JOURNAL OF NEW APPROACHES IN EDUCATIONAL RESEARCH
2022, VOL. 11, NO. 2, 209-225, e-ISSN: 2254-7339
https://doi.org/10.7821/naer.2022.7.930

ORIGINAL

Comparing Effect Sizes and their Confidence Intervals: A Primer on Equivalence Testing in Educational Research

Hector F. Ponce-Renova

Department of Humanities, Autonomous University of Ciudad Juarez, Mexico

ABSTRACT

This paper's objective was to teach the Equivalence Testing applied to Educational Research to emphasize recommendations and to increase quality of research. Equivalence Testing is a technique used to compare effect sizes or means of two different studies to ascertain if they would be statistically equivalent. For making accessible Equivalence Testing, this technique was explained with two examples by conducting manual calculations, using an online calculator, the software *R*, the software SPSS, and a *t* table. Furthermore, the software R with an Equivalence Testing code was used, and its results were graphed and discussed with details. Among other recommendations given, Equivalence Testing can be a useful tool for comparing means and effects within certain bounds that could hopefully imply a practical significance to provide meaning to findings. The results of Equivalence Testing can indicate that two treatments´effects are statistically equivalent or not. Thus, the Equivalence Testing can be a channel to replicate studies and observe if there is a possible pattern regarding the appearance of a phenomenon.

Keywords CONFIDENCE INTERVAL, COHEN'S D, EFFECT SIZE, EQUIVALENCE TESTING, PRACTICAL SIGNIFICANCE

Received 2021-09-23
Revised 2021-10-13
Accepted 2021-11-15
Published 2022-07-15

Corresponding Author
Hector F. Ponce-Renova,
hector.ponce@uacj.mx

Edificio W, Cubiculo D, Avenida Plutarco Elías Calles #1210, Fovissste Chamizal, CP 32310, Ciudad Juárez, Chihuahua, México.

DOI https://doi.org/10.7821/naer.2022.7.930
Pages: 209-225

1 INTRODUCTION

In experimental designs, researchers need references to compare their treatments' effect sizes (ES) and averages against those which others have encountered to provide meaning to their findings: *statistical equivalence* tests are the modus operandi. Researchers have already found the magnitude of their results' *practical significance* (i.e., "the extent to which a study result has meaningful applications in real-world settings" (VandenBos, 2015, p. 817) and they would also like to know where these results stand statistically. Nevertheless, Cohen (1965) warned against using fixed and general criteria to make inferences about results, mentioning three different effect sizes: small, medium and large as "a common conventional frame of reference which is recommended for use only when no better basis for estimating the ES index is available" (p. 25). Another critique for fixed and general criteria was that

"God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277). Thus, researchers should be responsible for basing their arguments on criteria (e.g., comparisons with others' results) and practical significance instead of using a fixed general standard.

To provide meaning to findings, this present paper demonstrates how to use *Equivalence Testing* (ET) for comparisons with *one single sample* and with *two independent groups*. Equivalence testing is defined as a simple statistical technique for determining whether one should reject the presence of an effect at least as extreme as the SESOI (i.e., Smallest Effect Size of Interest; Lakens et al., 2018). In other words, the objective of ET can be said to be demonstrating that two treatments have the same effect "or at least close enough to be considered similar beyond a reasonable doubt" (De Muth, 2019, p. 184).

This paper's *objective* is to exemplify how to execute an ET by manual calculations with a *t* table and an online calculator as well as with the *R* software. Additionally, the procedures and results of De Muth (2019) and Lakens et al. (2018), who utilized Minitab and *R* for ET respectively, are shown. Overall, this paper answers the question: How can we execute an ET?

Here, the gap in the relevant literature is identified and explained. First, a definition of the term *Educational Research* as a contribution to knowledge for improving the collective understanding of education was taken from (Gall, Gall, & Borg, 2007). Second, on December 29, 2020, the terms *Educational Research* and *Equivalence Testing* were searched in *Google Scholar* and approximately 16,700 results appeared. A detailed inspection of these documents showed that, strictly speaking, there were no publications covering the aforementioned terms together. Therefore, it can be concluded that there is be a gap in the literature that this paper tries to fill by using the definition given by Gall et al. (2007) about improving the collective understanding of education. As such, this study's contribution is to use ET in Educational Research. In contrast to this research area, pharmacology has seen the application and publications of ET for decades related to analysing the effects of different drugs (De Muth, 2019; Schuirmann, 1987). Furthermore, in psychology, Lakens et al. (2018) has published some articles on ET.

2 STATEMENT OF THE PROBLEM

Thompson (2008) reported that researchers use effect sizes to explain the *practical significance* of their findings as well as to evaluate the *replicability* of their results: i.e. comparison between their own effect sizes and others' to notice a possible pattern to generalize. Defining *replication*, Vandenbos (2015) wrote that it is, "the repetition of the original experiment or research study to verify or bolster confidence in its results" (p. 906). For practical significance, Cohen (1965) stated that the primary product of research is an *effect size* and it is not a *p*-value. Regarding a definition, an effect size can be described as "the degree in which a phenomenon is present in the population or the degree to which the null hypothesis is false" (Cohen, 1965, pp. 9-10). In addition, Cumming (2012) explained that an effect size might help to communicate, "to a wide range of readers, especially when the original units first used to measure the effect are not widely familiar" (p. 282). In addition, Cumming

and Calin-Jageman (2017) reported that, "We calculate from our data the sample *effect size* (*ES*) and use this as our estimate of the *population ES*, which is typically what we would like to know" (p. 111). These last authors (2017) added, "A population effect size is the true value of an effect in the population" (p. 2011). One of the additional reasons to use effect sizes is that they offer more precise information to theories and practice (cf. Plonsky & Oswald, 2014).

Additionally, a confidence interval can be described as: "An interval estimate calculated from sample data that indicates the precision of a point estimate" (Cumming, 2012, p. 439). Multiple authors (e.g., Wilkinson & the APA Task Force on Statistical Inference, 1999) have recommended and requested the reporting of effect sizes and confidence intervals alongside the Null Hypothesis Statistical Significance Testing results (NHSST). Even so, Cumming and Maillardet (2006) estimated that 95% *CI* can capture 84.3 % of the true value, while Kelley and Rausch (2006) commented that it would be misleading to report estimates without providing the uncertainty surrounding them.

