The benefits of reporting critical effect size values

Ambra Perugini[1], Filippo Gambarota[1], Enrico Toffalini[2], Daniël Lakens[3], Massimiliano Pastore[1], Livio Finos[4], Psicostat[1], and & Gianmarco Altoè[1]

[1] Department of Developmental and Social Psychology

University of Padova

Italy

[2] Department of General Psychology

University of Padova

Italy

[3] Eindhoven University of Technology

Netherlands

[4] Department of Statistics

University of Padova

Italy

Author Note

The authors made the following contributions. Ambra Perugini: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing, Methodology, Software; Filippo Gambarota: Writing – Original Draft Preparation, Writing - Review & Editing, Methodology, Software; Enrico Toffalini: Writing – Original Draft Preparation, Writing - Review & Editing, Software, Supervision; Daniël Lakens: Conceptualization, Writing - Review & Editing, Supervision; Massimiliano Pastore: Conceptualization, Writing - Review & Editing, Supervision; Livio Finos: Conceptualization, Writing - Review & Editing, Supervision; Psicostat: Conceptualization, Writing - Review & Editing, Supervision; Gianmarco Altoè: Conceptualization, Writing - Review & Editing, Supervision.

Correspondence concerning this article should be addressed to Gianmarco Altoè, Via Venezia 8, 35131, Padova, Italy. E-mail: gianmarco.altoe@unipd.it

Abstract

Critical effect size values represent the smallest detectable effect that can reach statistical significance given a specific sample size, alpha level and test statistic. It can be useful to calculate the critical effect size when designing a study, and evaluate whether such effects are plausible. Reporting critical effect size values may be useful when the sample size has not been planned a priori, or when there is uncertainty about the expected sample size that can be collected, when researchers plan to analyze the data with a statistical hypothesis test. To assist researchers in calculating critical effect size values we have developed an R package that allows researchers to report critical effect size values for group comparisons, correlations, linear regressions, and meta-analyses. Reflecting on critical effect size values could benefit researchers during the planning phase of the study by helping them to understand the limitations of their research design. Critical effect size values are also useful when evaluating studies performed by other researchers when a-priori power analyses were not performed, especially when non-significant results are observed.

Keywords: critical effect size values, statistical significance, hypothesis testing, effect sizes

Word count: 4745

The benefits of reporting critical effect size values

The "critical effect size" refers to the smallest effect size that can be statistically significant, given the test performed, the sample size, and the alpha level (Lakens, 2022). Consider a study that tests a bivariate correlation with $n = 20$ participants and uses $\alpha < 0.05$ as the threshold for statistical significance in a two-sided test. The critical effect size values for this test are Pearson's $r = -0.44$ and $r = 0.44$, which means that the statistical test will yield a significant result for effects larger than $r = -0.44$ or $r = 0.44$, while observed effect sizes between these two values will lead to a statistically non-significant result. In the present paper, we suggest that researchers should systematically report critical effect size values when they make inference using significance levels in hypothesis tests to make claims about the presence or absence of effects. Critical effect size values can serve as a useful complement for interpreting study results, communicate the informational value of a study design, and contextualize the interpretation of non-significant results, particularly when statistical significance is relied upon for inference. Reporting critical effect size values is especially important when the sample size was not (or could not) be predetermined on the basis of a smallest effect size of interest.

Null hypothesis significance testing (NHST) remains the most prominent approach for statistical inference in science, even though there are widespread concerns about the misuse of hypothesis tests (Chow, 1988; Cohen, 1994; Cortina & Dunlap, 1997; Gigerenzer et al., 2004; Hagen, 1997; Haig, 2017; Krueger, 2001; Lakens, 2021; Miller, 2017; Nickerson, 2000). Over the past decades, numerous proposals have emerged to improve the use of hypothesis testing. These include complementing hypothesis tests with effect sizes and their confidence intervals, preregistering hypotheses before the data is collected, and conducting power analyses to determine the sample size *a priori* based on a smallest effect size of interest. Researchers often publish studies with low power for medium to small effect sizes (Szucs & Ioannidis, 2017), with the main justification for the sample size being

resource limitations (Lakens, 2022). However, low power makes it challenging to distinguish signal from noise, and combined with the selective reporting of statistically significant results in the published literature leads to an overestimation of effect sizes (Altoè et al., 2020). Certain types of research questions can be studied by relying on online data collection, which has made it cheaper and more feasible to collect very large samples. Studies with very large samples are well-powered to detect even small effect sizes, but they also require researchers to carefully consider the possibility that an effect might be statistically significant, but practically insignificant.

