

# A review of effect sizes and their confidence intervals, Part I: The Cohen's $d$ family


Jean-Christophe Goulet-Pelletier <sup>a</sup>,  and Denis Cousineau <sup>a</sup>

<sup>a</sup>University of Ottawa

**Abstract** ■ Effect sizes and confidence intervals are important statistics to assess the magnitude and the precision of an effect. The various standardized effect sizes can be grouped in three categories depending on the experimental design: measures of the difference between two means (the  $d$  family), measures of strength of association (e.g.,  $r$ ,  $R^2$ ,  $\eta^2$ ,  $\omega^2$ ), and risk estimates (e.g., odds ratio, relative risk, phi; Kirk, 1996). Part I of this study reviews the  $d$  family, with a special focus on Cohen's  $d$  and Hedges'  $g$  for two-independent groups and two-repeated measures (or paired samples) designs. The present paper answers questions concerning the  $d$  family via Monte Carlo simulations. First, four different denominators are often proposed to standardize the mean difference in a repeated measures design. Which one should be used? Second, the literature proposes several approximations to estimate the standard error. Which one most closely estimates the true standard deviation of the distribution? Lastly, central and noncentral methods have been proposed to construct a confidence interval around  $d$ . Which method leads to more precise coverage, and how to calculate it? Results suggest that the best way to standardize the effect in both designs is by using the pooled standard deviation in conjunction with a correction factor to unbiased  $d$ . Likewise, the best standard error approximation is given by substituting the gamma function from the true formula by its approximation. Lastly, results from the confidence interval simulations show that, under the normality assumption, the noncentral method is always superior, especially with small sample sizes. However, the central method is equivalent to the noncentral method when  $n$  is greater than 20 in each group for a between-group design and when  $n$  is greater than 24 pairs of observations for a repeated measures design. A practical guide to apply the findings of this study can be found after the general discussion.

**Keywords** ■ Effect size, standard error, confidence intervals, Cohen's  $d$ , noncentral  $t$ -distribution.

 [jgoul014@uottawa.ca](mailto:jgoul014@uottawa.ca)

 JCGP: 0000-0002-2016-549X; DC: 0000-0001-5908-0402

 [10.20982/tqmp.14.4.p242](https://doi.org/10.20982/tqmp.14.4.p242)

**Acting Editor** ■  
Roland Pfister (University of Würzburg)

**Reviewers**  
■ Robert Calin-Jageman (Dominican University, River Forest, Illinois)

## Introduction

Researchers in social sciences are usually interested in answering two general questions: *Is there an effect in the population?* and *Is the effect big or small?* To answer the first question, a  $p$  value is often computed. This value represents the probability of obtaining the effect observed in the sample, or a more extreme one, if the population true effect is zero (Clay, 2014). Arguably, a  $p$  value conveys little information regarding the magnitude of an effect and the degree of error associated with this estimate (Clay,

2014; Cumming, Fidler, Kalinowski, & Lai, 2012; Thompson, 2002). Additionally, the hypothesis embedded in the  $p$  value is often a perfectly “nil” null effect; this unrealistic point of reference has made a strong case for the development of Bayesian analyses.

The second question, *Is the effect large or small?*, cannot be answered with  $p$  values because they fail to indicate the magnitude of a difference beyond its significance. Additional information must be provided, collectively called *effect sizes*. Effect sizes estimate the magnitude of an effect and often, but not necessarily, standardize this magni-



tude to facilitate comparison between studies. Those effect sizes are *estimates* because they infer the true population effect from the sample. Consequently, they should be accompanied by a confidence interval to assess their precision. One difficulty remains to select the right effect size measure, which depends on the experimental design and the desired interpretation. Kirk (1996) listed more than 40 effect size statistics and Huberty (2002), in a review, raised that number to 61. Some of those measures are redundant or too specific; some are more biased estimators than others; finally, many are based on approximations, not all very fortunate.

The present inquiry reviews commonly used effect sizes and their confidence intervals. This first part addresses the *d* family of effect size, which represents standardized measures of the difference between two means. More specifically, Cohen's *d* and Hedges' *g* for between and within-group designs are reviewed and tested using Monte Carlo simulations.

The reader is referred to Part II of this study for a review and evaluation of effect sizes and confidence intervals for analysis of variance (two or more groups), correlation (two related variables), and linear regression (one or more predictors; Goulet-Pelletier & Cousineau, *in preparation*).

### The Cohen's *d* family.

This measure of effect size was inaugurated by Cohen (1969). It is used to evaluate and to standardize the difference between two means.<sup>1</sup> It then allows comparing the impact of a treatment across studies that do not necessarily share the same units of measurement. As mentioned by Glass (1976) who popularized this measure, it is similar to a *Z* score. Hence, interpreting a Cohen's *d* is quite intuitive: a large Cohen *d* represents a large difference between two means. Cohen's *d* magnitude is expressed in a number of standard deviations that separate the two groups. Thus, a *d* of 0.5 can be understood as one group being located 0.5 standard deviations away from the other group. Cohen (1969) proposed guidelines to interpret the magnitude of this measure with 0.2 being "small", 0.5 being "medium", and 0.8 being "large". They are perhaps more meaningfully classify as "merely statistical," "subtle" and "obvious," respectively (Fritz, Morris, & Richler, 2012). These guidelines are not meant to be followed rigidly; the context of a study is crucial to interpret the magnitude of any effect (Cohen, 1988; Cumming, 2012; see also Pek & Flora, 2018, for a discussion). A *d* of 0.5 is said to be visible to the naked eyes of a careful observer (Cohen, 1992).

Cohen's *d* can be employed in four different scenarios:

a) in single group designs, where the sample mean is compared to a pre-specified target value; b) in two independent groups designs, where the interest is in the difference between two population means; c) in single group – two repeated measure designs (e.g., pretest-posttest or paired samples), where the interest is in the change in the population mean between the two measurements; finally, d) in designs where one of the groups is a baseline both in terms of mean and variability (Cumming & Finch, 2001).

The general formula is

$$d = \frac{M_1 - M_2}{S} \quad (1)$$

where  $M_1$  and  $M_2$  are the two means to be compared, and  $S$  is a measure of standard deviation.

Three decisions must be made before computing a Cohen's *d*. The first is to choose the right divider  $S$  in Eq. (1), i.e., the value that best estimates the standard deviation of the population in a given design. Depending on the design, a different standard deviation may be required in the denominator (hereafter called the divider). The second is to decide if an unbiased estimate of *d* is desired or not. The third decision, although optional, is to determine the correct standard error of the *d* statistic to establish its variability. At least six standard error estimates can be found in the literature. Finally, if confidence intervals are sought, decide whether a central or a noncentral *t* distribution is used for the coverage factor (often 95%).

Not all of the available options are equally potent. So, in addition to reviewing them, we evaluate them using Monte Carlo simulations and make recommendations if some options are less relevant than others.

### Review of the choices composing a Cohen's *d* effect size

#### Which divider to use?

The divider  $S$  in Eq. (1) depends on the design. Due to the alternative ways to calculate a Cohen's *d*, researchers need to carefully select the right formula and report unambiguously which one they have used (e.g., using a clear notation). Otherwise, misinterpretations are to be expected. For this reason, we added a subscript to *d* to identify which divider has been used to standardize the mean difference. Four designs are possible. (a) For a situation where the mean of a single sample is compared to a target value, the standard deviation of the sample is used and the effect size is noted  $d_1$ . (b) For two independent groups, the population variability ( $\sigma$ ) is indisputably best estimated by the pooled standard deviation,  $S_p$ , and the effect size is noted  $d_p$ . (c) For two repeated measures, there is no agreement on which divider to choose because different conceptual-

<sup>1</sup>Cohen (1969) originally proposed to divide the raw difference by an unusual standard deviation formula. The original Cohen's *d* formula is no longer used nowadays.



**Table 1** ■ Table 1 Cohen’s *d* effect size estimate for single group, two groups, two repeated measures and comparison to a baseline designs and the assumptions underlying the equations.

Notation and equation	Divider	Assumptions
Single group designs $d_1 = \frac{M_1 - \text{target value}}{S_1}$	$S_1 = \sqrt{\frac{1}{n_1 - 1} \sum (X_i - M_1)^2}$	Population is normal.
Two groups designs $d_p = \frac{M_2 - M_1}{S_p}$	$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$	Populations are normal; Homogeneity of variances.
Two repeated measures designs $d_p$ $d_{av} = \frac{M_2 - M_1}{S_{av}}$	(same as above) $S_{av} = \sqrt{\frac{S_1^2 + S_2^2}{2}}$	Populations are normal; Homogeneity of variances.
$d_D = \frac{M_2 - M_1}{S_D}$ $d_{Dc} = d_D / \sqrt{2(1 - r)}$	$S_D = \sqrt{S_1^2 + S_2^2 - 2r S_1 S_2}$	Populations are normal.
Comparison to a baseline designs $\Delta = \frac{M_2 - M_1}{S_{base}}$	$S_{base} = S_1$ or $S_2$ depending on which group is the baseline	Populations are normal.

*Note.* Note.  $M_1$ ,  $S_1$  and  $n_1$  are the mean, standard deviation and sample size, respectively, of Group 1 (or measurement time one).  $M_2$ ,  $S_2$  and  $n_2$  are the same for Group 2 (or measurement time two);  $r$  is the cross-measurement correlation in a repeated measure design. Note that  $d_p$ ,  $d_D$ , and  $d_{Dc}$  are sometimes called  $d_s$ ,  $d_z$ , and  $d_{rm}$  respectively.

ization of the variability can be made, resulting in different estimates. One can select amongst: the pooled standard deviation as in the between-group design ( $d_p$ ), a standard deviation of the differences ( $d_D$ ), a converted standard deviation of the differences ( $d_{Dc}$ ), and an averaged standard deviation ( $d_{av}$ ). Note that  $d_p$ ,  $d_D$ , and  $d_{Dc}$  are sometimes called  $d_s$ ,  $d_z$ , and  $d_{rm}$  respectively (e.g., Cumming, 2012; Lakens, 2013; Morris & DeShon, 2002). Finally, (d) the last case, when one group is compared to a baseline group, is referred to as Glass’s  $\Delta$  for unequal variances (Glass, McGraw, & Smith, 1981). It will be briefly covered in its own sub-section.

Equations for calculating these variations of Cohen’s *d*, with their divider, are given in Table 1.

**Biased or unbiased estimate?**

Unfortunately, and as will be confirmed in later simulations, the basic Cohen’s *d* formula (Eq. 1) is known to over-estimate the effect size in the population, more so when the sample size is small. It is therefore a biased estimate. A correction factor to unbiased *d* has been found by Hedges (1981). This correction factor, herein called *J*, is based on

the number of observations. The unbiased version of *d* is known as  $d_{unbiased}$ ,  $d_{unb}$ , Hedges *g*, sometimes as  $g^*$ , or even Hedges *h* (e.g., Nakagawa & Cuthill, 2007).<sup>2</sup> In this text, we use Hedges *g* as a synonym for  $d_{unbiased}$ . However, and as Cumming (2012) warns us, the term Hedges *g* has also been employed synonymously to Cohen’s *d* (biased). Consequently, be aware of the inconsistencies surrounding the term “Hedges *g*” in the literature. All the characteristics of Cohen’s *d* are also true of Hedges’ *g*, including the interpretation. The only difference is a correction to the result of Eq. (1). The correction applies similarly to all variants of *d*. To avoid confusion, the subscript for Hedges’ *g* is the same as for Cohen’s *d*.

Hedges’ correction factor can be computed with the following Equation (2a)

$$J(\nu) = \frac{\Gamma(\frac{1}{2}\nu)}{\sqrt{\frac{\nu}{2}} \Gamma(\frac{1}{2}(\nu - 1))} \tag{2a}$$

where  $\nu$  is the number of observations minus 2 ( $\nu = n_1 + n_2 - 2$  for a two-group design and  $2(n - 1)$  for a repeated measures design), and  $\Gamma$  is the Gamma function. For integer values of  $x$ , the Gamma function returns the factorial of

<sup>2</sup>Here we have another example of the misnomer law: Hedges found the unbiased estimate as well as its correct standard error and confidence intervals. Yet, it is named *g* after Glass (1976) who popularized this effect size.



**Table 2** ■ A few values of the correction factor  $J$  (both the exact formula, Eq. 2a, and its approximation, Eq. 2b) as a function of  $\nu$ .

$\nu$	$J(\nu)$	
	Exact	Approximation
2.	0.56	0.57
3.	0.724	0.727
4.	0.798	0.800
5.	0.841	0.842
10.	0.9227	0.9231
15.	0.9490	0.9492
20.	0.9619	0.9620
25.	0.9696	0.9697
50.	0.98491	0.98493
100.	0.992478	0.992481
200.	0.996245	0.996245
500.	0.998499	0.998499
1000.	0.999250	0.999250

$x$  minus 1,  $\Gamma(x) = (x - 1)!$ . For  $x$  having decimals, this formula is computationally complex and is not implemented in many programming languages. For that reason, the following approximation has been proposed (Hedges, 1981)

$$J(\nu) \approx \left(1 - \frac{3}{4\nu - 1}\right). \tag{2b}$$

Table 2 compares the real formula with its approximation, for a few degrees of freedom.