2.1 *p*-values

Another related issue of confidence intervals is the *p*-value. The American Psychological Association (2020) explained that "complete reporting of all tested hypotheses and estimates of appropriate effect sizes and confidence intervals are the minimum expectation for all APA journals" (p. 87). It is beyond this paper's objectives to discuss *p*-values in depth, but Wasserstein and Lazar (2016, pp. 131-132) addressed six related ideas and some are misconceptions about *p*-values.

In addition, *p*-values were conceptualized for the long run and not only for *one* occasion, as was done in several studies (cf. Greenland et al., 2016). The *p*-value is the probability of observing a test statistic's value (e.g., *t* or *F* calculated value) or smaller in a distribution, given a true null hypothesis (LeMire, 2010, p. 8). In support of using the *p*-value, Shi and Yin (2020) affirmed, "Although *p*-value is often regarded as an inadequate representation of statistical evidence, it has not stalled the scientific advancement in the past years" (p. 3). A further relevant view was that NHSST "is a small but important part of the entire research" (LeMire, 2010, p. 2).

2.2 Experimental Designs

Confidence intervals and *p*-values are part of experiments. Gall et al. (2007) stated that the experiment is the most powerful quantitative research method to establish cause and effect relationships between two or more variables. Many educational research studies do not comply with the strict definition of experiment, but they do have an *Experimental Design*. This is a plan of the procedures to be followed in scientific experimentation to reach valid conclusions, with considerations of such factors as selection of participants, manipulation of variables, data collection and analysis, and minimization of external influences (VandenBos, 2015, p. 397). Furthermore, Serlin and Lapsley (1993) explained that the typical point null hypothesis is false. For example, when two means are compared, it would be unlikely to find a difference equal to zero. Another characteristic of NHSST is that since the null

hypothesis is sensitive to sample size, large enough sample sizes will always result in its rejection (Serlin & Lapsley, 1993).

2.3 Important Aspects of Cohen's *d*

Cohen's *d* is, "A standardized ES expressed in units of some appropriate *SD*. It can often be considered a kind of a *z* score" (Cumming & Calin-Jageman, 2017, p. 532). Moreover, Cohen (1965) affirmed that the coefficient *d* responds to the question: "How large is the effect?" (p. 20). In an experimental design, *d* can be interpreted as the magnitude of the effect caused by a treatment, other things being equal. This difference can be between a calculated mean and a population's mean (i.e., target value), between two independent sample's means (between groups), and between the same groups (pre and post-test, within groups).

Regarding the sampling distribution of *t*, Cumming (2012) stated that, "*d* is also distributed as a non-central *t*" (p. 298). Incidentally, 1974 demonstrated how to calculate non-central *t* distribution, though it is beyond this study's objectives to cover it here. A warning should be given for this effect size; that Cohen's *d* is "measured with a rubber ruler that stretches and contracts as we take successive samples" (Cumming, 2012, p. 298). Moreover, Cumming (2012, p. 283) explained that while *d* is sensitive to the numerator, it is very sensitive to the denominator (*SD* used as the standardizer). Furthermore, Cohen's *d* depends on the mean and a non-robust measure of dispersion, which is an *SD*, so *d* is a non-robust measure of effect size (Garstats & Garstats, 2018) .

2.4 Equivalence Testing

Sometimes, the traditional NHSST has been used to infer that two means from different studies are statistically equivalent when the null hypothesis has not been rejected. This inference is *not* correct because NHSST was designed to test significance differences but not equivalence per se (De Muth, 2019). Moreover, Wellek (2010) explained that non-significant differences must not be confused with significance homogeneity or, as Altam and Bland (1995) put it, "absence of evidence is not evidence of absence" (p. 3). Thus, the appropriate technique to infer equivalence between two means or effects is Equivalence Testing, so when a replication of a study takes place, this is the correct procedure.

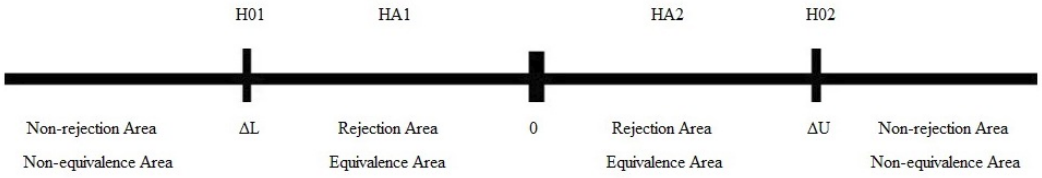
ET has been used in pharmacology to statistically establish the equivalence of an effect of treatments or drugs on patients (cf. De Muth, 2019). Furthermore, an effect size or a mean, considered best practice due to its *practical significance*, can be used as a criterion for comparison with another to observe whether there is a statistically significant difference between them.

Further to the above, Wellek (2010) provided a definition about this testing: "equivalence means here equality except for practically irrelevant deviations" (p. 1): i.e., the difference between two populations' means ($\mu_1 - \mu_2$) is within two limits: i.e., Lower Equivalence bound (Δ_L) and the Upper Equivalence bound (Δ_U). Briefly, the aforementioned definition of statistically equivalent means can be summarized thus: Lower Equivalence Bound (Δ_L) < $\mu_1 - \mu_2$ < Upper Equivalence Bound (Δ_U). If this happens, then equivalence is

proven (De Muth, 2019) . These Equivalence Bounds have been given different names in the literature, but the meaning has been the same. For example, Wellek (2010) called this range the *indifference zone* within the aforementioned bounds and named their limits critical bounds, where the lower one < 0 and the upper one > 0. Incidentally, Lakens (2017) named these limits the *smallest effect size of interest* (SESOI). Given these different names, and considering pharmacology's use of this testing and what it could be more intuitive, it seemed that the most appropriate name would be Lower Equivalence Bound (Δ_L) and Upper Equivalence Bound (Δ_U). Regarding Equivalence Bounds, there is a cautionary note that these limits were made by *experts* in the fields and not by statisticians (Hauck & Anderson, 1984). Thus, this can be interpreted as a range of practical significance, which is what is important.