We believe there are two clear benefits to reporting critical effect size values for a corresponding test. In the following sections we will illustrate examples of common scenarios and we will go through the main problems and possible solutions, highlighting the utility of critical effect size values in such contexts.

Small Sample Sizes and Uncertain sample size determination

First, when sample sizes are small, the critical effect size values inform readers about whether the effect sizes that could lead to rejecting the null hypothesis are in line with realistic expectations. If the sample size is small and only very large effects would yield a statistically significant result, and the underlying mechanism that is examined is unlikely to lead to such large effect sizes, researchers will realize they are not able to collect sufficient data to perform an informative hypothesis test. An a-priori power analysis would typically lead to a similar conclusion, but reporting critical effect size values will focus the attention more strongly on which effect sizes are reasonable to expect. In case of already conducted studies with small sample sizes it could be argued that it would be more informative to use retrospective design analysis, but this would require both knowledge on the plausible effect size and/or on the smallest effect size of interest (SESOI), which are not always easly to determine. The use of critical effect size values can be used in a simple an efficient way to evluate which findings could have not be found significant due to sample size limitations

and based on the complexity of the test. Power analysis provides information about how likely it is to detect a specific effect size if it truly exists in the population. However, it does not indicate the minimum effect size required to reach statistical significance, which could provide additional insights into the strengths and limitations of a study design.

Let us consider the following two scenarios as examples. First, imagine researchers who conduct a study involving a between-group comparison. Due to severe resource constraints, they are only able to collect a limited sample size of $n = 30$ ($n = 15$ per group). The authors did not perform an a-priori power analysis, and their sample size justification is purely based on how many participants they can collect. The statistical tests yields a non-significant result, $p > 0.05$. Given their prior expectation that the effect of interest may not be large, they acknowledge that their study was likely underpowered, though without a theoretically expected effect size, it remains uncertain how low the statistical power is. Subsequently, they compute the critical effect size values, revealing Cohen's $d = -0.75$ and $d = 0.75$. This indicates that any observed Cohen's d between -0.75 and 0.75 will certainly fail to reach statistical significance. By reporting these critical effect size values the researchers transparently convey that estimated effects much larger than most effects in the psychological literature will always fall short of significance. This means that the hypothesis test will not be informative.

To provide a more tangible illustration of the practical applications of the critical effect sizes, let us examine a real-world instance drawn from published research. The study of interest was conducted on musicians and non musicians to detect differences in working memory, a modest sample size of 57 participants (42 musicians and 15 controls) was collected (Weiss et al. (2014)). The two groups were compared on a variety of tasks, related to verbal working memory (VWM) and tasks related to auditory skills. Regarding the VWM, the syllable-span task was administered and scores were computed for the maximal span and for the total number of sequences they correctly repeated. A two-tailed

t-test was conducted to compare the groups. The resulting effect was statistically significant for the maximal span, $t(55) = -2.5$, $p = .017$, the Cohen's d calculated on summary statistics is of $-.74$, critical $d = \pm.62$. But for the total number of sequences the test did not reach statistical significance ($t(55) = -1.8$, $p = 0.076$), actual $d = -0.31$ and critical d again $\pm.62$. Thus, given the limited sample size, a result must be of rather large magnitude to be associated with $p < 0.05$, while medium to small effect sizes will simply fail to reach significance. This case makes it obvious that a failure to find a significant result does not imply absence of an effect, while a statistically significant effect may represent an overestimation. Non-significant effects in such scenarios are likely, even if there is a true effect of interest. If a significant effect is observed, but most effects in a field are smaller than a study can detect, it is probable the effect size is overestimated (Hedges, 1984), and this should be taken into account when interpreting the effect size.

