2021 ■ Vol. 17 ■ no. 1

A study of confidence intervals for Cohen's d_p in within-subject designs with new proposals

Denis Cousineau [a] and Jean-Christophe Goulet-Pelletier [a]

[a]Université d'Ottawa

Abstract ■ There exist many variants of confidence intervals for Cohen's d_p in within-subject designs. Herein, we review three past proposals (Morris, 2000; Algina & Keselman, 2003, Goulet-Pelletier & Cousineau, 2018) and examine five new ones, four of which are based on the recently discovered distribution of d_p in such design. We examine each method according to their accuracy in coverage rate (desired coverage is 95% in this study), symmetry (i. e., equal rejection rates from the left and from the right), and width of the interval. It is found that the past three proposals are pseudo confidence intervals, being too liberal under some circumstances (fortunately uncommon for the methods of Morris and Algina & Keselman). Additionally, they are not asymptotically accurate. Finally, they do not have symmetrical rejection rates on the left and on the right. Four of the five new techniques are asymptotically accurate but three of these are liberal for small samples. Finally, the relation of confidence intervals with inferential statistics testing is considered.

Keywords ■ Standardized mean difference; confidence interval; within-subject design; noncentral t distribution, noncentral Lambda distribution. **Tools** ■ R.

✉ denis.cousineau@uottawa.ca

10.20982/tqmp.17.1.p051

Acting Editor ■ Roland Pfister (Universität Würzburg)

Reviewers ■ One anonymous reviewer

Introduction

This study examines ways to obtain intervals for Cohen's d_p in within-subject designs. It is motivated by the numerous personal communications that we received following the publication of a review paper on the subject (Goulet-Pelletier & Cousineau, 2018) as well as by a follow-up article (Fitts, 2020). From these, it became apparent that there was to this day no satisfactory confidence interval for Cohen's d_p in within-subject designs (see Tothfalusi & Endrenyi, 2017; and Viechbauer, 2007, for explorations). We consequently wish to document the strengths and limitations of these past approaches. Following the recent discovery of the distribution of Cohen's d_p in within-subject designs (Cousineau, 2020b), we also explore new alternatives. We end with a central question: Is it really a confidence interval that we are looking for? There exists an alternative family of intervals that we call the *precision intervals*; we will examine these as well.

It needs to be clarified from the onset that in within-subject designs, a mean difference between two measures can be standardized in two different ways, resulting in two distinct Cohen's d that we call herein d_D and d_p. The problem is that these two standardized differences are on different scales. Therefore, they cannot be compared directly. As will be seen later, it is easy to convert a d_D into a d_p and vice-versa. However, their confidence intervals cannot be converted from one to the other. In d_D, the mean difference is standardized relative to the standard deviation of the *differences* between the scores (i.e., from the subtraction of pairs of scores). In d_p, the focus of the present text, the mean difference is standardized relative to the standard deviation of the scores. Assuming homogeneity of variances in the population, d_p estimates the standard deviation by pooling the standard deviation of the first with the second set of measurements, noted S_p, hence the name d_p.

An accurate confidence interval, with confidence level γ, should have a probability of rejecting the true population parameter of $1-\gamma$ (or at least not below $1-\gamma$). Hence, from samples to samples, a 95% confidence interval would have a non-rejection rate of at least 95%. As seen, the logic of null hypothesis testing is at the core of confidence intervals.

In what follows, we assess the non-rejection rates of three confidence interval methods that have been proposed for d_p in within-subject designs (Morris, 2000; Algina & Keselman, 2003; following Steiger & Fouladi, 1997; Goulet-Pelletier & Cousineau, 2018). We also explore one new proposal, called MAG, which mixes elements of the previous three methods (the acronym combines the first letter of each of the three methods). Finally, following the recent publication of the distribution of the Cohen's d_p in within-subject designs (Cousineau, 2020b), we explore four new proposals based on this distribution, called herein the *noncentral* t (noted t'), the *noncentral lambda* (noted Λ'), the *Pivotal of* t' and the *Adjusted* Λ' methods.

In a second Result section, we question the core assumption underlying confidence intervals. Indeed, basing a technique on null hypothesis statistical techniques is dubious in these times where estimation is favored over significance (e.g., Cumming, 2014; Amrhein, Greenland, & McShane, 2019, among numerous others). We will therefore examine an alternative type of intervals, that we call the *precision intervals*. As will be seen, they are nearly identical to confidence intervals although they aim to fulfill a different objective.

Attributes of a confidence interval method

Before assessing the confidence interval expressions using Monte Carlo simulations, we present formally the criteria that will be used to evaluate different methods. A confidence interval expression can be examined, in our opinion, from three quantitative attributes: *Accuracy*, *Symmetry* and *Width*. The first attribute of accuracy offers a way to classify the interval expression into three classes or types that we call *Exact*, *Valid* and *Pseudo*.

Accuracy. Accuracy is the difference between the actual rejection rate of the true population parameter and the desired rejection rate, noted γ. *Exact* confidence intervals are always perfectly accurate. On the other hand, *valid* confidence intervals have non-rejection rate that are never smaller than γ. This is the defining characteristic of a confidence interval (Neyman, 1934). Finally, *pseudo confidence intervals* are not truly confidence intervals because they do not respect this defining characteristic, being *liberal* in some circumstances (i. e., having a non-rejection rate of the true population parameter smaller than γ). Non-exact intervals might approach a non-rejection rate of γ when n is larger; this is called asymptotic accuracy. Thus, a non-exact confidence interval could be *asymptotically exact*. As will be seen, this is not the case for the existing confidence intervals that we tested.

Symmetry in rejection rates. Symmetry represents the balance between rejections from the left of the true population parameter and rejections from the right of the true population parameter. A rejection from the left occurs when the confidence interval of an observed statistic is above the true population value (with significant level $\alpha = (1-\gamma)/2$). We believe it to be advisable that the left rejection rate of a confidence interval is equal to its right rejection rate. When illustrated as an error bar, most readers assume that both extremities have an equal chance of occurrence. What would be their surprise if they were to discover that the lower limit is set to 0.5% and the upper limit to 4.5%? Yet, as will be seen in the results later on, this figure is actually typical of some of the methods examined.

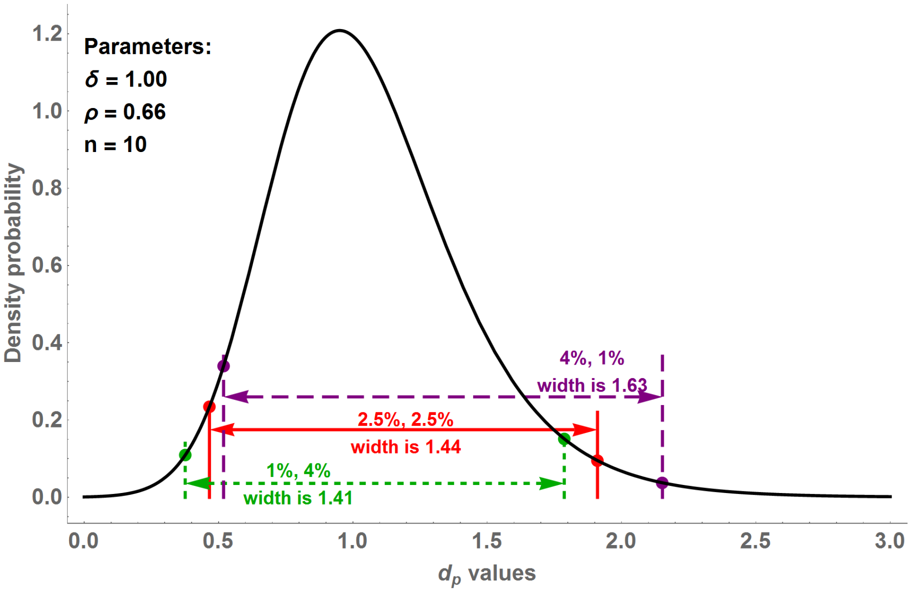
Width. An attribute of confidence intervals considered important by Neyman (1934, p. 563) was that if more than one method exists to generate a valid confidence interval, the one with the shortest width should be preferred. In the presence of symmetric rejection rates, this attribute is sensible. However, as illustrated in Figure 1, this attribute is problematic when the left and right rejection rates are allowed to be different. For a skewed distribution of effect size (we plotted a noncentral t distribution in Figure 1 but the argument goes for any skewed distribution), an easy way to obtain a shorter interval is to have unequal rejection rates. In the shortest tail (the left tail in Figure 1), a small decrease in the lower limit of the interval is accompanied by a large decrease in the upper limit of the interval where the tail is "flatter". In Figure 1, the interval width goes from 1.63 to 1.41 when the lower limit is moved left, *while keeping the exact same non-rejection rate*. Thus, this manipulation leads to an important reduction in the total width of the interval. In the presence of unequal left and right rejection rates, it is therefore difficult to weigh interval widths properly.

The confidence interval methods examined

In what follows, we examine eight methods to get a confidence interval. The first three have been published previously (Morris, 2000; Algina & Keselman, 2003; Goulet-Pelletier & Cousineau, 2018). The fourth one, introduced in this manuscript, is a morph between all three previous methods. The last four methods are all derived from the newly found distribution of d_p in within-subject designs (Cousineau, 2020b).

In Listings 0 to 8 detailed afterwards, we provide R scripts (also available on the journal's web site) which computes the confidence intervals from the methods described next given the observed statistics d_p, r, $S_{\mathbf{X}}$ and $S_{\mathbf{Y}}$ along with sample size n and confidence level γ (default 95%). It uses libraries from Genz et al. (2019), Kelley (2019), Pav (2017) and Revelle (2018).

Figure 1 ■ Illustration of allowing unequal left and right rejection rates using the quantiles of a noncentral t distribution on the interval width. Moving the interval to the left allows a considerable reduction of its total width.



Morris (2000)

This method, proposed by Morris (2000), is based on the fact that the noncentral t distribution tends to a normal distribution as the sample size tends to infinity (Becker, 1988). To accommodate the (slight) asymmetry in the noncentral t distribution, the standard deviation is artificially decreased so that shorter tails on the right is reasonably well approximated by the right tail of a normal distribution. This is achieved by estimating the variance of the population then multiplying it by the correction factor $J(n-1)$ squared (see Hedges, 1981; Becker, 1988, where this correction factor is noted c; and Goulet-Pelletier & Cousineau, 2018, for a description of this correction factor). As J is always below 1, it reduces the estimate of the population variance. Listings 0 and 1 summarize the steps to get the confidence limits with this method.

Algina and Keselman (2003)

This and the two subsequent methods are based on the noncentral t distribution with degree of freedom $(n-1)$, noted t'_{n-1}, which is the correct sampling distribution of d_p for single-group design and difference scores (Hedges, 1981; Becker, 1988) but not for within-subject designs (Cousineau, 2020b; Fitts, 2020).

The method put forth by Algina and Keselman (2003) uses the pivotal method adopted for between-group design by Steiger and Fouladi (1997; this method was explored for the first time in Clopper & Pearson, 1934). In order to estimate the lower limit up to where a plausible population δ may lie, it estimates the noncentrality parameter of a shifted distribution such that its 97.5% quantile corresponds to the observed d_p. The same is done for the upper limit. The lower and upper noncentrality parameters are subsequently scaled down to a within-subject d_p by a multiplication with $\sqrt{2(S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2-2S_{\mathbf{X}}S_{\mathbf{Y}}r)/(n(S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2))}$ where $S_{\mathbf{X}}^2$, $S_{\mathbf{Y}}^2$, $S_{\mathbf{X}}$ and $S_{\mathbf{Y}}$ are variances and standard deviations for $\mathbf{X}$ (measurements 1) and $\mathbf{Y}$ (measurements 2) and r is the Pearson correlation. Note that this formula can be simplified to

$$\begin{aligned}\sqrt{\frac{2(S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2-2S_{\mathbf{X}}S_{\mathbf{Y}}r)}{n(S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2)}} &= \sqrt{\frac{2}{n}}\times\sqrt{\frac{S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2-2S_{\mathbf{X}}S_{\mathbf{Y}}r}{S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2}}\\ &= \sqrt{\frac{2}{n}}\times\sqrt{\frac{2S_p^2(1-r_W)}{2S_p^2}}\\ &= \sqrt{\frac{2(1-r_W)}{n}}\end{aligned} \tag{1}$$

where S_p^2 is the pooled variance (equal to the mean variances in within-subject design so that $S_{\mathbf{X}}^2+S_{\mathbf{Y}}^2 = 2S_p^2$) and r_W is the rectified Pearson correlation ($r_W = r \times$ geometric.mean$(S_{\mathbf{X}}, S_{\mathbf{Y}})$/mean$(S_{\mathbf{X}}, S_{\mathbf{Y}})$; see Appendix A for details). The method is summarized in Listing 2.

