardized Mean Difference Effect Size

In many randomized studies in behavioral and educational research, data are clustered and the randomization happens at the upper level. Examples include students within schools (e.g., Jones, Brown, Hoglund, & Aber, 2010), participants in social groups (Bull, Levine, Black, Schmiege, & Santelli, 2012), and nurses within hospitals (Bolier et al., 2014), wherein schools, social groups, and hospitals are assigned to different intervention groups. A commonly used multilevel model for data from a two-level cluster-randomized trial consists of a level-1 model,

$$y_{ij} = \beta_{0j} + \varepsilon_{ij}, \tag{1}$$

and a level-2 model,

$$\beta_{0j} = \gamma_{00} + \gamma_{01}\text{TREAT}_j + u_{0j}, \tag{2}$$

where y_{ij} is the outcome of the ith individual in the jth cluster, TREAT is coded as 1 = treatment arm and 0 = control arm, β_{0j} is the cluster mean of the jth cluster, γ_{00} is the grand mean of the control arm, and γ_{01} quantifies the average treatment effect. The level-1 and level-2 error terms, ε_{ij} and u_{0j}, are assumed independent, with $\varepsilon_{ij} \sim N(0, \sigma^2_W)$ and $u_{0j} \sim N(0, \sigma^2_B)$.

The coefficient γ_{01} represents the unstandardized effect size on the raw metric of y. For many variables in social and behavioral sciences, however, the unit of y is not intrinsically meaningful, and measurement of the same construct across different studies may not be comparable (Blanton & Jaccard, 2006), so it is a common practice to compute effect size in standard deviation unit (Cohen, 1988), resulting in standardized mean difference effect size.

Under the above model, one way to define effect size is (Hedges, 2007, equation 3)

$$\delta_T = \frac{\gamma_{01}}{\sqrt{\sigma_W^2 + \sigma_B^2}}, \tag{3}$$

where δ_T is on the unit of the *total SD*.[1]

Obtaining CIs for Effect Size With Two-Level Data

There are various methods to compute effect size and obtain a CI under the model in (1) and (2). In this simulation study, I compared 17 procedures to construct CIs for multilevel effect size, including two analytic methods and 15 bootstrap procedures.

Analytic Method

For the two analytic (i.e., non-bootstrap) methods, they require the computations of a point (d) and a variance estimate (v) of effect size, after which one can obtain a symmetric $(1 - \alpha) \times 100\%$ CI as

$$[d - z_{1-\alpha/2}\sqrt{v}, d + z_{1-\alpha/2}\sqrt{v}], \tag{4}$$

where $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the standard normal distribution.

Analysis of variance (ANOVA) method. The ANOVA method (Searle, Casella, & McCulloch, 2006) uses the formula presented in Hedges (2007) to estimate the effect size, $d = d_T$, and the corresponding sampling variance, $v = V(d_T)$. When each cluster contains n observations, meaning that the cluster size is constant, a sample estimator of δ_T is (Hedges, 2007, p. 349)

$$d_T = \left(\frac{\bar{Y}^T_{\cdot\cdot} - \bar{Y}^C_{\cdot\cdot}}{S_T}\right)\sqrt{1 - \frac{2(n-1)\rho}{N-2}}, \tag{5}$$

with sampling variance

$$V(d_T) = \left(\frac{N^T + N^C}{N^T N^C}\right)[1 + (n-1)\rho] \\ + \delta_T^2 \left(\frac{(N-2)(1-\rho)^2 + n(N-2n)\rho^2 + 2(N-2n)\rho(1-\rho)}{2[(N-2) - 2(n-1)\rho]^2}\right), \quad (6)$$

where N^T and N^C are the number of level-1 units and $\bar{Y}^T_{..}$ and $\bar{Y}^C_{..}$ are the grand means of the outcome, respectively for the treatment and the control arms, $\rho = \sigma_B^2/(\sigma_W^2 + \sigma_B^2)$ is the intraclass correlation (ICC), and S_T is the pooled total sample standard deviation defined as

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

with J^T and J^C being the number of clusters in the treatment and the control arm respectively. In most situations, as δ_T is not known, one has to replace it with d_T when computing $V(d_T)$.

When cluster sizes are not constant, Hedges (2007) also derived a formula for d_T and $V(d_T)$. 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. Nevertheless, when raw data are available, the method based on linear mixed-effects model (LMM-based method), as described next, is generally more efficient and simpler to use, and thus should be preferred.

LMM-based approach. Using the linear mixed-effects model, when model estimates of γ_{01}, σ_W^2, and σ_B^2 (denoted as $\hat{\gamma}_{01}$, $\hat{\sigma}_W^2$, and $\hat{\sigma}_B^2$), and their sampling variances, are available, the sample effect size can be estimated as (Hedges, 2009, equation 18.24)

$$\hat{\delta}_T = \frac{\hat{\gamma}_{01}}{\sqrt{\hat{\sigma}_W^2 + \hat{\sigma}_B^2}} \quad (7)$$

with variance

$$V(\hat{\delta}_T) = \frac{V(\hat{\gamma}_{01})}{\hat{\sigma}_W^2 + \hat{\sigma}_B^2} + \frac{\hat{\gamma}_{01}^2[V(\hat{\sigma}_W^2) + V(\hat{\sigma}_B^2)]}{4(\hat{\sigma}_W^2 + \hat{\sigma}_B^2)^3}. \quad (8)$$

Note that, with maximum likelihood or restricted maximum likelihood estimation, $V(\hat{\gamma}_{01})$ is simply the squared value of the *SE* of $\hat{\gamma}_{01}$, whereas $V(\hat{\sigma}^2_B)$ and $V(\hat{\sigma}^2_W)$ need to be obtained from the asymptotic variance-covariance matrix of the random effects, which are available in SAS (Littell, 2006) and SPSS (Norusis, 2012).[2] For the `lme4` package (Bates, Mächler, Bolker, & Walker, 2015) in R, which was used in the current study, they can be obtained with the `vcov_vc()` function in the `bootmlm` package.

Bootstrap Methods

In the statistical and methodological literature, there are multiple bootstrap approaches even for single-level analyses, mainly depending on whether and what parametric assumptions are involved (Davison & Hinkley, 1997). Also, after a bootstrap sampling distribution of an estimate (e.g., an effect size estimate) is formed, there are multiple procedures to construct CIs. With multilevel data, additional care is needed to address the dependencies in the data when resampling. Below I first summarize the common multilevel bootstrap procedures studied in the current paper, and then discuss the various methods to obtain CIs. Readers interested in more detailed discussions on the bootstrap should consult Davison and Hinkley (1997) for the background statistical theory, and Goldstein (2011) and Van der Leeden, Meijer, and Busing (2008) for in-depth review on multilevel bootstrapping methods.

Parametric bootstrap (PBoot). In the parametric bootstrap for multilevel data (Goldstein, 2011), one first fit the multilevel model described in (1) and (2) to obtain estimates of the fixed and random effects, typically based on either maximum likelihood or restricted maximum likelihood estimation. Then, for each bootstrap sample, a new set of level-1 errors, ε^*_{ij}, is drawn from independent $N(0, \hat{\sigma}^2_W)$ distributions, and a new set of level-2 random effects, u^*_{0j}, is drawn from independent $N(0, \hat{\sigma}^2_B)$ to form a new set of responses, $y^*_{ij} = \hat{\gamma}_{00} + \hat{\gamma}_{01}\text{TREAT}_j + u^*_{0j} + \varepsilon^*_{ij}$. The multilevel model is then refitted to the new bootstrap data, and $\hat{\delta}^*_T$ is computed using equation (7). To obtain the studentized CI (discussed later), $V(\hat{\delta}^*_T)$ also needs to be estimated for each bootstrap sample, for example using equation (8). The process is then repeated for R (e.g., $R = 1,999$)

bootstrap samples to obtain a bootstrap sampling distribution of $\hat{\delta}^*_T$.

Residual bootstrap (RBoot). I followed the approach proposed in Carpenter, Goldstein, and Rasbash (2003) and Goldstein (2011) to implement the residual bootstrap for multilevel data. The residual bootstrap is similar to the parametric bootstrap, but instead of simulating new errors from the corresponding normal distributions, in residual bootstrap one obtains new errors by sampling *with replacement* the residuals based on a fitted model. In multilevel residual bootstrap, one first obtains the level-2 and level-1 residuals, $\tilde{u}_{0j}$ and $\tilde{\varepsilon}_{ij}$, from a fitted model. However, because the sampling variance of $\tilde{u}_{0j}$ is smaller than $\hat{\sigma}^2_B$, and similarly but to a lesser degree for $\tilde{\varepsilon}_{ij}$ and $\hat{\sigma}^2_W$, one needs to first reflate the raw residuals so that their sample variances match the estimated variance components. To reflate $\tilde{u}_{0j}$, one first computes its empirical variance, $S^2_{\tilde{u}} = \sum_{j=1}^J \tilde{u}^2_{0j}/J$, where J is the number of clusters, and then obtain the reflated level-2 residuals as $\sqrt{\hat{\sigma}^2_B/S^2_{\tilde{u}}}\tilde{u}_{0j}$. The same procedure is applied to obtain the reflated level-1 residuals. Then, for each bootstrap sample, new sets of level-2 and level-1 residuals, u^*_{0j} and ε^*_{ij}, are sampled with replacement from the reflated level-2 and level-1 residuals. After that, as in parametric bootstrap, a new set of responses y^*_{ij} is computed and $\hat{\delta}^*_T$ is computed, and the process is repeated R times to obtain a bootstrap sampling distribution of $\hat{\delta}^*_T$.[3] Vallejo Seco, Ato García, Fernández García, and Livacic Rojas (2013) showed that the residual bootstrap method produced more precise estimates, in terms of a smaller root mean squared error, for fixed effects than restricted maximum likelihood.