Beyond critical effect size values, researchers should carefully consider, before conducting a study, what conclusions can realistically be drawn from their experimental design. The more complex the study design, the higher the critical effect size value. Researchers should assess whether their available sample size allows them to address a simpler yet still relevant question. If this is not feasible, and sample size constraints make it impossible to detect significance for a plausible effect because it is smaller than the critical effect size value, alternative data collection strategies, such as multisite studies (Byers-Heinlein et al., 2020; Jarke et al., 2022; Moshontz et al., 2018; Sirois et al., 2023), should be considered. Lastly, if participating in or leading a multisite study is not an option, researchers should evaluate whether collecting the sample for exploratory analysis without making inferences might still be valuable. Such data could later be included in a meta-analysis.

Large Sample Sizes and Meta-analysis

When sample sizes are very large, the critical effect size values will make it clear that trivially small effect sizes will be statistically significant. Reporting critical effect size values will focus the attention of researchers on the difference between statistical significance and practical significance, and raises awareness of the importance to interpret the size of the effect. This is especially important in correlational studies with large sample sizes in psychology, where systematic but uncontrolled sources of variability may lead to small but theoretically meaningless non-zero effects, a phenomenon referred to as the 'crud factor' (Orben & Lakens, 2020).

Imagine researchers who gain access to a very large archival dataset ($n = 5,000$) and decide to explore bivariate correlations between variables. With such a large sample size, very small correlations will reach statistical significance, and the critical effect size values for a two-sided test are $r = \pm 0.03$. It is worth noting that such small effects may potentially reflect just minor artifacts (Wilson et al., 2020), such as subtle experimenter effects or slight non-independence among observations, even in meticulously designed studies.

Consider now another real-world instance involving the study by Kramer et al. (2014), which explored the impact of emotional content on Facebook users' experiences. With a notably large sample size of $n = 689.003$, the researchers observed a statistically significant increase in the number of negative words typed by participants when positive posts were reduced in their timeline, $t(310,044) = -5.63$, $p < 0.001$ (non-directional test). Based on this and other results from the study, the authors conclude that "the emotions expressed by friends, via online social networks, influence our own moods, constituting, to our knowledge, the first experimental evidence for massive-scale emotional contagion via social networks" (Kramer et al., 2014). The critical values for Cohen's d in this experimental design are -0.006 and $+0.006$. It should have been clear from the outset that

the researchers should have considered the effect size beyond merely testing for the statistical significance. The critical effect size should prompt reflection on which effect sizes are truly meaningful in this context, and whether the observed effect size ($d = 0.02$) is theoretically or practically interesting.

A similar scenario may arise in meta-analysis. Despite potential loss of precision due to substantial heterogeneity across effect sizes in different studies, meta-analyses typically synthesize a large amount of evidence, and as a consequence, even very small average effect sizes can reach significance. While the focus of meta-analysis is generally on estimating effect sizes with uncertainty, statistical significance is routinely reported and interpreted. Signaling the critical effect size values beforehand can serve as a clear warning that statistically significant results should not automatically be interpreted as practically significant, thus urging caution when interpreting the results.

Lastly, routinely reasoning about critical effect sizes, alongside the use of commonly known practices to enhance the quality of research (i.e. power analysis, design analysis, data simulation etc.), will bring benefits to researchers by making clear that significance is strictly related to the sample size, the alpha level and the statistic test used to analyse the data. Nevertheless, in educational settings, this concept will help students to better grasp such concepts and therefore give a more critical approach to published research and for their future studies.

How to Compute Critical Effect Size Values

In this section, we provide guidance and formulas for computing standardized critical effect sizes with examples for frequently encountered effect sizes including Standardized Mean Differences (Cohen's d), correlations (Pearson's r), and raw and standardized coefficients in linear models. The same equations can be used to calculated unstandardized effect sizes that can be preferred in some situations (e.g., linear models). These formulas have been incorporated into R functions of the package `criticalESvalue`

accessible at: https://github.com/psicostat/criticalESvalue, and are elaborated upon in the subsequent section.

t-test

For the t-test we considered the two-sample, paired, and one-sample tests. As a general approach the t statistic is computed as reported in Equation (1).

$$t = \frac{b}{\mathrm{SE}_b} \tag{1}$$

Where b is the unstandardized effect size that depends on the type of test. For example, in the two sample t-test the numerator is the difference between the two sample means while in the one-sample case is the difference between the sample mean and the population value. The denominator is the standard error of the numerator that depends on the sample size and the samples variances.