Listing 0 ■ Estimating Cohen's d_p; X is assumed to be a two-column, n-line array or dataframe

```
# Needed libraries
library(mvtnorm) # for rmvnorm in generating a random sample
library(MBESS)   # for conf.limits.nct
library(psych)   # for geometric.mean
library(sadists) # for qlambdap the lambda-prime distribution

# Correction factor
J <- function(df) {
    # compute unbiasing factor; works for small or large df;
    # thanks to Robert Calin-Jageman
    exp ( lgamma(df/2) - log(sqrt(df/2)) - lgamma((df-1)/2) )
}

# Get descriptive statistics
n  <- dim(X)[1]
Mx <- mean(X[,1])
My <- mean(X[,2])
sx <- sd(X[,1])
sy <- sd(X[,2])
r  <- cor(X[,1], X[,2])

# Get pairwise statistics Delta means and pooled SD
dmn    <- Mx-My
sdp    <- sqrt((sx^2 + sy^2)/2)

# Compute biased Cohen's d
dp  <- dmn / sdp
```

Goulet-Pelletier and Cousineau (2018)

This method opted for different degrees of freedom, $2(n-1)$, instead of $(n-1)$. It first estimates the noncentrality parameter by unbiasing the observed d_p into a quantity called Hedge's g_p with $g_p = d_p \times J(2(n-1))$ and scaled with $\sqrt{n/(2(1-r))}$. Then the 2.5% and 97.5% quantiles of the noncentral t distribution are computed to obtain the interval limits. See Listing 3 for the detailed steps.

MAG method (this manuscript)

This method is introduced here for the first time. Its algorithm is given in Listing 4. It borrows elements from the first three methods. First, as with the method of Goulet-Pelletier and Cousineau (2018) above, it uses quantiles from the noncentral t distribution but uses $(n-1)$ degrees of freedom. Second, it uses the method of Morris to decrease the noncentrality parameter by multiplying it an additional time by $J(n-1)$. Third, it uses Algina and Keselman (2003) downscaling method but not the pivotal method. Because of all these similarities, it was expected to behave in a similar fashion to these previous methods. In a sense, it is a morph of all the previous techniques, whence the name MAG which stands for Morris, Algina and Goulet-Pelletier.

t' method based on the true distribution (this manuscript)

This method and the subsequent ones are all based on the newly discovered distribution of the Cohen's d_p in within-subject designs (Cousineau, 2020b) given by

$$\sqrt{\frac{n}{2(1-\rho)}}\, d_p \sim t'_{2/(1+\rho^2)(n-1)}\left(\sqrt{\frac{n}{2(1-\rho)}}\,\delta\right) \qquad (2)$$

where ρ is the population correlation between the pair of measurements. This distribution is a noncentral t distribution as in the between-group case (see Fitts, 2020; Cousineau & Goulet-Pelletier, 2020). However, the degree of freedom is fractional, based on the correlation. The term $\sqrt{n/(2(1-\rho))}$ is used to scale the observed statistic onto the theoretical distribution.

This technique is called t' as it uses the t' distribution along with fractional degrees of freedom to get confidence limits. The t' method first estimates r to estimate the de-

Listing 1 ■ Morris (2000) function with arguments d_p, r, s_x, s_y and n from Listing 0.

```
morris2000 <- function(dp, r, sx, sy, n, gamma = .95) {
    vd <- (n-1)/(n-3) * 2*(1-r)/n * (1+dp^2 * n/(2*(1-r))) - dp^2/J(n-1)^2
    vd <- vd * J(n-1)^2

    dlow <- dp + qnorm(1/2-gamma/2) * sqrt(vd)
    dhig <- dp + qnorm(1/2+gamma/2) * sqrt(vd)

    limits <- c(dlow, dhig)
    limits
}
```

Listing 2 ■ Algina and Keselman (2003) function

```
alginakeselman2003 <- function(dp, r, sx, sy, n, gamma = .95) {
    W  <- geometric.mean(c(sx^2, sy^2)) / mean(c(sx^2, sy^2))
    rW <- r * W

    tCI <- conf.limits.nct(dp * sqrt(n/(2*(1-rW))), n-1, conf.level = gamma)
    tCI.low <- tCI$Lower.Limit
    tCI.hig <- tCI$Upper.Limit

    limits <- c(tCI.low, tCI.hig) / sqrt(n/(2*(1-rW)))
    limits
}
```

gree of freedom as $2/(1+\rho^2)(n-1)$. Then, the $1/2-\gamma/2$ and $1/2+\gamma/2$ quantiles of the t' distribution are obtained and used as the confidence limits. See Listing 5.

Λ' *method based on the dual of the true distribution (this manuscript)*

As shown in Lecoutre (1999) the t' distribution is the distribution of observed d_p when the population parameter is known. However, with confidence intervals, the problem is given the other way around: we are given an observed d_p and we seek the distribution of the population parameters that could have generated that observation. Thus, what is truly needed is the dual distribution, sometimes called the predictive distribution (Poitevineau & Lecoutre, 2010). This distribution, when the observed statistics follow a t' distribution, is known and called the noncentral lambda (Λ' or Λ-prime) distribution. It requires the same degrees of freedom $2/(1+\rho^2)(n-1)$ where ρ^2 is again estimated with r^2.[1]

The interval limits are then the $1/2-\gamma/2$ and $1/2+\gamma/2$ quantiles of that Λ' distribution. See Listing 6 for details.

Pivotal of the t' *method based on the true distribution (this manuscript)*

The Λ' distribution is the distribution of the parameters compatible with a given observation. When that distribution is not implemented, its quantiles can be estimated through a search over the t' distribution in which the noncentrality parameter λ is varied until one is found that makes the observation borderline valid. This is called the pivotal method and is present in the Algina and Keselman (2003) method where the incorrect degrees of freedom are used. Here, we implemented the pivotal process using the correct degrees of freedom $2/(1+r^2)(n-1)$. See Listing 7 for details.

Because the Λ' and the pivotal of the t' are both estimating the quantiles of the Λ' distribution, both methods are expected to behave in an identical fashion.

[1] Note that in all these methods, we also tried estimating r^2 with Olkin and Pratt (1958) unbiased estimator. However, we found no substantial improvement. In R, the Olkin-Pratt estimator of r^2 is obtained with

```
library(gls) #Hankin, 2006
r2OP <- 1 -(n-3)/(n-2)*(1-r^2) * hyperg_2F1(1,1,n/2,1-r^2)
```

We also explored the estimators proposed in Kubokawa, Marchand, and Strawderman (2017).

Listing 3 ■ Goulet-Pelletier and Cousineau (2018) function. Library psych required.

```
gouletpelletiercousineau2018 <- function(dp, r, sx, sy, n, gamma = .95) {
    gp <- dp * J( 2 * (n-1) )

    dlow = qt(1/2-gamma/2, df = 2*(n-1), ncp = gp * sqrt(n/(2*(1-r))) )
    dhig = qt(1/2+gamma/2, df = 2*(n-1), ncp = gp * sqrt(n/(2*(1-r))) )

    limits <- c(dlow, dhig) / sqrt(n/(2*(1-r)))
    limits
}
```

Listing 4 ■ MAG function (this manuscript). Library psych required

```
MAG <- function(dp, r, sx, sy, n, gamma = .95) {
    W  <- geometric.mean(c(sx^2, sy^2)) / mean(c(sx^2, sy^2))
    rW <- r * W

    #compute unbiased noncentrality parameter
    lambda <- dp * J(n-1)^2 * sqrt(n/(2*(1-rW)))

    #quantile of the noncentral t distribution
    dlow = qt(1/2-gamma/2, df = n-1, ncp = lambda )
    dhig = qt(1/2+gamma/2, df = n-1, ncp = lambda )

    limits <- c(dlow, dhig) / sqrt(n/(2*(1-rW)))
    limits
}
```

Adjusted Λ' method based on the converse of the true distribution (this manuscript)

The two previous methods are exact in between-group designs (where the correlation parameter ρ is non-existent; Cousineau & Goulet-Pelletier, 2020). However, in within-subject designs, the ρ parameter needs to be estimated from the data to get the correct degrees of freedom. Yet, pivotal techniques require that the distribution's degree of freedom be known. As it has to be estimated, the above two methods are likely to be inexact for small samples (it remains to be seen whether they are conservative or liberal).

We therefore propose an additional method in which the noncentrality parameter is adjusted. This parameter should be an observed d_p measure. However, its multiplication with the scaling parameter $1/2(1-r_W)$ turns d_p into a d_D (see Appendix A). Thus, the noncentrality parameter is not from the same distribution and is biased in a different way. As a solution, we propose to first unbias the observed d_p using the correction factor $J(n-1)$, then cancel this correction by dividing the obtained quantiles with $J(2/(1+r^2)(n-1))$. See the last listing, Listing 8.

Simulations

To evaluate the eight methods, we ran extensive Monte Carlo simulations. To avoid generalizing from a limited set of simulations, we ran 250 different scenarios by crossing (i) five magnitudes of true effect sizes, ranging from a null effect size to a huge effect size (δ = 0 to 1.333 by steps of 0.333); (ii) five sample sizes, from very small samples to moderate samples in addition to large samples (n = 10, 15, 20, 25 and 100); (iii) ten levels of correlations (ρ = -0.90 to +0.90 by steps of 0.20). We included negative correlations and extreme correlations even though these are probably implausible in the psychological sciences (in a review, Goulet and Cousineau, 2019, found the correlation of repeated measures in simple cognitive tasks to be around 0.2). In total, this represents 250 different scenarios ($5 \times 5 \times 10$), a much wider array of scenarios than in some previous studies (Morris, 2000, explored 27 scenarios; Algina & Keselman, 2003, reported 12 scenarios; but see Viechbauer, 2007, with 595 scenarios). Also, for stable results, we repeated each simulation 100,000 times, five times more than in some previous studies (but see Fitts, 2020; Viechbauer, 2007, who used 100, 000 replications per

Listing 5 ■ t' function based on the true distribution (this manuscript). Library psych required.

```
tprime <- function(dp, r, sx, sy, n, gamma = .95) {
    W  <- geometric.mean(c(sx^2, sy^2)) / mean(c(sx^2, sy^2))
    rW <- r * W

    #compute unbiased noncentrality parameter
    lambda <- dp * J(2/(1+r^2)*(n-1)) * sqrt(n/(2*(1-rW)))

    #quantile of the noncentral t distribution
    dlow = qt(1/2-gamma/2, df = 2/(1+r^2)*(n-1), ncp = lambda )
    dhig = qt(1/2+gamma/2, df = 2/(1+r^2)*(n-1), ncp = lambda )

    limits <- c(dlow, dhig) / sqrt(n/(2*(1-rW)))
    limits
}
```

Listing 6 ■ Λ' function based on the dual of the true distribution (this manuscript). Libraries psych and sadists required.

```
lambdaprime <- function(dp, r, sx, sy, n, gamma = .95) {
    W  <- geometric.mean(c(sx^2, sy^2)) / mean(c(sx^2, sy^2))
    rW <- r * W

    lambda <- dp * sqrt(n/(2*(1-rW)))

    #quantile of the noncentral t distribution
    dlow = qlambdap(1/2-gamma/2, df = 2/(1+r^2)*(n-1), t = lambda )
    dhig = qlambdap(1/2+gamma/2, df = 2/(1+r^2)*(n-1), t = lambda )

    limits <- c(dlow, dhig) / sqrt(n/(2*(1-rW)))
    limits
}
```

scenario as well). With this number of simulations, the standard error of the estimated rate of rejection was in some scenarios up to 0.075%, so that the current simulations are considered precise to twice this figure, ± 0.15%. Thus, the rejection rates are reported as percentage with one decimal (Cousineau, 2020a).