Case bootstrap (CBoot). In single-level analyses, with the case bootstrap the observations (as opposed to the residuals) are resampled with replacement (Davison & Hinkley, 1997). The case bootstrap assumes that each observation is independent, which is clearly violated with multilevel data (Goldstein, 2011; Hox, 2010; Roberts & Fan, 2004; Van der Leeden et al., 2008). While there are multiple proposals to adapt the case bootstrap for multilevel data (Roberts & Fan, 2004; Van der Leeden et al., 2008), the major ones are (a) to resample with replacement the level-2 clusters, and keep the level-1 units in a cluster intact, and (b) to resample first the clusters, and within each cluster resample the level-1 units. Both schemes result in bootstrap samples with level-1 sample size different from the original size when the cluster sizes are not equal. For simplicity, only scheme (a), which was recommended over scheme (b) by both Davison and Hinkley (1997) and Goldstein (2011), is included in the simulation study for comparison. It should be noted that because the case bootstrap makes fewer assumptions than the parametric and the residual bootstraps, it requires more information from the data. Not surprisingly, previous literature found that its performance was poor compared to the other two methods, even when the assumptions made in the parametric bootstrap and for the residual bootstrap were violated (Efron & Tibshirani, 1993; Van der Leeden et al., 2008). The exception is for designs where each cluster has the same number of observations, as Thai, Mentré, Holford, Veyrat-Follet, and Comets (2013) found that in longitudinal linear-mixed models where cluster size is constant, residual bootstrap and case bootstrap performed similarly when there were at least 100 individuals (i.e., $J = 100$).

Bootstrap confidence interval. There are numerous ways to construct CIs after obtaining a bootstrap sampling distribution (Davison & Hinkley, 1997; Efron & Tibshirani, 1993; Van der Leeden et al., 2008; Wu, 1986). For this study I focus on the five common procedures discussed in Davison and Hinkley (1997), which are available in the `boot` package (Canty & Ripley, 2017) in R (R Core Team, 2019). The five procedures for obtaining bootstrap CIs are briefly reviewed below.

Normal approximation (norm). With normal approximation, one simply replaces v from equation (4), the analytic estimation of the sampling variance of $\hat{\delta}_T$, with the bootstrap estimate $v^* = \sum_{r=1}^{R}(\hat{\delta}_T^{*(r)} - \bar{\hat{\delta}}_T^{*})^2/(R-1)$ to obtain an approximate $(1-\alpha) \times 100\%$ symmetric CI as

$$\left[\hat{\delta}_T - z_{1-\alpha/2}\sqrt{v^*}, \hat{\delta}_T + z_{1-\alpha/2}\sqrt{v^*}\right]. \tag{9}$$

Basic CI (basic). The basic CI (Davison & Hinkley, 1997), sometimes referred to as the percentile-*t* CI or the Hall's percentile CI (Van der Leeden et al., 2008), uses the bootstrap distribution of the bias, $\hat{\delta}_T^* - \hat{\delta}_T$ to obtain a CI defined as

$$\left[2\hat{\delta}_T - \hat{\delta}_{T,1-\alpha/2}^*, 2\hat{\delta}_T - \hat{\delta}_{T,\alpha/2}^*\right], \tag{10}$$

where $\hat{\delta}^*_{T,\alpha/2}$ and $\hat{\delta}^*_{T,1-\alpha/2}$ are the α and $1 - \alpha/2$ quantiles of the bootstrap distribution. Note that the bootstrap estimate of bias, $\bar{\hat{\delta}}^*_T - \hat{\delta}_T$, can also be used to correct for the bias in $\hat{\delta}_T$; for the current study, however, $\hat{\delta}_T$ is found to be approximately unbiased and so I focus mainly on the bootstrap CIs.

Studentized CI (stud). The studentized CI (Davison & Hinkley, 1997), also called the bootstrap-*t* CI, is a second-order accurate CI by using $z^* = \frac{\hat{\delta}_T - \delta_T}{\sqrt{V(\hat{\delta}^*_T)}}$ as a pivot, where $V(\hat{\delta}^*_T)$ is the approximate variance of $\hat{\delta}^*_T$ based on a bootstrap sample. This means that for each bootstrap sample one needs both the point and variance estimates of $\hat{\delta}_T$. The studentized CI is generally more accurate when the *SE* of an estimator depends on the estimated value, which is the case for $\hat{\delta}_T$ based on equations (7) and (8) for multilevel effect size. The studentized CI can be obtained as:

$$\left[\hat{\delta}_T - z^*_{1-\alpha/2}\sqrt{V(\hat{\delta}^*_T)}, z^*_{1-\alpha/2}\hat{\delta}_T + \sqrt{V(\hat{\delta}^*_T)}\right]. \tag{11}$$

Percentile CI (perc). The percentile CI (Efron, 1982) is a simple method to construct a $(1 - \alpha)$ CI from a bootstrap distribution by taking the $100\alpha/2$th and the $100(1 - \alpha/2)$th percentiles of the distribution (e.g., the 2.5th and the 97.5th percentiles for a 95% CI). It is denoted as

$$[\hat{\delta}^*_{T,\alpha/2}, \hat{\delta}^*_{T,1-\alpha/2}]. \tag{12}$$

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. However, Davison and Hinkley (1997) and Van der Leeden et al. (2008) both commented that the percentile CI tends to produce biased confidence limits, especially when using the case bootstrap, and especially when the original estimator is biased or when the original sample size is small (Efron & Tibshirani, 1993).

Bias-corrected and accelerated (BCa) CI. One can improve the performance of the percentile CI with the (BCa) method (Efron, 1987). Like the percentile CI, the BCa CI is constructed by taking two quantile values in the bootstrap sampling distribution. However, instead of using the $\alpha/2$ and the $1 - \alpha/2$ quantiles, one uses the α_L and α_U quantiles defined as (Efron & Tibshirani, 1993, p. 185)

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

where $\Phi(\cdot)$ is the cumulative density function (cdf) for the standard normal distribution. The two correcting factors are w and a, with w correcting for median bias and a correcting for the skewness (acceleration) of the distribution of the estimator, and can be estimated as

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

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).[4] The indicator function, $I_{\{\hat{\delta}_T^{*(r)} < \hat{\delta}_T\}}$, equals 1 when $\hat{\delta}_T^* < \hat{\delta}_T$ for the rth bootstrap sample and equals 0 otherwise. When the estimator is unbiased, $w = 0$; when the influence function is symmetric, $a = 0$. Therefore, for an unbiased estimator with a symmetric sampling distribution, $\alpha_L = \alpha/2$ and $\alpha_U = 1 - \alpha/2$, and the BCa CI is equivalent to the percentile CI. Like the studentized CI, the BCa CI is second-order accurate.

Simulation Study

The simulation study compares the performances of various procedures of constructing CIs for δ_T estimates.

Design Factors

There were seven design factors in the simulation study, as described below.

Intraclass correlation (ICC). The three ICC levels were chosen to reflect the common values in cross-sectional psychological and educational research based on a review of articles involving clustered data 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. The levels also match the range used in Flynn and Peters (2004). As found in Vallejo Seco et al. (2013), the performance of the residual bootstrap got worse with increasing ICC.

Distribution of $\boldsymbol{u_0}$ ($\boldsymbol{P[u_0]}$). Because one of the goal of the present study is to recommend procedures for constructing CIs when the normality assumption is violated, the level-2 random effects, u_0, followed either a normal distribution or a scaled skew-t distribution with four degrees of freedom and a slant parameter of 10. The slant parameter regulates the skewness of the variable, with slant = 0 resulting in a symmetric variable and larger values of slant giving more positive skewness.[5] For conditions with nonnormal u_0, a skewed-t variable was first simulated using the sn package in R (Azzalini, 2019), from which the theoretical expected value was subtracted to have a mean zero, and then rescaled so that the variance matched the desired σ^2_B value. The average sample skewness of the simulated u_0 values was 1.59 and the average sample kurtosis was 3.47. For all conditions, $\sigma^2_W = 1.0$, and so $\sigma^2_B = \text{ICC}/(1 - \text{ICC})$.

Distribution of $\boldsymbol{\varepsilon}$ ($\boldsymbol{P[\varepsilon]}$). Similar to $P(u_0)$, the distribution of level-1 errors, ε, was either normal or scaled skewed-t, with variance $\sigma^2_W = 1$. The same scaled skewed-t distribution and procedure to generate nonnormal u_0 was used to simulate nonnormal ε. The average sample skewness of the simulated u_0 values was 2.64 and the average sample kurtosis was 15.19. The inclusion of conditions with normal $P(u_0)$ and nonnormal $P(\varepsilon)$, and normal $P(\varepsilon)$ and nonnormal $P(u_0)$, allows examination of the relative impact of nonnormality at level 1 and at level 2.

Population effect size ($\boldsymbol{\delta_T}$). The simulation included $\delta_T = 0$ for the null treatment effect condition and $\delta_T = 0.5$ for the non-null treatment effect condition. An effect size of 0.5 was close to the recommended minimum effect size suggested by Ferguson (2009) and the minimally clinically important value suggested by Angst, Aeschlimann, and Angst (2017), and was similar to the values used in previous simulation studies (e.g., Feingold, 2015). Based on previous simulation results (e.g., Feingold, 2015; Kelley, 2005), δ_T was not expected to to have a big impact on the performance of the various procedures for constructing CIs.

Number of clusters (*J*). It is generally agreed that multilevel models require at least 30 clusters (e.g., Hox, 2010). In this study, J was set to either 20, 30, or 70 based on the literature review, and there was an equal number of clusters in the treatment and the control arms (i.e., $J^T = J^C$). The level with 20 clusters matched 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 was consistent with the small cluster size conditions in previous simulation studies (e.g., Maas & Hox, 2005; Thai et al., 2013), whereas the medium value was chosen to represent typical classroom size in the United States.

Imbalance of cluster sizes. Among the five methods for estimating δ_T, only the ANOVA method assumes equal cluster sizes. However, Ames (2013) showed that the LMM-based estimates of effect size can also be biased with unequal cluster sizes. Unequal cluster sizes also reduce the effective sample size (Candel & Breukelen, 2009), which may make both the analytic methods and the bootstrap methods less stable. The simulation thus included both balanced and unbalanced data conditions. For the unbalanced conditions, the J clusters were divided into five strata, each with $J/5$ clusters, and the cluster sizes were $\bar{n}/5$, $3\bar{n}/5$, $\bar{n}$, $7\bar{n}/5$, $9\bar{n}/5$, respectively, so that the ratio of the largest to the smallest cluster sizes was 9 to 1. For example, when $J = 30$ and $\bar{n} = 5$, for the unbalanced cluster size condition there were six clusters with only one observation, six clusters with three observations, six with five observations, six with seven observations, and the remaining six with nine observations.