Equation (2) describes a general formulation of a standardized effect size measure. Where b is the unstandardized effect size (e.g., the difference between two means) and s is the standardization term. For example, in the two-samples case, s is the pooled standard deviation between the two samples, while in the paired samples case, s is the standard deviation of the differences for paired samples.

$$d = \frac{b}{s} \tag{2}$$

The critical t_c is the test statistic associated with a p-value equal or lower to α. Substituting t_c into Equation (1) we realize that there is a b_c that with a certain SE_b produce the critical test statistics. The general form of the critical effect size d_c is defined in Equation (3). The b_c is derived solving Equation (1) for b.

$$\begin{aligned} d_c &= \frac{b_c}{s} \\ b_c &= t_c \times SE_b \end{aligned} \tag{3}$$

Another way of conceptualizing the critical effect size is by removing the sample size from the t statistics (see Equation (1)). In fact, SE_b can be generally defined as $\frac{s}{\sqrt{N}}$ where s is the denominator in Equation (2) and n is the sample size. Thus, $d_c = t_c \times \frac{1}{\sqrt{n}}$ ($\frac{1}{\sqrt{n}}$ because the standard deviation of a standardized effect size is one by definition) is equivalent to (3). The only caveat is that calculating d or d_c in this way assume that the standard deviation used to calculate the test statistics and the effect size is the same. While this is true most of the time, as explained in the paired t-test section, there are more than one methods to calculate the (critical) effect size using different standard deviations. Furthemore, this method assume that the two groups (in case of a two-sample t-test) have the same size.

One sample t-test

For the one-sample t-test b is the sample mean and s is the sample standard deviation and n is the sample size. $SE_b = \frac{s}{\sqrt{n}}$ and the t_c is based on $n-1$ degrees of freedom. Thus $b_c = t_c \times SE_b$ and $d_c = \frac{b_c}{s}$. Equivalently, $d_c = t_c \times \frac{1}{\sqrt{n}}$.

Two-sample t-test

We can apply the one-sample approach to the two-sample case. When assuming homogeneity of the variances between the two groups we have $n_1 + n_2 - 2$ degrees of freedom to calculate t_c and s is the pooled standard deviation (s_p) calculated using Equation (4). The s_p can be intepreted as the square root of the weighted average of the samples standard deviations. The effect size d is calculated as $\frac{b}{s_p}$ and the critical effect size can be calculated as $d_c = \frac{b_c}{s_p}$ with $b_c = t_c \times SE_b$. SE_b is calculated as $s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. As for the one-sample case, we can calculate the critical effect size from the t statistics as $d_c = t_c \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$.

$$s_p = \sqrt{\frac{s_1^2(n_1 - 1) + s_2^2(n_2 - 1)}{n_1 + n_2 - 2}} \tag{4}$$

When relaxing the assumption of equal variances (i.e., using the Welch's t-test) s is calculated as the square root of the average between the two samples variances, SE_b is calculated as $\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}$. The degress of freedom for t_c are calculated in a more complex way using the Welch-Satterthwaite equation (Satterthwaite, 1946; Welch, 1947). For the Welch's t-test we can calculate the critical effect size removing the sample size from t_c (as done above) only when $n_1 = n_2$.

Paired sample t-test

For the paired sample case the situation is less straightforward. The reason is that the t statistic is computed using b being the average of the paired differences between the two (paired) samples and s being the standard deviation of the differences. Using these values, we can just compute the critical effects size as for the one-sample case, in fact a paired t-test is a one-sample test on the vector of paired differences. Test statistics and effect size based on the standard deviation of differences (the so called d_z) are used for hypothesis testing and power analysis (Lakens, 2013). The problem is that the d_z cannot be directly compared to the effect size calculated using the pooled standard deviation (Morris & DeShon, 2002). If the correlation between the two paired samples is known there is a direct relationship between the pooled standard deviation s_p and the standard deviation of differences s_D as reported in Equation (5)

$$\begin{aligned} s_p &= \frac{s_D}{\sqrt{2(1 - \rho)}} \\ s_D &= s_p\sqrt{2(1 - \rho)} \end{aligned} \tag{5}$$

Even if the hypothesis testing is computed using s_D, we reported the critical effect size value both using the pooled standard deviation and the standard deviation of the

differences.