In a given simulation, we generated a bivariate sample of size n from a binormal distribution with a mean of μ (arbitrarily set to 0) increased by $\Delta/2$ for the first measure and decreased by the same quantity for the second measure. Thus, in the population, the difference in means is Δ. The variance, σ^2, is arbitrarily set to 1 so that the population standardized effect size is $\delta = \Delta/\sigma$. Finally, from the correlation ρ, the population variance-covariance matrix was

$$\Sigma = \begin{bmatrix} \sigma^2, \rho\sigma^2 \\ \rho\sigma^2, \sigma^2 \end{bmatrix} \qquad (3)$$

. None of these parameters (μ, δ, σ and ρ) are assumed known in the subsequent procedures.

For each simulation, we estimated d_p as the difference in means onto the pooled standard deviation and applied a confidence interval method in order to get the lower and upper bounds (we report only 95% confidence intervals). We recorded whether the true δ was included within the limits (a non-rejection), located below the lower bound (a left rejection) or located above the upper bound (a right rejection). The non-rejection, left and right rejection rates over the 100,000 reproductions are used for plots and analyses. This process was repeated for each method.

The random number generation was performed with Mathematica version 10.0 built-in MultinormalDistribution function (the default algorithm has a cycle length above 260; Tomassini, Sipper, & Perrenoud, 2000). Most analyses were performed with

Listing 7 ■ Pivotal function of the true distribution (this manuscript). Libraries `psych` and `MBESS` required.

```
pivotaltprime <- function(dp, r, sx, sy, n, gamma = .95) {
    W  <- geometric.mean(c(sx^2, sy^2)) / mean(c(sx^2, sy^2))
    rW <- r * W

    tCI <- conf.limits.nct(dp * sqrt(n/(2*(1-rW))), 2/(1+r^2)*(n-1),
                     conf.level = gamma)
    tCI.low <- tCI$Lower.Limit
    tCI.hig <- tCI$Upper.Limit

    limits <- c(tCI.low, tCI.hig) / sqrt(n/(2*(1-rW)))
    limits
}
```

Listing 8 ■ Adjusted Λ' function (this manuscript). Libraries `psych` and `sadists` required.

```
adjustedlambdaprime <- function(dp, r, sx, sy, n, gamma = .95) {
    W  <- geometric.mean(c(sx^2, sy^2)) / mean(c(sx^2, sy^2))
    rW <- r * W

    lambda <- dp * J(n-1) * sqrt(n/(2*(1-rW)))

    #quantile of the noncentral t distribution
    dlow = qlambdap(1/2-gamma/2, df = 2/(1+r^2)*(n-1), t = lambda )
    dhig = qlambdap(1/2+gamma/2, df = 2/(1+r^2)*(n-1), t = lambda )

    limits <- c(dlow, dhig) / sqrt(n/(2*(1-rW))) / J( 2/(1+r^2)*(n-1) )
    limits
}
```

Mathematica built-in functions except the following three for increased computational speed: the noncentral t cumulative distribution function was compiled from C++ source (algorithm asa243; Lenth, 1989; and its translation in C++ Burkhardt, 2020, this is the code used in R's equivalent function `pt`); two binary searches were programmed in C++ and compiled to get the noncentral t quantile function and the noncentral t inversion noncentrality parameter). Finally, the quantiles from the noncentral Λ distribution were obtained from R's `sadists` library (Pav, 2017). Whenever an estimate of the sample correlation is needed, we used the Pearson correlation. All code is available on OSF at osf.io/nwxsb.

Results

In a first subsection, we examine the confidence intervals estimator from the perspective of accuracy and symmetry of the rejection rates. In a second subsection, we concentrate on the bounds and the width of the methods.

Part I: Rejection rates

The results are displayed *in extenso* in Figures B.1 to B.8. The top panels of these figures show the proportion of nonrejection of the true δ as a function of the population correlation ρ (horizontal axis) separately for each sample size n (columns) and for each population effect size δ (rows). The bottom panel report the left (square) and right (triangle) rejection rates in the same format.

Types of confidence intervals

The most striking result is that the Goulet-Pelletier and Cousineau (2018) method is far too liberal in almost all scenarios examined (there are only four exceptions over the 250 scenarios). Thus, this method is not computing a confidence interval and should not be used as such. We will later highlight a strong point of this method but will not mention this method anymore in this Results section. Also, the Λ' and the Pivotal of t' methods are almost identical in all aspects, as expected, and their results will be presented jointly.

Table 1 ■ Statistics with regards to rejection rates across the 250 scenarios

	# invalid	Non-rejection rates			Rejection rates		Ratio right:left	
Method	(out of 250)	Average	Abs. dev.	Asymptotic	Left	Right	Average	SD
Morris (2000)	7	95.9%	0.9%	95.6%	1.6%	2.5%	2.47:1	1.56
Algina & Keselman (2003)	4	95.4%	0.5%	95.5%	1.8%	2.8%	1.81:1	0.44
Goulet-Pelletier & Cousineau (2018)	246	93.4%	1.6%	94.3%	3.9%	2.7%	0.74:1	0.07
MAG	0	95.9%	0.9%	95.6%	2.0%	2.1%	1.08:1	0.05
t'	223	94.2%	0.8%	94.9%	3.4%	2.4%	0.72:1	0.07
Λ'	198	94.6%	0.4%	94.9%	2.5%	2.9%	1.21:1	0.11
Pivotal of t'	200	94.6%	0.5%	94.9%	2.5%	2.9%	1.21:1	0.11
Adjusted Λ'	0	95.3%	0.3%	95.0%	2.1%	2.6%	1.37:1	0.19

Note. Note: SD is the standard deviation in the ratios across the 250 scenarios; Abs. dev. is the average absolute deviation to 95%.

Of the 250 scenarios, we located a few scenarios where the Morris (2000) and the Algina and Keselman (2003) methods returned a proportion of non-rejection smaller than the nominally tested 95% (more precisely, smaller than 94.85%, as we allowed a ± 0.15% random fluctuation in the simulation results). They are highlighted with a red arrow in Figures B.1 and B.2. This occurred seven times for the Morris (2000) method and three times for the Algina and Keselman (2003) method. In all these scenarios, the correlation was 0.7 or 0.9. We reproduced once more these 10 scenarios with a million simulations to check the robustness of these results; all were confirmed. Neyman (1934, p. 562-563) gave a single defining property for confidence intervals: that the proportion of non-rejection of the true population parameter be never less than γ (95% for a 95% confidence interval). Thus, although these occurred in implausible scenarios with regards to human behavior (correlation of .7 or above), it means *stricto sensus* that these methods are not valid confidence intervals.

The methods t', Λ' and Pivotal of t' were too liberal in a vast majority of the scenarios considered (more or less 200 times out of 250 scenarios). They are therefore pseudo-confidence intervals.

Finally, for the MAG method and the Adjusted Λ' method, we observed no rates below 94.96% in the scenarios considered (which is possibly not different from the nominal γ, owing to sampling error). We also tested these two methods in an additional, extreme, scenario (ρ = .99, δ = 1.00 and $n = 100$) and found non-rejection rates of 95.07% and 95.04% respectively over 100,000 simulation. Thus, MAG and Adjusted Λ' might be the only valid confidence interval methods for the Cohen's d_p in within-subject design known so far.

Table 1, column 2, provides the number of scenarios with non-rejection rates below 94.85%.

Accuracy

An examination of the non-rejection rates in Figures B.1, B.2, and B.4 shows that Morris (2000), Algina and Keselman (2003) and MAG methods all behave similarly: all three have near exact non-rejection rates when the population δ is null. Further, all three become conservative when δ is large, this last effect being lightly modulated by correlation. Looking from left to right in Figures B.1, B.2 and B.4, we see however that sample size has only a marginal effect. All three techniques are not more accurate as sample sizes increase. When sample size is restricted to $n = 100$ to get a glimpse of their asymptotic performance (listed in Table 1, fifth column), the non-rejection rates are 95.6%, 95.5% and 95.6% for Morris (2000), Algina and Keselman (2003), and MAG methods respectively. These asymptotic non-rejection rates are almost identical to the non-rejection rates across all sample sizes given in Table 1, column 3, confirming that they are insensitive to sample size. Thus, these three methods are not asymptotically exact.

Regarding the last four methods, they all tend to 95% when $n = 100$. Thus, these four techniques seem to be asymptotically exact. For the t', Λ' and Pivotal of t' methods, this was expected as all three tend to the same normal distribution. However, these methods tend towards 95% from below, making them invalid for smaller samples. By contrast, the Adjusted Λ' method tends towards 95% from above.

Table 2 provides the minimum and maximum non-rejection rates observed across 50 scenarios broken down by the sample sizes examined. As seen, the t' method is within 1% of the nominal γ when $n \geq 25$; the last three are within 1% of the nominal γ when $n \geq 15$.

Table 2 ■ Minimum and maximum rejection rates for each sample sizes across the 10 correlation scenarios and the 5 effect size scenarios

	$n=10$		$n=15$		$n=20$		$n=25$		$n=100$	
Method	Min	Max	Min	Max	Min	Max	Min	Max	Min	Max
Morris (2000)	94.9%	97.9%	94.6%	97.3%	94.6%	97.1%	94.7%	97.1%	94.9%	96.9%
Algina & Keselman (2003)	94.4%	96.3%	94.6%	96.5%	94.9%	96.6%	94.9%	96.7%	94.9%	96.9%
Goulet-Pelletier & Cousineau (2018)	87.5%	93.4%	87.7%	94.0%	87.8%	94.3%	87.9%	94.4%	88.3%	94.9%
MAG	95.5%	97.5%	95.3%	97.3%	95.3%	97.2%	95.2%	97.3%	95.0%	97.0%
t'	92.3%	94.7%	93.3%	94.8%	93.8%	94.9%	93.9%	94.8%	94.7%	95.2%
Λ'	93.6%	94.8%	94.1%	94.9%	94.3%	95.0%	94.4%	95.0%	94.8%	95.1%
Pivotal of t'	93.7%	94.9%	94.1%	94.8%	94.3%	94.9%	94.3%	95.0%	94.8%	95.1%
Adjusted Λ'	95.1%	96.4%	95.0%	95.9%	95.0%	95.7%	95.0%	95.5%	95.0%	95.2%

Note. Note: The t' method is accurate to within 1% when $n \geq 25$; the last three methods are accurate to within 1% when $n \geq 15$. These cases are highlighted in gray in the table.

Symmetry

The lower panels of the Figures B.1 to B.8 show the left and the right rejection rates. As seen, for most methods, there are wild differences between the left and the right rejection rates. This is mostly apparent when the population correlation is 0.1 or 0.3. This is particularly critical as it turns out to be the plausible range of correlations observed in simple cognitive tasks performed by human participants (Goulet & Cousineau, 2019). Table 1, sixth and seventh columns, indicates the observed left and right rejection rates averaged across the 250 scenarios.

The eighth column also shows the average ratio of the left rejection rate onto the right rejection rate. If both rates were roughly equal all the time, that ratio would be near 1:1. However, this is not the case for Morris (2000) and Algina and Keselman (2003) methods, with averaged ratios of 2.5 and 1.8 to 1, respectively. These average ratios hide more extreme differences. For example, for a very large effect size ($\delta = 1.333$) and a small sample size ($n = 15$), the Morris (2000) ratio is about 4 to 1 in the plausible range of correlations (0.1 and 0.3): The left rejection rate is less than 1% whereas the right rejection rate is above 3.5%. The Algina and Keselman (2003) method is barely less extreme.

As seen, the MAG method shows much smaller differences between the left and the right rejection rates (with an average ratio of 1.08:1). Still, the right rejection rates are influenced by the amount of correlation, showing a depression for correlations of about 0.5 to 0.9. The last column of Table 1 indicates the standard deviation in the ratios. It shows that this 1.08:1 ratio is closer to be a constant (standard deviation for MAG almost nine times smaller than for the Algina and Keselman (2003) method and 17 times more stable than the Morris (2000) method, respectively).

All the other methods had left and right rejection rates moderately equal with mean ratio ranging between 0.72:1 and 1.37:1 with standard deviations at least 2.5 times smaller than the Algina and Keselman (2003) method.