In other words, the difference between an observed effect and a target effect or difference between averages is tested to see if the difference is statistically larger than the Δ_L and smaller than the Δ_U. Additionally, these observed effects and differences between means are not only point estimates, but a confidence interval is also calculated (see De Muth, 2019; Lakens et al., 2018; Wellek, 2010). Confidence Intervals are explained below with examples.

An ET has two one-sided tests (TOST). In brief, Figure 1 shows what the rejection areas, which involve two alternative hypotheses (H_{A1} and H_{A2}) and the non-rejection areas (i.e., two null hypotheses: H_{01} and H_{02}), are. When the differences between effect sizes (i.e., in standard scores) or averages (in raw scores) of an original study and a replication study are bigger than the Lower Equivalence Limit (Δ_L), the null (H_{01}) is rejected. Moreover, when the difference is smaller than the Upper Equivalence Limit (Δ_U), the conclusion is that the other null (H_{02}) is rejected. Given these hypotheses, ET is designed to test whether a significant difference does *not* exist between two means. This is in contrast with the NHSST, which is designed to examine whether a significant difference exists.



H01: Difference between ES or mean original study and ES or mean replicated study ≤ Lower Equivalence Limit (ΔL). This implies $p \geq \alpha$.

HA1: Difference between ES or a mean original study and ES or a mean replicated study > Lower Equivalence Limit (ΔL): $p < \alpha$.

H02: Difference between ES or a mean original study and ES or a mean replicated study ≥ Upper Equivalence Limit (ΔU): $p \geq \alpha$.

HA2: Difference between ES or a mean original study and ES or a mean replicated study < Upper Equivalence Limit (ΔU): $p < \alpha$.

Figure 1 ET: Two One-Sided Tests (TOST)

Like the NHSST, ET starts from two different populations, from which two random samples are taken and tested to see if an inference can be drawn about statistically equivalent means.

2.5 Possible Outcomes of Equivalence Testing

Delacre et al. (2017) used Welch's *t*-test instead of the traditional Student's *t* test because they claimed the former analysis was better to control Type I Error rates when homogeneity of variance was not met, but this test loses robustness compared to the Student's *t*-test when the assumptions (homogeneity and normal distribution) were met. Using the *R* code by Lakens et al. (2018), ET and NHSST can be performed at the same time and, as part of the output, a graph with confidence intervals is provided. Lakens (2017) explained that there were *four possible outcomes* (Scenarios A, B, C, and D) when the NHSST and ET are used jointly, applying confidence intervals of 95% (thinner black lines) and 90% (thicker black Lines; Figure 2), respectively:

- Scenario A. The meanswere statistically equivalent ($\Delta_L < \mu_1 - \mu_2 < \Delta_U$) and, in the NHSST, the difference was not statistically different from 0: statistically equivalent (Rejection of the Null) and did not differ from 0 (Not Rejection of the Null).
- Scenario B. The meanswere not statistically equivalent ($\Delta_L > \mu_1 - \mu_2 < \Delta_U$), and, in NHSST, the difference was statistically different from zero: not statistically equivalent (not Rejection of the Null) and statistically different from 0 (Rejection of the Null).
- Scenario C. The meanswere statistically equivalent ($\Delta_L < \mu_1 - \mu_2 < \Delta_U$), and, in NHSST, the difference was statistically different from zero: statistically equivalent (Rejection of the Null), and statistically different from 0 (Rejection of the Null).
- Scenario D. The meanswere not statistically equivalent ($\Delta_L > \mu_1 - \mu_2 < \Delta_U$), and, in NHSST, the difference was not statistically different from zero: not statistically equivalent (not Rejection of the Null) and statistically different from 0 (not Rejection of the Null).

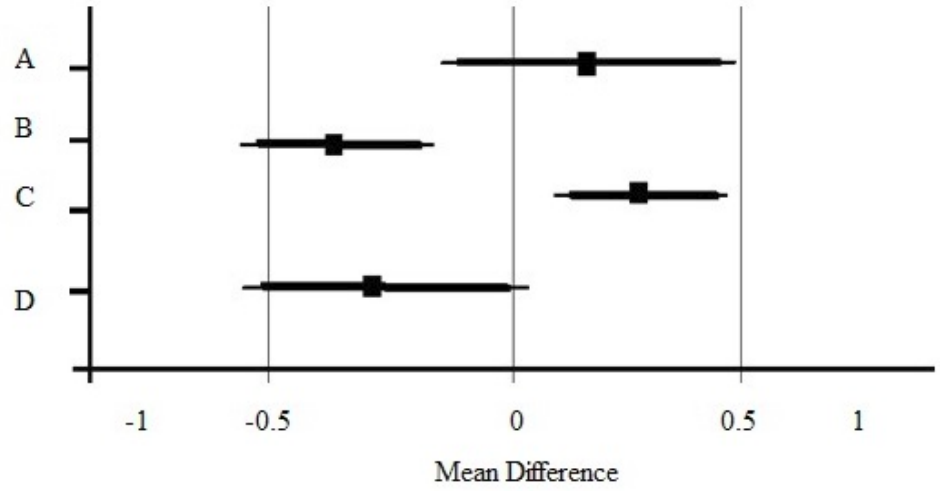
Lakens (2017) corrected the degrees of freedom for Welch's *t* test of ET according to Sattherthwaite (1946) and used Welch's *t* instead of Student's *t* test (see Ponce-Renova, 2021, for more about *t* test). In contrast to the traditional Student's *t*-test, Delacre, Lakens, and Leys (2017) claimed that Welch's *t* test would be a better option.