Hedges's correction

The effect size calculated as in the previous step is known to be inflated especially for small samples. For this reason, there is a corrected version of the effect size called Hedges's g Viechtbauer (2007). The Hedges's g can be calculated for all the t-test scenarios. The Hegdes's correction is implemented in Equation (6) where Γ is the gamma function and m the degrees of freedom.

$$c(m) = \frac{\Gamma\left(\frac{m}{2}\right)}{\sqrt{\frac{m}{2}}\Gamma\left(\frac{m-1}{2}\right)} \approx 1 - \frac{3}{4m-1} \tag{6}$$

The correction is applied as $g = c(m) \times d$. The hypothesis testing is not taking into account the Hedges's correction thus the critical g will be different from the critical d.

Correlation Test

Hypothesis testing for the Pearson's correlation coefficient is usually done using the t statistics or using a z statistics. The general approach presented for the t-test is still valid. The only difference is that the correlation is already a standardized effect size thus there is no need of the standardization term s. Equation (1) can be used substituing b with r (the sample correlation coefficient) as shown in Equation (7). The standard error is calculated as $SE_r = \sqrt{\frac{1-r^2}{n-2}}$ (Bowley, 1928).

$$\begin{aligned} t &= \frac{r}{SE_r} = r\sqrt{\frac{n-2}{1-r^2}} \\ r_c &= \frac{t_c}{\sqrt{n-2+t_c^2}} \end{aligned} \tag{7}$$

Another approach for hypothesis testing of the Pearson's correlation coefficient is using the Fisher's z transformation (Fisher, 1915, 1921) reported in Equation (8) and a z test statistics.

$$
\begin{aligned}
r_z &= \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \operatorname{artanh}(r) \\
r &= \frac{\exp(2r_z - 1)}{\exp(2r_z + 1)} = \tanh(r_z)
\end{aligned}
\tag{8}
$$

We can still use Equation (1) substituting b with r_z. The standard error of the r_z correlation is calculated as $SE_{r_z} = \frac{1}{\sqrt{n-3}}$. Equation (9) shows how to calculate the critical r_{z_c} value where z_c is the critical value of the standard normal distribution with a certain α level. Finally, r_{z_c} can be transformed back from the Fisher's z transformation using Equation (8)

$$
\begin{aligned}
z &= \frac{r_z}{SE_z} = r_z\sqrt{n-3} \\
r_{z_c} &= \frac{z_c}{\sqrt{n-3}}
\end{aligned}
\tag{9}
$$

Linear Regression

Hypothesis testing on regression coefficients (e.g., using the `lm` function in R) is performed using the so-called Wald test. The test can be considered a one-sample t-test where the t value is calculated dividing the regression parameter $\beta_0, \ldots, \beta_j$ by its standard error SE_{β_j}. The critical t value is calculated from a t distribution using $n - p - 1$ degrees of freedom where n is the number of observations and p is the number of coefficients beyond the intercept. We can substitute b with β_j and SE_b with SE_{β_j} in Equation (3) to find the critical unstandardized regression coefficient ($\beta_{c_j} = t_c \times SE_{\beta_j}$). There is no an unique way to calculate the standardized version of the critical regression coefficient. The main reason is that in regression modelling we can include categorical and numerical independent variables with main effects and interactions. For example, Groß and Möller (2023) proposed a way to calculate a generalized Cohen's d from linear regression with or without the presence of other variables dividing the estimated parameter with the residual standard deviation. Gelman (2008) proposed to standardize using two standard deviations because when both numeric and categorical variables are included the coefficients are not

on the same metric. Another complication is that in some cases, researchers standardize the dependent variable, the independent variables or both. In general, the unstandardized critical regression coefficient can be usually also be easily interpreted. For example, assuming we are regressing some scores on a cognitive tests using the age of participant,the critical regression coefficient is the minimum increase in the cognitive test for a unit age increase that would be significant.