Part II: Interval boundaries and interval widths

In this second Results section, we examine the bound positions. From these, it will be possible to examine the width. We can also compare the estimated positions with the true positions. Indeed, in all the scenarios, the true parameters of the simulations are known. Thus, we get access to two distributions. First, the sampling distribution of d_p is a noncentral t distribution with degree of freedom $2/(1+\rho^2)(n-1)$ and noncentrality parameter $\delta\sqrt{n/(2(1-\rho))}$ (Cousineau, 2020b). From it, the 2.5% and 97.5% quantiles of the observable d_p can be determined. Second, following Lecoutre (1999, 2007; see also Poitevineau & Lecoutre, 2010), the predictive distribution is known as well. This distribution describes the parameters that could have generated a given observation worth d_p. It is a Λ' distribution with the same degree of freedom as above and with noncentrality parameter $d_p\sqrt{n/(2(1-\rho))}$. Consequently, the theoretical 2.5% and 97.5% limits of all possible δ, that is, the 95% confidence interval, can be determined. This distribution was examined in the context of between-subject designs d_p in Cousineau and Goulet-Pelletier (2020).

In the simulations, the standard error of the average lower bounds within a scenario was 0.001 on average and the standard error of the upper bound was likewise 0.001. The results reported herein are therefore accurate to twice this figure, ± 0.002 (Cousineau, 2020a). The widths being the difference between the two bounds are accurate to ± 0.004.

Table 3 ■ Statistics with regards to bound positions and width across the 250 scenarios

	Left bound positions			Right bound positions			Mean
Method	Mean	Error w/r to t'	Error w/r to Λ'	Mean	Error w/r to t'	Error w/r to Λ'	width
Morris (2000)	0.020	-0.050 (-3.9%)	-0.025 (-2.1%)	1.353	-0.023 (-2.0%)	0.075 (6.1%)	1.333
Algina & Keselman (2003)	0.036	-0.034 (-2.6%)	-0.010 (-0.8%)	1.324	-0.052 (-4.2%)	0.046 (3.8%)	1.287
Goulet-Pelletier & Cousineau (2018)	0.078	0.008 (0.6%)	0.033 (2.6%)	1.351	-0.025 (-2.2%)	0.073 (5.9%)	1.273
MAG	0.017	-0.053 (-4.1%)	-0.029 (-2.3%)	1.404	0.028 (1.9%)	0.126 (10.2%)	1.387
t'	0.067	-0.004 (-0.3%)	0.022 (1.7%)	1.382	0.006 (0.2%)	0.104 (8.5%)	1.315
Λ'	0.058	-0.013 (-0.9%)	0.013 (1.0 %)	1.306	-0.070 (-5.6%)	0.028 (2.3%)	1.248
Pivotal of t'	0.058	-0.013 (-0.8%)	0.015 (1.2 %)	1.306	-0.070 (-5.8%)	0.026 (2.1%)	1.244
Adjusted Λ'	0.029	-0.041 (-3.2%)	-0.016 (-1.3 %)	1.313	-0.066 (-5.2%)	0.035 (2.7%)	1.284
Observed quantiles	0.0706			1.376			1.305
Theoretical t' quantiles	0.0704			1.3793			1.3089
Theoretical Λ' quantiles	0.0455			1.2777			1.2322

Note. Note: The percent of errors are relative to the width of the interval. Negative errors denote underestimation. Left is the 2.5% limit and right is the 97.5% limit. The theoretical t' quantiles are the bound positions of the 2.5% and 97.5% sampling distribution of d_p based on the true t' distribution averaged over the 250 scenarios. The theoretical Λ' quantiles are the bound positions of the 2.5% and 97.5% predictive distribution of δ based on the true Λ' distribution averaged over the 250 scenarios.

Left and right bounds

Table 3 shows the mean lower and upper bounds of the intervals obtained from each method, averaged across the 250 scenarios. It also shows, three lines before the end, the actual limits that contained the most central 95% of the 100,000 simulated d_p, again averaged across the 250 scenarios.

The actual limits match well the theoretical quantiles of the sampling distribution given on the line before last, which is not surprising. The t' method estimated these limits with less than 0.3% of error, even though it estimated these using an estimated r parameter rather than the true ρ parameter. Note that t' underestimated the lower bound and overestimated the upper bound so that it is a conservative estimate of the sampling distribution's quantiles.

The second method most apt to estimate the lower quantiles was the Goulet-Pelletier and Cousineau (2018) method. It is more than 5 times more precise than the Algina and Keselman (2003) and Morris (2000) methods. Regarding the upper quantiles, the Morris (2000), Goulet-Pelletier and Cousineau (2018), and MAG methods are similarly moderately precise. Thus, we have the paradoxical situation that the worst technique to determine a confidence interval, Goulet-Pelletier and Cousineau (2018), is the second best one to determine the range of possible values that the d_p statistic can take. These bounds depict sample-to-sample variations in the observed d_p. When focusing on estimation rather than significance (Cumming, 2014), what is desired is a range describing the plausible values of a statistic assuming that the one observed is representative of the population parameter. It therefore affords a measure of *precision*. Thus, we call intervals based on the sampling distribution *precision intervals*, as opposed to confidence intervals which are focused on significance.

Focusing on the theoretical quantiles of the predictive distribution (last line of Table 3), four techniques are reasonably good at estimating these bounds on average, the Algina and Keselman (2003; on the left side more than on the right side), the Λ' and Pivotal of t', and the Adjusted Λ' methods with errors of estimation of about 3.5% (considering both left and right errors of estimation). These bounds are the true confidence interval bounds. Thus, depending on what type of interval is sought, the best methods are not the same.

Interval width

Table 3 also gives the mean width across the 250 scenarios for each method. Λ' and Pivotal of t' both have the narrowest intervals, quite similar to the theoretical width predicted from the predictive distribution (last line of Table

3). Although the width is adequate, these two techniques have their intervals shifted up, overestimating both lower and upper bounds by about 1% and 2% respectively. Algina and Keselman (2003), Goulet-Pelletier and Cousineau (2018), and the Adjusted Λ' have the second best widths. Algina and Keselman's (2003) width is smaller owing to its right-to-left ratio much different from 1. Thus, its lack of symmetry benefited this method, as anticipated (see Figure 1). This shows that narrow width is a conflictual attribute relative to symmetry. Width is a fair attribute to consider only when comparing methods with similar symmetry.

Summary

First thing to note is that both Λ' and Pivotal of t' returned near identical results with all the indicators considered herein (always within the precision of the results ± 0.15% for the rates and ± 0.001 for the bound positions). Thus, these techniques are indeed the same, as shown by Lecoutre (1999). These two techniques return exact confidence intervals in between-subject designs (Cousineau & Goulet-Pelletier, 2020).

From MAG, the morph technique borrowing elements from Morris (2000), Algina and Keselman (2003) and Goulet-Pelletier and Cousineau (2018), we learn that modeling the symmetry in the tails is important to obtain valid non-rejection rates (i. e., above γ). Morris's (2000) method, using a symmetrical distribution is sometimes invalid and has the most important imbalance regarding left and right rejection rates. Further, we learn that using $2(n-1)$ as degrees of freedom was a mistake when the purpose is to get a confidence interval but was a reasonable recommendation when the purpose is to get a precision interval (Goulet-Pelletier & Cousineau, 2018).

In this text, we studied eight methods to get intervals. Only two seemed to be valid methods, MAG and Adjusted Λ'. We based our assessment on three attributes: Accuracy, symmetry in left and right rejection rates, and to a lesser extent, width. The results are summarized in Table 4. As seen, none of the methods is exact. The four new proposals were asymptotically exact confidence intervals but three were pseudo for smaller sample sizes, having liberal non-rejection rates.

The Adjusted Λ' method showed good properties as a confidence interval method (asymptotically accurate, symmetrical left and right rejection rates). It was valid (liberal or exact) in all the 250 scenarios examined. However, its two adjustments are weakly justified, so that it is an *ad hoc* method (we are still searching for a formally justified method). On the other hand, the t' method showed unsurpassed performance as a precision interval.

We mention before concluding that the expression "coverage level" often attached to confidence intervals is ambiguous (and we avoided it in this text, preferring confidence level). It could designate the non-rejection rate of the true population parameter. However, its meaning is literally "how well it covers the possible results". Thus, this expression is actually referring to boundary positions. It would actually describe very well the level γ of a precision interval. As was seen by our results, the two meanings are not interchangeable. Thus, great care should be exercised when using this expression.

General Discussion

What should we do with d_D?

Herein, we gave no attention to the second version of the Cohen's d, the d_D. This choice was motivated by two principles. (i) Effect size estimates should be universal and, in particular, independent from the experimental design so that they can be compared irrespective of how they were estimated. (ii) Intervals are used to describe the precision of the study; thus, interval estimation should be customized to match the experimental design, not the effect size estimation. It explains why effect sizes are independent from sample size, but intervals are adjusted using sample size. Likewise, effect sizes should be independent of the procedure used to get the sample (stratified, cluster, random, etc.) but intervals should be adjusted by the sampling procedure (Cousineau, 2017; Cousineau & Laurencelle, 2015).

Because d_D is only defined in within-subject design, it is dependent on the design and thus should be avoided. There is a non-null risk that both measures be confused and compared inadequately. Because within-subject designs generally afford more power and more precise estimations, these designs should return shorter intervals, which is the case with d_p as soon as correlation is positive (Cousineau, 2019). The fact that there exists an exact confidence interval for d_D is an important practical advantage. However, we believe that it does not outweigh the importance of the above two theoretical principles.

Two sorts of intervals

The results showed that interval widths can be examined with respect to two different theoretical perspectives, the sampling distribution or the predictive distribution. The second, underlying the confidence interval created by Neyman (1934), is defined under logic similar to null-hypothesis statistical testing (NHST). It seeks intervals which would contain the true population parameter a certain proportion of times. A crude distinction between confidence interval and NHST is that NHST seeks a region of rejection centered on the null hypothesis whereas confidence interval seeks a region of rejection centered on the

Table 4 ■ Attributes of a confidence interval and results with regards to five methods.

		Attributes			
	Type	Accurate?	Asympt. accurate?	Symmetry?	Short width?
Morris (2000)	Pseudo †	NO	NO	NO	NO
Algina and Keselman (2003)	Pseudo †	yes	NO	NO	yes ‡
Goulet-Pelletier & Cousineau (2018)	Pseudo	NO	NO	yes	yes
MAG	VALID	NO	NO	yes	NO
t'	Pseudo	NO	yes *	yes	NO
Λ'	Pseudo	yes	yes **	yes	yes
Pivotal of t'	Pseudo	yes	yes **	yes	yes
Adjusted Λ'	VALID	yes	yes **	yes	yes

Note. Note:
*: Accurate to within 1% when $n \geq 25$;
**: Accurate to within 1% when $n \geq 15$.
†: These methods are invalid in situations where ρ is close to 1. This is little plausible in the psychological sciences, but might be a plausible situation in other disciplines.
‡ : However, the left rejection rates are much smaller than the right rejection rates, favoring shorter intervals.

observed statistics. In both cases, the determination of the zone of non-rejection is defined through a critical p value.

Basing scientific research on statistical inference and on critical p values has been questioned in the recent decades. First, Cumming, with his famous *dance of the p value*, argued convincingly that the p value is a very unstable sample statistic which varies wildly across samples. Second, many ascribed the replication crisis to a blind reliance on thresholds to which are attached decisions to accept (sic) or reject a theory. Amrhein et al. (2019) argued that p-value thresholds should be abandoned and that confidence intervals should be renamed *compatibility intervals* to emphasis the fact that inferences must reflect a continuum of possibilities. Third, others have argued that a major problem with p values is that they are poorly understood (e.g. Gigerenzer, 2004; Haller & Krauss, 2002).

An alternative approach is to estimate, from the observed statistics, an interval which is likely to contain other, future, results. This effectively replaces a problem of inference with a problem of estimation: can we estimate the quantiles which delimit all the possible results compatible with a population inferred from the observed statistics and the experimental design? Analogous to confidence intervals, we set a desired coverage level for the interval, for example, a coverage of 95%, with (of course) equal left and right rejection rates.

We call such approach to interval estimation a precision interval. It is based on the assumption that the observed statistic is maximally representative of the population under scrutiny. Precision intervals are to estimation what confidence intervals are to inference.

The simulations showed that the t' method estimates *precision intervals* very precisely. One advantage of precision intervals is that they are unique: Their widths do not have to be the shortest; they must only match the statistic's spread. With respect to Cohen's d_p, the precision interval obtained from the t' method is exact.