2.6 Statistical Power

Ioannidis (2005) issued a warning by arguing that medical research was false due to lack of sufficient statistical power, first and foremost. Two *p* values for two *t* tests (TOST) are used in ET, so statistical power should be taken into consideration. Statistical Power is the probability of rejecting a false null hypothesis (see Cohen, 1965; Ponce-Renova, 2019). Statistical Power has a positive relationship with effect size, alpha level and sample size. Therefore, a recommendation for researchers who are planning to use ET, would be to find out whether 80% power is feasible in their field of study (see Cohen, 1965; Ponce-Renova, 2019).

3 EXAMPLES

This section presents several examples of how to run equivalence testing. Given the equations for estimating the Standard Error for the difference between means ($SE_{difference}$),



Note: A, B, C, and D represent the scenarios.

Figure 2 Scenarios of ET and NHSST. Source: Lakens (2017, p. 357).

a bigger sample size would reduce this $SE_{difference}$ (e.g. Equation 1 because of a bigger denominator) and as a consequence the tails of confidence intervals would be reduced (e.g. Equations 2 and 3 due to a smaller multiplier) and the *t* calculated values would be increased (e.g. Equations 4 and 5 as a result of a smaller denominator). The end result is that a bigger sample size would have higher probabilities of producing *statistically significant results*, other things being equal. Moreover, as the sample increases, the effect size's variation is less (Schönbrodt & Perugini, 2013). With small sample sizes, it is not possible to conclude an absence of an effect size when $p > \alpha$ because of low power to detect a true effect (Lakens, 2017, p. 355). For simplicity, the following examples did not consider statistical power and the sample sizes were relatively small.

3.1 Example 1: One Single Sample Test After a Pre and Post Test

For the following *One single sample example*, the data and Equivalence Bounds were taken from De Muth (2019), who performed this ET with Minitab (his results contrasted with the present paper's). The present example has contextualized ET in educational research with an experimental design. This problem was solved through manual calculations using Critical Values of the *t* distribution table, as well as using an online calculator to estimate *p* values from *t* values (Science Statistics, 2020) and SPSS 24 for the NHSST.

After treatment of the experimental design with pre and post-tests to improve mathematical information, the researcher wanted to compare the findings (post-test scores) to a target value (Population's mean: μ = 100 points). In brief, a sample's mean ($\bar{x}$= 99.85) was obtained from a set of six students' post-test scores (99.6, 100.2, 98.3, 99.9, 100.4, and 100.7 with a *SD* = 0.85).

Before the Equivalence testing, a One sample *t* test was performed in SPSS 24 for the NHSST (μ =100 vs. $\bar{x}$= 99.85): the results showed $t_{calculated}$ (df = 5) = -0.432, p = .684, standard error of the mean ($SE_{difference}$) = 0.347, with a mean difference = -0.15 and $CI_{95\%}$ for the difference was [-1.0423, 0.7423]. Given an alpha = .05, there was not a statistically significant difference between the sample's mean and the population's one. However, this result of no statistically significant difference does *not* imply statistical equivalence per se (cf. De Muth, 2019).

For practical significance (this would be established by experts and not by statisticians), it was considered that a range of 1.5 points from the population's mean of 100 points would be statistically equivalent: Δ_L = -1.5 and Δ_U = 1.5. The research question was: Does statistically significant equivalence exist between $\bar{x}$ and μ? The first step was to establish the hypotheses and the alpha = .05 for each *t* test (the TOST procedure has two *t* tests):

1.- Hypotheses:

$$\mathrm{H}_{01} : \bar{x} - \mu \leq \Delta\mathrm{L}$$
$$\mathrm{H}_{\mathrm{A}1} : \bar{x} - \mu > \Delta\mathrm{L}$$
$$\mathrm{H}_{02} : \bar{x} - \mu \geq \Delta\mathrm{U}$$
$$\mathrm{H}_{\mathrm{A}2} : \bar{x} - \mu < \Delta\mathrm{U}$$

$\bar{x} - \mu =$ observed average $-$ Target average (the Target average can be found in the literature or other reliable source).

De Muth (2019) has showed several steps to develop an ET that were followed to a certain extent in this present study.

2.- Establish the difference: Difference = $\bar{x} - \mu$; $(99.85 - 100 = -0.15)$.

3.- Find or calculate the Lower and the Upper Equivalence Bounds (Lakens et al., 2017, explained several methods to estab- lish these limits, the main goal of which is to achieve practical significance for certain field). These can be expressed in raw or standardized form. For the simplicity of the example, here these bounds are:

$$\text{Lower} - \text{Equivalency Bound} : \Delta_{\mathrm{L}} = -1.5$$
$$\text{Upper} - \text{Equivalency Bound} : \Delta_{\mathrm{U}} = 1.5$$

4.- Degrees of freedom: $n - 1; df = 6 - 1 = 5$

5.- Standard error of the difference (*SE*):

$$SE_{d\,\text{ifference}} = SD/\text{ square root of } n \qquad \text{(Equation 1)}$$

Substituting, $SE_{d\,\text{ifference}} = 0.85/\text{ square root of } 6 = 0.35$

6.- Reliability Coefficient, also known as critical t : $100(1-\alpha)$; Given the aforementioned alpha = .05, $t_{\text{critical}}\,(df = 5) = 2.015$ (Taken from a t-table for one tail with a significance level of 5%).