Meta-analysis

Meta-analysis allows to pool information from multiple studies related to a specific research question. The main advantage of meta-analysis is pooling multiple studies to obtain a more precise and powerful estimation of the effect. From a statistical point of view, a meta-analysis can be considered as a weighted linear regression with heterogeneous variances. Similarly to standard linear regression, hypothesis testing is performed using Wald t or z tests (Borenstein et al., 2009). For this reason, we can apply the same equations used for the linear regression. The main difference and advantage is that the meta-analysis model usually works directly with standardized effect size measures and regression coefficients are already standardized.

Other models

Despite the fact that we discussed only linear models, the same approach could be applied to other types of models such as generalized linear models. In fact, we simply need to multiply the critical value of the chosen distribution (e.g., t or z) by the standard error of the regression coefficient. For example, in a logistic regression with a binary predictor the estimated regression coefficient is the (log) odds ratio. We can obtain the critical odds ratio by multiplying the critical z statistics (the default in R with the `glm` function) with the standard error of the regression coefficient.

Examples in R

In this section, we introduce a user-friendly implementation of the aforementioned mathematical computations as functions of the package “criticalESvalue” in R. The

complete package can be accessed at: https://github.com/psicostat/criticalESvalue. Here, we demonstrate its application through two examples: one example of a t-test on real data and a computation of the critical effect size values for a correlation from sample size. It is important to note that, in general, depending on the researcher's hypotheses, the package allows for the calculation of either two critical effect size values for two-tailed tests or a single critical effect size value for one-tailed tests, for which the direction must be specified. In the Supplementary online materials additional examples can be found for correlation, t-test, paired t-test, linear models and regression coefficients.

First the package should be downloaded and opened with the library function:

```
# require(remotes)
# remotes::install_github("psicostat/criticalESvalue")
library(criticalESvalue)
```

For our examples on real data we used from the package 'psych' the dataset "holzinger.swineford" which has a series of demographics and scores of different subtests measuring intelligence on 301 subjects. Once the package is retrieved with 'library', the dataset can be opened using 'data("name of the dataset")'. For simplicity we decided to rename it with a shorter name.

```
library(psych)
library(psychTools)
data("holzinger.swineford")
Holz <- holzinger.swineford
```

We want to know the critical value for a t-test comparing boys and girls on a cognitive variable of visual perception. In this case, it can be easily done using the 't.test' function:

```
tt <- t.test(Holz$t01_visperc[Holz$female == 1],
             Holz$t01_visperc[Holz$female == 2], var.equal = T)
```

```
critical(tt)
```

```
#>
#>  Two Sample t-test
#>
#> data:  Holz$t01_visperc[Holz$female == 1] and Holz$t01_visperc[Holz$female == 2]
#> t = 1.4062, df = 299, p-value = 0.1607
#> alternative hypothesis: true difference in means is not equal to 0
#> 95 percent confidence interval:
#>  -0.06472157  0.38876046
#> sample estimates:
#> mean of x mean of y
#>  4.314090  4.152071
#>
#> |== Effect Size and Critical Value ==|
#> d = 0.1621764 dc = ± 0.2269606 bc = ± 0.226741
#> g = 0.1617692 gc = ± 0.2263908
```

The output gives a wide range of values: the Cohen's d calculated on the data (d), the critical Cohen's d (dc), the numerator of the formula for the critical Cohen's d (bc), the Cohen's d adjusted for small samples (g) and the critical Cohen's d adjusted for small samples (gc). The variance is set to be equal, but if that is not the case of your data, you can also run Welch two sample t-test and obtain the critical effect size values for that.

In the next example we will show the use of the package's function `critical_cor` to calculate the critical effect size values for a correlation in a prospective framework.

```
n <- 60
critical_cor(n = n, hypothesis = "two.sided", test = "z")
```

```
#> $rc
#> [1] 0.2539247
#>
```

```
#> $rzc
#> [1] 0.2596036
#>
#> $df
#> [1] 58
#>
#> $se_r
#> numeric(0)
#>
#> $se_rc
#> [1] 0.1270027
#>
#> $se_rzc
#> [1] 0.1324532
#>
#> $test
#> [1] "z"
```

The direction of the hypothesis and the test to apply, either t-test or z-test, should be specified. The output will return the critical correlation value(s), the degrees of freedom and the type of test used.