As seen in Table 2, the approximation is not different from the true value up to 2 decimals places for tiny groups (a total of 5 participants if two groups are measured, so that  $\nu = 3$ ) and to 5 decimal places for samples totalizing 100 participants or more. Regarding the correction factor itself, the correction is important for small sample sizes, but the multiplication tends rapidly towards 1 as sample size (and, consequently, degrees of freedom) increases. For example, for two groups with a total of 12 participants (hence,  $\nu = 10$ ), the correction factor decreases  $d$  by about 8% ( $J(10) \approx 0.92 = 92\%$ ). The correction is negligible when  $\nu$  exceeds 100.

Once the correction factor is obtained, Hedges'  $g$  is obtained with

$$g = J(\nu) \times d \tag{3}$$

where  $J(\nu)$  is the correction factor of Eq. (2a) or (2b),  $d$  is the biased Cohen's  $d$  of Eq. (1), and Hedges'  $g$  is the unbiased Cohen's  $d$ .

As a rule of thumb, unbiased estimates should always be used and consequently, Hedges'  $g$  should always be preferred over Cohen's  $d$ . However, for medium sample size and above ( $n$  per group larger than 20), the difference between the two estimators is negligible, as will be seen in the simulations next.

### Which standard error to compute?

The distribution of Cohen's  $d$  (or Hedges'  $g$ ) results from the subtraction of two distinct population means. Hence, it follows the noncentral  $t$  distribution instead of the (central)  $t$  distribution for a single mean (or the normal distribution when samples are very large). Therefore, the correct standard error needs to be derived from the non-central  $t$  distribution. Unfortunately, this distribution is computationally demanding as it does not exist in closed form. This explains why many standard error approximations have been proposed. Reviewing the literature in the search for the most commonly used approximation unfolded a high level of confusion regarding which equation should be employed, not to mention various errors in the re-transcription of symbols (e.g., the harmonic mean of the sample sizes is frequently replaced incorrectly by the arithmetic mean). One mission of the present text is to reduce this uncertainty regarding the calculation of a standard error for Cohen's  $d$ . Hereafter, we review 7 expressions for which we could locate their origin. They are listed in Table 3 and described below.

**True standard error.** Hedges (1981) was the first to report the true formula for the variance of the Cohen's  $d$  for two independent means, and consequently, after taking the square root, the standard error of the Cohen's  $d$ . It is given by

$$SE_{\text{between group}} = \sqrt{\frac{\nu}{\nu - 2} \frac{2}{\tilde{n}} \left(1 + \delta^2 \frac{\tilde{n}}{2}\right) - \frac{\delta^2}{(J(\nu))^2}} \tag{4}$$

where  $J(\nu)$  was given in Eq. (2a),  $\tilde{n}$  is the harmonic mean of  $n_1$  and  $n_2$ , the size of the two samples,  $\nu$  is the number of measurements minus 2, and  $\delta$  is the true standard-



**Table 3** ■ Table 3: Formulas for the variance of the Cohen's *d*.

Name	Formula	Reference	Note
True formula	$\frac{\nu}{\nu-2} \times \frac{2}{\bar{n}} \left(1 + \delta^2 \frac{\bar{n}}{2}\right) - \frac{\delta^2}{J(\nu)^2}$	Hedges, 1981, eq. 6b, p. 111.	
True *	(same as above)	Morris (2000)	Uses Eq. (2b) instead of (2a)
Hedges Approximation	$\frac{2}{\bar{n}} \left(1 + \frac{\delta^2 \frac{\bar{n}}{2}}{2\nu}\right)$	Hedges, 1981, p. 117.	For $N > 50$
Hedges & Olkin approx. <sup>a</sup>	$\frac{2}{\bar{n}} \left(1 + \frac{\delta^2 \frac{\bar{n}}{2}}{2N}\right)$	Hedges & Olkin, 1985, Eq. 15, p. 86.	For $N > 50$
MLE approximation	$\frac{\nu+2}{\nu} \times \frac{2}{\bar{n}} \left(1 + \frac{\delta^2 \frac{\bar{n}}{2}}{2\nu}\right)$	Hedges & Olkin, 1985, Eq. 11, p. 82	For $N > 50$
Large N approximation <sup>b</sup>	$\frac{2}{\bar{n}} \left(1 + \frac{\delta^2}{8}\right)$	Hedges, 1981, Corollary 1, p. 112	For $N > 50$ and balanced group sizes
Correction for small N <sup>c</sup>	$\frac{\nu+1}{\nu-1} \times \frac{2}{\bar{n}} \left(1 + \frac{\delta^2}{8}\right)$	Hunter & Schmidt, 1990	For balanced group sizes

*Note.*  $\delta$  is the population standardized effect size or an estimate of it (preferably Hedges' *g*);

$$J(\nu) = \frac{\Gamma(\frac{1}{2}\nu)}{\sqrt{\frac{\pi}{2}}\Gamma(\frac{1}{2}(\nu-1))} \approx \left(1 - \frac{3}{4\nu-1}\right)$$

$\nu$  is the degree of freedom ( $\nu = n_1 + n_2 - 2$ );  $\bar{n}$  is the harmonic mean of  $n_1$  and  $n_2$ ;  $N$  is the total sample size in two-group designs ( $n_1 + n_2$ ).

Some of the formulas are more commonly seen as: <sup>a</sup> :  $\frac{n_1+n_2}{n_1n_2} + \frac{\delta^2}{2(n_1+n_2)}$

<sup>b</sup> :  $(4/N) \left(1 + \frac{\delta^2}{8}\right)$ ; note that this formulation and the next one is accurate only when  $n_1 = n_2$ , in which case  $\bar{n} = N/2$ .

<sup>c</sup> :  $(N - 1) / (N - 3) (4/N) \left(1 + \frac{\delta^2}{8}\right)$

ized effect size in the population. This formula requires the Gamma function in order to calculate  $J(\nu)$ , a function that is not available in all programming languages, as said earlier. In practice, the true effect size  $\delta$  is unknown. Hence, the parameter  $\delta$  in Eq. (4) is replaced by an effect size estimate, preferably the unbiased effect size *g*.

**True\*.** This approximation substitutes  $J(\nu)$  from the true formula by its approximation given by Eq. (2b). This allows avoiding the  $\Gamma$  function while keeping the same formula. This standard error approximation is the most recent in the literature, first proposed by Morris (2000).

**Hedges approximation.** This approximation was proposed in Hedges (1981) but is accurate for large  $N$  only (where  $N$  is the total sample size).<sup>3</sup>

**Hedges and Olkin approximation.** For large  $N$ ,  $\nu$  is similar to  $N$  so that one can be replaced by the other, as was proposed in Hedges and Olkin (1985).

**MLE approximation.** Hedges and Olkin (1985) noted that the maximum likelihood estimate (MLE) of the pooled standard deviation suggests to divide the sum of squares by  $N$  (i.e.,  $\nu+2$ ), the total number of observations, rather than by  $N - 2$  (i.e.,  $\nu$ ). Consequently, this approximation restores

the division by  $N$ . It is not guaranteed that it returns a better estimate as MLE are often biased.

**Large N approximation.** The Hedges approximation above can be further simplified if we assume that  $\nu \approx N = 2 \times \bar{n}$  and that  $\bar{n} \approx \bar{n}$  (which is roughly exact when  $n_1$  and  $n_2$  are about the same size, i.e., there is no major imbalance between the groups). Hence  $\frac{\bar{n}}{2} / (2\nu) \approx \frac{\bar{n}}{2} / (2 \times 2 \bar{n}) = 1/8$ . The formula had a typo in the original article which was corrected in Hunter and Schmidt (1990).

**Correction for small N to the large N approximation.** Stacking one approximation on top of another, Hunter and Schmidt (1990) suggested a further correction  $(\nu + 1) / (\nu - 1)$  to the previous formula to accommodate  $N$  smaller than 50.

To find which approximation is the best estimate of the standard error for a between-group design, we explored three scenarios: (a) small sample sizes ( $3 \leq n_1 \leq 20$  and  $3 \leq n_2 \leq 20$ ); (b) medium sample sizes ( $20 \leq n_1 \leq 100$  and  $20 \leq n_2 \leq 100$ , by steps of 10); and (c) imbalanced sample sizes ( $3 \leq n_1 \leq 20$  and  $100 \leq n_2 \leq 1000$  by steps of 100). In all scenarios, we tested all true effect sizes  $\delta$  from 0.1 to 1.0 by steps of 0.1. The average estimated

<sup>3</sup>This formula encapsulates elegantly all the components of a standard error: the base variance of a standardized normal variate is 1. This baseline is increased as the noncentrality parameter increases. Also, the average sample size  $\bar{n}$  reduces the variance as usual by  $1/\sqrt{\bar{n}}$ . Finally, because there are two groups, the variance is multiplied by 2 when it is assumed that both groups have homogeneous variances.





**Table 4** ■ Comparisons of the seven existing formulas for the variance of the Cohen's *d*.

Formula	Mean variance	bias in percent
Scenario 1: Small <i>ns</i>		
True	0.2877	–
True*	0.2878	.0463%
Correction for small <i>N</i>	0.2810	–2.350%
MLE Approximation	0.2773	–3.733%
Large <i>N</i> Approximation	0.2442	–17.75%
Hedges Approximation	0.2439	–17.93%
Hedges & Olkin Approximation	0.2426	–18.57%
Scenario 2: Medium <i>ns</i>		
True	0.04575	–
True*	0.04575	.00583%
MLE Approximation	0.04568	–0.1400%
Correction for small <i>N</i>	0.04592	0.3832%
Large <i>N</i> Approximation	0.04492	–1.835%
Hedges Approximation	0.04470	–2.342%
Hedges & Olkin Approximation	0.04466	–2.434%
Scenario 3: Imbalanced <i>ns</i>		
True	0.1207	–
True*	0.1207	0.0002757%
MLE Approximation	0.1207	–0.009815%
Hedges Approximation	0.1200	–0.5888%
Hedges & Olkin Approximation	0.1200	–0.5931%
Large <i>N</i> Approximation	0.1252	3.597%
Correction for small <i>N</i>	0.1259	4.1537%

*Note.* Note: A positive bias means that the variance is overestimated; resulting in longer confidence intervals. Negative biases imply underestimated variances and shorter confidence intervals, which must be avoided.

variance (across all  $n_1$ , all  $n_2$  and all  $\delta$ ) returned by all the equations above are listed in Table 4, as well as the relative deviation to the true value (Eq. 4), sorted from the smallest deviation to the worst.

As seen, the correction for small *N* does approximate well the true variance only when the sample sizes are indeed small. For medium *ns* and imbalanced *ns*, this method is not so good or the worst. The MLE approximation does well in all but the small *ns* scenario. The other three approximations are just plain bad for small *ns*, where underestimation is over 15%. Thus, none of these techniques can be used in all circumstances. Using the approximation True\* did, however, wonderfully well in all circumstances, with an error of estimation relative to the true variance less than 0.5 ‰ (per mil) in the least favorable scenario (small *ns*). Identical results were found for repeated measures design. Hence, if the  $\Gamma$  function is not implemented, or if computational speed is an issue, we strongly recommend using the approximation for  $J(\nu)$

within the true formula, i.e. the True\* SE.

**Which method to estimate a confidence interval?**

There are, at least, three different methods to build confidence intervals for Cohen's *d* effect sizes. The three methods involve: (1) a noncentral *t* distribution, (2) a central *t* distribution, or (3) bootstrap. Even though it is applicable, the bootstrap method will not be further considered in this text. An additional method by Steiger and Fouladi (1997, also see Steiger, 2004) is discussed in Appendix C.

Often based on the sample characteristics (e. g., sample size and standard deviation), a CI is characterized by  $\gamma$ , the confidence level, which is often 95%. A good confidence interval is such that 95% (a proportion  $\gamma$ ) of the intervals, if many were collected, do contain the true effect size (see Steiger & Fouladi, 1997; and Cumming & Finch, 2001, for more).

**CIs obtained from a noncentral *t* distribution.** As shown by Hedges (1981), this method returns exact CIs for



Hedges *g*. It requires the noncentral *t* distribution having two parameters  $\nu$  and  $\lambda$  (the degree of freedom and the noncentrality parameter respectively).<sup>4</sup> The noncentral *t* distribution is asymmetrical because a number divided by an estimated standard deviation will either be magnified or attenuated, but these two outcomes are not equivalent as magnification results in a wider range of estimates. Consequently, the noncentral *t* distribution is skewed with a longer right tail.<sup>5</sup>

The noncentral method to construct a CI first requires to estimate a noncentrality parameter ( $\lambda$ ) which is based on the observed effect size. This parameter is then used to derive a noncentral *t* distribution centered at  $\lambda$  with  $\nu$  degrees of freedom. From this distribution, two values are taken at each end of the distribution, in order to constitute a confidence interval around  $\lambda$  so that 95% (or a proportion  $\gamma$ ) of the distribution is covered by the interval. The interval around  $\lambda$  is then transformed back into an interval for the effect size.