Data Generation and Analyses

Computation for the work described in this paper was supported by the University of Southern California's Center for High-Performance Computing (hpcc.usc.edu). In this study there were a total of 3 (ICC) × 2 ($P[u_0]$) × 2 ($P[\varepsilon]$) × 2 (δ_T) × 3 (J) × 2 ($\bar{n}$) × 2 (imbalance of n) = 288 conditions. For each of the 288 condition, there were 1,000 replication data sets, and in each of which J values of u_0 and N values of ε were generated independently in R (R Core Team, 2019) according to the predefined distribution (i.e., normal or nonnormal) for each condition. As previously explained, for skewed u_0 and ε, the `rst()` function in the `sn` package was used to simulate nonnormal data, and they were transformed to have a population mean of zero and a population variance of desired value. For normal u_0 and ε, the `rnorm()` function in R was used. Then the outcome (y) values were computed based on the model described in equation (1). For each data set, I first obtained d_T and $V(d_T)$ based on the ANOVA method, and LMM-based $\hat{\delta}_T$ and $V(\hat{\delta}_T)$, in order to compute the analytic 95% CIs. Based on the LMM-based result of each data set, I used the R package `bootmlm` to run the parametric, residual, and case bootstrap, each with 1,999 bootstrap samples. For each type of bootstrap methods, I then obtained five 95% CIs (norm, basic, stud, perc, and BCa). Therefore, for each data set there were 17 CIs to be compared (e.g., RBoot-stud for the residual bootstrap with studentized CI, and PBoot-basic for the paramteric bootstrap with normal approximation CI). All multilevel model fitting were performed using the `lmer()` function in the R package `lme4` (Bates et al., 2015). All 17 CIs were successfully computed for all replications across all simulation conditions, although in some replications with small sample sizes, $\hat{\sigma}^2_B$ were estimated to be zero, in which case $V(\hat{\sigma}^2_W)$ was used to substitue $V(\hat{\sigma}^2_W) + V(\hat{\sigma}^2_B)$ in equation (8).

Evaluation Criteria

Across simulation conditions, the point estimates of ES based on both the ANOVA and LMM-based methods were approximately unbiased, with biases between -0.018 and 0.039 across simulation conditions. The following outcomes were obtained regarding the performance of the 95% CIs.

Coverage of 95% CI. If the methods for constructing CI are valid, then it is expected that in 950 out of the 1,000 replications the CIs constructed would include δ_T, with a standard error of $\sqrt{.95 \times .05/1,000} \approx 0.69\%$. Let $\hat{\delta}_{T,L}$ and $\hat{\delta}_{T,U}$ be the lower and upper limits of a sample 95% CI for δ_T. The empirical coverage percentage was calculated as

$$\frac{\sum_{r=1}^{R} I_{\{\delta_T \in [\hat{\delta}_{T,L}^{(r)}, \hat{\delta}_{T,U}^{(r)}]\}}}{R} \times 100\%$$

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

Lower- and upper-tail error rates. In addition to maintaining a close to 95% coverage rate, it is generally preferable for a CI construction procedure to not systematically be above or below the parameter. Thus, I also examined the upper- and lower-tail error rates as

$$\text{Error rate}_{\text{up}} = \frac{\sum_{r=1}^{R} I_{\{\delta_T > \hat{\delta}_{T,U}^{(r)}\}}}{R}$$
$$\text{Error rate}_{\text{lo}} = \frac{\sum_{r=1}^{R} I_{\{\delta_T < \hat{\delta}_{T,L}^{(r)}\}}}{R}$$

For any two procedures of constructing 95% CIs with 95% coverage rates, the one with 2.5% error rates for both upper and lower tails would be better than one that is consistently below (or above) the population parameter with a 5% upper-tail (or lower-tail) error rate.

To better interpret the symmetry of the intervals, I also computed a shape index as

$$\text{Shape} = \frac{\text{Error rate}_{\text{up}} - \text{Error rate}_{\text{lo}}}{\text{Error rate}_{\text{up}} + \text{Error rate}_{\text{lo}}}, \tag{13}$$

where shape > 0 indicates that the CI has a larger error rate on the upper tail.

Width of 95% CI. The width of a confidence interval was defined as $\hat{\delta}_{T,U} - \hat{\delta}_{T,L}$. If two CIs have similar empirical coverage percentages but with different widths, the one with a shorter average length should be preferred as it is more precise. Because the CI width was highly dependent on design factors such as J, $\bar{n}$, and ICC, for each condition I also obtained the relative CI width of the bootstrap CIs as the width ratio of the bootstrap CI to the LMM-based CI, with values larger than 1 indicating that the width of the bootstrap CI is longer.

Although a total of 17 procedures for constructing CIs were studied, the CIs based on the

ANOVA method and the case bootstrap showed suboptimal performance (see Table 1), which was consistent with previous literature. Therefore, the Result section only focuses on the comparison of the remaining 11 procedures, namely the LMM-based (linear-mixed-effects-model-based) approach and the parametric and the residual bootstrap procedures, and the full results can be obtained from the author.

To summarize the influence of the seven between-condition design factors (ICC, $P[u_0]$, $P[\varepsilon]$, δ_T, J, $\bar{n}$, imbalance), as well as two within-condition factors (i.e., method: LMM-based, RBoot, etc and CI type: basic, stud, BCa, etc), I conducted a split-plot ANOVA for each evaluation criterion. Because coverage rates and left-tail and right-tail error rates are intrinsically bounded between 0 and 1, I used a logit transformation, $\text{logit}(x) = \log[x/(1-x)]$, for these measures before running the ANOVAs. For each ANOVA, I included all main, two-way, and three-way interaction effects and computed the η^2 effect size for each effect (i.e., the sum of squares of each effect divided by the total [within- and between-condition] sum of squares), as preliminary analyses showed that none of the 4-way or higher-order interaction effects yielded η^2 larger than 0.5%.

Table 1 should appear here

Results

Coverage

For coverage, the design factors with the highest η^2 were CI type (η^2 = 29.2%), method (η^2 = 12.9%), method × $P(\varepsilon)$ interaction (η^2 = 12.2%), CI type × $\bar{n}$ interaction (η^2 = 11.3%), and CI type × J interaction (η^2 = 4.45%). The η^2 values for terms involving ICC (< 3%), δ_T (< 1.7%), $P(u_0)$ (< 1.3%), and imbalance (< 0.5%) were small. As shown in Table 1, among the 11 procedures for constructing CIs, residual bootstrap with studentized CI (RBoot-stud) gave the best coverage (M_{coverage} = 94.8%, SD = 0.82%), followed by RBoot with basic CI (RBoot-basic; 94.5%, SD = 0.92%), parametric bootstrap with stud CI (PBoot-stud; 94.4%, SD = 1.02%), and RBoot with normal approximation CI (RBoot-norm; 94.3%, SD = 0.89%). The performance of RBoot-stud, RBoot-basic, RBoot-norm, and PBoot-stud were also better than other procedures and had similar coverage for conditions with nonnormal $P(\varepsilon)$ and/or nonnormal $P(u)$. On the other hand, the LMM-based analytic CI was suboptimal with nonnormal $P(\varepsilon)$ and/or nonnormal $P(u)$ (M_{coverage} = 93.7%), as elaborated more below.

Figure 1 should appear here

Normal data. Figure 1 shows the coverage percentages with normal $P(u_0)$ and normal $P(\varepsilon)$. Overall, the LMM-based CI showed undercoverage (M_{coverage} = 93.8%), but improved with large $\bar{n}$ and J. Even when both u_0 and ε were normally distributed, due to the non-normality of the sampling distribution of $\hat{\delta}_T$, many of the bootstrap CIs still had better coverage percentages than LMM-based CI, especially with small J and large $\bar{n}$. For example, when $J = 20$ and $\bar{n} = 25$, the average coverage percentage for LMM-based CI was 92.8%, while RBoot-stud and PBoot-stud CIs had average coverage percentages of 94.8% to 94.9%. On the other hand, when $J = 70$ and $\bar{n} = 5$, LMM-based CI was similar to PBoot-stud (94.3% vs. 94.7%) and was close to optimal. Furthermore, although in most situations PBoot-stud and RBoot-stud had the best coverage percentages, for conditions with $\bar{n} = 5$, PBoot-basic and RBoot-basic (94.4% average) were slightly better than PBoot-stud and RBoot-stud (94.3% average), especially when J was also small (94.3% to 94.4% vs. 94.0%). Therefore, in conditions with small cluster sizes, a simpler procedure like the basic CI tends to perform better.

Figure 2 should appear here

Nonnormality at level 1. Figure 2 shows the coverage percentages with nonnormal $P(\varepsilon)$ and normal $P(u_0)$. Most CI construction procedures that assume normality showed worse coverage, such as LMM-based CI (M_{coverage} = 93.0%, *SD* = 1.1%) and PBoot CIs (M_{coverage} = 93.0% to 93.9%). On the other hand, as expected the coverage percentages for most RBoot CIs were not affected by nonnormal $P(\varepsilon)$; specifically, RBoot-stud (94.6%) and RBoot-basic (94.5%) had the highest average coverage percentages.

Figure 3 should appear here

Nonnormality at level 2. Figure 3 shows the coverage percentages with normal $P(\varepsilon)$ and nonnormal $P(u_0)$. The pattern in general was similar to that with normal $P(\varepsilon)$ and normal $P(u_0)$, indicating that level-2 nonnormality had less impact on coverage rates than level-1 nonnormality. The LMM-based CI showed reasonable coverage (M_{coverage} = 94.4%). Overall, PBoot-stud (M_{coverage} = 94.9%) and RBoot-stud (M_{coverage} = 94.9%) performed the best.