Discussion

With the present article, we propose that researchers compute and report the "critical effect size value(s)" in their empirical articles. This is not intended to replace other strategies aimed at enhancing the NHST approach to inference. Such strategies, such as the emphasis on estimating effect sizes with confidence intervals (Transue, 2019) or the *a priori* planning for statistical power are valuable in their own right. Instead, our proposal serves as a complementary tool, especially beneficial for facilitating the interpretation of results when statistical power deviates from an optimal level (typically falling below, but occasionally exceeding it). Interestingly, critical effect size values can be retrospectively

applied even to already published studies. This possible application facilitates potential reframing of the original interpretations. Serving as a tool for retrospective analysis, critical effect size values may enable a reconsideration of the relevance of previously reported findings.

An advantage of reporting critical effect size values is that they can be precisely computed in any scenario, without requiring assumptions about the expected effect size, as is the case with power calculations. The critical effect size value represents a directly interpretable benchmark that is especially useful in situations where statistical power is below the desired level and researchers are left otherwise uncertain about how to proceed with the interpretation of a study findings. For example, let us say that we read a published article reporting some effects as statistically significant, while others as not: we suspect that the study may be underpowered, but we are widely uncertain about the magnitude of possible true effects. To what extent can the reported results be interpreted, precisely? Knowing the critical effect size value provides us with a clear benchmark. Conversely, let us say that an effect achieves significance in a very large sample: researchers tend to draw substantive conclusions based on this. But is it of real theoretical relevance? If in comparing two groups, such as controls versus treatment, any Cohen's $d > 0.07$ would reach significance, is statistical significance enough to signal a "successful" treatment? Maybe yes, even if effects are tiny (e.g., Funder & Ozer, 2019), but knowing the critical effect size value certainly prompts some appropriate interpretive caution.

Reporting the critical effect size value(s) can also be an efficient way to allow researchers to evaluate which findings are statistically significant. For example, in a correlation table researchers customarily add an asterisk to all statistically significant correlations. But as long as all correlations are based on the same sample size, researchers can simply remark 'the critical effect size is $r = 0.3$' and readers will know that all correlations larger than this value are statistically significant.

Beyond enhancing study design and statistical inferences based on hypothesis tests, reporting critical effect size values can also serve an educational purpose. It underscores how the distinction between a significant and non-significant result is not solely determined by the presence or absence of a true effect, but also by the sample size. By highlighting a critical effect size value, researchers can become more aware of the possibility of Type 2 errors when results are non-significant. Conversely, in studies with exceedingly large samples and in many meta-analyses, the critical effect size value(s) may serve as a reminder that any observed effect larger than a trivially small value will likely achieve significance. This emphasizes that the mere attainment of statistical significance in a test is not particularly surprising, especially in non-experimental studies.

Real-case scenarios may not always be that simple. Hence, we chose to expand the application of computing critical effect size values beyond Cohen's d and correlation to include linear regression with both raw and standardized coefficients and meta-analysis. This serves as a first step in computing critical effect size values for a wider array of effects encountered in practical scenarios, where linear models and their extensions are commonly utilized for modeling purposes. A prerequisite is that researchers must be able to identify what parameters in their statistical models reflect the effect sizes of interest, and that they are able to assess their relevance. Notably, however, this prerequisite aligns with the requirements of APA style guidelines concerning the reporting of effect sizes. For further illustration and application, additional examples are provided in the Supplementary online material.

We suggest that reporting critical effect size values is particularly valuable when sample size planning was not feasible or did not occur *a priori*. In cases where optimal power can be attained with a sufficiently large sample size for an effect of a specific magnitude of interest, and this is truly determined *a priori*, the interpretation of both significance and non-significance becomes straightforward. However, when power analysis

did not inform the sample size or when power is likely but undeterminedly low, reporting critical effect size values for the obtained sample can help provide context for interpretation. Critical effect size values can be computed and interpreted even retrospectively or for studies that have already been published.