The noncentrality parameter  $\lambda$  for two independent groups design is obtained with

$$\lambda_{\text{between groups}} = d\sqrt{\frac{\bar{n}}{2}} \quad (5a)$$

where *d* is the effect size (estimated using, e. g., Cohen's *d* or Hedges' *g*) and  $\bar{n}$  is the harmonic mean of both  $n_1$  and  $n_2$ .<sup>6</sup> For repeated measures design, the parameter  $\lambda$  is obtained with

$$\lambda_{\text{repeated measure}} = d\sqrt{\frac{n}{2(1-r)}} \quad (5b)$$

where *n* is the number of pairs of observations and *r* is the correlation between the pairs (Algina & Keselman, 2003; Morris, 2000). Incidentally, the correct  $\lambda$  value is also the *t* value returned if the correct *t* test is performed. Then, the noncentral *t* distribution, derived with the parameters  $\nu$  and  $\lambda$ , provides the lower and upper bounds of the confidence interval around  $\lambda$ , at quantiles  $1/2 - \gamma/2$  and  $1/2 + \gamma/2$  (e.g., .025 and .975 to form a 95% CI):

$$CI_{\lambda} = [t_L = t_{\nu, \lambda(0.025)}, t_U = t_{\nu, \lambda(0.975)}] \quad (6)$$

<sup>4</sup>Noncentral distributions exist for all commonly used (central) distribution (e.g., *t*, *F*,  $\chi^2$ ). They all have an extra noncentrality parameter (ncp; sometimes symbolized as  $\lambda$ ,  $\Delta$ , or  $\delta$ ), that indicates how much non-central the distribution is; when its value is zero, noncentral and central distributions are identical. Although the true population ncp is often unknown, it can be estimated from the observed effect size and the sample size. When non zero, the ncp shifts the distribution towards the estimated effect, creating more and more asymmetry as the effect increases in magnitude. Central distributions (e.g., central *t*, central *F*) are appropriate for null hypothesis testing, where the results of a single study are "tested given the premise that the population results are known" (encapsulated in the null hypothesis), and compared to this particular distribution (Fidler & Thompson, 2001). However, the construction of a confidence interval for a non-null effect size is better approached by noncentral distributions.

<sup>5</sup>The noncentral *t* distribution equations are not available in closed form so that no simple formulas can compute its characteristics directly. Instead, iterative search or piecewise approximations are employed to find the relevant quantiles needed to obtain the interval (Smithson, 2001; Kennedy & Gentle, 1980). Up until recently, the computational power required to compute noncentral distribution functions was insufficient and consequently, noncentral distributions were not implemented on commonly used software. For this reason, the noncentral method was somewhat impractical, which explains why it was left behind despite its exact interval estimations (Steiger, 2004). Fortunately, this has changed and current computers can solve these functions quickly and accurately.

<sup>6</sup>The quantity  $2/\bar{n}$  is often written with a different notation as  $(n_1 + n_2)/(n_1 \times n_2)$ .

The CI for  $\lambda$  is then transformed back into a CI for the effect size with

$$CI_d = [d_L = t_L/\frac{\lambda}{d}, d_U = t_U/\frac{\lambda}{d}] \quad (7)$$

where the square brackets denote the extremity of the interval,  $t_L$  is the lower limit of the  $\lambda$  interval,  $t_U$  the upper limit, and  $\lambda$  is the result of equation (5a) or (5b) depending on the design.

**CIs obtained from a central *t* distribution.** The second method builds CIs from a generic equation (Cumming & Finch, 2001; Harding, Tremblay, & Cousineau, 2014):

$$CI = \text{Observed statistic} \pm SE \text{ of that statistic} \times \text{coverage factor}$$

where the observed statistic can be a simple descriptive statistic or an effect size estimate, SE represents the standard error of the observed statistic, and the coverage factor (obtained via the central *t* distribution at quantiles  $1/2 - \gamma/2$  and  $1/2 + \gamma/2$ ) increases the SE width so that the interval can be assigned a desired level of confidence  $\gamma$  (often 95%). This generic method is appropriate to build CIs around the mean, median, median deviation, the interquartile range, and many more, and is very accurate when *n* is above 20 and the assumptions are met (Harding et al., 2014).

Constructing a CI for *d* is easily done via this generic method. However, this method implicitly assumes that the distribution is symmetrical. This assumption is incorrect for a non-null effect size and consequently, a confidence interval around *d* built upon this distribution yields incorrect coverage. Furthermore, the correct distribution of the effect changes as a function of the magnitude of the true effect and the sample size, with more asymmetry for larger effects and smaller sample sizes (Steiger & Fouladi, 1997). Thus, a confidence interval constructed from a central distribution should be more and more incorrect as the true effect increases in magnitude. For more in-depth explanations, see e.g., Cumming and Finch (2001), Fleishman (1980), Smithson (2003), and Steiger and Fouladi (1997).



The central confidence interval around  $d$  is obtained by multiplying the SE with a  $t$  score (from a central  $t$  distribution) at quantile .025 and .975 (assuming a  $\gamma$  of 95%). The interval is obtained by adding the two resulting values (one positive and one negative) to the effect size estimate,

$$CI_{\gamma} = d \pm SE_d \times t_{\nu}, \quad (8)$$

where  $d$  is the observed effect size estimate (e.g., Cohen's  $d$  or Hedges'  $g$ ),  $SE_d$  is the standard error of  $d$  (evaluated in Table 4) and  $t_{\nu}$  is found using the  $t$  distribution with  $\nu$  degree of freedom,  $\nu = n_1 + n_2 - 2$  for independent groups and  $\nu = n_{pairs} - 1$  for a repeated measure design.

Note that the relevant assumptions to be respected with both methods are that (a) the data are sampled independently of one another (simple randomized sampling); (b) the data are sampled from normally distributed populations; and (c) the homogeneity of variance is respected for all groups under study (Kelley, 2005).

### Comparing the methods via simulations

As the introduction highlighted, choices must always be made when computing effect sizes and their confidence intervals. Crossing these choices (divider to use; biased or unbiased; what method to construct a CI, and if the central  $t$  distribution is used, which SE to select) results in many different estimators. Therefore, we will compare them hereafter.

The next sections examine the choices as a function of the design. Monte Carlo simulations are performed to compare a) the Cohen's  $d$  and Hedges'  $g$  in estimating the true effect size, b) the central and the noncentral method in estimating the correct CI, and c) the four different Cohen's  $d$  in repeated measures design to determine which one should be preferred.

Note that constructing a CI about an effect size implies that the effect size is correctly estimating the real value in the population without a systematic bias. When the estimate is biased, as it is with Cohen's  $d$ , the confidence interval will also be biased, irrespective of the method used. Comparing Hedges'  $g$  with Cohen's  $d$ , using a simulated population, provides an appreciation of the amount of bias  $d$  can possibly have.

As said earlier, the notation utilized here to differentiate the various Cohen's  $d$  is based on which standard deviation equation is used to divide the difference. We report only 95% confidence intervals but the results reported next generalize to other confidence levels.

The general methodology is given in Appendix A. We begin with the two independent groups design.

### Two independent groups design.

In a two-group design, one group of participants is compared to another group of participants. The purpose is to assess the difference between the two groups with respect to their mean performances. Possible measures in this case are the raw difference in means, the Cohen's  $d$  and the Hedges'  $g$ .

**The correct divider.** Assuming that both groups are from populations with the same variance, the best estimate of the population standard deviation  $\sigma$  is a weighted average of the standard deviations of both samples. Therefore,  $S_p$ , the pooled standard deviation is the best estimator. The formula for  $S_p$  is given in Table 1. Consequently, Cohen  $d_p$  for two independent groups is given by  $d_p = (M_2 - M_1)/S_p$ .

**Biased or unbiased?** Cohen's  $d$  being a biased estimator, a comparison of  $d_p$  and  $g_p$  will help determine in which situation  $g_p$  should be preferred over  $d_p$ . Hedges  $g_p$  is given by  $g_p = d_p \times J(\nu)$  with  $\nu = n_1 + n_2 - 2$ .

**Which SE if the central  $t$  distribution is used?** Confidence intervals for the  $d$  family should be estimated via the noncentral method for exact coverage rate. However, in practice, the difference between the noncentral and the central methods might be too small to change the interpretations, especially when the sample size is large. Four standard error formulas, given in Table 3, will serve to compute four central CIs for  $d_p$ : (1) the True\* approximation, (2) the MLE approximation, (3) the Correction-for-small-N approximation and (4) the Hedges approximation; the first three were amongst the best estimates in Table 4. Central and noncentral CIs will be computed with  $g_p$  for comparison purposes.

In all the simulations reported, the population means were arbitrarily set to 95 and 105 (a difference of 10) with a standard deviation of 15 so that the true population effect size is  $10/15 = 0.666$ . The intervals width will be compared for sample sizes going from  $n = 4$  (tiny) to  $n = 64$  (moderate to large). As previously noted, the quality of the central method of CI estimation will decrease as the effect size increases (and improve when the effect size decreases). This trend is not investigated in the present simulations, because only a single effect size is specified. We return to this in the discussion.

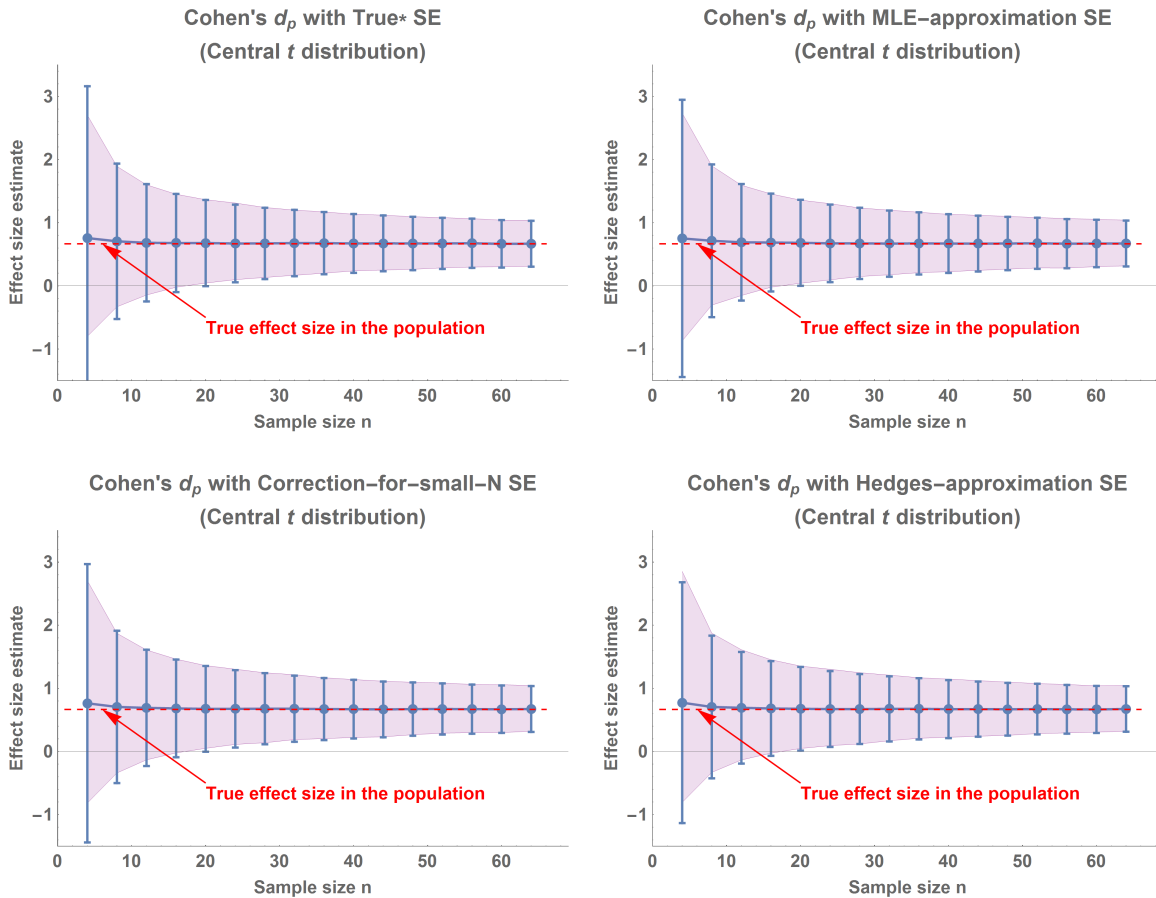
### Results for two independent groups design.