Figure 4 should appear here

Nonnormality at level 1 and level 2. Figure 4 shows the coverage percentages with nonnormal $P(\varepsilon)$ and nonnormal $P(u_0)$. With nonnormality at both levels, three RBoot CIs showed the best average coverage percentages: 95.1% for RBoot-stud, 95.0% for RBoot-basic, and 94.8% for RBoot-norm. All other procedures had average coverage percentages of 94.3% or lower. As also shown in the figure, there were more variability in the coverage rates across conditions, and additional analyses revealed that, for conditions with nonnormal $P(\varepsilon)$ and nonnormal $P(u_0)$, the population effect size δ_T explained a substantial proportion of variance (η^2 = 10.3%). Specifically, for all 11 procedures, the average coverage percentages were lower for $\delta_T = 0.5$ than $\delta_T = 0$ by 1.0 (RB-perc) to 2.1 (PBoot-BCa) percentage points. Despite the difference, RBoot-stud (M_{coverage} = 95.7% and 94.4%), RBoot-basic (M_{coverage} = 95.6% and 94.3%), and RBoot-norm (M_{coverage} = 95.4% and 94.2%) were still the best procedures for both $\delta_T = 0$ and $\delta_T = 0.5$.

Lower- and Upper-Tail Error Rates

Table 1 shows the average lower and upper-tail error rates (in percentages) of the 11 procedures for constructing CI for δ_T. As expected, the error rate was higher on the upper tail than the lower tail, especially with a larger δ_T (M_{shape} = .08 for $\delta_T = 0.50$ and .06 for $\delta_T = 0$), because when the normality assumptions hold, the sampling distribution of $\hat{\delta}_T$ has heavier density on the upper tail with increasing δ_T (Hedges, 2007).

Given that balanced error rates are not as important when the coverage rate is suboptimal, I only focused on the error rates of the three best methods in terms of coverage: RBoot-stud (studentized CI with residual bootstrap), RBoot-basic (basic CI with residual bootstrap), and

PBoot-stud (studentized CI with parametric bootstrap). The mean lower-tail error rates were 2.62%, 2.77%, and 2.66%, respectively, whereas the upper-tail error rates were 2.63% (shape = .003), 2.69% (shape = −0.01), and 2.93% (shape = 0.04). Therefore, the error rates were relatively balanced for RBoot-stud, RBoot-basic, but not for PBoot-stud.

Width of 95% CI

An ANOVA analysis on the relative CI widths (i.e., the ratio to the width of LMM-based CI) showed that the terms with highest η^2 were CI type (η^2 = 22.22%), CI type × $\bar{n}$ (η^2 = 9.21%), methods × $P(\varepsilon)$ interaction (η^2 = 9.06%), methods (η^2 = 8.24%), $P(\varepsilon)$ (η^2 = 8.08%), and J (η^2 = 5.98%). Comparing the width of the 11 procedures, on average the bootstrap CIs had slightly longer width than that of LMM-based CI ($M_{\text{Rel CI width}}$ between 1.01 and 1.04), which partly explained the undercoverage of the LMM-based CI in some simulation conditions. The studentized procedures ($M_{\text{Rel CI width}}$ = 1.03 for PBoot-stud and 1.04 for RBoot-stud) generally yielded the longest intervals, followed by the BCa procedures ($M_{\text{Rel CI width}}$ = 1.02 and 1.03).

Figure 5 should appear here

Figure 5 showed the distributions of the width of the 95% CI of the 10 bootstrap procedures relative to that of the LMM-based CI across conditions of J, $\bar{n}$, and normality of the level-1 error term ($P[\varepsilon]$). Here I only focus on methods that generally show good coverage across conditions, namely RBoot-stud, RBoot-basic, PBoot-stud, and PBoot-basic. For conditions with normal $P(\varepsilon)$ and smaller sample sizes, studentized CIs had worse coverage and longer intervals than basic CIs when $\bar{n}$ = 5 and J = 20 or 30. For example, when J = 20, PBoot-basic had an average coverage percentage of 94.4% and an average CI width of 0.98, compared to 94.0% and 1.00 for PBoot-stud. The coverage and width of the studentized and the basic CIs were similar when $\bar{n}$ = 5 and J = 70 (difference in average width = .004). When $\bar{n}$ = 25, PBoot-basic and RBoot-basic had suboptimal coverage (M_{width} = 93.6%), so PBoot-stud and RBoot-stud should be preferred even with wider intervals. For conditions with nonnormal $P(\varepsilon)$ or $P(u_0)$, PBoot CIs generally had suboptimal coverage, and among RBoot CIs, RBoot-stud had similar coverage and width as

RBoot-basic when $\bar{n} = 5$ (M_{coverage} = 94.9% vs. 95.2%; M_{width} = 0.78 for both), but better coverage percentages when $\bar{n} = 25$ (M_{coverage} = 95.0% vs. 94.3%).

Overall, when sample size was large (i.e., $\bar{n} \geq 25$ or $J \geq 70$), RBoot-stud had good coverage and was robust to nonnormality. With small sample size but with violations of normality, RBoot-stud was still on par with RBoot-basic. When sample size was small (e.g., $\bar{n} = 5$ and $J \leq 30$) and normality assumption was met, PBoot-basic and RBoot-basic should be preferred.

Illustrative Example

To illustrate the use of the bootstrap methods to obtain CIs for effect size for cluster-randomized trials, I used a simulated data set based on the results and parameter estimates in Haug et al. (2017). Both the simulated data and the R script are included in the supplemental material so that readers can reproduce the results. In the original study, 1,041 students from 80 schools ($N^T = 547$, $J^T = 43$) in Switzerland were randomly assigned, at the school level, to either receive a web- and text messaging-based intervention aiming to reduce problem drinking or an assessment-only control. The intervention lasted 3 months and provided normative feedback on alcohol consumption to participants, such as heavy drinking occasions and maximum number of drinks per week based on a prior study. One of the outcome variable at 6-month follow up was the estimated peak blood alcohol concentration (BAC). Whereas Haug et al. did not report the skewness and kurtosis of this variable, other studies have indicated that the distribution of blood alcohol concentration is usually positively skewed (e.g., White, Mun, & Morgan, 2008), so I simulated the random effects and level-1 errors to be nonnormal, similar to the ones in the simulation study, with sample skewness and kurtosis of 2.07 and 4.82.

To obtain effect size and 95% CI estimates, I first fitted a multilevel model with `lme4` on the simulated data, including a fixed effect for intervention and a random intercept for school, the same model as in (1) and (2). The parameter estimates were: $\hat{\gamma}_{00}$ (i.e., mean estimated peak BAC of the control arm) = 1.05, SE = 0.06; $\hat{\gamma}_{10}$ (i.e., difference in estimated peak BAC between intervention and control) = -0.09, SE = 0.09; $\hat{\sigma}^2_W$ (i.e., within-school variance of estimated peak

BAC) = 0.81; and $\hat{\sigma}^2_B$ (i.e., between-school variance of estimated peak BAC) = 0.08. The ICC was estimated to be 0.094. The estimates were similar to those reported in Haug et al. (2017).

Using the R scripts in the supplemental material, the LMM-based (linear-mixed-effects-model-based) effect size estimate was $\hat{\delta}_T$ = -0.092, indicating an intervention effect of roughly 0.1 *SD* in reducing estimated peak BAC. Based on the simulation results in this paper for nonnormal data, RBoot-stud and RBoot-basic CIs (i.e., residual bootstrap with studentized and basic CIs) are recommended; with the simulated data, 95% RBoot CIs can be obtained using the `ComputeES()` function provided in the supplemental material, with the following input

```
ComputeES(dat, "y", "treat", "gp_id", type = "model",
          boot = "residual", ci = "all",
          nsim = 1999L)
```

which took a few minutes to run and returned all five types of CIs:

```
         d     d_boot  normal.ll  normal.ul   basic.ll   basic.ul
    -0.092     -0.092     -0.278      0.095     -0.274      0.097
student.ll student.ul percent.ll percent.ul     bca.ll     bca.ul
    -0.281      0.097     -0.281      0.089     -0.263      0.119
```

so the RBoot-basic CI was [−0.274, 0.097], and the RBoot-stud CI was similar but with a slightly wider interval [−0.281, 0.097]. The CIs provided information on the uncertainty of the intervention effect size estimate: it could be as low as a reduction by 0.3 *SD* or a small increase by 0.1 *SD*.

It is also instructive to compare the bootstrap intervals to one that ignored the uncertainty due to clustering, the latter being the one reported in Haug et al. (2017). Using the simulated data but ignoring the clustering, the Cohen's *d* was estimated as 0.086, which was similar to the $\hat{\delta}_T$ estimate that accounted for clustering; however, the 95% CI was found to be [-0.036, 0.207] (width = 0.243) when the clustering was ignored, which was substantially narrower than the ones I obtained using RBoot-basic (width = 0.371) and RBoot-stud (width = 0.378), even when the ICC and cluster sizes of the data was relatively small (ICC = .094; $\bar{n}$ = 13). Therefore, it is important for applied researchers to report valid CIs that accurately account for the data dependency due to clustering. It should also be pointed out that the ANOVA-based and LMM-based procedures for obtaining CIs had not been discussed until Hedges (2007) and Hedges (2009), so it is likely that CIs for multilevel effect sizes reported in earlier literature were too narrow.

Discussion

Despite recommendations from methodologists and professional organizations (e.g., AERA, 2006; APA, 2010; Hedges, 2008), very rarely do researchers in the behavioral sciences report CIs for effect size in single level studies, and little attention has been given to assist substantive researchers in computing CIs for multilevel effect size. The present study compares the performance of the bootstrap methods to the analytic methods in obtaining CIs, and results supported various bootstrap procedures, especially the residual bootstrap, as viable alternatives when analytic methods are not available, which makes CIs for multilevel effect size more accessible to substantive researchers in interpreting research findings.

In addition, although some of the bootstrap techniques in this paper have been implemented in previous software packages and macros,[6] to my knowledge no previous package has implemented all the bootstrap methods discussed in this paper; specifically, the studentized CIs, which were shown to perform better than the other CIs for multilevel effect size, had not been implemented. Therefore, the implementations of the various bootstrap methods and CI construction procedures will be useful for researchers to obtain standard error and interval estimates for not just multilevel effect size, but also other derived quantities such as R^2, intraclass correlation, and indirect effect, as well as to examine the robustness of the point and interval estimates of model parameters in the presence of assumption violations.

