Discussion Papers in Economics

No. 22/2017

Finite Relative Efficiency of Confidence Interval Methods around Effect Sizes

Monika Doll
University of Erlangen-Nürnberg

ISSN 1867-6707

Friedrich-Alexander-Universität Erlangen-Nürnberg
Institute for Economics
https://www.iwf.rw.fau.de/research/iwf-discussion-paper-series/

Finite Relative Efficiency of Confidence Interval Methods around Effect Sizes

Monika Doll[a,]

[a] *University of Erlangen-Nürnberg, Lange Gasse 20, 90403 Nürnberg, Germany*

Tuesday 21st November, 2017

Abstract

Reporting effect sizes and corresponding confidence intervals is increasingly demanded, which generates interest to analyze the performance of confidence intervals around effect sizes. As effect sizes take on the value zero in case of no effect per definition, not only the inclusion of the population effect, but also the exclusion of the value zero are therefore performance criteria for these intervals. This study is the first to compare the performance of confidence interval methods applying these two criteria via determining their finite relative efficiency. Computing the quotient of two methods' minimum required sample sizes to achieve pre-specified levels of both criteria allows to account for the problem of limitations in available observations, which often occurs in the educational, behavioral or social sciences. Results indicate that confidence intervals based on a noncentral t-distribution around the robust effect size proposed by Algina et al. (2005) possess high relative efficiency.

Keywords: Effect Size, Confidence Interval, Minimum Required Sample Size, Finite Relative Efficiency

Email address: monika.doll@fau.de (Monika Doll)

1. Introduction

An increasing number of journals require research's findings to be presented by effect sizes in addition to or even instead of the reporting of null hypothesis tests' p-values (Trafimow and Marks, 2014,Wasserstein and Lazar, 2016). Various limitations of p-values, as being influenced by sample size, encouraging dichotomous decisions on (not) rejecting null hypotheses or being incapable of informing about the magnitude of an existing population effect can be coped with by reporting effect sizes. These statistics are dimensionless quantities, independent of sample size, show the size of an estimated population effect in standard deviations, and take on the value 'zero' in case of no population effect (Cohen, 1969). In the 'Methods for the Behavioral, Educational, and Social Sciences' the measurement of the extent of an effect is of special interest, but accompanying confidence intervals around effect sizes is demanded to reveal the level of uncertainy in the estimation of population effects to add information to these single point estimates. In contrast to the communication of findings in terms of effect sizes, the simultaneous reporting of confidence intervals around these statistics has not shown an equally strongly increasing trend, even though this practice is perceived as being research's future, and being especially important for the MBESS (e.g. Bird, 2002, Thompson, 2002, Kelley, 2007a, Peng et al., 2013). Thus, emphasis should be given to spreading the benefits and importance as well as fostering the understanding of the statistical methods of computing confidence intervals around effect sizes.

This paper focusses on the performance comparison of confidence intervals around standardized mean difference effect sizes for two identically and independently distributed groups. In the field of location difference measures, the most commonly used and known effect sizes are Cohen's d and Hedges's g (Cohen, 1969, Hedges, 1981). When assuming normally distributed populations, confidence intervals around these effect size statistics can be calculated using the noncentrality interval estimation approach based on a noncentral t-distribution (see Steiger and Fouladi, 1997, Cumming, 2001). When the population is not assumed to be normally distributed, the application of robust measures of population effects has been stressed and a robust effect size estimator has been proposed by Algina et al. (2005). In case of nonnormally distributed parent populations recommendations were made to compute confidence intervals around effect sizes via nonparametric bootstrap methods (Micceri, 1989,

Algina et al., 2005). Thus, facing a variety of effect size estimators and confidence interval calculation methods that differ in underlying assumptions, it is essential to examine and contrast the performance of these methods especially in situations when assumptions are violated.

Literature so far compared the performance of the noncentrality interval estimation approach, the percentile bootstapping method and the bootstrap corrected and accelerated method for computing confidence intervals around effect sizes estimated by Cohen's d, Hedges's g, Algina et al.(2005)'s (AKP's) d_R, and variations of trimmed effect sizes at underlying (non-)normal distributed parent populations. When parent populations are normally distributed, the exact method, thus the noncentrality interval estimation approach using Cohen's d as effect size estimator is suggested. When parent populations were nonnormally distributed, the usage of the bootstrap corrected and accelerated method using Hedges's g as effect size estimator (Kelley, 2005) or applying the percentile bootstrap method using a robust effect size estimator d_R (Algina et al., 2005, Algina et al., 2006a , Algina et al., 2006b) was recommended. Approximate confidence interval methods around variations of effect sizes were shown to provide accurate results, as long as sample sizes were moderately large (Viechtbauer, 2007).

However, all of these recommendations for confidence interval calculation methods around effect sizes were made based on the criterion of precisely meeting the nominal level of coverage probability, thus on the probability of including the population effect. For some recommendations, the criterion of an interval possessing small width was regarded as well. Both criteria, coverage probability and interval width are well established in analyzing the performance of confidence intervals around different statistics. Nevertheless, this paper claims that ignoring the criterion 'power', which is the probability that a confidence interval does not include the value 'zero', yields an incomplete performance comparison of confidence intervals around effect sizes, since effect sizes are defined as taking on the value 'zero' if no population effect exists.