Figure 1 shows the results of the simulations for Cohen's  $d_p$  and central CI based on four standard error approximations. In all four panels, the same estimate  $d_p$  is used and so the mean  $d_p$  and the observed spread of  $d_p$  across simulations (the shaded areas) are identical. What differs among these panels are the CI limits that are estimated using different SE, all with the central method. In Figure 2,





**Figure 1** ■ Results of the mean estimated  $d_p$  and its confidence intervals as a function of sample size (4 to 64) in a between-group design. The four panels represent four different methods to compute CIs. The red dashed line is the true  $d$  in the population. Samples size  $n$  refers to the number of observations within each group.



Hedges'  $g_p$  is shown along with two types of CI, the central method with the Correction for small N SE and the noncentral method.

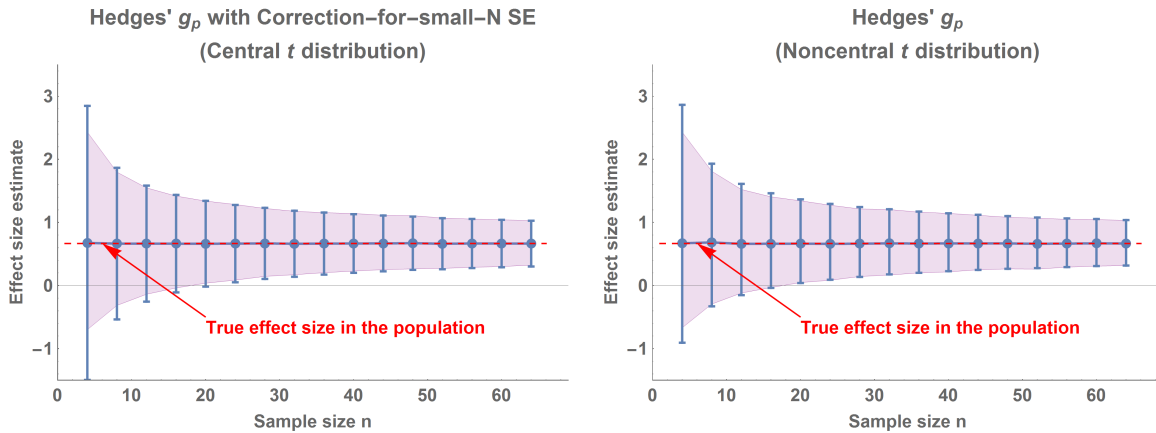
**Effect size estimates.** Looking at the means (central dots), we see that  $d_p$ , ranging from 0.73 (for small  $n$ ) to 0.67 (for large  $n$ ), is always overestimating the true effect (0.666). Overestimation is important for smaller sample sizes ( $n = 4$  and  $n = 8$ ) but negligible past  $n = 16$ . On the other hand, the mean  $g_p$  is really closely estimating the true effect, ranging from 0.662 to 0.670.

**Mean CI estimates.** The central  $t$  distribution CIs are fairly reliable to estimate the real CI. Figure 1 shows the average CI estimates (the error bars), to be compared to the shaded areas which shows the real extent of 95% of the estimates. For smaller sample sizes ( $n \leq 20$ ), the lower bounds of the confidence interval with the central methods are severely underestimating the real lower bounds.

This is a frequent finding (Harding et al., 2014). It is also to be expected that theoretical CIs are not perfect for small sample sizes as they are derived using asymptotic (large  $n$ ) theory. Regarding which SE to choose for the central CI, if we focus on the smallest  $n$  only ( $n$  of 4; for large  $n$ , there is no sizeable differences between the methods), the True\* approximation is 37% too long relative to the observed spread which is 3.50. More importantly, the CIs of this approximation extend well below zero; this is unfortunate as often, we want to assess differences relative to zero. The MLE approximation and the correction-for-small-N approximation CIs are comparable, being 24% too long. These two shows less overestimation compared to True\* because, as seen in Table 4, these SE underestimate the variance by 2% to 4%. Lastly, in the last panel, Hedges approximation has too short CIs in the upper limit (Table 4 indicated close to 18% underestimation of variance).



**Figure 2** ■ Results of the mean estimated  $g_p$  and its confidence intervals as a function of sample size (4 to 64) in a between group design. The two panels represent two different methods to compute CIs. The red dashed line is the true  $d$  in the population. Sample size  $n$  refers to the number of observations within each group.



Hence, this last approximation should be avoided.

Whereas True\* SE should be better, the CI from this SE are the largest (for very small sample size). This is caused by the distribution used. The central  $t$  distribution poorly captures the spread of the estimates. This is evidenced in Figure 2 where the central and the noncentral methods are compared directly. As seen in the latter figure, the noncentral method returns very adequate CIs: (1) For very small  $n$ , its length is 8% too long relative to the observed spread; (2) underestimation of the lower limit is roughly similar to overestimation of the upper limit.

Therefore, the noncentral CI is without a doubt the most reliable method especially when  $n$  is smaller than 20. For the central CI methods, the MLE approximation and the correction-for-small-N approximation are equivalent, and equivalent to the non-central method when  $n > 20$ . Hedges approximation should be avoided for small  $n$  whereas True\* SE is too conservative for very small  $n$ .

**Variability in CI estimates.** Another way to assess the quality of the estimate is to examine the precision of the end-points of the confidence intervals. The precision can be examined by looking at the variance of the lower and upper bounds based on the 10,000 simulated CIs. The True\* approximation SE leads to CI estimates that are about 2.8 times more variable than the noncentral CIs, while the MLE approximations are about 1.4 times more variable. The correction-for-small-N approximations and the noncentral CI's are about the same, with the lowest variances (variances are 0.23 and 0.24 respectively when  $n = 4$ ).

### Discussion

Comparing  $d_p$  with  $g_p$  allows to conclude that  $g_p$  is preferable (least biased) for estimating the magnitude of the effect when  $n < 20$ , and equivalent to Cohen's  $d_p$  when  $n \geq 20$ . Likewise, the noncentral CI had the most reliable coverage rate, followed by the (central distribution) correction-for-small-N standard error approximation. The MLE approximation is adequate on average only, being more variable from one dataset to the other. The True\* is less precise in very small sample sizes and more variable. However, as seen in the next section, it is superior to other methods in repeated measure designs. There was no sizeable difference between all the CI methods when  $n > 20$ .

In the above simulations, we tested a single true effect of medium size. Smaller effect sizes make the central method overestimate the upper and the lower bounds of the CI equivalently. By comparison, larger effect sizes exacerbate the fact that the central methods overestimate the lower limit more. This is caused by the fact that the positive asymmetry of the noncentral  $t$  distribution increases in the presence of larger effect sizes. Hence, one side of the distribution will be more correctly estimated than the other, a problem inversely proportional to sample size. This limitation of the central method should be kept in mind for all subsequent simulations, for which the effect size is kept constant at 0.666.



### Single group - Two repeated measures

In the case of a repeated measures design (pretest-posttest or paired samples), the interest lies in the difference between the two measurement times. The question of interest, when using such designs, is directed on the “change within a person, relative to the variability of change scores” (Morris & DeShon, 2002, p. 107). For the effect size calculation, this could imply to take into account the correlation between the two measures of each individual.

**The choice of divider.** Two scenarios apply, based on two different conceptualizations of the standard deviation, reviewed next. First, just like a between groups design, the population standard deviation can be estimated via a combination of the standard deviation of both measurement times (the pre- and post-measures, instead of the two groups). This can be done using the weighted pooled standard deviation found in Cohen’s  $d_p$ . Another way to obtain an estimate that combines the standard deviations is to use  $d_{av}$ , which is based on a regular (unweighted) average of the two standard deviations,  $S_{av}$  (given in Table 1). However, because there is always an equal number of observations between the two measurement times in a repeated measures design, in such cases,  $S_{av}$  is totally identical to  $S_p$  (Grissom & Kim, 2012, p. 87).

Second, unique to the repeated measures design, the standard deviation can be estimated via the standard deviation of the differences. This solution is inspired from the  $t$ -test for repeated measures:

$$t = \frac{M_{diff}}{S_D / \sqrt{n}} \tag{9}$$

in which  $S_D$  is the standard deviation of the differences,  $n$  is the number of pairs,  $M_{diff}$  is the difference between the post- and pretest means, and

$$S_D = \sqrt{\frac{\sum (X_{diff,i} - M_{diff})^2}{N - 1}}, \tag{10}$$

in which  $X_{diff,i}$  is the difference between the two measurement times for participant  $i$  (Lakens, 2013). This formula can be shown to be equivalent to

$$S_D = \sqrt{S_1^2 + S_2^2 - 2 \times r \times S_1 \times S_2}. \tag{11}$$

One major problem with  $d_D$  is that this estimate is not directly comparable to  $d_p$ . More specifically, the variability within subjects is often less than the variability between subjects. Hence,  $d_D$  is divided by a smaller number compared to  $d_p$ , which leads to a bigger effect magnitude. More specifically, assuming the homogeneity of variances (which is not needed for  $S_D$ , but needed for  $S_p$ ),

with  $S_1$  and  $S_2$  being roughly equal (to say,  $S$ ),  $S_D$  simplifies to  $\sqrt{2S^2(1-r)}$  and  $S_p$  to  $S$ . Thus, the result of  $d_D$  is  $\sqrt{2(1-r)}$  times larger than Cohen’s  $d_p$  (Cohen, 1988; Morris, 2000). Therefore, the factor  $\sqrt{2(1-r)}$  can be used to convert  $d_D$  into  $d_p$  estimates (Lakens, 2013).

The conversion is obtained as follow

$$d_D = d_p \times \sqrt{2(1-r)} \tag{12a}$$

or equivalently with

$$d_p = d_D / \sqrt{2(1-r)}. \tag{12b}$$

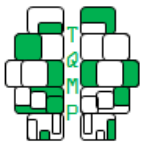
The method suggested by Eq. (12b) will be called  $d_D$  converted, or  $d_{Dc}$ .

Guidelines for what constitute small, medium or large  $d_D$  have been suggested in Eid, Gollwitzer, and Schmitt (2017). However, these indications have to be taken with cautions as they merge two distinct effect sizes, namely the mean separation (embodied in  $d_p$ ) and the degree of association (embodied in  $r$ ). A "large"  $d_D$  could be found because of a very large separation and a correlation close to zero, because of a small separation and a strong correlation, or because of any in-between results. Thus,  $d_D$  lacks specificity as a measure of effect size. Note that  $g^*$ Power and other power computation software use  $d_D$  for repeated measures designs (also called  $d_z$ ).

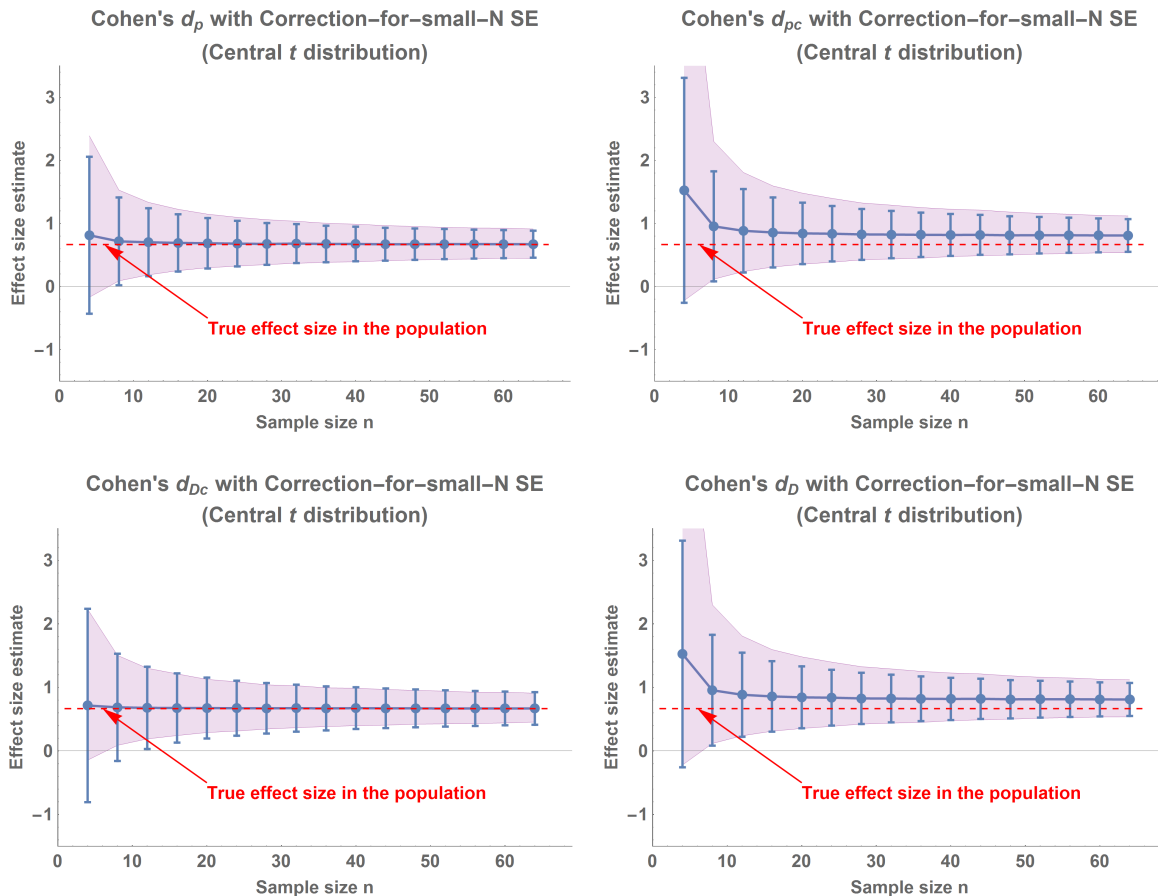
For a given effect, the experimental design should not influence the effect size reported, otherwise comparisons between designs is no longer possible, a position also supported by Dunlap, Cortina, Vaslow, and Burke (1996) and Lakens (2013). Since  $d_D$  for repeated measures leads to larger number compared to the same effect in a between-group design this could create confusion when data from different designs are compared. Thus, joining our voice to the above authors and Cumming (2012), we recommend that a common measure of effect size,  $d_p$ , be used irrespective of the design. Experimental designs should be inscribed in the computations only when statistical significance is at stake.