7.- Two one-sided tests (confidence intervals with 5% per tail):

$$\text{DifferenceLower} = (\bar{x} - \mu) - \text{t}_{l-\alpha,v}\text{X}SE_{\text{difference}} \quad \text{(Equation 2)}$$

Substituting, –0.15 – 2.015 (0.35) = –0.855

$$\text{DifferenceUpper} = (\bar{x} - \mu) + \text{t}_{1-\alpha,v}\text{X}SE_{\text{difference}} \quad \text{(Equation 3)}$$

Substituting, –0.15 + 2.015 (0.35) = 0.555

8.- Ratios of t-statistic provide two calculated *t* values and their corresponding *p* values: one for the probability of exceeding the Δ_{L} and one for the probability of exceeding the Δ_{U} (*t* calculated):

$$t_{\text{lower limit}} = [(\bar{x} - \mu) - \Delta_{\text{L}}]\,/SE_{\text{difference}} \quad \text{(Equation 4)}$$

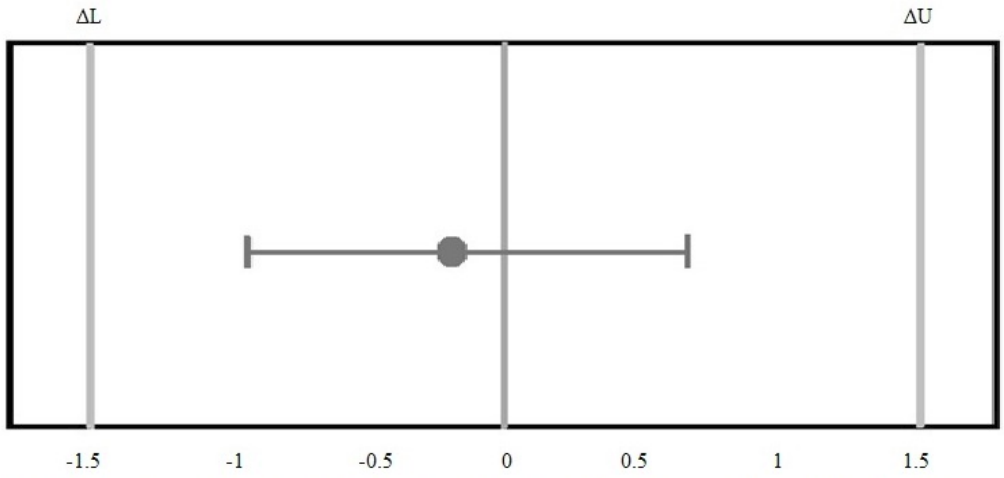
Substituting, t lower limit $[-0.15 - (-1.5)]/0.35 = 3.86$;*p* = .006 (with an $\alpha = .05$ and one tail, for calculating the *p*-value; see the calculator on Social Science Statistics, 2020).

$$t \text{ upper limit} = [(\bar{x} - \mu) - \Delta\text{U}]/SE_{\text{difference}} \quad \text{(Equation 5)}$$

Substituting, t upper limit $[-0.15 - (1.5)]/0.35 = -4.71; p = .003$.

Using Minitab, De Muth (2019, p. 182) reported the $CI_{95\%}$ = [-0.849486, 0.549486], but this $CI_{95\%}$ corresponded to the $CI_{90\%}$ [-0.8495, 0.5495], which was calculated in SPSS 24, as well as the one calculated here manually. Thus, there was a contradiction in calling this $CI_{95\%}$ when it was actually a $CI_{90\%}$. That is, the $t_{critical}$ (df = 5) was 2.015 (i.e., Level of significance for two tailed tests at 10%, and Level of significance for one tailed test at 5%), and this level of confidence resulted in a $CI_{90\%}$, which had the same values as the $CI_{95\%}$ from De Muth (2019). To calculate a $CI_{95\%}$, the $t_{critical}$ (df = 5) had to be 2.5706 (Level of significance for two tailed tests at 5%, and Level of significance for on tailed test at 2.5%), so the *CI* would have been 95%. This was the reasoning to opt for the $CI_{90\%}$: since the TOST implies two t tests at the same time and each one has an alpha = .05, the lower tail of the *CI* has an alpha of 5% plus the upper one that has another 5%, so if they are added up the result is 10% (cf. Lakens et al., 2018). However, in order to be consistent with De Muth's (2019) results, the $CI_{90\%}$ was observed.

In conclusion, the null hypotheses were rejected because *p* values < α. Therefore, there was a statistically significant equivalence with a 90% confidence interval between the sample and population means (cf. De Muth, 2019). The $CI_{90\,\%}$ for the difference was [-0.855, +0.555], with t_{lower} = 3.86 (p = .006) and t_{upper} = –4.71 (p = .003). In other words, the sample's $CI_{95\%}$ mean difference [–0.855, 0.555], which represented a population, was within Δ_L (–1.5) and Δ_U (+1.5), so it can be considered as statistically equivalent to the target of 100 points (see Figure 3; cf. De Muth, 2019).



Note: ΔL = Lower-Equivalence Limit, ΔU = Upper-Equivalence Limit. 90% CI for Equivalence [-0.85, 0.55], which is within the equivalence interval of (-1.5, 1.5).

Figure 3 $CI_{90\%}$ for the Difference.

3.2 Example 2: For Two Independent Groups' Tests After a Pre and Post Test

The ET for *Two independent samples* was intended thus: "To overcome the disadvantage of failing to reject the null hypothesis with *t*-test, the two-sample equivalence was developed to identify the similarity or equivalence between two methods or treatments" (De Muth, 2019, p. 185). When both nulls are rejected, the alternatives are proven, which means that the results are within equivalence limits (De Muth, 2019). Once again, because a null was not rejected, it does not imply equivalence. The data of this last author (Table 1) was used for the following example, which was solved partly by using SPSS 24 and manual calculations with Critical Values of the *t* distribution table and online calculators (**Science Statistics, 2020**), as well as using an *R* code created by Lakens (2017) for Equivalence Testing.

Table 1 Data for the sample of independent groups.

Study A	Study B
96,5	101,1
101,1	100,6
99,1	98,8
98,7	99
97,8	98,7
99,5	100,8

In the experimental design, the scenario was that Researcher A was trying to replicate a study carried out by Researcher B to observe if a treatment for improving *reading* had the

same significance in both studies. Researcher A followed the steps of Researcher B, such as collecting a random sample representative of a population and applying the treatment. Furthermore, Researcher A had access to Researcher B's post-test data to test for equivalence. For practical effect size, two units above and below the difference were considered as the Equivalence Bounds. The research question was: Does statistically significant equivalence exist between these two means?