Consequently, this paper is the first to compare the performance of confidence interval methods around effect sizes by simultaneously regarding coverage probability and power. This contrasting by simultaneously considering two criteria was accomplished by determining

the finite relative efficiency between two confidence interval methods, which is computing the quotient of these methods' minimum required sample sizes to achieve prespecified levels of both criteria. Inserting the parameter 'minimum required sample size' as central part of this paper's efficiency comparisons allows to deal especially with the behavioral and educational research's often occuring restrictions on available observations as well as meeting exploratory and primary studies' needs. Moreover, performance is contrasted at normally distributed as well as nonnormally distributed parent populations, to evaluate confidence interval methods in situations when assumptions are met as well as when assumptions are violated.

Findings showed that the finite relative efficiency of methods for computing confidence intervals around effect sizes is strongly influenced by the length of the tails of the parent population's distribution. At normal distribution, thus symmetrically distributed parent populations with rather short tails, the best finite relative efficiency was observable for the exact confidence interval construction method based on the noncentral t-distribution using Cohen's d as effect size estimator. At violation of the assumption of normal distributed parent populations the main result is that for some conditions no sample size could be found for any method at which both criteria coverage probability and power are met. Moreover, the confidence interval method based on the noncentral t-distribution showed high finite relative efficiency compared to both bootstrapping methods considered when the robust effect size estimator d_R is used.

The overall structure of this study takes the form of six sections, including this introduction. The second section presents the theoretical framework on considered effect size estimators, confidence interval methods and generated distributions. The third section further discusses and defines the performance criteria. The fourth section describes the simulation procedure, while the fifth one presents the results. The sixth section discusses the findings and concludes.

2. Theoretical Framework

First the considered effect size estimators, then the compared confidence intervals methods and finally the distribution generating method will be presented.

Effect Size Estimators

This paper focusses on standardized mean difference effect sizes

$$\delta = \frac{\mu_1 - \mu_2}{\sigma},$$

with μ_1 being the population mean of the first group, μ_2 being the population mean of the second group and σ being the population standard deviation, which is assumed to be identical for both groups.

Estimators of effect sizes considered in this study are the most commonly known and used estimator Cohen's d, its unbiased version Hedges's g, as well as the robust estimator AKP's d_R.

Cohen (1969) defined an effect size, called Cohen's d as

$$d = \frac{\overline{X_1} - \overline{X_2}}{S},$$

with $\overline{X_1}$ being the first group's mean, $\overline{X_2}$ being the second group's mean and S being the pooled standard deviation

$$\sqrt{S^2} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}},$$

with n_1 being the first group's sample size, n_2 being the second group's sample size, S_1^2 being the first group's variance, S_2^2 being the second group's variance (Cohen, 1969).

Hedges (1981) showed that this effect size is positively biased for small sample sizes and proposed the unbiased effect size estimator Hedges's g:

$$g = d\left(\frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\sqrt{\frac{(n_1+n_2)}{2}}\,\Gamma\left(\frac{n_1+n_2-1}{2}\right)}\right),$$

with $\Gamma(\cdot)$ referring to the gamma function (Hedges, 1981).

The bias correction term can as well be approximated:

$$g' = d\left(1 - \frac{3}{4(n_1 + n_2 - 2) - 1}\right).$$

However, effect size measures can be inaccurate when samples are drawn from nonnormal distributions. Thus, both estimators, Cohen's d and Hedges's g are influenced by outliers

due to the standard deviation's sensitivity to the tails of a distribution. Therefore, Algina et al. (2005) proposed a robust mean difference effect size for trimmed samples' means, standardized by Winzoried variances, which are less sensitive to the tails of a distribution:

$$\delta_R = 0.642 \, \frac{\mu_{1,trim} - \mu_{2,trim}}{\sigma_{win}},$$

with $\mu_{1,trim}$, and $\mu_{2,trim}$ being the population trimmed mean of group one, respectively group two, and σ_{win} being the population Winsorized standard deviation. As by a trimming percentage of 20% the estimated trimmed mean has a relatively small standard error among commonly occuring situations while simultaneously little accuracy is lost when sampling from a normal distribution, this percentage was proposed for general use by Wilcox (2005) and accordingly implemented for the effect size estimator d_R by Algina et al. (2005):

$$d_R = 0.642 \, \frac{\overline{X}_{1,trim} - \overline{X}_{2,trim}}{S_{win}},$$

with $\overline{X}_{1,trim}$ being the 20% trimmed mean of group one, $\overline{X}_{2,trim}$ being the 20% trimmed mean of group two, thus the mean calculated after dropping 20% of the data on each side of the distribution. S_{win} is calculated as the square root of

$$S^2_{win} = \frac{(n_1 - 1)S^2_{win,1} + (n_2 - 1)S^2_{win,2}}{(n_1 + n_2 - 2)},$$

where $S^2_{win,1}$, and $S^2_{win,2}$ denote the 20% Winsorized variances in group one, respectively group two, thus the variance computed after replacing the 20% lowest (highest) values by the one that remains as lowest (highest) value after 20% trimming (Dixon, 1960). This way, δ_R measures the mean difference between the middle 60 percent of both groups in units of the standard deviations of the groups' middle 60 percent. By using the multiplier $0.642 = \sqrt{0.4121}$ it is ensured that δ_R equals δ when samples are drawn from normally distributed parent populations.