Despite their conceptual differences, the boundary positions obtained from confidence intervals and from precision intervals are rather similar. For example, in the situation where $\delta = 1$, $\rho = .3$ and samples of size 25, the confidence interval of the theoretical predictive distribution puts the limit at 0.468 and 1.504 whereas the theoretical sampling distribution puts them at 0.526 and 1.574. We have to examine the second decimal to see differences between the methods. A recent examination of measurement precision suggests that psychology experiments do not have this amount of precision (Cousineau, 2020a). Consequently, the difference between precision intervals and confidence intervals is immaterial in practice.

Symmetry vs. width

Neyman, who laid down the theory of confidence interval, favored minimal width.[2] We argue that comparing methods using width makes sense only if all have the same left and right rejection rates. When the rejection rates are unequal, how can we evaluate their width fully?

Instead, we believe that Neyman should have added the attribute of equal left and right rejection rates. We consider desirable that rejections should occur equally

[2] Neyman (1934) mentioned another attribute not considered herein (p. 563): that the confidence interval be based on a table or an easily computed function. Whereas in the 1930, a room filled with computers working day and night estimating the noncentral t distribution, the hypergeometric function or the beta incomplete function was a concern; this concern is receding rapidly with electronic computers and more efficient algorithms.

frequently for too-small estimates than for too-large estimates. Imagine a manuscript in which the researcher performs a t test with an alpha level of 5% but decides that this alpha is split 0.5% on the left and 4.5% on the right. A reviewer reading that would definitely raise a red flag, and if published, the results would raise suspicions. This is actually not illegal and no written rule forbids this unequal division of alpha; however, the general acceptance is that rejection rates should be evenly distributed on either side. Similarly, in the psychological sciences, we are generally interested in testing effects against null effect. Thus, an interval will more probably be used to check its lower limit (actually, the limit pointing towards zero, which would be the lower limit when the effect is positive). As such, any limits that are not evenly distributed unknowingly transmit a biased picture of the result. If the lower limit has, say, a rejection rate of 0.5% (instead of the expected 2.5% for a 95% confidence interval), then this limit is actually far more conservative than anticipated. It would correspond to the limit of a 99% confidence interval whose limits are evenly divided between extremities.

Conclusion

What we retain from this extensive study of confidence intervals is that we still have not found an exact confidence interval method for the Cohen's d_p in within-subject design. If validity is essential, then MAG or Adjusted Λ' are two adequate methods; if asymptotic accuracy is desired, then the t', the Λ' and the Adjusted Λ' methods are adequate. In contrast, for between-subject designs, exact methods are known, the Λ' (Lecoutre, 1999) and the Pivotal of t' (Steiger & Fouladi, 1997; explored in Cousineau & Goulet-Pelletier, 2020). As argued by Neyman (1941), a method whose nonrejection rates would be exactly 95% is in principle possible for continuous distributions. Thus, additional work may improve or replace the current methods. As the distribution of d_p was published only recently, new proposals may be on their way. We are excited to see what other proposals will emerge in the upcoming years.

Authors' note

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*, 537–553. doi:10.1177/0013164403256358

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*, 305–307. doi:10.1080/00031305.2018.1543137

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

Burkhardt, J. (2020). Asa243.c and asa243.h. Retrieved, from https://people.sc.fsu.edu/asa243/asa243.html

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. doi:10.1093/biomet/26.4.404

Cousineau, D. (2017). Varieties of confidence intervals. *Advances in Cognitive Psychology*, *13*, 140–155. doi:10.5709/acp-0214-z

Cousineau, D. (2019). Correlation-adjusted standard errors and confidence intervals for within-subject designs: A simple multiplicative approach. *The Quantitative Methods for Psychology*, *15*, 226–241. doi:10.20982/tqmp.15.3.p226

Cousineau, D. (2020a). How many decimals? Rounding descriptive and inferential statistics based on measurement precision. *Journal of Mathematical Psychology*, *97*, 102362–102370. doi:10.1016/j.jmp.2020.102362

Cousineau, D. (2020b). The distribution of Cohen's dp in within-subject designs with a demonstration. *The Quantitative Methods for Psychology*, *16*(4), 418–421. doi:10.20982/tqmp.16.4.p418

Cousineau, D., & Goulet-Pelletier, J.-C. (2020). A review of five techniques to derive confidence intervals with a special attention to the Cohen's dp in the between-group design. *PsyArXiv*, *2597*, 1–44. doi:10.31234/osf.io/s2597

Cousineau, D., & Laurencelle, L. (2015). A correction factor for the impact of cluster randomized sampling and its applications. *Psychological Methods*, *21*, 121–135. doi:10.1037/met0000055

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

Fitts, D. (2020). Commentary on "a review of effect sizes and their confidence intervals, part I: The Cohen's d family": The degrees of freedom for paired samples designs. *The Quantitative Methods for Psychology*, *16*(4), 250–261. doi:10.20982/tqmp.16.4.p250

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2019). Mvtnorm: Multivariate normal and t distributions [R package] (Version 1.0.11). Retrieved from http://CRAN.R-project.org/package=mvtnorm

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*, 587–606.

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*, 242–265. doi:10.20982/tqmp.14.4.p242

Goulet, M.-A., & Cousineau, D. (2019). The power of replicated measures to increase statistical power. *Advances in Methods and Practices in Psychological Sciences*, *online*, 1–15. doi:10.1177/2515245919849434

Haller, H., & Krauss, S. (2002). Misinterpretations of signification: A problem students share with their teachers? *Methods of Psychological Research Online*, *7*, 1–20.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128. doi:10.2307/1164588

Kelley, K. (2019). Mbess: The mbess r package [R package] (Version 4.6.0). Retrieved from https://CRAN.R-project.org/package=MBESS

Kendall, M. G., & Stuart, A. (1979). *The advanced theory of statistics*. New York: Macmillan.

Kubokawa, T., Marchand, É., & Strawderman, W. E. (2017). A unified approach to estimation of noncentrality parameters, the nultiple correlation coefficient, and mixture models. *Mathematical Methods of Statistics*, *26*, 134–148. doi:10.3103/S106653071702003X

Lecoutre, B. (1999). Two useful distributions for Bayesian predictive procedures under normal models. *Journal of Statistical Planning and Inference*, *79*, 93–105. doi:10.1016/S0378-3758(98)00231-6

Lecoutre, B. (2007). Another look at confidence intervals from the noncentral t distribution. *Journal of Modern Applied Statistical Methods*, *6*, 107–116. doi:10.22237/jmasm/1177992600

Lenth, R. (1989). Algorithm as 243: Cumulative distribution function of the non-central t distribution. *Applied Statistics*, *38*, 185–189.

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*, 17–29. doi:10.1348/000711000159150

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society A*, *97*, 558–625. doi:10.2307/2342192

Neyman, J. (1941). Fiducial arguments and the theory of confidence intervals. *Biometrika*, *21*, 128–150. doi:10.1093/biomet/32.2.128

Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *The Annals of Mathematical Statistics*, *29*, 201–211. doi:10.1214/aoms/1177706717

Pav, S. E. (2017). Sadists: Some additional distributions [R package] (Version 0.99). Retrieved from https://github.com/shabbychef/sadists

Poitevineau, J., & Lecoutre, B. (2010). Implementing Bayesian predictive procedures: The K-prime and K-square distributions. *Computational Statistics and Data Analysis*, *54*, 724–731. doi:10.1016/j.csda.2008.11.004

Revelle, W. (2018). Psych: Procedures for personality and psychological research [R package] (Version 1.8.12). Retrieved from https://CRAN.R-project.org/package=psych

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

Tomassini, M., Sipper, M., & Perrenoud, M. (2000). On the generation of high-quality random numbers by two-dimensional cellular automata. *IEEE Transactions on Computers*, *49*, 1146–1151. doi:10.1109/12.888056

Tothfalusi, L., & Endrenyi, L. (2017). Algorithms for evaluating reference scaled average bioequivalence: Power, bias and comsumer risk. *Statistics in Medecine*, *36*, 4378–4390. doi:10.1002/sim.7440

Viechbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics*, *32*, 39–60. doi:10.3102/1067998606298034

Appendix A: The distributions of the standardized difference in within-subject design

Consider a population $\mathcal{D}$, a normal population whose mean is μ and standard deviation is $\sigma_{\mathbf{D}}$. Let $\overline{D}$ denote the difference between a sample mean (obtained from n observations) taken from that population and a constant that could be different from μ, say $\mu - \Delta$. Following the work of Hedges (1981), it is known that the distribution of that mean difference divided by the sample standard deviation $S_{\mathbf{D}}$ is a variate, call it d_D, which —when scaled by $\sqrt{n}$— follows a noncentral t distribution,

$$\sqrt{n}\, d_{\mathbf{D}} = \sqrt{n}\, \frac{\overline{D}}{S_{\mathbf{D}}} \sim t_{n-1}\left(\lambda_{\mathbf{D}} = \frac{\Delta}{\sigma_{\mathbf{D}}}\sqrt{n}\right) \tag{4}$$

in which $\overline{D}$ is the observed mean difference, $n-1$ is the degree of freedom parameter, and $\lambda_{\mathbf{D}}$ is the noncentrality parameter. When there are two repeated measures, say $\mathbf{X}$ and $\mathbf{Y}$, both normally distributed with means that are Δ

apart, with a common standard deviation of σ, and with a correlation ρ, we can compute the difference $\mathbf{D} = \mathbf{X} - \mathbf{Y}$ and we are back to the result of Eq. (1). Because $\sigma^2_\mathbf{D} = \sigma^2_\mathbf{X} + \sigma^2_\mathbf{Y} - 2\sigma_\mathbf{X}\sigma_\mathbf{Y}\rho$ Kendall and Stuart (1979) and because variances were assumed homogeneous in the population (i.e., $\sigma_\mathbf{X} = \sigma_\mathbf{Y} = \sigma$), we obtain

$$\sigma_\mathbf{D} = \sqrt{2}\,\sigma\sqrt{1-\rho} \tag{5}$$

so that the following equation, identical to Eq. (1), is the main result of Becker (1988; also see Morris, 2000):

$$\sqrt{n}\,d_\mathbf{D} = \sqrt{n}\frac{\overline{D}}{S_\mathbf{D}} \sim t_{n-1}\left(\lambda_\mathbf{D} = \frac{\Delta}{\sigma}\sqrt{\frac{n}{2(1-\rho)}}\right) \tag{6}$$

; in what follow, we use δ to denote the population's true standardized different Δ/σ. In a sample, the relation $S^2_\mathbf{D} = S^2_\mathbf{X} + S^2_\mathbf{Y} - 2S_\mathbf{X}S_\mathbf{Y}r$ is still valid. However, it is implausible to maintain that the two sample standard deviations are precisely equal. Cousineau (2019), and Goulet-Pelletier and Cousineau (2018) ignored this next step (but note that the difference is generally small). Observing that $S_\mathbf{X} \times S_\mathbf{Y}$ is the geometric mean of the two variances $S^2_\mathbf{X}$ and $S^2_\mathbf{Y}$, we can rewrite

$$\begin{aligned} S^2_\mathbf{D} &= S^2_\mathbf{X} + S^2_\mathbf{Y} - 2S_\mathbf{X}S_\mathbf{Y}r \\ &= 2\,S^2_p - 2\,S^2_g\,r\,\frac{S^2_p}{S^2_p} \\ &= 2\,S^2_p\left(1 - r\,\frac{S^2_g}{S^2_p}\right) \\ &= 2\,S^2_p\,(1 - r_W) \end{aligned} \tag{7}$$

where S^2_p is the pooled variance (i.e., the arithmetic mean of $S^2_\mathbf{X}$ and $S^2_\mathbf{Y}$), S^2_g is the geometric mean of the same two variances and $r_W = r\ S^2_g/S^2_p$ is a rectified Pearson correlation which takes into account fluctuations in the sample variances. The ratio S^2_g/S^2_p is very close to —but never exceeds— 1 as the two sample variances should be roughly equal and as the geometric mean is always smaller than the arithmetic mean. Thus, $|r_W|$ never exceeds 1, as expected from a correlation. From (4), we see that it is straightforward to convert a d_D to a d_p as

$$d_\mathbf{D} = \frac{\overline{D}}{S_\mathbf{D}} = \frac{\overline{D}}{\sqrt{2}S_p\sqrt{1-r_W}} = d_p\frac{1}{\sqrt{2(1-r_W)}} \tag{8}$$