Performances of the Analytic and Bootstrap Methods to Construct CI

The results from the current study are mostly in line with the expectations. The ANOVA-based method is considered model-free in the sense that it does not require fitting a multilevel model to the data first, so it is the only procedure among the ones tested in this paper that can be used for secondary data analyses when only summary statistics, but not the raw data, are available. However, its coverage was suboptimal for conditions with unbalanced data (91.6% on average) and when the normality assumption at level 1 was violated (92.0% on average). The LMM-based (linear-mixed-effects-model-based) method, on the other hand, does not require balanced data and gave 95% CIs with reasonable coverage for multilevel effect size.

However, when raw data are available, several bootstrap CIs had better coverage properties than the LMM-based procedure. Specifically, even when the normality assumptions hold, the studentized and basic CIs with parametric and residual bootstraps generally outperformed the LMM-based method. This is likely because the bootstrap methods accommodate the nonnormal sampling distribution of $\hat{\delta}_T$. When the normality assumptions do not hold, residual bootstrap CIs generally had the best coverage properties.

It was also found that methods that assumed normality (i.e., ANOVA-based, LMM-based, and parametric bootstrap) were more affected by nonnormality at level 1 than at level 2. It is possibly due to the higher sample skewness and kurtosis of the simulated data at level 1 than at level 2, even though the level-1 errors and the level-2 random effects were drawn from the same skewed-*t* distribution. As a post hoc analysis, I resimulated the data from the condition with nonnormal $P(\varepsilon)$ and normal $P(u_0)$, $J = 30$, $\bar{n} = 25$, $\delta_T = 0$, ICC = 0.1, and balanced data, but reduced the sample skewness and kurtosis of ε (to 1.33 and 2.69 on average across replications) to match the sample skewness (average = 1.50) and kurtosis (average = 2.72) of u_0 for the counterpart condition with normal $P(\varepsilon)$ and nonnormal $P(u_0)$. If the impact of nonnormality on the coverage of the ANOVA-based, LMM-based, and parametric bootstrap CIs is merely a function of the degree of nonnormality but not the level, I should expect similar coverage percentages for these two conditions with matching skewness and kurtosis. I found that, however, level-1 nonnormality still had a bigger impact than level-2 non-normality, as the coverage rates in the former condition were 94.9%, 94.9%, and 95.8% for ANOVA-based, LMM-based, and PBoot-stud, but were 93.6%, 93.6%, and 94.7% in the latter condition. Therefore, I do not recommend ANOVA-based and LMM-based CIs for multilevel effect size especially when the level-1 error term is not normal, but future studies can further compare the effect of level-1 and level-2 nonnormality in multilevel modeling.

Finally, CIs based on the case bootstrap, while not requiring parametric assumptions, generally had coverage less than nominal level (around 93.7% across conditions) and performed worse than the residual bootstrap. This is consistent with the finding in Huang (2018) that the case bootstrap showed underestimated standard errors of the fixed effects especially when number of clusters was small (e.g., by 6.1% to 8.5% with 20 clusters). Even for conditions with $J = 70$ and $\bar{n} = 25$, the performance of CIs based on the case bootstrap were generally similar to that of the LMM-based CI. Therefore, I recommend the LMM-based CI and the CIs based on the residual bootstrap for multilevel effect size over those based on the case bootstrap.

Comparisons of the Five Bootstrap CIs

Among the five types of bootstrap CIs, the studentized and the basic CIs performed best. Given that the studentized CI was proposed as an improved procedure over simpler bootstrap intervals such as normal CIs and basic CIs, its better performance is not surprising. The basic CI, on the other hand, is the best option for smaller sample sizes such as with $\bar{n} = 5$ and $J = 20$, where point and variance estimates of $\hat{\delta}_T$ may be noisy so that a simpler procedure is better.

On the contrary, the percentile and the BCa CIs generally had the worst coverage for multilevel effect size. Also, the improvement of BCa CI over percentile CI was minimal. It is unclear why the BCa CI, which is second-order accurate theoretically, performed worse than the basic CI and the normal CI. That said, I am not alone in getting mixed results with BCa CIs; for example, Biesanz, Falk, and Savalei (2010) also found that the BCa CI had inconsistent coverage for indirect effects. Future studies can perhaps provide more conclusive evidence on whether and when BCa is a good option in constructing CIs.

Whereas studentized CI overall was shown to be the best method for constructing CIs for multilevel $\hat{\delta}_T$, the method itself has two major limitations. First, it requires that a reasonable sampling variance estimate of the estimator to be available, which is specific to a given multilevel structure. Although formulas for computing the variance of multilevel effect size have been derived for three-level data (Hedges, 2011), cross-classified data (Lai & Kwok, 2014), and partially nested designs (Lai & Kwok, 2016), there are other multilevel structures with more than three levels and with specific features such as the multiple membership structure (Beretvas, 2011), where variance formulas have not been derived. One alternative is to use nested bootstrap (Hinkley & Shi, 1989) so that for each bootstrap sample, an additional layer of bootstrapping with a few resamples is used to obtain the variance estimate, but such a procedure is very computationally intensive in practice, especially when sample size is large. Second, the studentized CI, as well as the basic and the normal CIs, assumes that the parameter space is unbounded, as is the case for δ_T which can take any value in the real line. For bounded quantities, such as R^2 or intraclass correlation, direct use of these procedures may give CIs with impossible upper or lower bound values (e.g., $R^2 < 0$). A solution is to first transform the target quantity to an unbounded space (see Ukoumunne, Davison, Gulliford, & Chinn, 2003, for an example), obtain bootstrap CI for the transformed quantity, and then back transform the confidence limits. On the other hand, percentile and BCa CIs can be directly used for bounded quantities. Future studies are needed to evaluate various bootstrap CI procedures for other quantities of interest and data with alternative multilevel structures.

Recommendations of Procedures for Constructing CIs for Multilevel Effect Size

While researchers regularly report measures of uncertainty, such as *SE* or CI, for sample statistics such as means and regression coefficients, the same reporting practice has not been but should be adopted for effect size reporting. After all, effect size aims to quantify the effect of interest in an interpretable way. A significant barrier for applied researchers to adhere to the reporting guidelines is the complexity associated with CI computations for effect size, especially with multilevel data, and I hope the discussion and illustration in the present study will familiarize researchers with the analytic and bootstrap methods for obtaining CIs for multilevel effect size.

Based on the results of the present study, I have a few suggestions for reporting CIs for multilevel studies (with data similar to the ones in the current simulation):

1. When possible, use the studentized CI with the residual bootstrap for moderate to large sample sizes (i.e., 70 clusters or more, or an average cluster size of 25 or more) and the basic CI with the residual bootstrap for smaller sample sizes;
2. When normality holds approximately, the LMM-based CI, which requires the asymptotic variance estimates of the random effect variances and some computations, is also reasonable when available;
3. If neither the LMM-based CI nor the residual bootstrap CIs are available, obtain the studentized or the basic 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 a skewed-*t* distribution 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. Second, the bootstrap methods studied in the present study, namely the parametric, residual, and case bootstrap, only represent three multilevel bootstrap procedures that are better known to researchers (Van der Leeden et al., 2008). Other bootstrap procedures have been developed for clustered data (e.g., Deen & de Rooij, 2019; Field, Pang, & Welsh, 2010; Owen & Eckles, 2012), and future research is needed to examine their performance for multilevel effect size. Third, I only considered violations of the normality assumption in the current study, but violations of other assumptions should be tested in future research, such as heterogeneous variances of level-1 errors and/or level-2 random effects, to which only the case bootstrap is theoretically robust. Fourth, in this paper I mainly focused on the performance of bootstrap CIs as the point estimates of effect size were approximately unbiased in the present simulation conditions. However, the bootstrap can also be used to correct for biases in point estimates of various statistical quantities of interest, and future studies can explore the use of it in multilevel analyses (see also Vallejo Seco et al., 2013).

In addition, due to the intensive nature of the bootstrap methods, I was only able to obtain

1,000 replications per condition, which was not ideal for precisely comparing the various procedures. In future studies, the bootstrap methods can be evaluated with more replications (and perhaps with fewer simulation conditions) so that coverage rates can be more accurately estimated. Finally, simulations done in the present study only pertain to the basic two-level strictly hierarchical structure with no effect heterogeneity (i.e., no random slopes). Future research should utilize more complex structures to verify whether the bootstrap procedures, especially the residual bootstrap, still perform well in those designs.

References

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*, 33–40. https://doi.org/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. https://doi.org/10.1080/00273171.2013.802978

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Angst, F., Aeschlimann, A., & Angst, J. (2017). The minimal clinically important difference raised the significance of outcome effects above the statistical level, with methodological implications for future studies. *Journal of Clinical Epidemiology*, *82*, 128–136. https://doi.org/10.1016/j.jclinepi.2016.11.016

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology. *American Psychologist*, *73*(1), 3–25. https://doi.org/10.1037/amp0000191

Azzalini, A. (2013). *The skew-normal and related families (with the collaboration of Antonella Capitanio)*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139248891

Azzalini, A. (2019). The R package `sn`: The skew-normal and related distributions such as the skew-*t* (version 1.5-4). [Computer software manual]. Retrieved from `http://azzalini.stat.unipd.it/SN`

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

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.