Confidence Interval Methods

This study considered three different methods to calculate confidence intervals around the previously mentioned effect sizes for performance analysis. The first one is an exact method based on the assumption of normally distributed parent populations and homogeneity of

variance (NCT), the second one is a bootstrapping method based on the assumption of randomly drawn samples that are representatives for a larger population (Perc), as is the third method (BCa), while the latter corrects for skewness as well as bias in the bootstrap distribution. Moreover, only two-sided confidence intervals were considered, as one-sided intervals would not add any indication of level of uncertainty in the effect size estimation.

For normally distributed parent populations and independently drawn samples, the exact method to build a confidence interval around a standardized mean difference effect size is the 'noncentrality interval estimation approach', which is based on the noncentral t-distribution with noncentrality parameter

$$\lambda = \frac{\mu_2 - \mu_1}{\sigma}\sqrt{\frac{n_1 n_2}{n_1 + n_2}}.$$

The lower and upper confidence interval limits for this method, here called 'NCT', are calculated by using the confidence interval transformation principle, the inversion confidence interval principle and the connection between the effect size δ and the noncentrality parameter λ as $\delta = \lambda\sqrt{\frac{n_1+n_2}{n_1 n_2}}$ (Steiger and Fouladi, 1997, Cumming, 2001). Thus, first it is to identify the lower, respectively upper confidence interval limit for the noncentrality parameter λ by calculating the samples' two-group t-statistic :

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\left(\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)}}$$

, then to find those values of the noncentrality parameter λ that generate the noncentral t-distributions with $(n_1 + n_2 - 2)$ degrees of freedom, in which the observed value of the samples' t-statistic has cumulative probability of $1 - \alpha/2$, respectively $\alpha/2$.

Second, the confidence interval limits for the effect size δ have to be obtained by multiplying the confidence interval limits of λ by $\sqrt{\frac{n_1+n_2}{n_1 n_2}}$.

For constructing a confidence interval around the effect size δ_R, the NCT method has to be modified by replacing the samples' t-statistic by the 'trimmed t-statistic':

$$t_R = \frac{\overline{X}_{1,trim} - \overline{X}_{2,trim}}{\sqrt{\left(\frac{1}{h_1} + \frac{1}{h_2}\right)\left(\frac{(n_1+n_2-2)S_{win}^2}{h_1+h_2-2}\right)}}$$

(Yuen and Dixon, 1973). This trimmed t-statistic's distribution can satisfactorily be approximated up from a sample size of 7 by Student's t distribution with $(h_1 + h_2 - 2)$ degrees

of freedom, where $h_i = n_i - 2g_i$ with g_i being the number of values that were replaced by Winsorizing and $i = 1, 2$ referring to the sample of group 1 or group 2 respectively. At last, the confidence interval limits for the noncentrality parameter have to be found and multplied by $0.642\sqrt{\frac{(h_1 h_2)(n_1+n_2-2)}{(h_1 h_2)(h_1+h_2-2)}}$ (Algina et al., 2005).

No assumptions on the underlying parent distributions are necessary for using nonparametric bootstrap methods to compute confidence intervals. These approaches only assume data to be a random and representative sample from some larger population (Efron, 1979).

The percentile bootstrap method for calculating confidence intervals (Perc) around δ proceeds in three steps. First, from both groups' original samples random samples of a prespecified size n are drawn with replacement, thus bootstrapped, B times. Second, the effect size estimator (for Cohen's d: d^*) is caluclated for each of the B sub-samples of both groups. Third, the confidence interval's lower and upper limits are found as being the respective $\frac{\alpha}{2}$ or 1-$\frac{\alpha}{2}$ -quantiles of the ranked, B times calculated, estimators d^*. This procedure can as well be applied using the bootstrapped effect size estimator Hedges's g: g^*.

To use the Perc method for calculating confidence intervals around δ_R, the first step has to be slightly modified such that after a sample size of n is randomly selected with replacement from the first, respectively second group's sample, the 20% trimmed mean and the 20% Winsorized variances are calculated for both samples accordingly and d_R^* is computed using these subsamples' trimmed means and Winsorized variances (Algina et al., 2005). The bootstrap bias-corrected and accelerated method (BCa) accounts for the Perc method's substantial coverage error if the distribution of the estimated parameter is not nearly symmetric by allowing for asymmetry and a change in skewness of the distribution function of the estimated parameter, thus effect size. Moreover, the BCa method assumes the existence of a monotonic increasing function g and constants z_0 and a, such that $\hat{\phi} = g(\hat{\delta})$ and $\phi = g(\delta)$ satisfy $\hat{\phi} = \phi + \sigma_\phi(Z - z_0), \quad \text{with} \quad Z \sim N(0,1)$ and $\sigma_\phi = 1 + a\phi$ (Efron and Tibshirani, 1993). Here, a denotes the skewness of the score transformation of $\hat{\delta}$, called the acceleration constant, whereas z_0 is the bias-correction value. The BCa method automatically selects a transformation g that transforms the regarded effect size to normality, computes an exact confidence interval, and then transforms backwards to the effect size's original scale. Therefore, as for the Perc method, first, B random subsamples of

size n each are drawn from both groups' original samples with replacement. Then, the effect size estimator d^* (respectively g^*) is estimated for each of the B subsamples. Next, the bias-correction value $\hat{z_0}$ is obtained by calculating the proportion of the B bootstrapped effect sizes d^* that take on a smaller value than that of the original samples' effect size estimator d (respectively g), then finding the quantile of the normal distribution with the respective cumulative probability:

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#(d^* < d)}{B}\right).$$

The acceleration value

$$\hat{a} = \frac{\sum_{i=1}^{n_1+n_2}(\tilde{d} - d_{(-i)})^3}{6\left(\left(\sum_{i=1}^{n_1+n_2}(\tilde{d} - d_{(-i)})^2\right)^{3/2}\right)}$$

is obtained by first performing a jack-knife procedure on the original sample, whereby $d_{(-i)}$ is the value of the effect size estimator d calculated after the i^{th} data point out of the combined sample $(n_1 + n_2)$ has been deleted and $\tilde{d}$ is the mean of the $n_1 + n_2$ times jack-knifed $d_{(-i)}$ values. Once the bias correction value z_0 and the acceleration parameter a have been calculated, the limits of the confidence interval are determined by finding those values from the bootstrap sample that correspond to the lower and upper confidence interval quantiles of the ranked observed bootstrap distribution at

$$\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(\alpha/2)}{1 - \hat{a}(\hat{z}_0 + \Phi^{-1}(\alpha/2))}\right),$$

respectively

$$\Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + \Phi^{-1}(1-\alpha/2)}{1 - \hat{a}(\hat{z}_0 + \Phi^{-1}(1-\alpha/2))}\right),$$

with $\Phi(\cdot)$ referring to the standard normal distribution.

The advantage of this BCa method is its second order accuracy, thus the rate of $\frac{1}{min(n_1,n_2)}$ with which its coverage error approaches zero, being faster then the Perc method's rate of $\frac{1}{\sqrt{min(n_1,n_2)}}$ (Efron and Tibshirani, 1993).

The BCa method is applied for constructing confidence intervals around δ_R the same way as is done for the Perc method.

The g-and-h Distributions

This paper's analysis focusses on confidence interval methods' performances around standardized mean differences of two independent and identically distributed samples. These samples were drawn from normally distributed parent populations as well as from nonnormally distributed parent populations, so that efficiency could be evaluated in case of meeting as well as violation of the normality assumption.

In line with Algina et al. (2005), Algina et al. (2006b), and Keselman et al. (2005), four distributions were considered that cover main types of nonnormality. Thus, parent populations were (1) normally, thus symmetrically distributed with short tails, (2) symmetrically distributed with long tails, (3) asymmetrically distributed with short tails and (4) exponentially distributed. All distributions were generated from the family of g-and-h distributions, where the parameter h controls the elongation of the tails and the parameter g controls the amount and direction of the skewness. Standard normally distributed random variables (Z_i) are transformed via $Y_i = Z_i \exp\left(\frac{hZ_i^2}{2}\right)$, when the parameter g is chosen to equal zero, respectively via $Y_i = \frac{\exp(gZ_i-1)}{g} \exp\left(\frac{hZ_i^2}{2}\right)$, when the parameter g is chosen to be non-zero (Hoaglin et al., 1985).

Thus, to generate the considered distributions, the parameters were set the following way:

(1) Symmetric distribution with short tails (normal)	g=0	h=0
(2) Symmetric distribution with long tails (ltsym)	g=0	h=0.225
(3) Asymmetric distribution with long tails (ltskew)	g=0.225	h=0.225
(4) Exponential distribution (exp)	g=0.76	h=-0.098

The normally distributed random variable Z_i was drawn n_1 times and transformed into Y_{i1} for forming the first group's sample. As this analysis focuses on independently and identically distributed samples that differ by a fixed population effect while having equal variance and equal sample size, the sample for the second group was generated by $Y_{i2} = Y_{i1} + \delta\sigma_{dist}$, where σ_{dist} is the standard deviation of the regarded parent population, which is defined by the chosen parameter values.

At $g = 0$: $\sigma_{dist} = (1-2h)^{-\frac{3}{2}}$, whereas at $g \neq 0$:

$$\sigma_{dist} = \frac{1}{g^2\sqrt{1-2h}}\left[\exp\left(\frac{2g^2}{1-2h}\right) - 2\exp\left(\frac{g^2}{2(1-2h)}\right) + 1\right] - \frac{1}{g^2(1-h)}\left[\exp\left(\frac{g^2}{2(1-h)}\right) - 1\right]^2.$$

δ is the fixed size of the standardized mean difference population effect, that was set to equal 0.2, 0.5 or 0.8 (see Cohen (1969)'s suggestions for interpreting an effect size as being a small, medium, or large effect). Note that thus the distributions were designed to differ by 0.2, 0.5, or 0.8 times the standard deviation, which is called the population effect.

Thus, overall 108 conditions (3 effect sizes, 3 confidence interval methods, 4 distributions, 3 sizes of population effects) were analyzed.

3. Performance Criteria and Finite Relative Efficiency

Performances of the methods to compute confidence intervals around an effect size are compared regarding their minimum required sample size to simultaneously achieve a coverage probability of at least 95% and a power of at least 80%, displayed as finite relative efficiency (FRE).