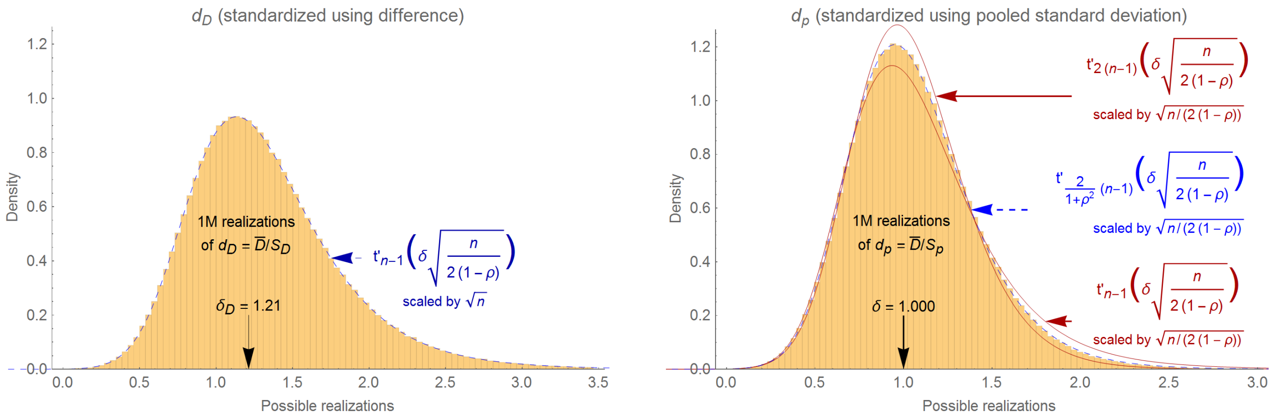
(apart from the rectified correlation used in lieu of the regular correlation, this is the formula reported in Goulet-Pelletier and Cousineau (2018), Goulet and Cousineau (2019). Whereas we can convert d_D into d_p (or vice versa) from Eq. (5), we cannot convert their confidence intervals. Observe that

$$\begin{aligned} \sqrt{n}\,d_p &= \sqrt{n}\,\frac{\overline{D}}{S_p} = \sqrt{n}\,\frac{\overline{D}}{S_\mathbf{D}}\sqrt{2(1-r_W)} \\ &= \begin{cases} \sqrt{n}\,\frac{\overline{D}}{S_\mathbf{D}} & \sim t_{n-1}\left(\delta\sqrt{\frac{n}{2(1-\rho)}}\right) \\ \times & \\ \sqrt{2(1-r_W)} & \sim\ ? \end{cases} \end{aligned} \tag{9}$$

This formulation makes the problem apparent. First, we do not know the distribution of $\sqrt{2(1-r_W)}$. Olkin and Pratt (1958) reported the distribution of r so that the distribution of $\sqrt{2(1-r)}$ could in principle be derived but here the function is based on the rectified Pearson correlation (to derive the distribution, it might be easier to use the relation $2(1-r_W) = S^2_\mathbf{D}/S^2_p$). Second, and more critical, both terms are correlated so that integrating them as if they were independent is not legitimate.

An illustration. We show in Figure A.1 an example from simulated scores. We generated one million samples composed of pairs of normally distributed scores using these parameters: true difference between the means $\Delta = 15$, $\sigma = 15$ (so that the true population standardized difference δ is 1), correlation ρ = 0.25 and sample size $n = 10$.

Figure A.1 ■ One million simulated samples from a population with parameter δ = 1, ρ = 0.25, and $n = 10$ and the distribution of d_D (left) and d_p (right) computed from these.



In the left panel, we show the frequency distribution (yellow histograms) of d_D. The dashed line is the theoretical distribution. The true variance of the differences (Eq. 2) is $\sqrt{2}\,\sigma\sqrt{1-\rho} = 18.37$ so that $\delta_{\mathbf{D}} = 15/18.37 = 0.816$.

In the right panel, we show the frequency distribution of d_p. Also shown is the theoretical distribution (in blue dashed line). The other two lines shows noncentral t distributions with integer degree of freedom $n-1$ and $2(n-1)$ which are the limiting distribution when ρ = ± 1 and ρ = 0 respectively.

In sum. The distribution of d_p is the distribution of d_D rescaled by $\sqrt{2(1-\rho)}$. However, in actual application, we do not know ρ and using an estimate of it introduces variability, distorting the distribution.

Open practices

The *Open Material* badge was earned because supplementary material(s) are available on osf.io/nwxsb.

Citation

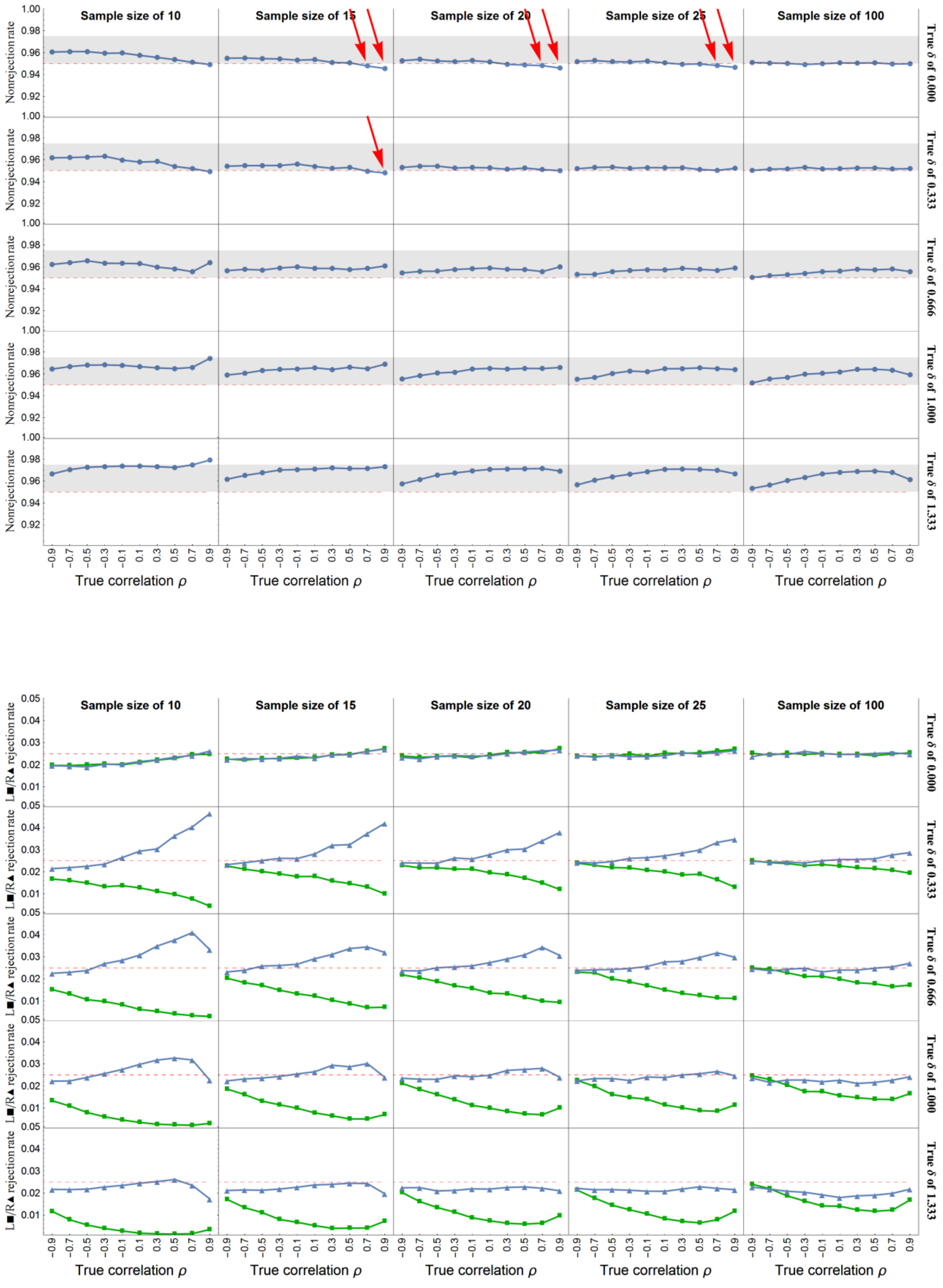
Cousineau, D., & Goulet-Pelletier, J.-C. (2021). A study of confidence intervals for Cohen's dp in within-subject designs with new proposals. *The Quantitative Methods for Psychology*, *17*(1), 51–75. doi:10.20982/tqmp.17.1.p051

Received: 06/10/2020 ∼ Accepted: 22/02/2021

Figures B.1 to B.8 follow.

Figure B.1 ■ Morris (2000) method. (a): nonrejection rate of the true population parameter for 95% confidence intervals as a function of the population correlation (horizontal axis), the population true effect size (the rows) and the sample sizes (the columns). The three red arrows indicates scenarios where the proportion of nonrejection is smaller than desired. (b) left (green square) and right (blue triangle) rejection rates in the same simulations.

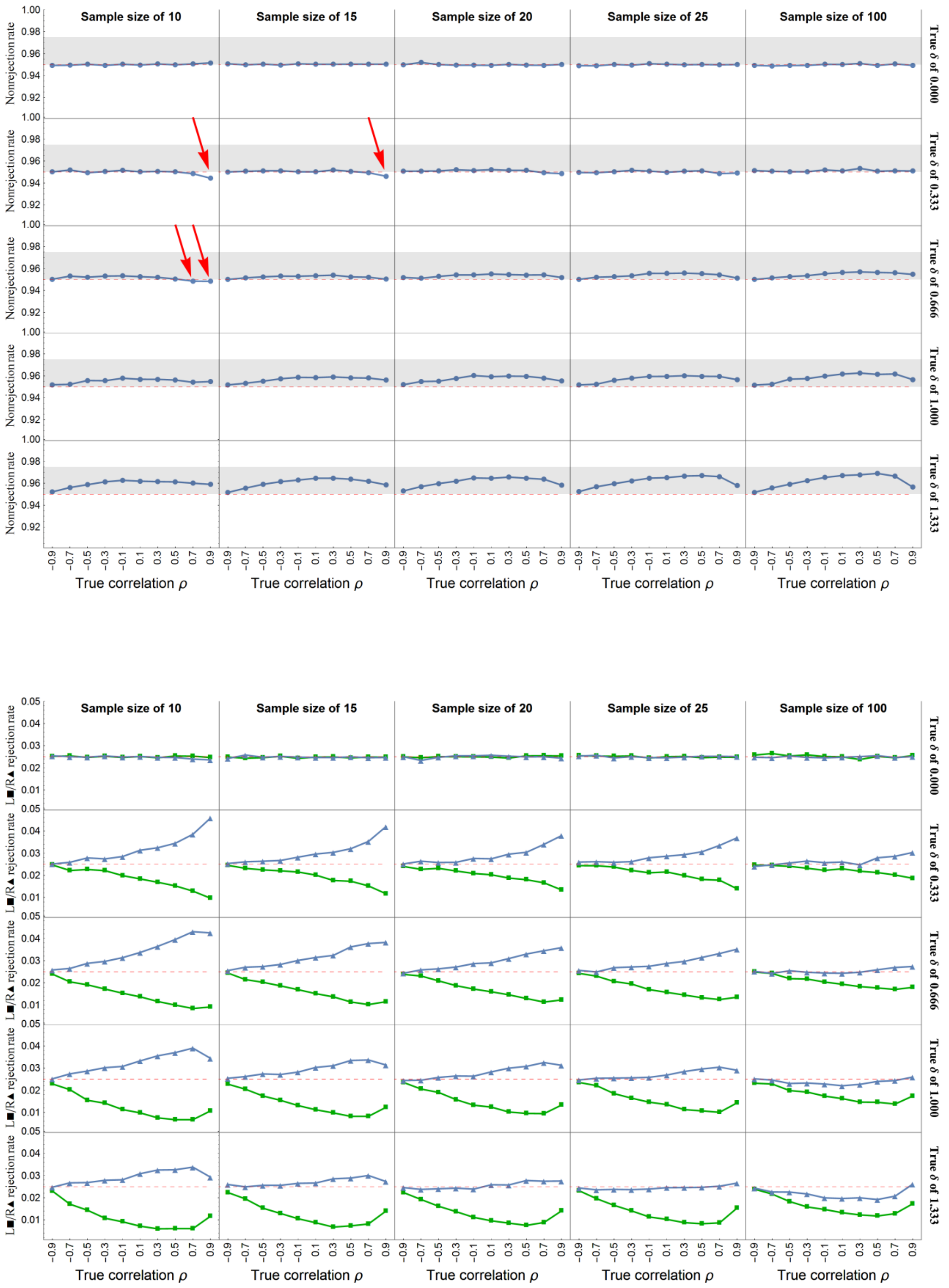
(a)