Biesanz, J. C., Falk, C. F., & Savalei, V. (2010). Assessing mediational models: Testing and interval estimation for indirect effects. *Multivariate Behavioral Research*, *45*(4), 661–701. https://doi.org/10.1080/00273171.2010.498292

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

Bolier, L., Ketelaar, S. M., Nieuwenhuijsen, K., Smeets, O., Gärtner, F. R., & Sluiter, J. K. (2014). Workplace mental health promotion online to enhance well-being of nurses and allied health professionals: A cluster-randomized controlled trial. *Internet Interventions*, *1*(4), 196–204. https://doi.org/10.1016/j.invent.2014.10.002

Bull, S. S., Levine, D. K., Black, S. R., Schmiege, S. J., & Santelli, J. (2012, nov). Social media-delivered sexual health intervention: A cluster randomized controlled trial. *American Journal of Preventive Medicine*, *43*(5), 467–474. https://doi.org/10.1016/j.amepre.2012.07.022

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

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. https://doi.org/10.1002/sim.3620

Canty, A., & Ripley, B. D. (2017). boot: Bootstrap R (S-Plus) functions [Computer software manual]. (R package version 1.3-20)

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. https://doi.org/10.1111/1467-9876.00415

Chen, L.-T., & Peng, C.-Y. J. (2015). The sensitivity of three methods to nonnormality and unequal variances in interval estimation of effect sizes. *Behavior Research Methods*, *47*, 107–126. https://doi.org/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. https://doi.org/10.1037/0003-066X.45.12.1304

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

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

Deen, M., & de Rooij, M. (2018). Clusterbootstrap: Analyze clustered data with generalized linear models using the cluster bootstrap [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=ClusterBootstrap` (R package version 1.0.0)

Deen, M., & de Rooij, M. (2019). ClusterBootstrap: An R package for the analysis of hierarchical data using generalized linear models with the cluster bootstrap. *Behavior Research Methods*. https://doi.org/10.3758/s13428-019-01252-y

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

Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, *82*, 171–185. https://doi.org/10.1080/01621459.1987.10478410

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

Feingold, A. (2015). Confidence interval estimation for standardized effect sizes in multilevel and latent growth modeling. *Journal of Consulting and Clinical Psychology*, *83*(1), 157–168. https://doi.org/10.1037/a0037721

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, *40*, 532–538. https://doi.org/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. https://doi.org/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. https://doi.org/10.1186/1472-6963-4-33

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. https://doi.org/10.1037/a0024338

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

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

Harrell, F. E., Jr. (2019). rms: Regression modeling strategies [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rms` (R package version 5.1-3.1)

Haug, S., Paz Castro, R., Kowatsch, T., Filler, A., Dey, M., & Schaub, M. P. (2017, feb). Efficacy of a web- and text messaging-based intervention to reduce problem drinking in adolescents: Results of a cluster-randomized controlled trial. *Journal of Consulting and Clinical Psychology*, *85*(2), 147–159. https://doi.org/10.1037/ccp0000138

Hayes, A. F., & Scharkow, M. (2013). The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis. *Psychological Science*, *24*(10), 1918–1927. https://doi.org/10.1177/0956797613480187

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. https://doi.org/10.3102/10769986006002107

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and*

Behavioral Statistics, *32*, 341–370. https://doi.org/10.3102/1076998606298043

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*, 167–171. https://doi.org/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. https://doi.org/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. https://doi.org/10.3102/0162373707299706

Hinkley, D. V., & Shi, S. (1989). Importance sampling and the nested bootstrap. *Biometrika*, *76*(3), 435–446. https://doi.org/10.1093/biomet/76.3.435

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

Jones, S. M., Brown, J. L., Hoglund, W. L. G., & Aber, J. L. (2010). A school-randomized clinical trial of an integrated social–emotional learning and literacy intervention: Impacts after 1 school year. *Journal of Consulting and Clinical Psychology*, *78*(6), 829–842. https://doi.org/10.1037/a0021383

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. https://doi.org/10.1177/0013164404264850

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

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

Lai, M. H. C. (2019). bootmlm: Bootstrap resampling for multilevel models [Computer software manual]. (R package version 0.0.1) https://doi.org/10.5281/zenodo.1879127

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. https://doi.org/10.3102/1076998614532950

Lai, M. H. C., & Kwok, O.-m. (2016). Estimating standardized effect sizes for two-and three-level partially nested data. *Multivariate Behavioral Research*, *51*(6), 740–756. https://doi.org/10.1080/00273171.2016.1231606

Littell, R. C. (2006). *SAS for mixed models*. Cary, NC: SAS Institute.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86–92. https://doi.org/10.1027/1614-1881.1.3.86

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. https://doi.org/10.1207/s15327906mbr3901

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. https://doi.org/10.1037/1082-989X.9.2.147

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

Norusis, M. J. M. J. (2012). *IBM SPSS statistics 19 advanced statistical procedures companion*. Upper Saddle River, NJ: Prentice Hall.

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

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

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against

Imprecise Power Estimates. *Perspectives on Psychological Science*, *9*(3), 319–332. https://doi.org/10.1177/1745691614528519

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. https://doi.org/10.3758/BF03206553

R Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.1) [Computer software]. Vienna, Austria: R Foundation for Statistical Computing.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2019). *A user's guide to MLwiN, v3.03.* Retrieved from `http://www.bristol.ac.uk/cmm/media/software/mlwin/downloads/manuals/3-03/manual-web.pdf`

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.

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.

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. https://doi.org/10.1371/journal.pmed.1000251

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

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*(4), 605–632. https://doi.org/10.1177/00131640121971392

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

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. https://doi.org/10.1002/pst.1561

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

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

Ukoumunne, O. C., Davison, A. C., Gulliford, M. C., & Chinn, S. (2003). Non-parametric bootstrap confidence intervals for the intraclass correlation coefficient. *Statistics in Medicine*, *22*(24), 3805–3821. https://doi.org/10.1002/sim.1643

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. https://doi.org/10.7334/psicothema2013.58

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.

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, *20*(3), 293–309. https://doi.org/10.1037/met0000025

Wang, J., Carpenter, J. R., & Kepler, M. A. (2006, may). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine*, *82*(2), 130–143. https://doi.org/10.1016/j.cmpb.2006.02.006

White, H. R., Mun, E. Y., & Morgan, T. J. (2008). Do brief personalized feedback interventions work for mandated students or is it just getting caught that works? *Psychology of Addictive Behaviors*, *22*(1), 107–116. https://doi.org/10.1037/0893-164X.22.1.107

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis.

The Annals of Statistics, *14*, 1261–1295. https://doi.org/10.1214/aos/1176350142

Footnotes

[1] Note that, whereas the definition of *SD* causes no confusion for single-level studies, it is ambiguous in multilevel studies, because several different *SD*s can be used. These include the within-cluster *SD* (σ_W), the between-cluster *SD* (σ_B), and the total *SD*. Hedges (2007) viewed the issue in a 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 choose the total *SD* in this study because it uses more information in the data and theoretically can be converted to a variance accounted for effect size (Snijders & Bosker, 1994).

[2] In Hedges (2009), $V(\hat{\sigma}^2_W)$ in the numerator of the second term of equation (8) was dropped with the assumption that it was negligible, but in this manuscript it is kept for more accurate estimation.

[3] Technically speaking, the sampling variance of each individual residual depends on their hat values, which is a measure of the influence of each individual case (or cluster) on the model estimates. Therefore, rescaling all residuals by the same value is not mathematically correct, and one should instead rescale the level-2 and level-1 residuals according to their hat values, as recommended in Davison and Hinkley (1997). In single-level regression, the variance of a residual is $V(\hat{e}_i) = \sigma^2(1 - h_{ii})$, which depends on the leverage (i.e., hat value, h_{ii}) of the $i$the observation. Therefore, a theoretically better residual bootstrap procedure is to transform the residuals differentially as $\hat{e}^*_i = \hat{e}_i/\sqrt{1 - h_{ii}}$ so that $V(\hat{e}^*_i) = \sigma^2$ for all i. The results based on this improved residual bootstrap procedure are not presented in this paper, however, as they were essentially the same as those for the procedure by Carpenter et al. (2003). In the `bootmlm` package, Carpenter et al.'s procedure can be called using the argument `type = 'residual_cgr'`, and the hat-value-reflated procedure can be called using the argument `type = 'residual'`.

[4] In a conventional jackknife estimator, $l_i = T(\mathbf{x}) - T_{(i)}(\mathbf{x})$ 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 the BCa CIs were obtained using the grouped jackknife as described in Van der Leeden et al..

[5] In this paper I adopt the parameterization of skew-*t* by Azzalini (2013, chapter 4), in which a variable $Z = Z_0/\sqrt{V}$ has a skew-*t* distribution with slant parameter α and degrees of freedom ν, if Z_0 follows a skew-normal distribution with slant = α and V follows a χ^2/ν distribution with degrees of freedom = ν. The skew-normal density function is $2\phi(x)\Phi(\alpha x)$, where $\phi(x)$ is the normal density function and $\Phi(x)$ is the cumulative normal density function.

[6] For example, the parametric and the residual bootstrap with percentile and bias-corrected CIs in MLwiN (Rasbash, Steele, Browne, & Goldstein, 2019), the bootstrap procedures in the R packages `rms` (Harrell, 2019) and `ClusterBootstrap` (Deen & de Rooij, 2018), the case bootstrap routines in SPSS and in Stata, and the SAS macros

by Roberts and Fan (2004) and Wang, Carpenter, and Kepler (2006).

Table 1

Performance Summary of the Procedures for Constructing 95% Confidence Intervals.

Procedure	Coverage (%)	Up (%)	Lo (%)	Shape	Relative CI Width[a]
ANOVA-based	92.70 (1.83)	3.44 (0.96)	3.87 (1.16)	5.41	0.98
LMM-based	93.70 (1.10)	2.98 (0.67)	3.31 (0.80)	4.95	—
PBoot-norm	93.88 (1.07)	2.91 (0.66)	3.21 (0.78)	4.68	1.01
PBoot-basic	94.06 (1.08)	2.88 (0.67)	3.06 (0.75)	2.63	1.01
PBoot-stud	94.41 (1.02)	2.66 (0.61)	2.93 (0.74)	4.40	1.03
PBoot-perc	93.60 (1.10)	2.93 (0.67)	3.47 (0.84)	8.17	1.01
PBoot-bca	93.53 (1.20)	3.09 (0.73)	3.37 (0.84)	4.09	1.02
RBoot-norm	94.35 (0.89)	2.79 (0.64)	2.86 (0.62)	1.21	1.02
RBoot-basic	94.53 (0.92)	2.77 (0.65)	2.69 (0.63)	−1.37	1.02
RBoot-stud	94.75 (0.82)	2.62 (0.61)	2.63 (0.56)	0.32	1.04
RBoot-perc	93.87 (0.86)	2.82 (0.63)	3.31 (0.71)	7.72	1.02
RBoot-bca	93.93 (0.97)	2.99 (0.66)	3.08 (0.70)	1.49	1.03
CBoot-norm	93.69 (0.85)	2.99 (0.66)	3.32 (0.74)	4.94	1.01
CBoot-basic	93.84 (0.93)	3.11 (0.70)	3.05 (0.67)	−0.75	1.01
CBoot-stud	93.72 (1.20)	3.07 (0.73)	3.21 (0.78)	2.13	1.06
CBoot-perc	92.94 (1.02)	2.91 (0.88)	4.15 (1.21)	16.64	1.01
CBoot-bca	93.16 (1.16)	3.08 (0.76)	3.76 (1.01)	9.36	1.03

Note. CI = confidence interval. Up = upper-tail error rate. Lo = lower tail error rate. Shape = (Up - Lo) / (Up + Lo). PBoot = parametric bootstrap. RBoot = residual bootstrap. CBoot = case bootstrap. normal = normal approximation CI. stud = studentized CI. perc = percentile CI. bca = bias-corrected and accelerated CI. Numbers represented means (and standard deviations in parentheses) across conditions.