Coverage probability is the percentage of confidence intervals whose limits correctly bracketed the population value, hence the percentage of coverage. The nominal coverage probability level was set to be 95%. Thus, as a benchmark level, in all but at most 5% of all computed confidence intervals per condition the size of the population effect has to be included to achieve this criterion's benchmark. In this paper, the population effect size was fixed to equal 0.2, 0.5, or 0.8 times the respective population's standard deviation. Note that this inclusion of the true population effect was not only chosen as coverage probability criterion for confidence intervals around δ, but also around δ_R. This is due to the fact that at the baseline situation with normally distributed parent populations δ_R is defined to equal δ by being multiplied with the factor 0.642. This approach of inducing the same criteria on coverage probability on all confidence intervals even at violation of the normality assumption attempts to decrease the uncertainty in applying, interpreting, and contrasting the robust effect size δ_R in comparison to the commonly known standardized mean difference effect size δ.

Power is defined as the percentage of confidence intervals whose limits correctly did not bracket the value 'zero'. The nominal power level was set to be 80%. Thus, as a benchmark level, in at least 80% of all computed confidence intervals per condition, 'zero' has to be excluded to achieve this criterion's benchmark. All of this study's considered conditions were designed to obtain a non-zero population effect. By exluding the value 'zero', in case of an existing underlying population effect, power is in this manner the chance to find a real effect if there is one (Maxwell et al., 2008, Cumming, 2001). Although this paper is the first to explicitely take power into account when comparing the performances of confidence interval methods that are built around effect sizes, Kelley (2005) and Algina et al. (2005) displayed empirical power of different confidence interval methods at fixed sample sizes and fixed population effect sizes. At different underlying nonnormal parent populations, they found that the NCT method showed slightly higher power than the BCa method when built around δ estimated by Cohen's d (Kelley, 2005). Moreover, they foun that the NCT method also shows higher power than the Perc method when built around δ_R, estimated by d_R (Algina et al., 2005). No conclusions on the confidence intervals' performances were drawn based on these findings.

In this paper, two confidence interval methods' performances were compared regarding the quotient of each one's minimum sample size m_0 required to simultaneously meet both benchmarks (achieving a coverage probability of at least 95% and power of at least 80%). This quotient is called finite relative efficiency (FRE) following Büning and Trenkler (1978). Thus, $FRE = \frac{m_{0;a}}{m_{0;b}}$, where $m_{0;a}$ and $m_{0;b}$ are those minimum sample sizes at which method a, respectively method b meet the prespecified performance benchmarks, with $a \neq b$. Accordingly, if $FRE_{a,b} = 1$, both confidence interval methods a and b require the same amount of observations to meet both benchmarks and have the same relative efficiency. If $FRE_{a,b} < 1$, method a requires a lower minimum sample size and is therefore more efficient than method b. If $FRE_{a,b} > 1$, method a requires more observations and is therefore less efficient than method b.

4. Simulation Procedure

The minimum sample size (m_0) at which the coverage probability was not smaller than 95%, while the power was not smaller than 80% was determined via simulation study using an exponential search algorithm (Knuth, 1998, Bentley and Yao, 1976).

Thus, for each of the 108 conditions this iterative procedure consisted of two parts. The first part determined a range in which the searched for minimum sample size resides in four steps:

(1) A small sample size m_i was specified, (2) 1000 confidence intervals were computed, and (3) coverage probability and power were calculated. (4a) If performance benchmarks were not met at sample size m_i steps (1) to (3) were repeated, while sample size $m_{(i+1)}$ was calculated by increasing m_i by 2^x with $x = x + 1$ with each repetition of steps (1) to (3), $i \in \mathbb{N}$. (4b) If performance benchmarks were met at m_I, the interval $[m_{I-1}, m_I]$ was identified as including the minimum sample size m_0, with I being the number of repetitions of steps (1) to (3). At this point, the second part of the simulation started.

Now, a binary search was conducted, thus, the meeting of the performance benchmarks was tested at sample size $((m_I - m_{I-1})/2 + m_{I-1})$ and the exponent x was reduced by 1. If benchmarks were (not) met, the performance was tested at the middle of the lower (upper) half of the interval $[m_{I-1}, m_I]$, while $((m_I - m_{I-1})/2 + m_{I-1})$ was set to be the upper (lower) limit of the interval that had to be tested. This second part of the procedure continuted until x equalled 'one', whereas then the minimum sample size m_0 was found being the smallest value at which both criteria were met. Minimum required sample sizes presented in chapter 'Results' are averages of 1000 datasets generated by the simulation procedure per condition. The number of bootstrap repetitions was set at $B = 1000$.[1]

[1] All simulations were conducted with the statistical software R with usage of the packages 'boot' (Angelo and Ripley, 2016, Davison and Hinkley, 1997), 'MBESS' (Kelley, 2007b), 'bootES' (Kirby and Gerlanc, 2013), and 'psych' (Revelle, 2016).

5. Results

The main finding was that not for all conditions a sample size was determinable at which both performance criteria were met simultaneously (m_0). Therefore, results are discribed in two parts. First, the focus will be on findings of those conditions in which an m_0 was determinable. Second, the focus will be on those conditions, in which no m_0 was determinable.