Before performing the ET, a two-independent-sample *t* test was applied in SPSS 24 for NHSST. Given homogeneity of variance, the results were: Levene's test $F = 0.084$, and $p = .78$. In summary, there was not a statistically significant difference between the means of both groups and the: $t\ (10) = -1.343$, $p = .209$, with a mean difference = -1.05, *SE* difference = 0.7819, and $CI_{95\%}$ [-2.792, 0.692]. Once again, this lack of statistically significant difference does not mean statistical equivalence. In contrast, De Muth (2019) and Lakens et al. (2018) did not assume homogeneity of variance and calculated coefficients accordingly including a $CI_{90\%}$. These authors used $df = 9$ and $df = 9.03$ respectively instead of $df = 10$ supported by homogeneity of variance. Given this, the SPSS 24 was used to calculate results under no assumption of variance homogeneity: $t\ (9.035) = -1.343$, $p = .212$, mean difference = -1.05, *SE* difference = 0.7819 and the $CI_{90\%}$ [-2.4827, 0.38271].

1.- Hypotheses:

$$H_{01} : \mu_A - \mu_B \leq \Delta_L$$
$$H_{A1} : \mu_A - \mu_B > \Delta_L$$
$$H_{02} : \mu_A - \mu_B \geq \Delta U$$
$$H_{A2} : \mu_A - \mu_B < \Delta U$$

If the Nulls are rejected, equivalence is proven.

2.- Establish the difference. Difference =

$\bar{x}_A$ = mean of the replication study by Researcher A: 98.78 (*SD* = 1.56)

$\bar{x}_B$ = mean of study B by Researcher B: 99.83 (*SD* = 1.11)

3.- Find or calculate:

Lower-Equivalence Bound: Δ L = –2. This value was translated into a *d* value to be included in the *R* code of Lakens et al. (2018) by dividing the Lower Equivalence-Bound, which was taken as a difference between two means, by the pooled *SD**: –2 / 1.35 = –1.48 Upper-Equivalence Bound: Δ U = 2 translated into a *d* value as the previous one: 2 / 1.35 = 1.48

4.- Degrees of freedom. $n\ 1 + n\ 2 - 2$: 12 – 2 = 10 (This was under the assumption of equal variance, but De Muth, 2019, and Lakens et al., 2018, used $df = 9$ and $df = 9.03$, respectively, so for coherence with their approaches, the rounded degrees of freedom were 9 for the present manuscript.

5.- Standard error of the difference ($SE_{difference}$):
$SE_{difference}$ = square root of (2 X S s^2 / n) (for equal sample sizes):

$$SE_{\text{difference}} = \text{square root of } \left[\left(s_s^2/n_1 + s_s^2/n_2\right)\right] \text{ (for equal sample sizes)}$$

$$s_s^2 = \text{ variance of both samples.}$$

That is,

$$s_s^2 = \left[\Sigma\left(X_{i1} - \overline{X}_1\right)^2 + \Sigma\left(X_{i2} - \overline{X}_2\right)^2\right]/n_1 + n_2 - 2 = 1.83$$

$$\text{Substituting, } SE_{\text{difference}} = \text{ square root of } (1.83/6 + 1.83/6) = 0.78$$

6.- Reliability Coefficient, also known as critical t: 100(1 – α) = α = .05, $t_{critical}$ (df = 9) = 1.8331 (Taken from the t-table for two tails, significance level = 10%)

7.- Two one-sided tests (confidence interval for the difference: TOST confidence interval 90%):

$$\text{Difference Lower Limit } = \left(\overline{X}_A - \overline{X}_B\right) - t_{l-\alpha,v} X SE_{\text{difference.}}$$

$$\text{Substituting, } = -1.05 - 1.8331(0.78) = -2.4798$$

$$\text{Difference Upper Limit } = \left(\overline{X}_A - \overline{X}_B\right) + t_{l-\alpha,v} X SE_{\text{difference.}}$$

$$\text{Substituting, } = -1.05 + 1.8331(0.78) = 0.3798$$

Thus, the $CI_{90\%}$ = [–2.4798, 0.3798], which differs to a certain extent from the ones calculated in SPSS 24 $CI_{90\%}$ [–2.4827, 0.38271], De Muth $CI_{95\%}$ [–2.4833, 0.3833] and Lakens et al. (2018) $CI_{90\%}$ = [–2.482, 0.382]. These authors differed in calling the CI 95% and 90%, which was discussed previously. The four TOST CI coincided in that the lower limit (i.e. ≈ –2.4798) went beyond the Lower Equivalence Bound (–2), so there was no equivalence.

8.- Ratios t-statistic (t calculated):

$$t \text{ Lower Limit } = \left(\overline{X}_A - \overline{X}_B - \Delta_L\right)/SE \text{ difference.}$$

Substituting,

[–1.05 – (-2)] / 0.78 = 1.22; $p_{Value\ Lower\ Limit}$ = .25 (see Social Science Statistics, 2020, for calculating a p val- ue from a t value).

$$t \text{ Upper limit } = \left(\overline{X}_A - \overline{X}_B - \Delta U\right)/SE_{\text{difference}}$$

$$\text{Substituting,} [-1.05 - (2)]/0.78 = -3.91; p \text{ Value Upper Limit } = .003$$

De Muth (2019) reported $t_{Value\ Lower\ Limit}$ (9) = 1.215 (p = .128) and $t_{Value\ Upper\ Limit}$ (9) = –3.9007 (p = .002). Moreover, using the *R* code of Lakens et al. (2018): $t_{Value\ Lower\ Limit}$ (9.03) = 1.22 ($p_{Value\ Lower\ Limit}$ = 0.127; and $t_{Value\ Upper\ Limit}$ (9.03) = –3.91 ($p_{Value\ Upper\ Limit}$ = 0.002).

Using Lakens et al.'s (2018) *R* code, the results presenting the NHSST with $CI_{95\%}$ were very similar to the ones calculated manually here, as well as to De Muth's (2019). In comparison to the manually calculated results, this code's results were easy to operate, and required (Table 2 in Bold) the means, standard deviations, and sample sizes of both independent groups, as well as *d* coefficients that worked as the Lower Equivalence Bounds. For this example, the Lower Equivalence Bounds (|2|) were converted into *d* coefficients by dividing $-\Delta_L$ by pooled standard deviation and Δ_U by pooled standard deviation.