[a] Relative CI width was the ratio of the CI width based on one procedure relative to that based on the LMM-based CI.

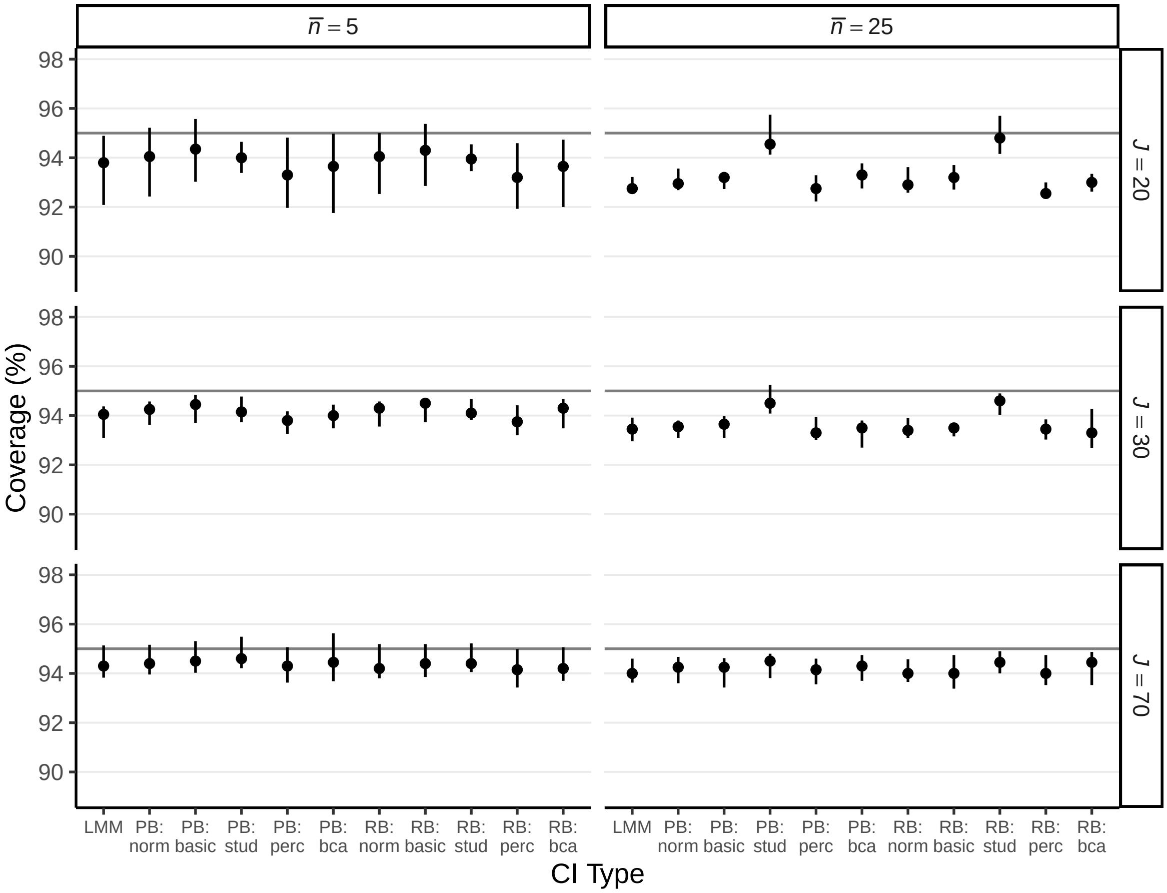

Figure 1. Empirical 95% confidence interval (CI) coverage with normal level-1 errors and normal level-2 random effects across number of clusters (J) and average cluster size ($\bar{n}$). The points (line segments) show the medians (ranges) across conditions. LMM = linear-mixed-model-based method. PB = parametric bootstrap. RB = residual bootstrap. norm = normal approximation CI. stud = studentized CI. perc = percentile CI. bca = bias-corrected and accelerated CI.

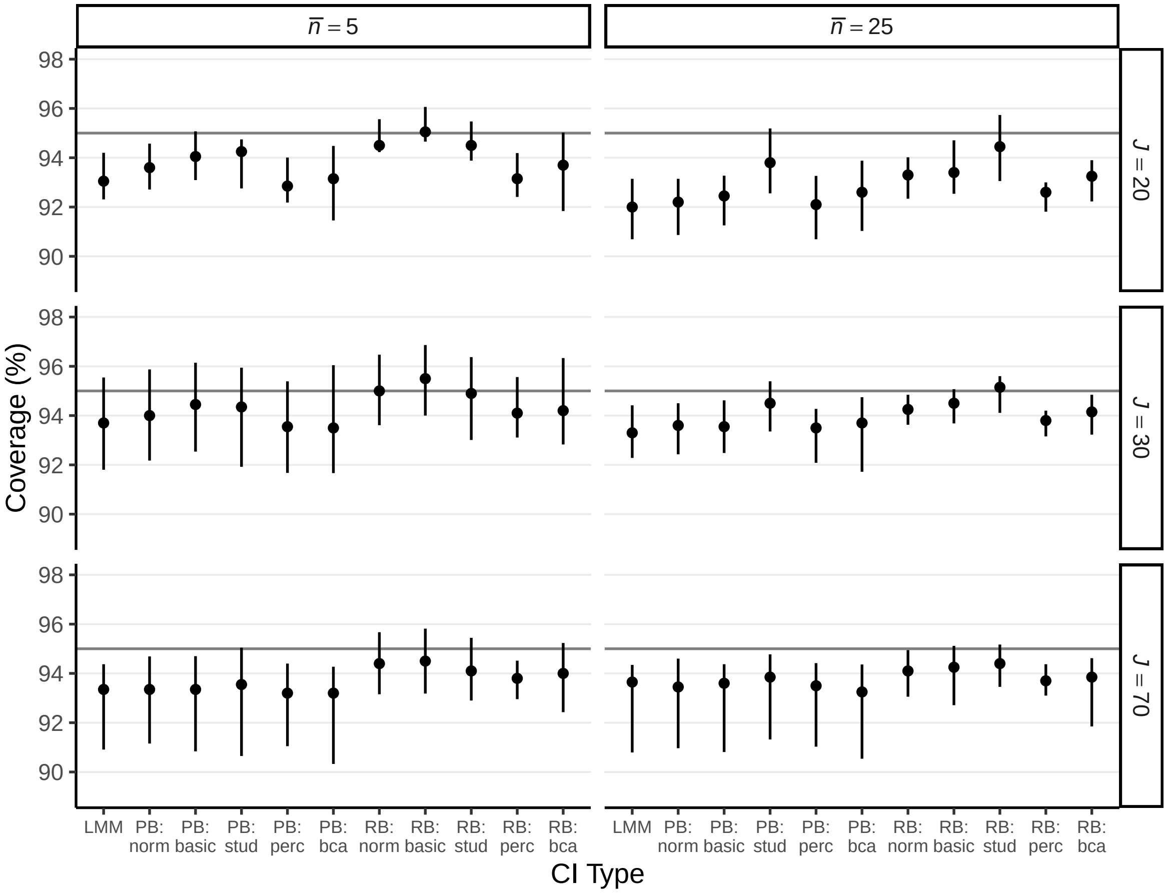

Figure 2. Empirical 95% confidence interval (CI) coverage with nonnormal level-1 errors and normal level-2 random effects across number of clusters (J) and average cluster size ($\bar{n}$). The points (line segments) show the medians (ranges) across conditions. LMM = linear-mixed-model-based method. PB = parametric bootstrap. RB = residual bootstrap. norm = normal approximation CI. stud = studentized CI. perc = percentile CI. bca = bias-corrected and accelerated CI.