With focus on those conditions for which m_0 was determinable, it is to say that for some of these conditions m_0 was not determinable for all of the 1000 datasets generated by the simulation procedure. These conditions are highlighted by underlined values in the following tables. Tables 1 to 3 present the finite relative efficiency of the regarded methods per condition.[2]

Table 1: Finite Relative Efficiency of NCT to BCa

		0.2	0.5	0.8
normal	Cohen's d	0.9458	0.9277	0.8718
	Hedges's g	0.9526	0.9759	1
	AKP's d_R	0.9607	0.9255	0.8605
ltsym	Cohen's d	0.8882	0.767	0.7083
	Hedges's g	0.8942	0.8155	0.7917
	AKP's d_R	0.9608	<u>0.9074</u>	
ltskew	Cohen's d	0.8348		
	Hedges's g	0.8498	<u>0.7397</u>	
	AKP's d_R	0.9388		
exp	Cohen's d	0.9311	0.8451	0.7536
	Hedges's g	0.9369	0.8732	0.8261
	AKP's d_R	0.9597	0.9346	0.8235

The noncentrality interval estimation approach NCT showed a higher relative efficiency

[2] 0.2, 0.5 and 0.8 define the population effect's sizes; normal= normally distributed parent population, ltsym= long tailed symmetrically distributed parent population, ltskew= long tailed asymmetrically distributed parent population; exp= exponentially distributed parent population.

than the bootstrap bias-corrected and accelerated approach BCa, as can be seen in table 1, which displays for all conditions $\text{FRE}_{NCT,BCa} \leq 1$. Thus, over all conditions, NCT requires at most the same amount of observations as BCa to as few as 70.83% of the BCa's minimum required sample size to meet the prespecivied benchmarks.

Table 2: Finite relative efficiency of Perc to BCa

		0.2	0.5	0.8
normal	Cohen's d	0.9661	0.8916	0.7949
	Hedges's g	0.9661	0.8916	0.7949
	AKP's d_R	0.9862	0.9574	0.8837
ltsym	Cohen's d	0.9002	0.7476	1.2708
	Hedges's g	0.9002	0.7476	1.0833
	AKP's d_R	0.9961	0.9815	1.0357
ltskew	Cohen's d	0.8348		
	Hedges's g	0.8348		
	AKP's d_R	1		
exp	Cohen's d	0.9484	0.831	0.7101
	Hedges's g	0.9484	0.831	0.7101
	AKP's d_R	0.9817	0.9533	0.9216

The percentile bootstrap method Perc showed a higher relative efficiency than the bootstrap bias-corrected and accelerated approach BCa, as can be seen in table 2, which displays for almost all conditions $\text{FRE}_{Perc,BCa} \leq 1$. Thus, over almost all conditions, Perc requires at most the same amount of observations as BCa to as few as 71.01% of the BCa's m_0. As exception in the condition with symmetric, long tailed populations' distributions and a large population effect of $\delta = 0.8$, Perc showed to require between 1.0357 to 1.2708 times the sample size than BCa to meet the benchmarks. The overall findings on the relative efficiency of BCa show to be in contrast to results based on performance analysis by coverage probability as the only criterion, which recommended the use of BCa around Hedges's g when parent populations are nonnormal (Kelley, 2005).

Table 3: Finite relative efficiency of NCT to Perc

		0.2	0.5	0.8
normal	Cohen's d	0.979	1.0405	1.0968
	Hedges's g	0.986	1.0946	1.2581
	AKP's d_R	0.9741	0.9667	0.9737
ltsym	Cohen's d	0.9867	1.026	0.5574
	Hedges's g	0.9933	1.0909	0.7308
	AKP's d_R	0.9646	0.9245	
ltskew	Cohen's d	1		
	Hedges's g	1.018		
	AKP's d_R	0.9388		
exp	Cohen's d	0.9818	1.0169	1.0612
	Hedges's g	0.9879	1.0508	1.1633
	AKP's d_R	0.9776	0.9804	0.8936

As can be seen in table 3, $\text{FRE}_{NCT,Perc}$ does not tend to be either greater or smaller than 1, thus none of these two methods clearly requires overall more observations than another to meet both criteria simultaneously. However it is worth highlighting that for the condition with symmetric, long tailed populations' distributions and a large population effect $\delta = 0.8$, Perc showed to require as extremely few as 0.5574 to 0.7308 times the sample size to meet the benchmarks than does NCT, where determinable.

Tables 4 to 6 present one method's finite relative efficiency contrasting its performance at different used effect size estimators.[2]

Table 4: FRE of a confidence interval method using Cohen's d to using Hedges's g as effect size estimator

	0.2			0.5			0.8		
	NCT	BCa	Perc	NCT	BCa	Perc	NCT	BCa	Perc
normal	0.9929	1	1	0.9506	1	1	0.8718	1	1
ltsym	0.9933	1	1	0.9405	1	1	0.8947	1	1.1731
ltskew	0.9823	1	1		0.9863				
exp	0.9939	1	1	0.9677	1	1	0.9123	1	1

Table 4 displays that the NCT method needs fewer observations to meet the criteria when using Cohen's d instead of Hedges's g as effect size estimator, as $\text{FRE}_{NCT(d),NCT(g)} < 1$, regardless of the underlying distribution. Moreover, BCa as well as Perc show that the minimum required sample size to meet the benchmarks is similar at almost all conditions for using Cohen's d and using Hedges's g, as finite relative efficiencies are 1 for almost every condition.