In sum, the best estimate of the population  $\sigma$  is possibly either: a)  $S_p$ , a pooled standard deviation (equivalent to  $S_{av}$ , an averaged standard deviation) or b)  $S_{Dc}$ , a standard deviation of the differences converted to enable comparisons (Cohen, 1988; Cumming, 2012; Grissom & Kim, 2012; Morris & DeShon, 2002). Equations for the dividers are given in Table 1.

**Biased or unbiased?** Cohen’s  $d$  for repeated measure is also biased. Therefore, Hedges correction applies (Eq.3). The correction factor is the same for all variations of  $d$ . However, the correction factor  $J$  must be based on the total number of observations minus 2 instead of the number of subjects minus one, as seen in Appendix D. This is an



**Figure 3** ■ Results of the mean estimated  $d$  and its confidence intervals as a function of sample size (4 to 64) in a repeated measure design. The four panels represent four different methods to compute  $d$ , all CIs being based on the Correction for small  $N$  approximation central  $t$  distribution method. The red dashed line is the true  $d$  in the population. Sample size  $n$  refers to the number of participants measured twice.



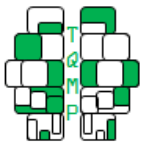
error seen in, e.g., ESCI (Cumming, 2016), and in a recent book on meta-analysis, which led some authors to wrongly conclude that the correction does not completely un-bias  $d$  in repeated measures design (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cumming, 2012).

**Which SE if the central  $t$  distribution is used?** The distribution of a repeated measure Cohen's  $d$  is  $\sqrt{2(1-r)}/n$  times the noncentral  $t$  distribution (Becker, 1988; Morris, 2000). The true SE for this design can be calculated using the formula identified by Hedges (1981) times this correction factor (Morris, 2000). Hence, the within-group equivalent of Eq. (4) is

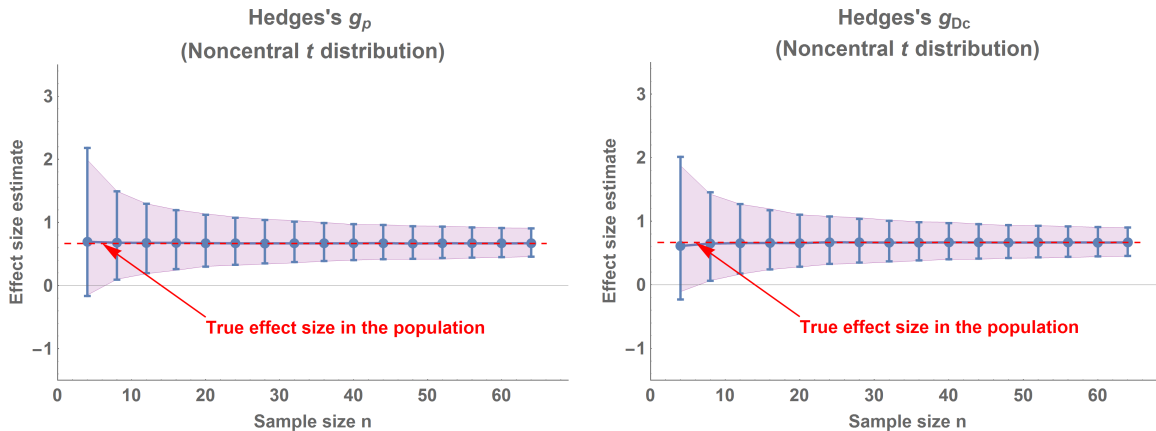
$$SE_{\text{repeated measure}} = \sqrt{\frac{v}{v-2} \frac{2(1-r)}{n} \left(1 + \delta^2 \frac{n}{2(1-r)}\right) - \frac{\delta^2}{(J(v))^2}} \quad (4b)$$

where  $n$  is the number of pairs (Algina & Keselman, 2003). This change is also valid for the True\* approximation. There is no documented equivalent of the other approximations for the repeated-measure design. However, replacing in Table 3 all occurrences of  $2/\tilde{n}$  with  $2(1-r)/n$  results in SE approximations adapted to repeated measures design. Also, the standard error of  $d_D$  is indirectly obtained from the relation found in Eq. (12a) by multiplying (4b) by  $\sqrt{2(1-r)}$ .

To find which Cohen's  $d$  for repeated measure is preferable, simulations has been conducted, estimating the true effect using  $d_p$  and  $d_D$ , along with their converted expression  $d_{pc}$  and  $d_{Dc}$  (Eqs. 12a and 12b). In all the simulations, binormal scores are generated with means 95 and 105, standard deviations of 15, and a correlation of 0.65. Confidence intervals were estimated using the noncentral



**Figure 4** ■ Results of the mean estimated  $g$  and its confidence intervals as a function of sample size (4 to 64) in a repeated measures design. The red dashed line is the true  $d$  in the population. Sample size  $n$  refers to the number of participants measured twice.



method and the central method with the Correction for small N SE approximation adapted for repeated measure. This specific SE was selected because it had the most adequate coverage when we examined between-group designs.

**Results for repeated measures design.**

The results for the various Cohen’s  $d$  with the central CIs are presented in Figure 3. Likewise, the noncentral method with the estimators  $g_p$  and  $g_{Dc}$  are depicted in Figure 4.

**Effect size estimates.** Comparing with a true effect size set at  $10/15 = 0.666$ , the results confirm clearly that  $d_D$  and the converted version of  $d_p$  ( $d_{pc}$ ) are poor estimators, with a very strong overestimation of the true effect. As previously argued, these effect sizes should be avoided because they lack specificity. On the other hand, both  $d_p$  and the converted version of  $d_D$  ( $d_{Dc}$ ) performed well. The advantage goes to  $d_{Dc}$  who was markedly less biased for small  $n$ , with estimations ranging from 0.716 to 0.669 (0.687 when  $n$  is 8). The estimations of  $d_p$  were ranging from 0.814 to 0.672 (0.717 when  $n$  is 8). Biases in percent are 7.4% for  $d_{Dc}$  vs. 22.1% for  $d_p$  when  $n$  is 4.

Regarding Hedges estimators, Hedges’  $g_p$  largely reduced the bias found in small sample size, ranging from 0.693 to 0.670 (0.676 when  $n$  is 8). Hedges’  $g_{Dc}$ , on the other hand, underestimated the effect in small sample sizes ( $n$  below 12) with estimates ranging from 0.612 to 0.666 (0.649 when  $n$  is 8), but completely eliminated the bias in sample size above 12. Nonetheless, the difference between the two Hedges’ estimators is immaterial (less than 0.004) for  $n$

above twelve. When  $n$  is 4, the bias for  $g_{Dc}$  is -8%, whereas the bias for  $g_p$  is +3.9%, half the bias of the former. Hence  $g_p$  is the better choice for this design as well.

**Mean CI estimates.** The central method to construct a CI with the Correction for small N standard error was not exact even in larger sample size. With  $d_p$ , underestimation of the upper bound is noticeable for all sample sizes. This is not desirable and therefore this method of constructing CI should be avoided either by using a different SE approximation or by using the noncentral method, as seen in Figure 4. On the other hand,  $d_{Dc}$  CI length underestimated the true lower bounds (too conservative) for about all sample sizes examined, something never seen for the other methods. By comparison, this combination exceeded the true interval by 24% for  $n$  of 4, whereas the noncentral CI with  $g_p$  exceeded it by 3.9%. Thus, even if the Correction for small N approximation with  $d_{Dc}$  is not underestimating the true interval, it is quite wide.

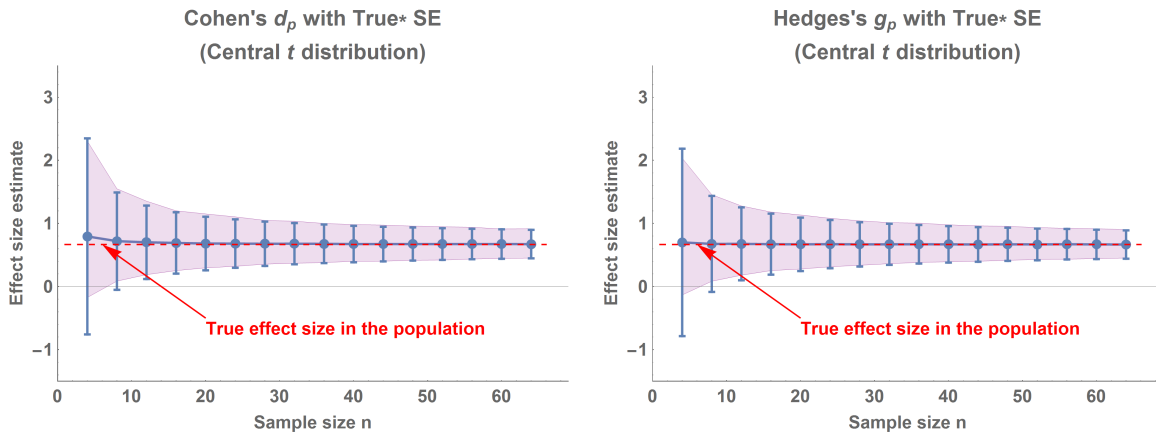
Given the generally poor performance of the correction for small N SE, we tested another central approximation, namely  $d_p$  and  $g_p$  with the True\* method. The results are shown in Figure 5. As seen, this standard error provides much better coverage than the correction for small N. However, in conjunction with Cohen’s  $d_p$ , it shows some underestimation of the upper bounds for ns of 8 to 24. Underestimation is reduced when used in conjunction with Hedges’  $g_p$ , visible only for ns of 16 and 20.

Therefore, the noncentral CI is the most reliable method, whereas the central method with the True\* SE provides equivalent coverage for sample sizes above 24.





**Figure 5 ■** Results of the mean estimated  $d$  (left) and  $g$  (right) and its confidence intervals as a function of sample size (4 to 64) in a repeated measures design. The two panels represent two different estimates with the True\* method to compute CIs. The red dashed line is the true  $d$  in the population. Sample size  $n$  refers to the number of participants measured twice.



The Correction for small N SE is not exact even in large samples ( $n = 64$ ), thus should be abandoned for this design.

**Variability in CI estimates.** Looking at the variability at the two ends of the intervals revealed that the Correction for small N and the True \* SE with  $d_p$  had the highest variabilities (2.39 and 2.34 respectively for  $n$  of 4), closely followed by  $d_{Dc}$  (2.26 for  $n$  of 4). The noncentral method with  $g_p$  had the lowest variability of all (1.31 for  $n$  of 4), followed by the noncentral method with  $g_{Dc}$  (1.42 for  $n$  of 4).

**Discussion**

Comparing four different Cohen's  $d$  for repeated measures, two SE approximations and two methods to build a CI, along with Hedges' unbiased estimators, allows concluding that the noncentral method is superior to the central method and Hedges  $g_p$  is superior to the other estimators. The difference between the central method with the True\* SE and the noncentral was noticeable only for sample size below 24. Regarding the best estimate of effect size,  $g_p$  was found slightly superior to  $d_{Dc}$  and vastly superior to the other estimators,  $d_p$ ,  $d_D$  and  $d_{pc}$ . However, the difference between the estimators quickly resorbs with sample size greater than 20. Hedges correction considerably reduced the bias found with small sample size ( $n < 20$ ) in all  $d$  estimators except  $d_{Dc}$  which was additionally biased after the correction.