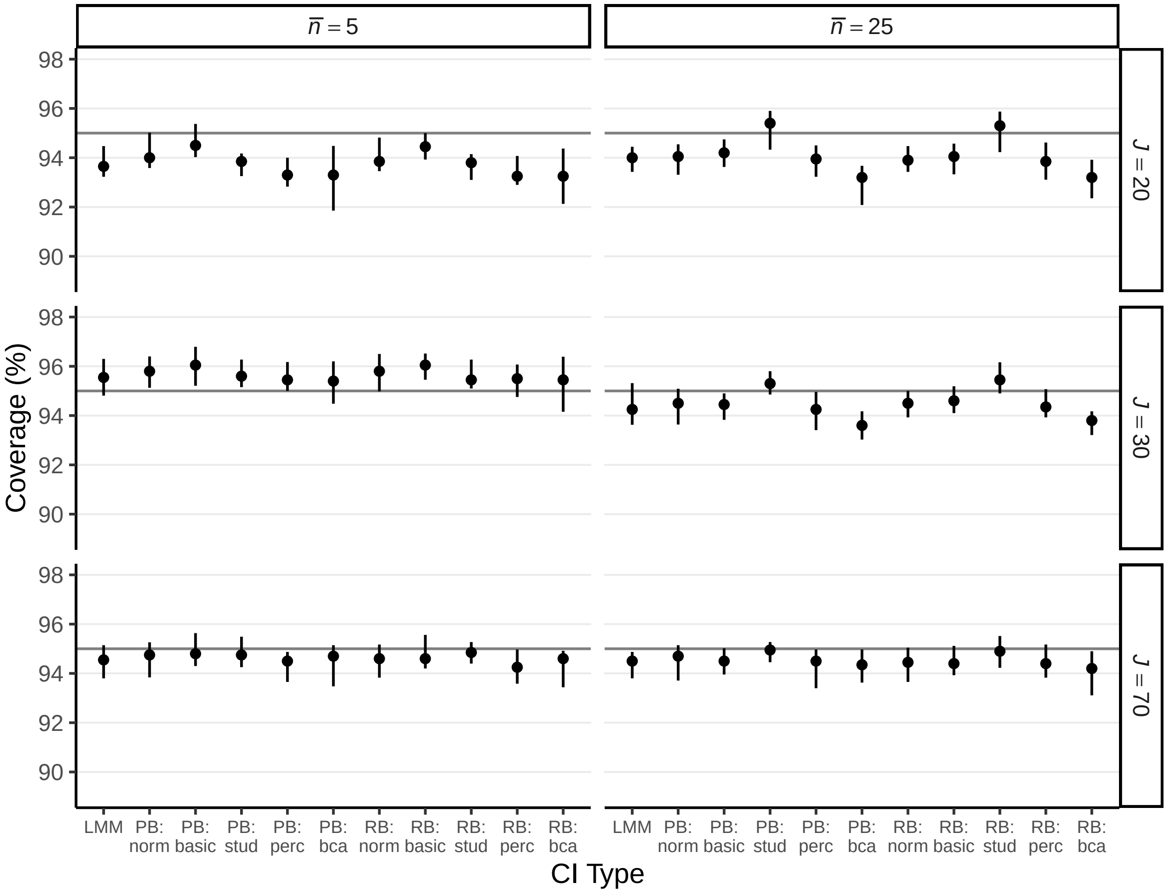

Figure 3. Empirical 95% confidence interval (CI) coverage with normal level-1 errors and nonnormal level-2 random effects across number of clusters (J) and average cluster size ($\bar{n}$). The points (line segments) show the medians (ranges) across conditions. LMM = linear-mixed-model-based method. PB = parametric bootstrap. RB = residual bootstrap. norm = normal approximation CI. stud = studentized CI. perc = percentile CI. bca = bias-corrected and accelerated CI.

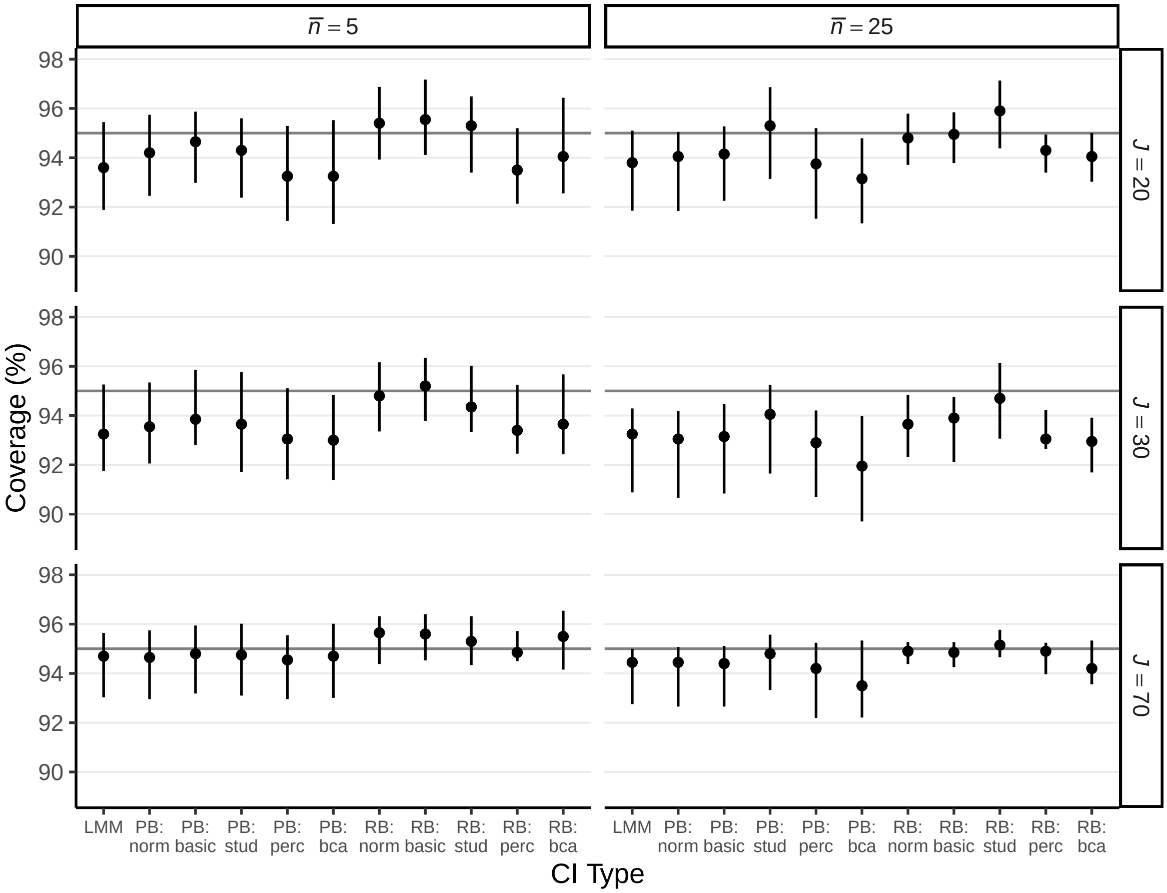

Figure 4. Empirical 95% confidence interval (CI) coverage with nonnormal level-1 errors and nonnormal level-2 random effects across number of clusters (J) and average cluster size ($\bar{n}$). The points (line segments) show the medians (ranges) across conditions. LMM = linear-mixed-model-based method. PB = parametric bootstrap. RB = residual bootstrap. norm = normal approximation CI. stud = studentized CI. perc = percentile CI. bca = bias-corrected and accelerated CI.

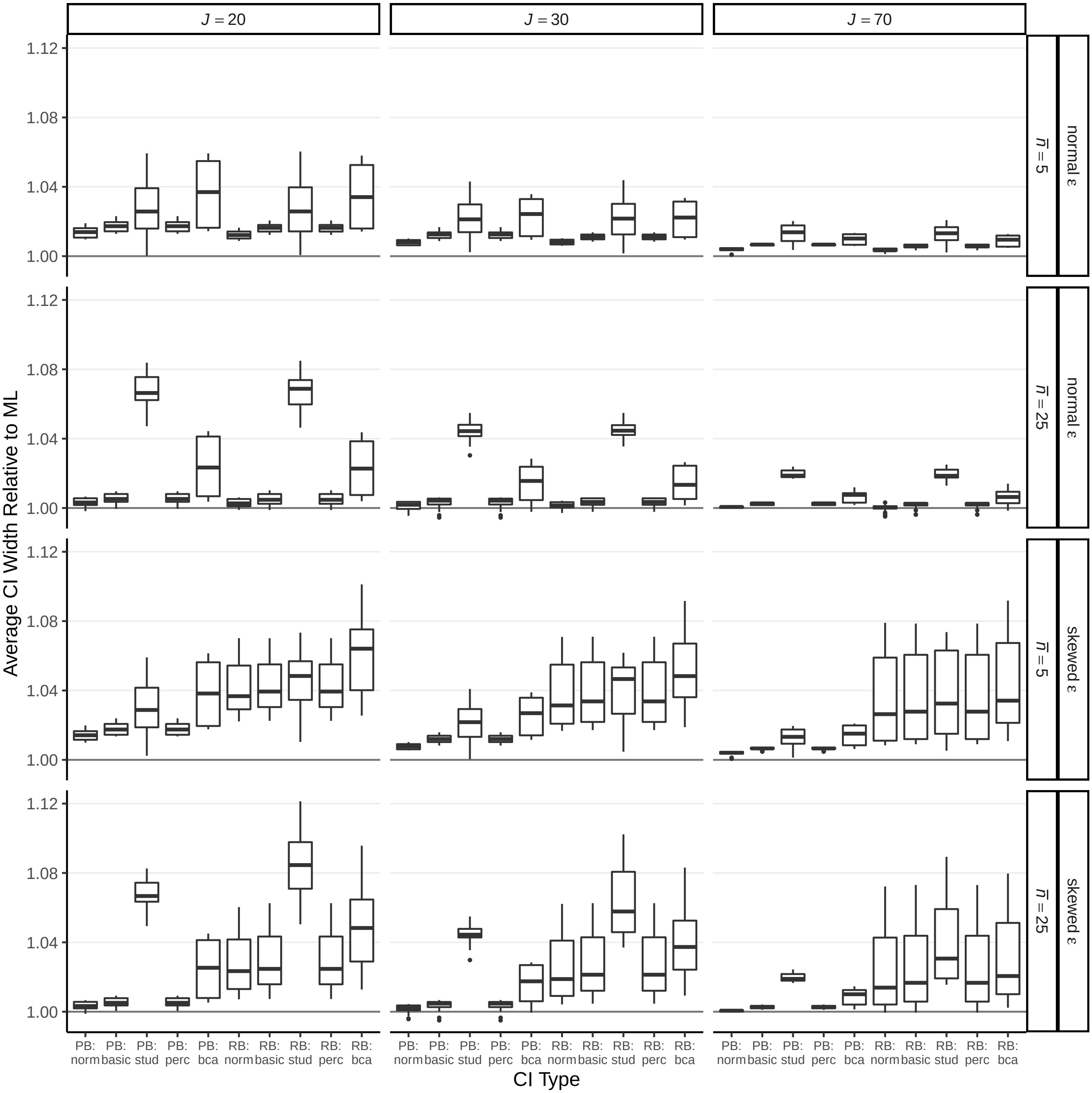

Figure 5. Width of 95% confidence interval (CI) coverage across number of clusters (J), average cluster size ($\bar{n}$), and normality of level-1 error term (ε). PB = parametric bootstrap. RB = residual bootstrap. norm = normal approximation CI. stud = studentized CI. perc = percentile CI. bca = bias-corrected and accelerated CI.