Table 5: FRE of a confidence interval method using Cohen's d to using AKP's d_R as effect size estimator

	0.2			0.5			0.8		
	NCT	BCa	Perc	NCT	BCa	Perc	NCT	BCa	Perc
normal	0.8569	0.8703	0.8526	0.8851	0.883	0.8222	0.9189	0.907	0.8158
ltsym	1.8163	1.9647	1.7756	1.6122	1.9074	1.4528		1.7143	2.1034
ltskew	2.0145	2.2653	1.8912						
exp	1.2385	1.2766	1.2332	1.2	1.3271	1.1569	1.2381	1.3529	1.0426

Table 6: FRE of a confidence interval method using Hedges's g to using AKP's d_R as effect size estimator

	0.2			0.5			0.8		
	NCT	BCa	Perc	NCT	BCa	Perc	NCT	BCa	Perc
normal	0.863	0.8703	0.8526	0.931	0.883	0.8222	1.0541	0.907	0.8158
ltsym	1.8286	1.9647	1.7756	1.7143	1.9074	1.4528		1.7143	1.7931
ltskew	2.0507	2.2653	1.8912						
exp	1.2462	1.2766	1.2332	1.24	1.3271	1.1569	1.3571	1.3529	1.0426

Tables 5 and 6 display that the finite relative efficiency of one method using different effect size estimators varies widely depending on the population's distribution. At normally distributed parent populations, all three analyzed confidence interval constructing methods show higher relative efficiency when Cohen's d or Hedges's g were used as effect size estimators instead of AKP's d_R, as at almost all conditions $FRE < 1$. However, at nonnormally distributed parent populations all methods show to need a higher sample size to meet the benchmarks when Cohen's d or Hedges's g were used as effect size estimators instead of AKP's d_R. Here, extreme values of finite relative efficiencies are observable, such that for

some conditions about twice the sample size was needed when using Cohen's d or Hedges's g.

As highlighted by underlined values in tables 1 to 6, at six conditions the search algorithm did not find a sample size at which both criteria are met (m_0) for all datasets generated by the simulation procedure.

In case of symmetric distributed parent populations with long tails, when the size of the effect was 0.5 for only 99.5% of the generated datasets for the NCT method using AKP's d_R as effect size estimator and for 99.9% of generated datasets for the BCa method using AKP's d_R, an m_0 could be found. When the size of the effect was 0.8 for the latter at 12.7% of the generated datasets an m_0 was found by the search algorithm. At asymmetrically distributed parent populations with long tails, when the effect size was $\delta = 0.5$ for the NCT method using Hedges's g as effect size estimator, only for 23.1% of all generated datasets an m_0 was determinable. In this condition, for the BCa method an m_0 was only determinable in 1.4% of all generated datasets. By using Cohen's d as effect size estimator only in 0.1% of all generated datasets an m_0 was determinable for this method in this condition. For no method at all a sample size at which the benchmarks were met was found when the parent population's distribution was skewed with long tails with effect size $\delta = 0.8$. For the percentile bootstrap method, no m_0 was determinable when the size of the effect was 0.5 regardless of the effect size estimator used.

In these cases, for certain conditions the computed confidence intervals achieved a coverage probability of 0.95 at a very low sample size, at which however the intervals were too wide to exclude the value zero with high probability, while at other conditions the computed confidence intervals did not achieve a coverage probability of 0.95 at all. These findings on confidence intervals methods' performances on coverage probability are in line with those by Algina et al. (2005) and Algina et al. (2006b), which are demonstrating coverage probability's rapidly increasing divergence from nominal levels as the population effect size increases.

6. Discussion

This paper is the first to compare the performance of methods to compute confidence intervals around effect sizes based on finite relative efficiency as quotient of two methods'

minimum sample sizes required to include the population effect with coverage probability of at least 95% and simultaneously to exclude the value 'zero' with power of at least 80%. As the usage of effect sizes and corresponding confidence intervals is highly recommended and increasingly demanded for reporting research's findings, this study seeks to add to the understanding of these methods, the reduction of uncertainty with their application and the further dissemination of these methods.

The main finding was that for certain conditions no sample size could be found at which coverage probability was at least 95% and power was at least 80% for any of the considered methods. Hence, results showed that confidence interval methods' performances are strongly influenced by the parent populations' distributions. The higher a population's distribution's variance, the larger m_0 as a high standard deviation inflates a confidence interval's width, whereas the broader a confidence interval, the more observations are needed to increase estimation precision and thus reduce the interval's width such that the value 'zero' is not included. These findings are even more intensivied the larger the size of the population effect as the mean difference found for both samples has to be even larger at a large population variance to detect a large effect size.