**Design with a comparison group: Glass's  $\Delta$ .**

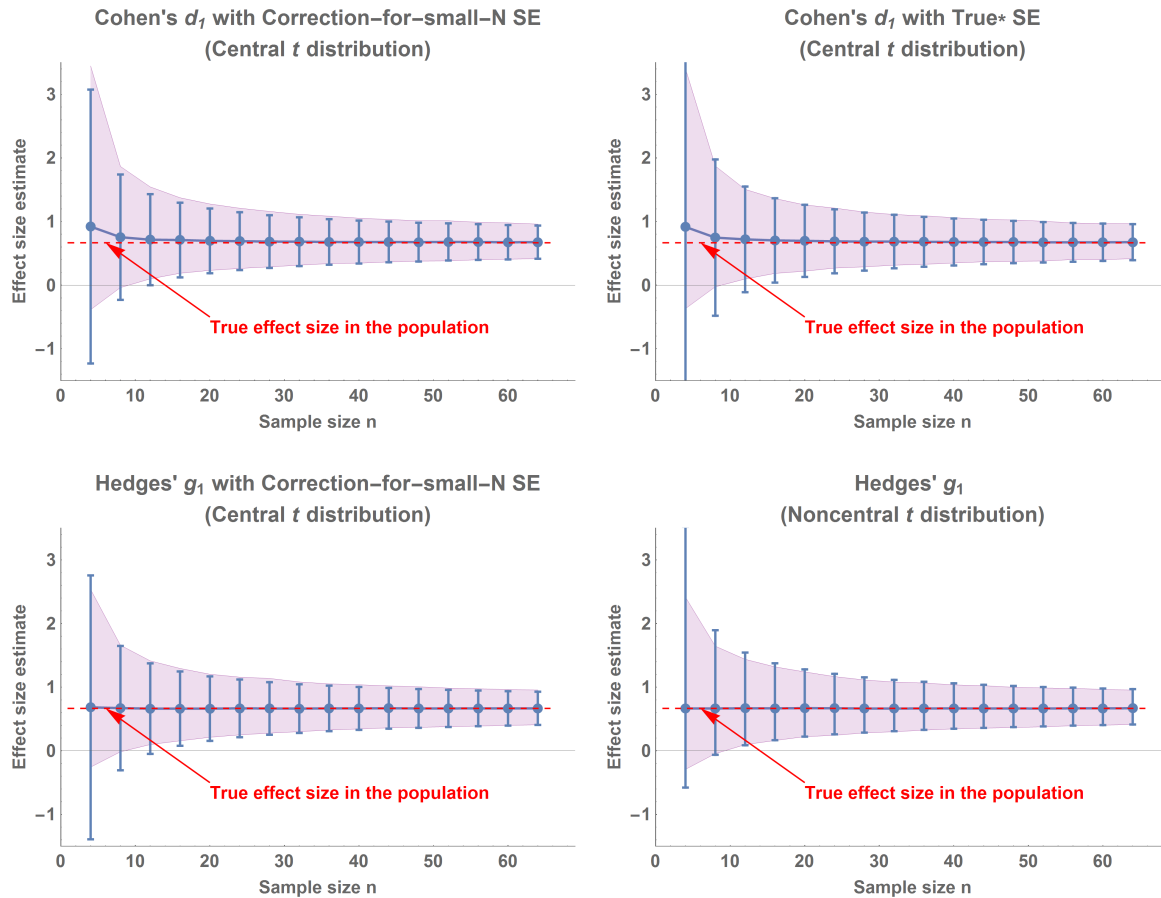
In situation of unequal variances, the standard deviations of both groups (or both measurement times) cannot be pooled together. When the assumption of homogeneity of variances is not respected, or when a control group is present, the divider that best estimates the population variation is the unaffected group's standard deviation. This approach is referred to as Glass's  $\Delta$  (Glass et al., 1981). With a pre- and post-test design, the pre-group is usually thought to better represent the population variations since it has not been affected by the intervention. Similarly, when a control condition is available, the standard deviation of this group is taken without combining it with other standard deviations. These cases will not be further tested using simulations considering that the heterogeneity of variances can be modeled in numerous ways, which would exceed the scope of this article (see Algina, Keselman, & Penfield, 2006, for a robust CI; see Morris, 2008, for effect sizes with a control group). We simply note that Glass'  $\Delta$ , being based on a single measure of standard deviation, has all the properties of a single-group Cohen's  $d$ .

**Single group  $d$**

A last experimental design in which Cohen's  $d$  can be used is when researchers are interested in comparing a single mean to a specific value. This value is taken as the population mean ( $\mu$ ) or a reliable estimate of it; it can also be arbitrarily chosen as a point of reference. For example,



**Figure 6 ■** Results of the mean estimated  $d$  and  $g$  and its confidence intervals as a function of sample size (4 to 64) in a single-group design. The panels represent two different estimates ( $d$  and  $g$ ) and three methods to compute CI. The red dashed line is the true  $d$  in the population. Sample size  $n$  refers to the number of participants.



it could come from a normalized value, specified after a meta-analysis, or based on a theoretical prediction.

**The correct divider.** In the absence of a better estimate (e.g., a meta-analytic estimate or a normalized measurement), the standard deviation of the group is chosen as the only estimate of the population standard deviation available.

**Biased or unbiased?** For any value constructed upon an estimate of the population standard deviation, the result is likely biased (as confirmed in the subsequent simulations). Therefore, Hedges correction also applies to single group  $d_1$ , with  $J$  (Eqs. 2a or 2b) based on the number of observations minus 1.

**Which SE if the central  $t$  distribution is used?** Similar to the other designs, the noncentral method should offer the best coverage rate. However, if the central method is desired, the True\* or the Correction-for-small-N standard

error approximations are two potential candidates. Both are tested here.

In another set of simulations, we compared four methods, Cohen's  $d_1$  with a central CI based on the True\* and the Correction-for-small-N standard error, Hedges'  $g_1$  with the Correction for small N standard error, and finally, Hedges estimate with the noncentral method.

The results seen in Figure 6 indicate that  $d_1$  is biased upward whereas  $g_1$  completely eliminate this bias. Regarding the intervals,  $d_1$  in conjunction with the Correction for small N poorly estimated the upper bounds of the confidence interval for all sample sizes tested. The same SE approximation in conjunction with Hedges  $g_1$  also led to systematic underestimation of the upper bounds for all sample size tested. The True\* method with  $d_1$  does far much better with no underestimation of the upper bounds; the lower bounds are however more conservatively esti-



mated for about all sample sizes. The inexact estimations of the lower bounds in larger samples is not attributable to the absence of Hedges correction. Finally, the noncentral method is clearly superior with reliable estimates of the lower bound in small samples, but conservative estimates of the upper bound for  $n$  below 32. In brief, the noncentral method should be prioritized, in conjunction with  $g_1$  for small sample sizes ( $n \leq 24$ ). The central method with  $d_1$  and the True\* has conservative lower bounds, but do well with sample sizes above 30.

### General discussion

By reporting an effect magnitude and confidence interval, researchers ensure that the results are translated while preserving their nuance. In fact, interpreting this information allows judging to what extent the research hypothesis is supported. This can be achieved by reporting the magnitude of an effect in its original unit, displaying graphical representations of the results, or via standardization of the effect. All these options share the common challenge that they infer the population characteristics based on the sample. Certainly, one of the most efficient ways to judge the precision of such an inference is to examine the confidence interval around the reported statistic. This paper was dedicated to the  $d$  family of effect sizes, which is used to standardize the difference between two means for designs with two independent groups, two repeated measures, or relative to a specified value. To differentiate the various Cohen's  $d$ , a notation has been proposed based on the divider employed. Two methods to construct a confidence interval around  $d$  have been evaluated, either relying on the noncentral  $t$  or the central  $t$  distribution. To report the correct confidence intervals and to properly estimate the effect size with a Cohen's  $d$  estimator, the right divider must be identified, an unbiased estimator must be considered, and a reliable method for building confidence intervals must be selected. The aim of the present paper was to shed light on those concerns. The methodology employed Monte Carlo simulations to compare the various estimators with a simulated population composed of 10,000 samples for a given sample size. A summary of the results is presented below and in Table 5.

For design with two-independent groups, the best divider is a pooled standard deviation. Since all Cohen's  $d$  are systematically overestimating the effect, Hedges' correction should be applied for small sample size ( $n < 20$  per group). This leaves us with  $g_p$  as the best estimator for this design. The construction of a confidence interval around  $g_p$  should be done with the noncentral distribution. However, the central method with the Correction-for-small-N (or the True\*) standard error approximation was similar to the noncentral method when  $n_{\text{per group}} > 20$ .

For a repeated measure design, the best divider is a pooled standard deviation as well, found in  $d_p$ , or equivalently in  $d_{av}$ . Here again, Hedges' correction should be applied for small sample sizes ( $n < 20$ ), which leaves us with  $g_p$  as the best estimator for this design. Confidence intervals should be constructed with the noncentral distribution. However, the central method with the True\* did equally well with sample sizes above 24. The Correction for small N standard error approximation for repeated measures performed poorly and should be avoided.

Therefore, a major recommendation emerges: the noncentral method with  $g_p$  should be employed in all situations. If calculations are a concern, the central method with the True\* SE, in conjunction with  $d_p$ , performed well in a between-group design with more than 20 participants per group, and in a repeated measure design with more than 24 participants measured twice. Implementation of the noncentral method to construct CI around Hedges  $g$ , using R software, is given in the following subsection. Also, a comparison of the formula used in the MBESS and the metafor packages for R and the ESCI spreadsheets for Excel is available in Appendix B.

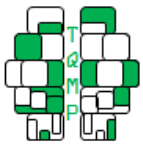
One novel contribution of this article was to rectify the Hedges  $g$  correction factor in a repeated measure design, which was thought to rely on the degree of freedom (different according to the design), while it actually relies on the number of observations minus two for both between and within subject designs. Hedges  $g$  was found to completely un-bias the  $d$  estimator in all three designs. More on that can be found in Appendix D.

Lastly, the formula that best estimates the standard error is the True formula, given in Table 3, or its approximation, True\*. However, only in between-group design and for  $n$  below 12, the latter approximation is less appropriate to estimate the central distribution standard error compared to the Correction-for-small-N formula.

### Application Guide

In view of the current results, we propose this guide to help researchers apply the best practices regarding the  $d$  family of effect size. Standardized effect sizes are great tools to communicate research results in a comparable scale. However, un-standardized effect sizes are just as meaningful to interpret, especially when the measurement tools are well known or have readily interpretable units (e.g., Beck Depression Inventory, responses time). Hence, we encourage researchers to report un-standardized, as well as standardized effect sizes. In all cases, accurate confidence intervals should always accompany effect size estimates to assess their variability and precision.

Acknowledging that an important determinant of good statistical practices are their accessibility and ease of im-



**Table 5** ■ Main results of the Monte Carlo simulations.

Design	Effect size estimation	Confidence interval estimation
Two independent groups	<ul style="list-style-type: none"> <li><math>g_p</math> is the best estimator overall.</li> <li>Hedges <math>g</math> correction is necessary when <math>n_{\text{per group}} &lt; 16</math>.</li> <li>The pooled standard deviation, denoted <math>S_p</math>, is the best divider.</li> </ul>	<ul style="list-style-type: none"> <li>The noncentral CI is the most reliable method, especially when <math>n_{\text{per group}} &lt; 20</math>.</li> <li>The central CI method is equivalent to the noncentral method when <math>n_{\text{per group}} &gt; 20</math>, no matter the SE approximation.</li> <li>The MLE and the correction-for-small-N SE approximations led to the best central estimations in small sample sizes.</li> </ul>
Repeated measures	<ul style="list-style-type: none"> <li><math>g_p</math> is the best estimator overall. Hedges <math>g</math> correction is necessary for <math>S_p</math> when <math>n_{\text{subjects}} &lt; 16</math>.</li> <li>The pooled standard deviation, <math>S_p</math>, and the standard deviation of the differences converted, <math>S_{Dc}</math>, are the best dividers.</li> <li><math>g_p</math>, <math>d_p</math> and <math>d_{Dc}</math> are equivalent when <math>n_{\text{subjects}} &gt; 16</math>.</li> </ul>	<ul style="list-style-type: none"> <li>The noncentral CI is the most reliable method, especially when <math>n_{\text{per group}} &lt; 24</math>.</li> <li>The central CI method with the True* SE approximation is equivalent to the noncentral method when <math>n_{\text{subjects}} &gt; 24</math>.</li> </ul>
Single mean relative to a target value	<ul style="list-style-type: none"> <li><math>g_1</math> is better overall compared to <math>d_1</math>.</li> <li>Hedges <math>g</math> correction is necessary when <math>n_1 \leq 24</math>.</li> </ul>	<ul style="list-style-type: none"> <li>The noncentral CI is the most reliable method, especially when <math>n_1 &lt; 32</math>. However, conservative upper bounds are expected with this method and those sample sizes.</li> <li>The central CI method with the True* SE approximation is equivalent to the noncentral method when <math>n_1 &gt; 30</math>.</li> </ul>