Table 2 *R* results for the ET and the NHSST.

***R* Code**	
> TOSTtwo(m1 = **98.78**, m2 = **99.83**, sd1 = **1.56**, sd2 = **1.11**, n1 = **6**, n2 = **6**, low_eqbound_d = -**1.48**, high_eqbound_d = **1.48**)	
TOST results:	t -value lower bound: 1.22 p-value lower bound: 0.127
	t-value upper bound: -3.91 p-value upper bound: 0.002
	degrees of freedom: 9.03
Equivalence bounds (Cohen's d):	low eqbound: -1.48
	high eqbound: 1.48
Equivalence bounds (raw scores):	low eqbound: -2.0037
	high eqbound: 2.0037
TOST confidence interval:	lower bound 90% CI: -2.482
	upper bound 90% CI: 0.382
NHST confidence interval:	lower bound 95% CI: -2.817
	upper bound 95% CI: 0.717
Equivalence Test Result:	The equivalence test was non-significant, t(9.03) = 1.220, p = 0.127, given equivalence bounds of -2.004 and 2.004 (on a raw scale) and an alpha of 0.05.
	Null Hypothesis Test Result: The null hypothesis test was non-significant, t(9.03) = -1.343, p = 0.212, given an alpha of 0.05.
	Based on the equivalence test and the null-hypothesis test combined, we can conclude that the observed effect is statistically not different from zero and statistically not equivalent to zero.

Table 2 and Figure 4 were products of Lakens' (2017) *R* code, which performs equivalence tests for independent and dependent *t* tests, correlations, and meta-analysis.

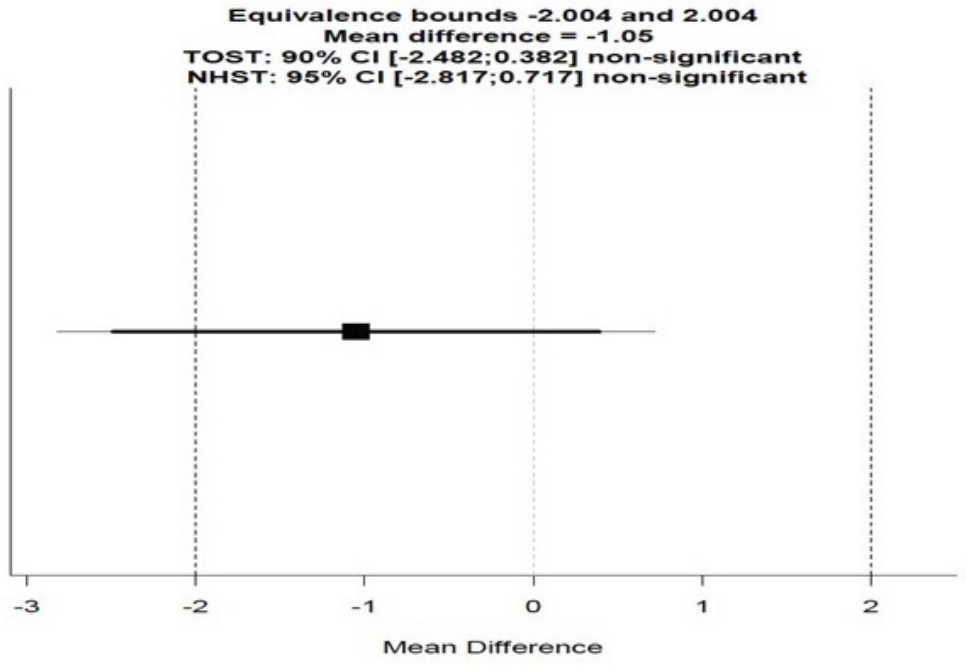


Figure 4 Output from the *R* code: Equivalence Testing and Null Hypothesis Significance Statistical Testing.

In conclusion, given a target of 100, the lower limit was set at –2 and the upper one at +2. The null hypothesis (H_{01}) was not rejected because of the *CI* [–2.4638, 0.3638], which exceeded the range of |2|. Furthermore, the *t*-ratio failed to reject the hypothesis for the lower limit t (9) = 1.22, p = .25. Thus, there was a failure to show equivalence.

4 DISCUSSION AND CONCLUSIONS

The present paper accomplished the objective and answered the research question by exemplifying how to execute an ET by manual calculations with a *t* table and an online calculator and *R*, in the area of educational research. This kind of analysis can help researchers to provide meaning to their results (means and effect sizes) by comparing the means of two different studies and testing whether they are statistically equivalent. Additionally, the procedures and results of other authors, De Muth (2019) and the *R* code of Lakens et al. (2018), who utilized Minitab and *R* for ET respectively, were demonstrated with one sample and two independent samples.

4.1 Recommendations

As a summary of the topics covered in this manuscript, here is a list of recommendations for future research with experimental designs about effect sizes and general research:

- Keeping in mind that p-values were conceptualized for the long run and not only for one study because it would be necessary to see if a pattern emerges from a series of studies before drawing conclusions (see Greenland et al., 2016).
- Verifying the Statistical Power level of data (see Cohen, 1988; Faul et al., 2007).
- Calculating an effect size, showing how it was calculated, and estimating its confidence interval (see Cumming, 2012)
- Interpreting the practical significance of the findings (see Kirk, 1996; Kraft, 2020).
- If there is a target or identified best practice (i.e. mean or effect size), using ET to see if they are statistically equivalent (De Muth, 2019; Lakens et al., 2018).
- Writing as many as possible of these recommendations on paper (i.e. because a proper inference requires reporting transparency, Wasserstein and Lazar (2016) and using Open Science practices for future replications and advancement of science (see Cumming & Calin-Jageman, 2017).

It could be that researchers might ask themselves: why should I care as a researcher about effect sizes, confidence intervals and equivalence testing? The benefits are that effect sizes and their confidence intervals can help to measure *practical significance* and see if a treatment seems to be improving a situation. By the same token, using ET can provide meaning for findings to see whether one study's findings are statistically equivalent to another's, given some Equivalence Bounds. Regarding Equivalence Bounds, there is a cautionary note: these limits are set by experts in the fields and not by statisticians (cf. Hauck & Anderson, 1984, p. 658).