Moreover, this finding highlights the importance of not only analyzing coverage probability, but also power when examining the performance of confidence intervals around effect sizes. Overall, confidence interval methods showed to require more observations to meet the benchmark of 80% of them not including the value 'zero', as they do for reaching a coverage probability of 95%. Exceptions are those cases in which at no sample size a coverage probability of 95% could be found. This finding is in contrast to Algina et al. (2005)'s argumentation on the performance of confidence interval methods stating that 'it will often require larger sample sizes to achieve adequate accuracy than it does to achieve adequate power' (Algina et al., 2005). In addition, this result suggests that Kelley (2005)'s consideration on confidence intervals decreasing power in case its coverage probability is increasing to meet the nominal level can be refused.

Another finding was that the BCa method showed the worst performance compared to NCT and Perc over almost all considered conditions. This observation is due to the fact that confidence intervals constructed using the BCa method are relatively broader than those

built using the methods NCT or Perc, inducing a worse power at smaller sample sizes. This weakness regarding the performance of confidence intervals around effect sizes built using the BCa method was already detected by Kelley (2005), who nevertheless recommended it.

Mixed results were observable for the relative efficiency of NCT compared to Perc over all conditions (see table 3). Considering that confidence intervals built using AKP's d_R as effect size estimator provide high finite relative efficiency (see tables 5 and 6), the results of this paper show that the NCT method performs better than the Perc method when AKP's d_R is used as effect size estimator over all conditions (see table 3). This observation is in contrast to findings from Algina et al. (2006b) that recommend to build confidence intervals via the Perc method when their effect size estimator d_R is used. This recommendation however was not based on the power criterion but only on the coverage probability criterion. This paper argues that the high relative efficiency of confidence intervals computed with the NCT method using d_R as effect size estimator justifies the recommendation of its usage even at situations in which the assumption of normality is violated. As this recommendation is based on a comparison of minimum required sample size, it could be especially useful for the behavioral, educational or social sciences, which regularly face restrictions on available observations.

Bibliography

Algina, H., Keselman, H. J., and Penfield, R. D. (2005). An alternative to cohen's standardized mean difference effect size: A robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, 10(3):317–328.

Algina, H. J., Keselman, H. J., and Penfield, R. D. (2006a). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods*, 5(2):2–13.

Algina, J., Keselman, H. J., and Penfield, R. D. (2006b). Confidence interval coverage for cohen's effect size statistic. *Educational and Psychological Measurement*, 66(6):945–960.

Angelo, C. and Ripley, B. (2016). boot: Bootstrap r (s-plus) functions. *R package version 1.3-18*.

Bentley, J. L. and Yao, A.-C. (1976). An almost optimal algorithm for unbounded searching. *Information Processing Letters*, 5(3):82–87.

Bird, K. (2002). Confidence intervals for effect sizes in analysis of variance. *Educational and Psychological Measurement*, 62(2):197–226.

Büning, H. and Trenkler, G. (1978). *Nichtparametrische Statistische Methoden*. Walter de Gruyter Verlag, Berlin.

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Psychology Press, New York, USA.

Cumming, F. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement*, 61:385–391.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap methods and their applications*. Cambridge University Press, Cambridge.

Dixon, W. J. (1960). Simplified estimation from censored normal samples. *Annals of Mathematical Statistics*, 31(2):385–391.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *Annals of Statistics*, 7:1–26.

Efron, B. and Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapmen & Hall, New York.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6:107–128.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., editors (1985). *Exploring data tables, trends, and shapes*. John Wiley & Sons, Inc., Hoboken, USA.

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(1):51–69.

Kelley, K. (2007a). Confidence intervals for standardized effect sizes: theory, application, and implementation. *Journal of Statistical Software*, 20(8).

Kelley, K. (2007b). Methods for the behavioral, educational, and social sciences: An r package. *Behavior research methods*, 39:979–984.

Keselman, H. J., Algina, J., and Fradette, K. (2005). Robust confidence intervals for effect size in the two-group case. *Journal of Modern Applied Statistical Methods*, 4(2):353–371.

Kirby, K. N. and Gerlanc, D. (2013). Bootes: An r package for bootstrap confidence intervals on effect sizes. *Behavior research methods*, 45(4):905–927.

Knuth, D. E. (1998). *The art of computer programming: 3: Sorting and Searching*. Addison-Wesley Professional, Mass.

Maxwell, S. E., Kelley, K., and Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual review of psychology*, 59:537–563.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105:156–166.

Peng, C.-Y., J., Chen, L.-T., Chiang, H.-M., and Chiang, Y.-C. (2013). The impact of apa and aera guidelines on effect size reporting. *Educational Psychology Review*, 25(2):157–209.

Revelle, W. (2016). *psych: Procedures for personality and psychological research: Version 1.6.12*. Northwestern University, Evanston.

Steiger, J. H. and Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In Harlow, L., Mulaik, S., and Steiger, J. H., editors, *What if there were no significance tests?* Erlbaum, Hillsdale, USA.

Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher*, 31:25–32.

Trafimow, D. and Marks, M. (2014). Editorial. *Basic and Applied Social Psychology*, 37(1):1–2.

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *The Annals of Statistics*, 32(1):39–60.

Wasserstein, R. and Lazar, N. (2016). The asas statement on p-values: context, process, and purpose. *The American Statistician*, 70(2).

Wilcox, R. R. (2005). *Introduction to robust estimation and hypothesis testing*. Elsevier Acad. Press, Amsterdam.

Yuen, K. K. and Dixon, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed t. *Biometrika*, 60:369–374.