*Note.* The simulations have been realized with a true effect size of  $d = .666$ . The quality of the central CI method decreases as the true effect size increases, due to the increase in asymmetry. The results are based on simulations which comply with the assumptions of normality, homogeneity of variances, and independence of data sampling.

plementation, we give a function in R to compute Hedges  $g_p$  with its CI based on the noncentral method, available on the journal’s web site and given in Listing 1 at the end. The function returns the most accurate estimates for a between and a within group designs, according to the present study. The function first computes a Cohen’s  $d_p$  based on the pooled standard deviation of the two groups or the two measurement times. Then, an unbiased Hedges  $g$  is computed by multiplying  $d_p$  with the correction factor  $J$ , which is based on the gamma function and the number of observations minus two (Eq. 2a). The CI is obtained from the noncentral  $t$  distribution with the degree of freedom and the noncentrality parameter. The command

```
gethedgesg(x1, x2)
```

does all the required computations for a between group Hedges  $g$  with its CI, assuming that the vectors “x1” and “x2” have been defined with for example

```
x1 <- c(53, 68, 66, 69, 83, 91)
x2 <- c(49, 60, 67, 75, 78, 89)
```

For a within group Hedges  $g_p$ , the command is the following

```
gethedgesg(x1, x2, design = "within")
```

The between group and the within group commands differ only in terms of the calculation of the noncentrality parameter. Finally, a coverage level  $\gamma$  different from the default 95% can be added to the argument list, with for example

```
gethedgesg(x1, x2, coverage = 0.9)
```

Three programs commonly used to compute a Cohen’s  $d$  and its CI are the MBESS and the metafor packages for R and the ESCI spreadsheets for Excel. In Appendix B, we compare the results of those programs with the formulas suggested in this text. We found that MBESS and ESCI rely on a CI estimation method described by Steiger and Fouladi



(1997) that is less appropriate than the noncentral method described in the present text to construct a CI around  $d$  or  $g$ . Unless revisions are made to the above programs, the code given above provides an easy way to obtain the correct gp and noncentral CI.

### Limits and conclusion

This study is limited in its approach, relying on the assumptions of a normally distributed population, homogeneous variances and independent data points. The findings of this study, and the applicability of the noncentral and the central CIs, are contingent on the respect of those assumptions. On the other hand, bootstrapping procedures, which do not rely on those assumptions, have been occulted from the present paper, for reasons of parsimony. However, comparing the bootstrap method with the noncentral one would be of great value. Lastly, it is always preferable to operate on real data to reach ecologically valid conclusions. Nevertheless, Monte Carlo methodology allows investigating situations which would require an impressive amount of resources otherwise.

In conclusion, this study reviewed the  $d$  family of effect sizes and their confidence intervals. Some ambiguity regarding all the possible combinations have been addressed and resolved. The literature on commonly-used effect sizes was found to be very confusing, with many different names for the same constructs and various contradictions. We hope this work correctly addressed some misunderstandings in a way that will promote the use of effect sizes and confidence intervals within the field of social sciences.

### Authors' note

We would like to thank Daniel Lakens, Geoff Cumming, Robert Calin-Jageman, Vincent Leblanc, Marc-André Goulet, and the reviewers, for their comments and suggestions on an earlier version of this text. This research was supported in part by the Conseil pour la recherche en sciences naturelles et en génie du Canada.

### References

- Algina, J., & Keselman, H. J. (2003). Approximate confidence intervals for effect sizes. *Educational and Psychological Measurement, 63*(4), 537–553. doi:10.1177/0013164403256358
- Algina, J., Keselman, H. J., & Penfield, R. D. (2006). Confidence interval coverage for Cohen's effect size statistic. *Educational and Psychological Measurement, 66*(6), 945–960. doi:10.1177/0013164406288161
- Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257–278. doi:10.1111/j.2044-8317.1988.tb00901.x
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: John Wiley & Sons.
- Clay, R. (2014). *Confidence intervals for effect sizes from non-central distributions*. Open Science Collaboration. Retrieved from <http://osc.centerforopenscience.org/2014/03/06/confidence>
- Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioural sciences (2nd ed.)* Hillsdale, NJ: Lawrence Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Methods, 1*(2), 155–159. doi:10.1037/0033-2909.1.2.155
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.
- Cumming, G. (2016). *Esci (exploratory software for confidence intervals)* (Version nov.2017). Retrieved from <https://thenewstatistics.com/itns/esci/>
- Cumming, G., Fidler, F., Kalinowski, P., & Lai, J. (2012). The statistical recommendations of the American psychological association publication manual: Effect sizes, confidence intervals, and meta-analysis: The apa publication manual and statistical change. *Australian Journal of Psychology, 64*(3), 138–146. doi:10.1111/j.1742-9536.2011.00037.x
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*(4), 532–574. doi:10.1177/0013164401614002
- Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol. Methods, 1*, 170–177. doi:10.1037/1082-989X.1.2.170
- Eid, M., Gollwitzer, M., & Schmitt, M. (2017). *Statistik und forschungsmethoden*. Berlin: Beltz.
- Fidler, F., & Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed-and random-effects effect sizes. *Educational and Psychological Measurement, 61*(4), 575–604.
- Fleishman, A. I. (1980). Confidence intervals for correlation ratios. *Educational and Psychological Measurement, 40*(3), 659–670. doi:10.1177/001316448004000309
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General, 141*(1), 2–18. doi:10.1037/a0024338





- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational researcher*, 5(10), 3–8. doi:[10.3102/0013189X005010003](https://doi.org/10.3102/0013189X005010003)
- Glass, G. V., McGraw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.
- Goulet-Pelletier, J.-C., & Cousineau, D. (in preparation). A review of effect sizes and their confidence intervals, part ii: Analysis of variance, correlation, and regression.
- Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications (2nd ed.)* New York, NY: Routledge.
- Harding, B., Tremblay, C., & Cousineau, D. (2014). Standard errors: A review and evaluation of standard error estimators using monte carlo simulations. *The Quantitative Methods for Psychology*, 10(2), 107–123. doi:[10.20982/tqmp.10.2.p107](https://doi.org/10.20982/tqmp.10.2.p107)
- Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6(2), 107–128. doi:[10.2307/1164588](https://doi.org/10.2307/1164588)
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Huberty, C. J. (2002). A history of effect size indices. *Educational and Psychological measurement*, 62(2), 227–240. doi:[10.1177/0013164402062002002](https://doi.org/10.1177/0013164402062002002)
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
- 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. doi:[10.1177/0013164404264850](https://doi.org/10.1177/0013164404264850)
- Kelley, K. (2007). Methods for the behavioral, educational, and social sciences (mbess) (Version 4.4.3). Retrieved from <http://www.cran.r-project.org/>
- Kelley, K. (2017). The mbess r package (Version 4.4.3). Retrieved from <http://nd.edu/~kkelley/site/MBESS.html>
- Kennedy, W. J., & Gentle, J. E. (1980). *Statistical computing*. New York: Marcel Dekker inc.
- Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, 56(5), 746–759. doi:[10.1177/0013164496056005002](https://doi.org/10.1177/0013164496056005002)
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and anovas. *Frontiers in Psychology*, 4, 863–875. doi:[10.3389/fpsyg.2013.00863](https://doi.org/10.3389/fpsyg.2013.00863)
- Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, 53(1), 17–29. doi:[10.1348/000711000159150](https://doi.org/10.1348/000711000159150)
- Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods*, 11(2), 364–386. doi:[10.1177/1094428106291059](https://doi.org/10.1177/1094428106291059)
- Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, 7(1), 105–125. doi:[10.1037/1082-989X.7.1.105](https://doi.org/10.1037/1082-989X.7.1.105)
- Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biological Reviews*, 82(4), 591–605. doi:[10.1111/j.1469-185X.2007.00027.x](https://doi.org/10.1111/j.1469-185X.2007.00027.x)
- Pek, J., & Flora, D. B. (2018). Reporting effect sizes in original psychological research: A discussion and tutorial. *Psychological Methods*, 23(2), 208–225. doi:[10.1037/met0000126](https://doi.org/10.1037/met0000126)
- Smithson, M. J. (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. doi:[10.1177/0013164012197139](https://doi.org/10.1177/0013164012197139)
- Smithson, M. J. (2003). *Confidence intervals*. Thousand Oaks, CA: Sage.
- Steiger, J. H. (2004). Beyond the f test: Effect size confidence intervals and tests of close fit in the analysis of variance and contrast analysis. *Psychological Methods*, 9, 164–182. doi:[10.1037/1082-989X.9.2.164](https://doi.org/10.1037/1082-989X.9.2.164)
- Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
- Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, 31(3), 25–32. doi:[10.3102/0013189X031003025](https://doi.org/10.3102/0013189X031003025)
- Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of statistical software*, 36(3), 1–48. doi:[10.18637/jss.v036.i03](https://doi.org/10.18637/jss.v036.i03)

## Appendix A: General methodology

To evaluate the reliability of an estimate, Monte Carlo simulations can be used to compare the estimation with the “real” value of a simulated population (Harding et al., 2014). The simulated population is based on a set of true parameters which are known so that the estimates can be compared. One simulation consists of a very large number of samples

**Table 6** ■ Estimated effect size [95% confidence intervals] for a fictitious data set (given in Appendix B) for the noncentral method and from three software packages.

Method or software	Independent groups	Repeated measures
Noncentral method	$d_p = 0.1449$ [-1.1068, 1.4704] $g_p = 0.1338$ [-1.1204, 1.4561]	$d_p = 0.1449$ [-0.1497, 0.5129] $g_p = 0.1338$ [-0.1624, 0.4977]
MBESS	$d_p = 0.1449$ [-0.9919, 1.2748] $g_p = 0.1338$ [-1.0026, 1.2636]	No direct commands available.
Metafor	[ $d_p$ is not computed] $g_p = 0.1338$ [-0.999, 1.267]	No direct commands available.
ESCI	$d_p = 0.145$ [-0.992, 1.275] $g_p = 0.134$ [no interval available.]	$d_{av} = 0.145$ [-0.161, 0.438] $g_{av} = 0.122$ [no interval available.]

generated randomly. In this study, we assumed that simulated groups of participants are taken from normal distributions with known means and a common standard deviation. The groups have unequal means so that the magnitude of effect is not zero. Hence, all simulations comply with the normality assumption, the homogeneity of variances assumption and the independence of data points assumption, that are taken for granted in most formulas.

We chose to simulate 10,000 samples within each group sizes varying from  $n = 4$  to  $n = 64$  (increasing by increments of 4). A distribution of the 10,000 effect sizes is then obtained, and the quantiles (e.g., .025 and .975 for a 95% coverage) from this second-order sample of effect sizes are taken as the true confidence interval. Concurrently, for each sample, a confidence interval is estimated via the central or noncentral estimation method. The mean values of the estimated CI's is then compared to the true confidence interval. This operation is repeated for all sample sizes. If the estimated CI's yield similar results than the true CI, we conclude that the estimation method is reliable.

#### Appendix B: Existing software to compute $d$ and its confidence interval

There are three commonly used software to compute Cohen's  $d$  and Hedges'  $g$  along with their confidence intervals, the MBESS and metafor packages within R (Kelley, 2007, 2017; Viechtbauer, 2010) and ESCI within Excel (Cumming, 2016). We explore their results based on a tiny sample with two sets of scores: Set 1: 53, 68, 66, 69, 83, 91; Set 2: 49, 60, 67, 75, 78, 89. These scores can either be from two independent groups or from a repeated measures design.

With these data, the estimators  $d_p$  and  $g_p$ , with their 95% confidence intervals, based on the noncentral method, are presented in Table 6. Because  $g_p$  is the same for both within and between-subject designs, the only difference is in the interval estimations, due to the correlation between scores in the repeated measure design which improves estimations; as a consequence, the noncentrality parameter is calculated differently. The results of the noncentral method, MBESS, metafor and ESCI are presented in Table 6 and discussed afterwards.

#### MBESS within R

We tested MBESS version 4.4.1. A simple set of commands is the following (the lines beginning with # are the output triggered by the instructions):

```
library(MBESS)
x1 <- c(53, 68, 66, 69, 83, 91)
x2 <- c(49, 60, 67, 75, 78, 89)
smd(Group.1=x1, Group.2=x2)
smd(Group.1=x1, Group.2=x2, Unbiased=TRUE)
# [1] 0.1449935
# [1] 0.1337921
```

The first two commands create a two-group dataset, x1 and x2, composed of six participants in each group. The first smd (standardized mean difference) command return  $d_p$  whereas the second return  $g_p$ , the unbiased version of  $d_p$ . Next, we compute the CIs.

```
ci.smd(ncp=0.1449935 * sqrt(6/2), n.1=6, n.2=6, conf.level=0.95)
ci.smd(ncp=0.1337921 * sqrt(6/2), n.1=6, n.2=6, conf.level=0.95)
```



```
# $Lower.Conf.Limit.smd: -0.9918915; $Upper.Conf.Limit.smd: 1.274752
# $Lower.Conf.Limit.smd: -1.002561; $Upper.Conf.Limit.smd: 1.263567
```

The above instruction computes the 95% confidence interval of  $d_p$  and  $g_p$  respectively, in a between group design.