(b)

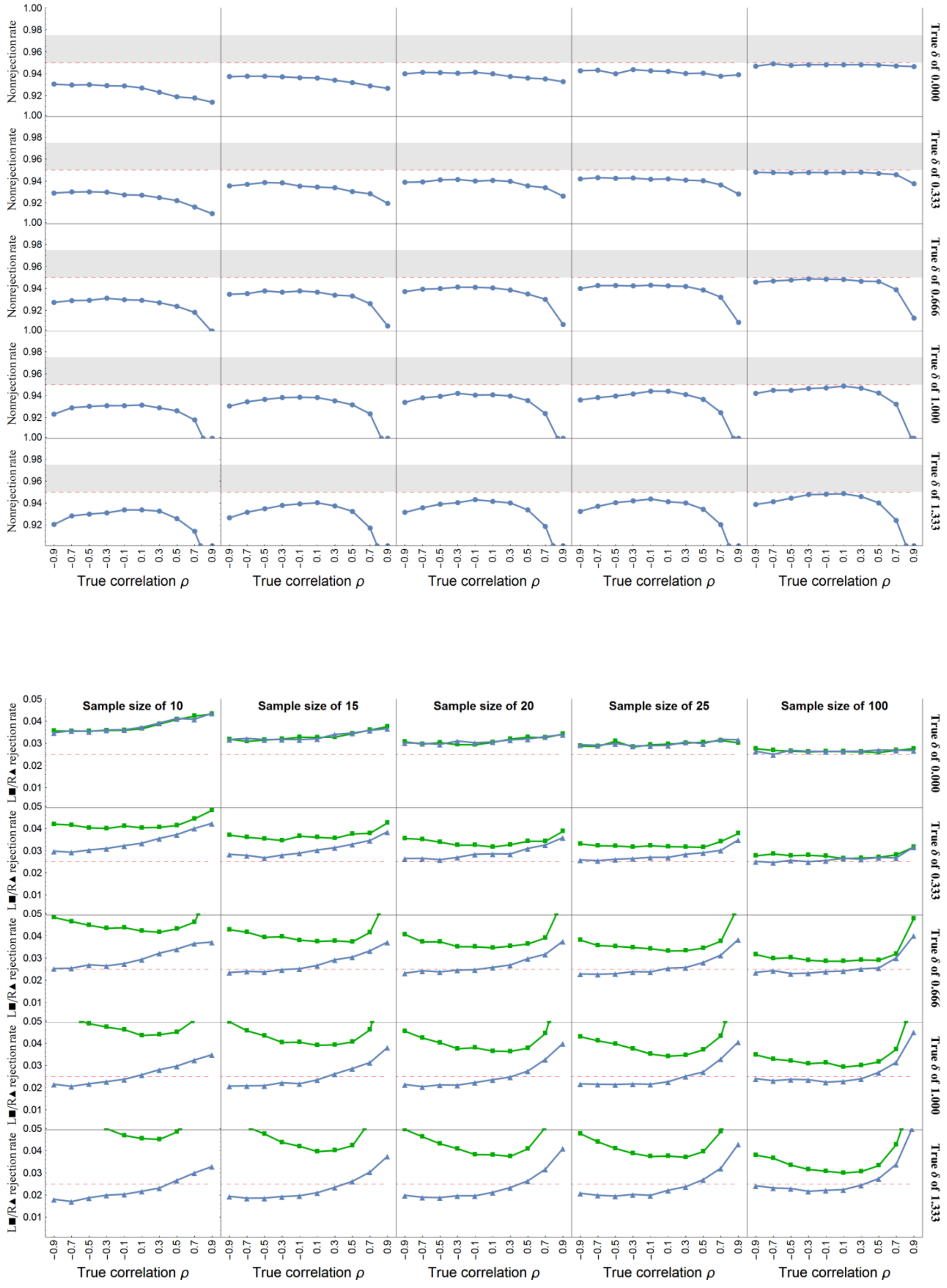
Figure B.2 ■ Algina and Keselman (2003) method in the same format as Figure B.1.

(a)



(b)

Figure B.3 ■ Goulet-Pelletier and Cousineau (2018) method in the same format as Figure B.1.

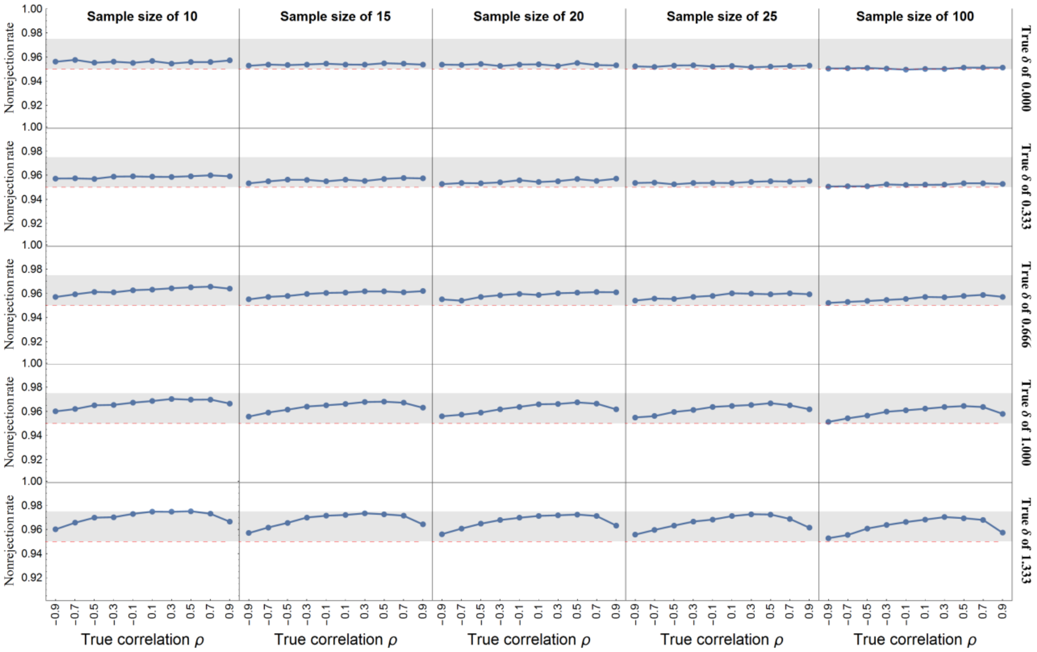


(a)

(b)

Figure B.4 ■ MAG method in the same format as Figure B.1.

(a)



(b)

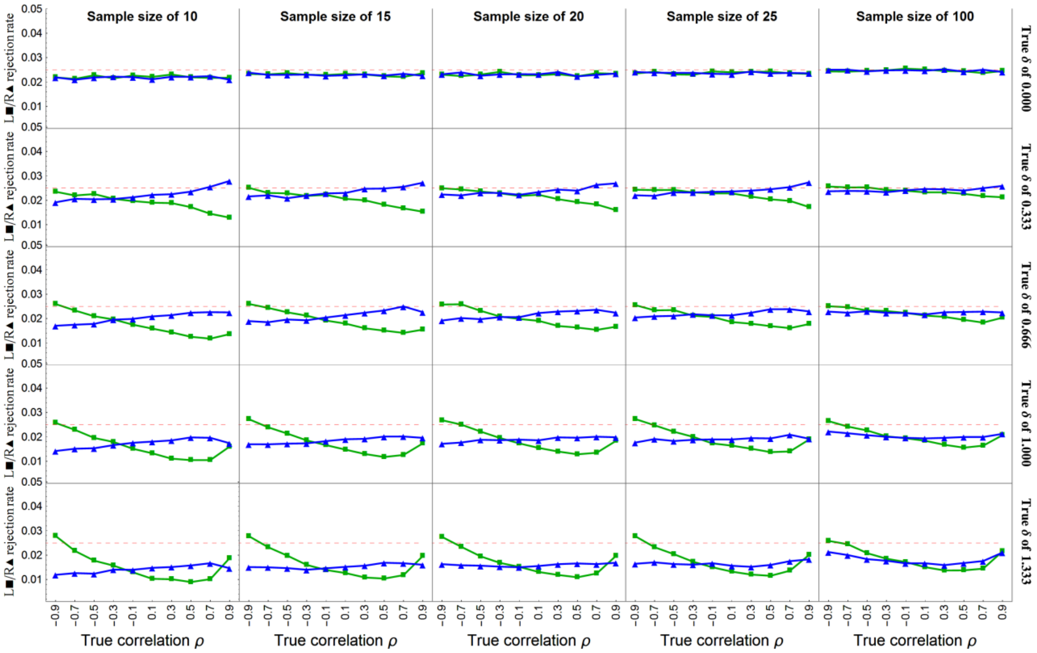
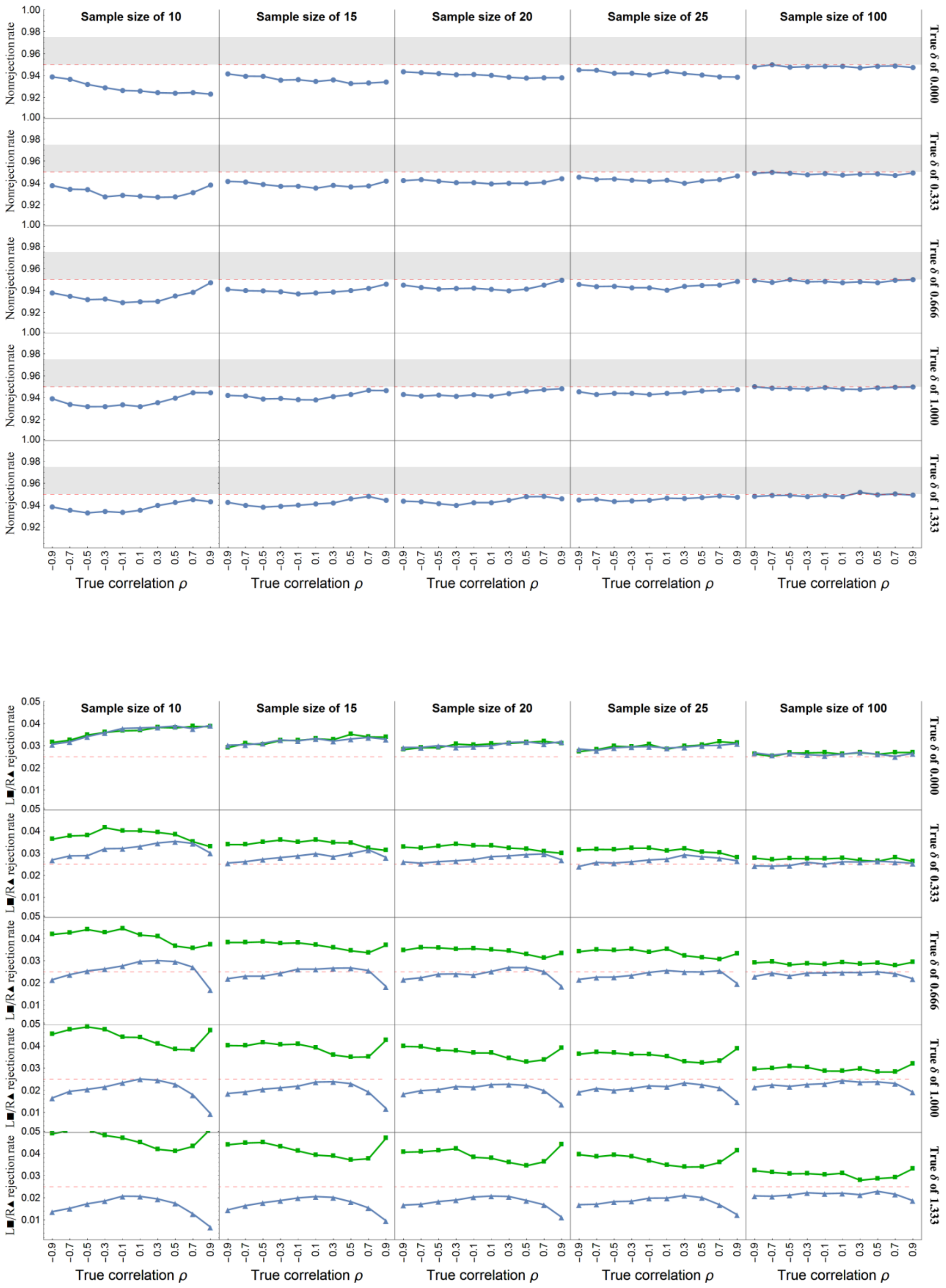