4.2 Limitations

One limitation was that Lakens et al. (2018) discussed some epistemological implications of ET in relation to *empirical falsification,* as explained by Sir Karl R. Popper. Given the likely complexity of *epistemological issues*, they would warrant the creation of a whole article dealing with ET and empirical falsification (see Popper, 2002). Another limitation was the lack of coverage of *non-inferiority testing*, which is closely related to ET and deals similarly with comparisons, so Wellek (2010) would be a good source for this test.

4.3 Conclusions

The contribution of the present paper was to fill a gap regarding Equivalence Testing and Educational Research because there was no paper in *Google Scholar* with these two terms. By filling this gap in the literature, as Gall et al. (2007) might have said, there was an improvement in the collective knowledge of education. Besides this, there was a practical element in the context of educational research, as it how to use an ET to see if two means were statistically equivalent given two equivalent bounds was shown.

REFERENCES

Altam, D. G., & Bland, J. M. (1995). Absence of evidence is not evidence of absent. *Br Med J*, 485–486. https://doi.org/10.1136/bmj.311.7003.485

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Cohen, J. (1965). Some statistical issues in psychological research. In B. Wolman (Ed.), *Handbook of clinical Psychology* (pp. 95–121). McGraw-Hill.

Cumming, G. (2012). *Understanding the New Statistics: Effect sizes, Confidence intervals, and Meta-Analysis*. Routledge. https://doi.org/10.4324/9780203807002

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the New Statistics: Estimation, Open Science, & Beyond*. Routledge. https://doi.org/10.4324/9781315708607

Cumming, G., & Maillardet, R. (2006). Confidence intervals and replication: Where will the next mean fall? *Psychological methods*, *11*(3), 217. http://doi.org/10.1037/1082-989X.11.3.217

De Muth, J. E. (2019). *Practical Statistics for Pharmaceutical Analysis with Minitab Applications*. Springer. https://doi.org/10.1007/978-3-030-33989-0

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's t-test Instead of Student's t-test. *International Review of Social Psychology*, *30*(1), 92–101. http://doi.org/10.5334/irsp.82

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. https://doi.org/10.3758/BF03193146

Gall, M. D., Gall, J. P., & Borg, W. R. (2007). *Educational Research: An introduction* (8th ed.). Pearson.

Garstats, & Garstats. (2018, April 4). *Cohen's d is biased. Basic statistics.* Retrieved from https://garstats.wordpress.com/2018/04/04/dbias/

Greenland, S., Senn, S. J., Rothman, K., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *Eur J Epidemiol*, *31*, 337–350. https://doi.org/10.1007/s10654-016-0149-3

Hauck, W. W., & Anderson, S. (1984). A new statistical procedure for testing equivalence in two-group comparative bioavailability trials. *Journal of Pharmacokinetics and Biopharmaceutics*, *12*(1), 83–91. http://doi.org/10.1007/BF01063612

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701. https://doi.org/10.1371/journal.pmed.0020124

Kelley, K., & Rausch, J. R. (2006). Sample Size Planning for Standardized Mean Difference: Accuracy in Parameter Estimation Via Narrow Confidence Intervals. *Psychological Methods*, *11*(4), 363–385. http://doi.org/10.1037/1082-989X.11.4.363

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *57*, 746–759. https://doi.org/10.1177/0013164496056005002

Kraft, M. A. (2020). Interpreting effect sizes of educational interventions. *Educational Researcher*, *49*(4), 241–253. https://doi.org/10.3102/0013189X20912798

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations and meta-analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lemire, S. D. (2010). An argument framework for the application of null hypothesis statistical testing

in support of research. *Journal of Statistical Education*, *18*(2), 1–23. https://doi.org/10.1080/10691898.2010.11889492

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 Research. *Language Learning*, *64*, 878–912. https://doi.org/10.1111/lang.12079

Ponce-Renova, H. F. (2019). *Conceptos básicos de estadística inferencial aplicados a la investigación educativa*. Universidad Autónoma de Ciudad Juárez.

Ponce-Renova, H. F. (2021). *Estadística para comparaciones básicas de grupos: con uso de SPSS y calculadoras en línea*. Universidad Autónoma de Ciudad Juárez.

Popper, K. (2002). *The Logic of Scientific Discovery*. New York: Routledge.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical Procedures and the justification of Knowledge in Psychological Science. *American Psychologist*, *44*(10), 1276–1284. https://doi.org/10.1037/0003-066X.44.10.1276

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114. https://doi.org/10.2307/3002019

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize. *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680. https://doi.org/10.1007/BF01068419

Serlin, R. A., & Lapsley, D. K. (1993). Rational appraisal of psychological research and good-enough principle. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 199–228). Lawrence Erlbaum Associates.

Shi, H., & Yin, G. (2020). Reconnecting p-value and posterior probability under one- and two-sided tests. *The American Statistician*, *0*(0), 1–11.

Thompson, B. (2008). Computing and interpreting effect sizes, confidence intervals, and confidence intervals for effect sizes. In J. W. Osborne (Ed.), *Best Practices in Quantitative methods* (pp. 246–262). Sage. https://doi.org/10.4135/9781412995627.d21

Vandenbos, G. R. (2015). *APA Dictionary of Psychology* (2nd ed.). American Psychological Association. https://doi.org/10.1037/14646-000

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129–133. https://doi.org/10.1080/00031305.2016.1154108

Wellek, S. (2010). *Testing statistical hypotheses of equivalence and noninferiority* (2nd ed.). CRC Press. https://doi.org/10.1201/EBK1439808184

Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals. *American Psychologist*, *54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594

Young, J. C., & Minder, C. E. (1974). Algorithm AS76. An integral used in calculating non-central t and bivariate normal probabilities. *Appl. Stat*, *23*, 455–457. https://doi.org/10.2307/2347148