As seen in Table 6, MBESS returns the correct values for the estimators. However, it returns a shorter confidence interval for  $d_p$  and  $g_p$  than the noncentral method. The difference is due to the method used by this package to construct the confidence interval, which is based on the interval estimation approach by Steiger and Fouladi (1997). This alternative approach is discussed in Appendix C.

For a repeated measure design, there exists no command to compute a confidence interval in MBESS. There is a workaround with the following instructions, which gives a CI around  $g_p$  also based on the estimation approach of Steiger and Fouladi, resulting in a shorter CI compared to the noncentral method:

```
r <- cor(x1,x2)
ci <- ci.smd(ncp = 0.1337921 * sqrt(6/2) / sqrt(1-r), n.1=6, n.2=6,
  conf.level = .95)
ci$Lower.Conf.Limit.smd * sqrt(1-r)
ci$Upper.Conf.Limit.smd * sqrt(1-r)
# [1] -0.1590722
# [1] 0.4203255
```

### metafor within R

We tested the metafor package version 2.0-0 in R, a package made for calculating various effect sizes and specifying meta-analytic models. For a between-group design, the following lines of codes return the unbiased Hedges  $g_p$  (under “yi”) and its variance of error (under “vi”). The last command returns the summary statistics, which includes the CI (under “ci.lb” and “ci.ub”):

```
library(metafor)
x1 <- c(53, 68, 66, 69, 83, 91)
x2 <- c(49, 60, 67, 75, 78, 89)
res <- escalc(measure = "SMD", vtype = "UB", mli = mean(x1), m2i = mean(x2),
  sdli = sd(x1), sd2i = sd(x2), n1i = length(x1), n2i = length(x2))
summary.escalc(res)
# yi vi sei zi ci.lb ci.ub
# 0.1338 0.3344 0.5783 0.2314 -0.9996 1.2672
```

The results from this package can be compared to the other packages in Table 6. As seen, the CI from this package is really close, yet not the same, as the CI built upon the Steiger and Fouladi method with MBESS and ESCI. In fact, we found out that this package returns a CI using the central method with the Hedges approximation SE, but instead of multiplying the SE with a  $t$ -value, it multiplies it with a  $z$ -value. There exists no direct command for a repeated measures design.

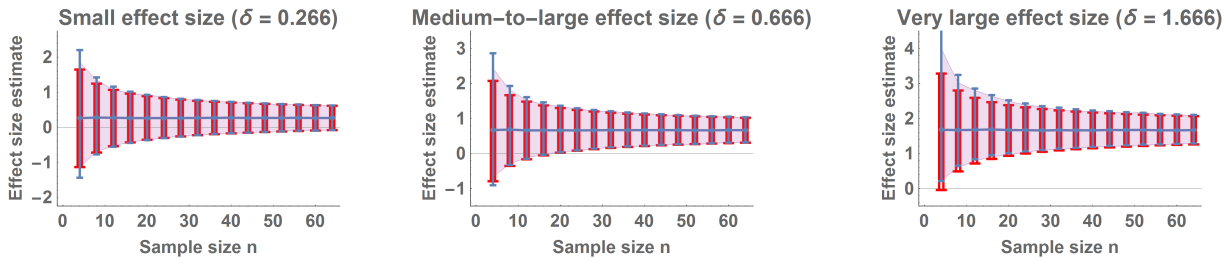
### ESCI

The spreadsheet ESCI version Nov-17-2016 was tested. In the "Data two" sheet, dedicated to the two-group design, data points can be entered under "Group 1" and "Group 2" columns. ESCI returns the correct Cohen's  $d_p$  and unbiased (Hedges)  $g_p$ . However, the confidence interval returned for  $d_p$  is the same as MBESS, which is shorter than the noncentral method described in this text. Hence, ESCI is also based on the Steiger and Fouladi (1997) estimation approach. There is no confidence interval for  $g_p$ .

For a repeated measures design, the data are entered in the "Data paired" sheet. ESCI returns the correct (biased) Cohen's  $d_p$  (ESCI reports  $d_{av}$  but it is identical to  $d_p$ ), but the Hedges (unbiased)  $g_p$  is 0.122. This erroneous value comes from the correction factor  $J$  being based on  $\nu = n_{\text{pairs}} - 1$  instead of the correct  $2(n_{\text{pairs}} - 1)$  parameter. Also, the 95% confidence interval for  $d_p$ , which returned [ 0.161, 0.438], is based on the Steiger and Fouladi (1997) estimation approach and the incorrect  $\nu$ . Note that ESCI estimates  $\sqrt{2(1-r)}$  directly from the ratio of the standard deviation of the differences onto the pooled standard deviation (rather than estimating  $r$ ; see Eq. 11) and the noncentrality parameter from the paired-sample  $t$  test (rather than using Eq. 5b). These two variations cause negligible differences in the CI (compare with the 95% CI without these variations: [-.159, 0.420]). Also, ESCI requires at least 6 pairs of observations to



**Figure 7** ■ Comparison of the Steiger and Fouladi (1997) noncentral method (in red, thick error bars), with the noncentral method discussed in this text (in blue, thin error bars), in a between-group design. Sample size  $n$  refers to the number of observations within each group.



work in a repeated measure design.

**Conclusion.**

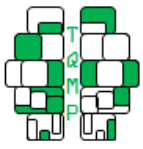
In brief, MBESS and ESCI both reported the same confidence intervals for a between-group design based on the method described by Steiger and Fouladi (1997), which leads to a different result compared to the noncentral method described in the present text. MBESS has no direct command for repeated measure designs whereas ESCI do not report the confidence interval around Hedges  $g$ . Furthermore, ESCI incorrectly applies Hedges  $g$  correction for repeated measure based on the degree of freedom instead of the total number of data points minus 2. Finally, metafor uses a  $z$  score for the coverage of the confidence interval, making it less valid for smaller samples and for larger effects where asymmetry is more pronounced.

**Appendix C: The interval estimation approach by Steiger and Fouladi**

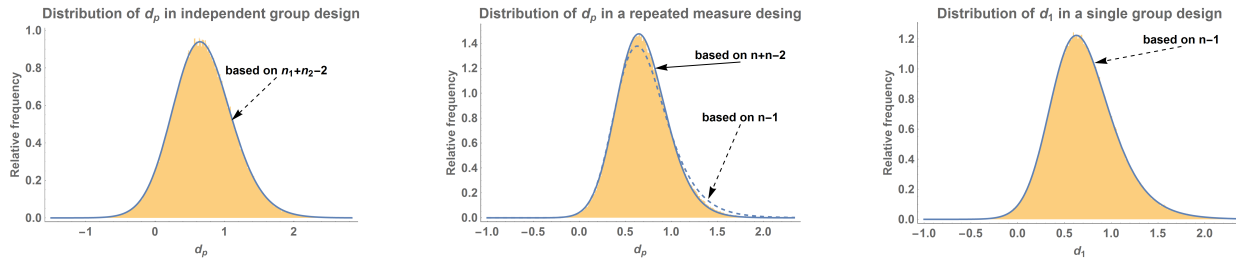
Steiger and Fouladi (1997) popularized a method to construct confidence intervals that is similar to the noncentral method used and described in the present paper, and also relies on the noncentral distribution, yet varies on important aspects. Their method, that we will further refer to as the pivoting method, has been implemented in some popular software, e.g., ESCI and MBESS, which is why it is briefly addressed here. In both methods, a non-centrality parameter (ncp) is estimated from the observed effect size. In the pivoting method of Steiger and Fouladi, two noncentral distributions are built instead of one. The first noncentral distribution is built around a value that is yet to be obtained from the distribution that positions the observed ncp at the 0.025 quantile. In a similar manner, the second noncentral distribution is obtained by placing the observed ncp at the 0.975 quantile. Then, the interval between these two distributions' ncp is taken to form the upper and lower bound of the 95% CI respectively. Thus, the pivoting method requires two noncentral distributions and assumes that the observed ncp is an extreme value, being either exceptionally small or exceptionally large. By comparison, the noncentral method discussed in this text assumes that the observed ncp is the most likely and best estimate of the population ncp. Only one noncentral distribution is constructed, centered at the observed ncp, and the upper and lower bounds are taken from this distribution to form the 95% CI. Hence, the difference between the two methods is subtle, but quite substantial in theory, in computation time, and possibly also in practice. Algina and Keselman (2003) have contributed to the validation of the pivoting method for a repeated measures design,  $d_{av}$ , scenario. However, we suspect that this method is less appropriate than the noncentral method presented in this text in most, if not all, scenarios.

To support our suspicions, we ran three additional simulations. Because the asymmetry of the noncentral distribution is an issue, and the asymmetry increases with the true effect size, we manipulated the effect size from small ( $\delta = 0.266$ , which corresponds to a difference of 4 points when standard deviation is 15); medium (0.666) and very large ( $\delta = 1.666$ , a difference of 25 points).

As seen in Figure 7, the pivoting method is always too short in the upper limit. This failure is increasing with increasing effect sizes as expected because asymmetry is larger for the left tail of the upper end, whereas the upper ncp identified by the pivoting method has short right tail. This problem is vanishing only slowly such that for medium  $n$ ,



**Figure 8** ■ Distribution of 500,000 Cohen’s  $d$  values in three different designs. Sample sizes are 12, the true  $d$  is 0.666 and in the repeated measures design, the true correlation is 0.65.



underestimation is still visible, more so if the true effect size is large. The lower limits, on the other hand, are accurately estimated for all  $n$ . Because a CI should always have coverage of at least  $\gamma$ , underestimated upper bounds are problematic and thus, the pivoting method should be abandoned.

**Appendix D: The sampling distributions of Cohen’s  $d$  and the parameter  $\nu$  of the correction factor  $J$ .**

We illustrate the results of 500,000 simulated Cohen’s  $d$  with its theoretical distribution (full line) in Figure 8. Regarding the repeated measure design, we illustrate the noncentral  $t$  distribution with both  $2(n - 1)$  and  $n - 1$  degrees of freedom. The results show unambiguously that the value  $2(n - 1)$  must be employed in the correction factor  $J$ .

**Open practices**

📄 The *Open Material* badge was earned because supplementary material(s) are available on the [journal’s web site](#).

**Citation**

Goulet-Pelletier, J.-C., & Cousineau, D. (2018). A review of effect sizes and their confidence intervals, part I: The Cohen’s  $d$  family. *The Quantitative Methods for Psychology, 14*(4), 242–265. doi:10.20982/tqmp.14.4.p242

Copyright © 2018, Goulet-Pelletier and Cousineau. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 26/08/2018 ~ Accepted: 04/12/2018

Listing 1 follows.





**Listing 1** ■ A R function that computes Hedges'  $g$  and its confidence interval for within and between subject designs.

```
gethedgesg <-function( x1, x2, design = "between", coverage = 0.95) {  
  # mandatory arguments are x1 and x2, both a vector of data  
  
  require(psych) # for the function harmonic.mean.  
  
  # get basic descriptive statistics  
  ns <- c(length(x1), length(x2))  
  mns <- c(mean(x1), mean(x2))  
  sds <- c(sd(x1), sd(x2))  
  
  # get pairwise statistics  
  ntilde <- harmonic.mean(ns)  
  dmn <- abs(mns[2]-mns[1])  
  sdp <- sqrt( (ns[1]-1) *sds[1]^2 + (ns[2]-1)*sds[2]^2) / sqrt(ns[1]+ns[2]-2)  
  
  # compute biased Cohen's d (equation 1)  
  cohend <- dmn / sdp  
  
  # compute unbiased Hedges' g (equations 2a and 3)  
  eta <- ns[1] + ns[2] - 2  
  J <- gamma(eta/2) / (sqrt(eta/2) * gamma((eta-1)/2) )  
  hedgesg <- cohend * J  
  
  # compute noncentrality parameter (equation 5a or 5b depending on the design)  
  lambda <- if(design == "between") {  
    hedgesg * sqrt( ntilde/2)  
  } else {  
    r <- cor(x1,x2)  
    hedgesg * sqrt( ntilde/(2 * (1-r)) )  
  }  
  
  # confidence interval of the hedges g (equations 6 and 7)  
  tlow <- qt(1/2 - coverage/2, df = eta, ncp = lambda )  
  thig <- qt(1/2 + coverage/2, df = eta, ncp = lambda )  
  
  dlow <- tlow / lambda * hedgesg  
  dhig <- thig / lambda * hedgesg  
  
  # all done! display the results  
  cat("Hedges' g□□", hedgesg, "\n", coverage*100, "%□CI□□", dlow, dhig, "]\n")  
}
```