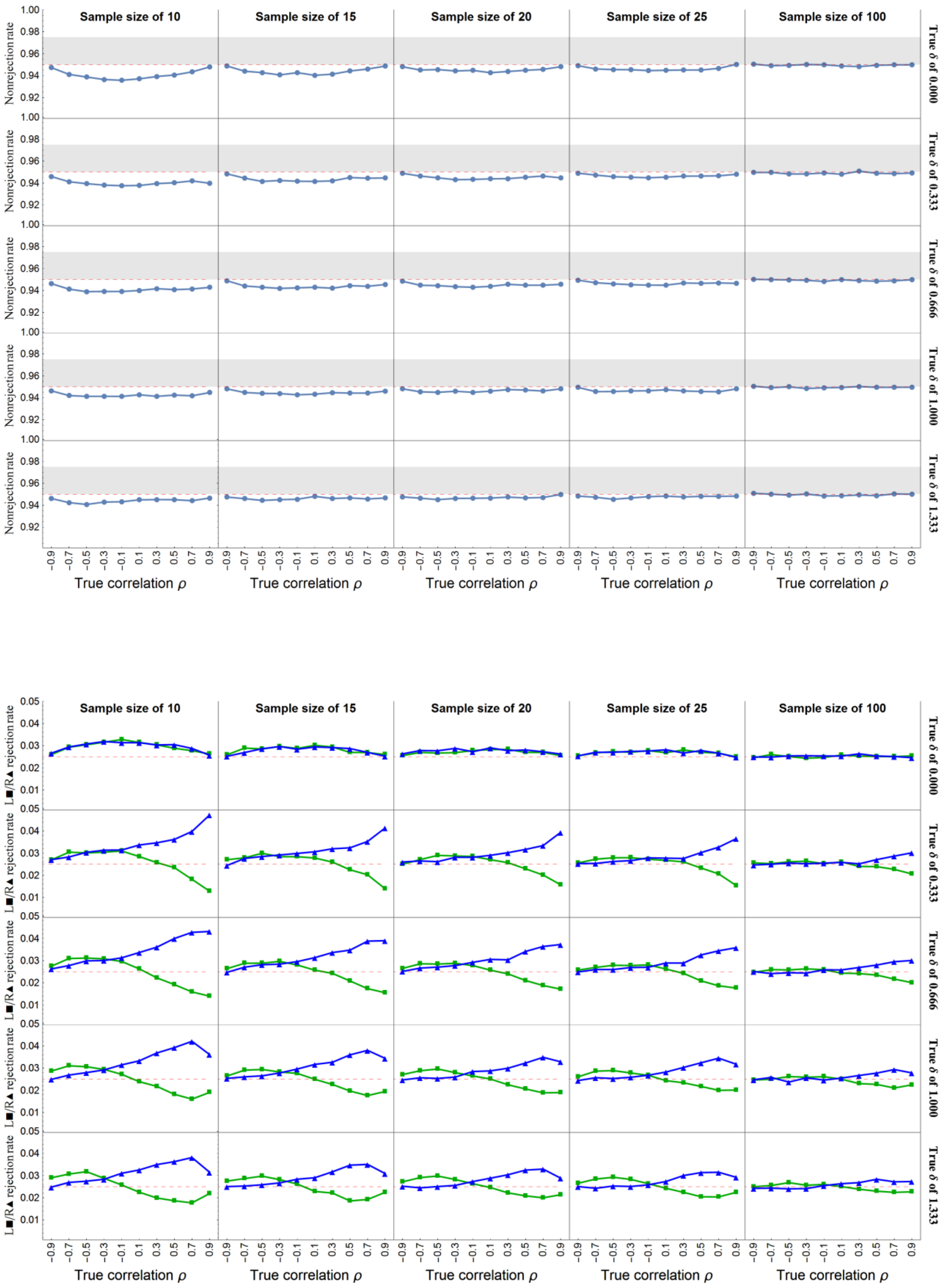
Figure B.5 ■ t' method in the same format as Figure B.1.



(a)

(b)

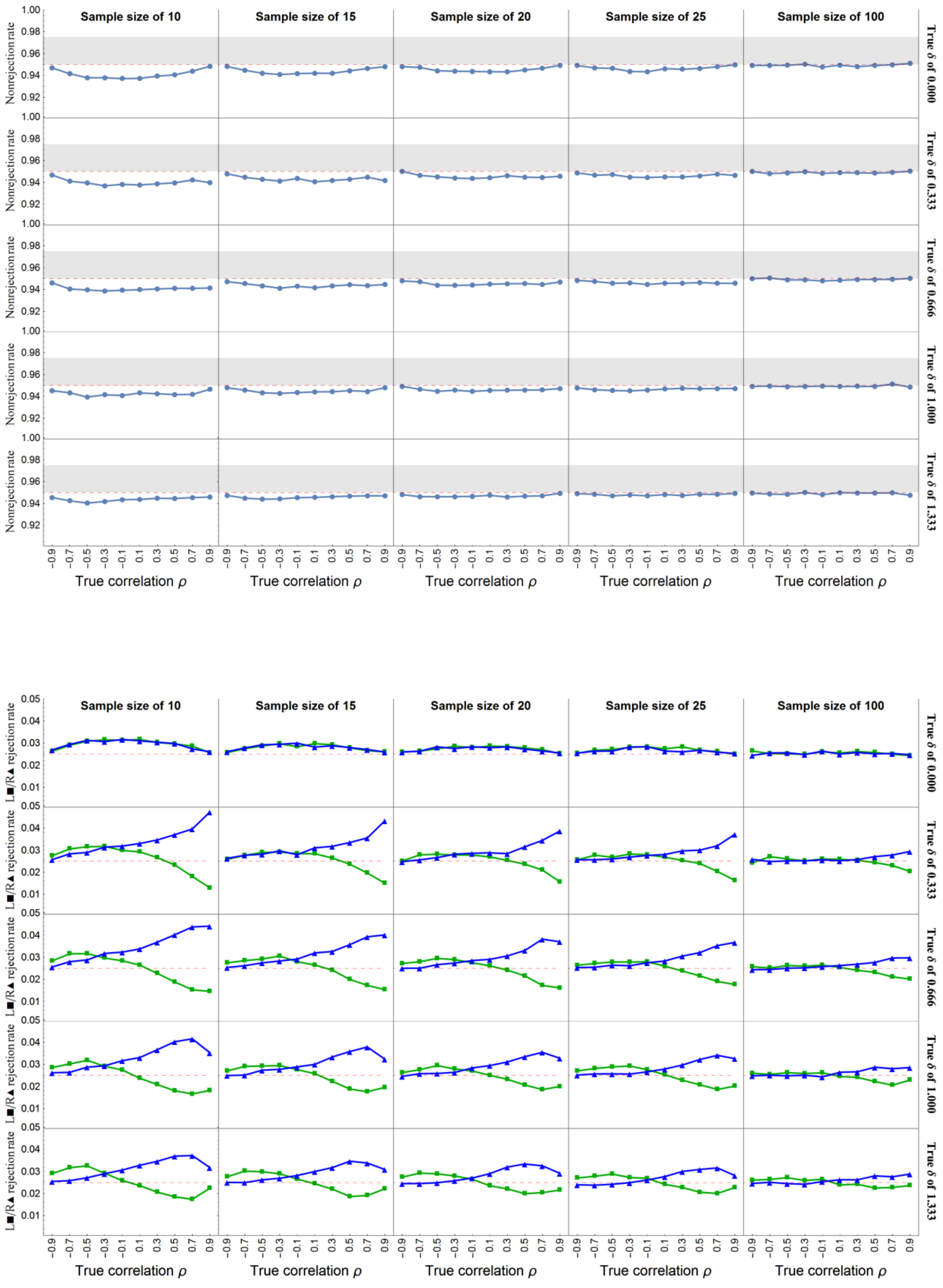
Figure B.6 ■ Λ' method in the same format as Figure B.1.



(a)

(b)

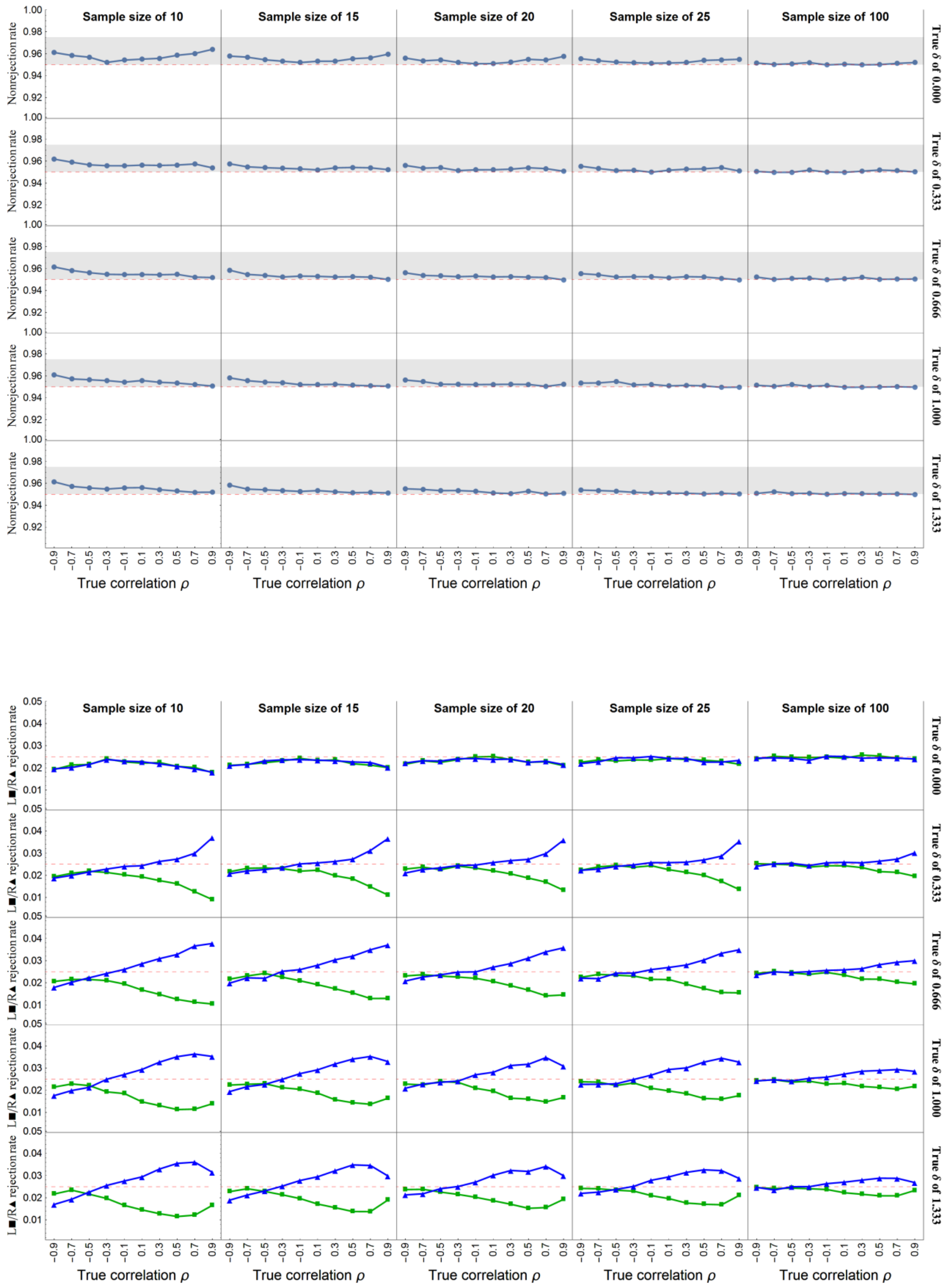
Figure B.7 ■ Pivotal of t' method in the same format as Figure B.1.



(a)

(b)

Figure B.8 ■ Corrected Λ' method in the same format as Figure B.1.



(a)

(b)