In conclusion, reporting critical effect size values in empirical articles serves as a valuable addition to researchers' toolkit, aimed at augmenting transparency and facilitating the interpretability of their findings. While not designed to supplant existing practices, it provides a useful aid in interpreting newly presented and previously published results, thus advancing the understanding of research outcomes.

References

Altoè, G., Bertoldo, G., Zandonella Callegher, C., Toffalini, E., Calcagnì, A., Finos, L., & Pastore, M. (2020). Enhancing statistical inference in psychological research via prospective and retrospective design analysis. *Frontiers in Psychology*, *10*, 499756.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. https://doi.org/10.1002/9780470743386

Bowley, A. L. (1928). The standard deviation of the correlation coefficient. *Journal of the American Statistical Association*, *23*, 31. https://doi.org/10.2307/2277400

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., Kominsky, J. F., Kosie, J. E., Lew-Williams, C., Liu, L., et al. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 349.

Chow, S. L. (1988). Significance test or effect size? *Psychological Bulletin*, *103*(1), 105.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*(2), 161.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507. https://doi.org/10.2307/2331838

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*.

Funder, D. C., & Ozer, D. J. (2019). Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, *2*(2), 156–168.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine*, *27*, 2865–2873. https://doi.org/10.1002/sim.3107

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual. *The Sage Handbook of*

Quantitative Methodology for the Social Sciences, 391–408.

Groß, J., & Möller, A. (2023). A note on cohen's d from a partitioned linear regression model. *Journal of Statistical Theory and Practice*, *17*. https://doi.org/10.1007/s42519-023-00323-w

Hagen, R. L. (1997). *In praise of the null hypothesis statistical test.*

Haig, B. D. (2017). Tests of statistical significance made sound. *Educational and Psychological Measurement*, *77*(3), 489–506.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational Research Association and the American Statistical Association*, *6*, 107–128. https://doi.org/10.3102/10769986006002107

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61–85.

Jarke, H., Anand-Vembar, S., Alzahawi, S., Andersen, T. L., Bojanić, L., Carstensen, A., Feldman, G., Garcia-Garzon, E., Kapoor, H., Lewis, S., et al. (2022). A roadmap to large-scale multi-country replications in psychology. *Collabra: Psychology*, *8*(1), 57538.

Kramer, A. D., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(24), 8788.

Krueger, J. (2001). Null hypothesis significance testing: On the survival of a flawed method. *American Psychologist*, *56*(1), 16.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 62627.

Lakens, D. (2021). The practical alternative to the p value is the correctly used p value. *Perspectives on Psychological Science*, *16*(3), 639–648.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), 33267.

Miller, J. (2017). Hypothesis testing in the real world. *Educational and Psychological Measurement*, *77*(4), 663–672.

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.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., et al. (2018). The psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241.

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110. https://doi.org/10.2307/3002019

Sirois, S., Brisson, J., Blaser, E., Calignano, G., Donenfeld, J., Hepach, R., Hochmann, J.-R., Kaldy, Z., Liszkowski, U., Mayer, M., et al. (2023). The pupil collaboration: A multi-lab, multi-method analysis of goal attribution in infants. *Infant Behavior and Development*, *73*, 101890.

Szucs, D., & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology*, *15*(3), e2000797.

Transue, B. (2019). *APA style 7th edition.*

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. *Journal of Educational and Behavioral Statistics: A Quarterly Publication Sponsored by the American Educational*

Research Association and the American Statistical Association, *32*, 39–60. https://doi.org/10.3102/1076998606298034

Weiss, A. H., Biron, T., Lieder, I., Granot, R. Y., & Ahissar, M. (2014). Spatial vision is superior in musicians when memory plays a role. *Journal of Vision*, *14*(9), 18–18.

Welch, B. L. (1947). The generalization of "student's" problem when several different population variances are involved. *Biometrika*, *34*, 28. https://doi.org/10.2307/2332510

Wilson, B. M., Harris, C. R., & Wixted, J. T. (2020). Science is not a signal detection problem. *Proceedings of the National Academy of Sciences*, *117*(11), 5559–5567.