Invited Forum: Providing Evidence Against the Presence of a (Meaningful) Effect
Tutorial

Advances in Methods and
Practices in Psychological Science
2018, Vol. 1(2) 259–269

DOI: 10.1177/2515245918770963
www.psychologicalscience.org/AMPPS

Equivalence Testing for Psychological Research: A Tutorial

Daniël Lakens, Anne M. Scheel, and Peder M. Isager
Human-Technology Interaction Group, Eindhoven University of Technology

Abstract

Psychologists must be able to test both for the presence of an effect and for the absence of an effect. In addition to testing against zero, researchers can use the two one-sided tests (TOST) procedure to test for *equivalence* and reject the presence of a smallest effect size of interest (SESOI). The TOST procedure can be used to determine if an observed effect is surprisingly small, given that a true effect at least as extreme as the SESOI exists. We explain a range of approaches to determine the SESOI in psychological science and provide detailed examples of how equivalence tests should be performed and reported. Equivalence tests are an important extension of the statistical tools psychologists currently use and enable researchers to falsify predictions about the presence, and declare the absence, of meaningful effects.

Keywords

frequentist, null hypothesis, power, equivalence testing, null-hypothesis significance test, TOST, falsification, open materials

Received 11/17/17; Revision accepted 3/1/18

Psychologists should be able to falsify predictions. A common prediction in psychological research is that a nonzero effect exists in the population. For example, one might predict that American Asian women primed with their Asian identity will perform better on a math test compared with women who are primed with their female identity. To be able to design a study that allows for strong inferences (Platt, 1964), it is important to specify which test result would *falsify* the hypothesis in question.

Equivalence testing can be used to test whether an observed effect is surprisingly small, assuming that a meaningful effect exists in the population (see, e.g., Goertzen & Cribbie, 2010; Meyners, 2012; Quertemont, 2011; Rogers, Howard, & Vessey, 1993). The test is a simple variation of widely used null-hypothesis significance tests. To understand the idea behind equivalence tests, it is useful to realize that the null hypothesis tested can be any numerical value. When researchers compare two groups, they often test whether they can reject the hypothesis that the difference between these groups is zero (see Fig. 1a), but they may sometimes want to reject values other than zero. Imagine a researcher who is interested in voluntary participation in a national program to train young infants' motor skills. The researcher wants to test whether more boys than girls are brought into the program by their parents. The researcher should not expect 50% of the participants to be boys because, on average, 103 boys are born for every 100 girls (Central Intelligence Agency, 2016). In other words, approximately 50.74% of applicants should be boys, and 49.26% should be girls. If boys and girls were exactly equally likely to be brought into the program by their parents, the expected difference between boys' and girls' application rates would be not 0 but 1.5 percentage points (50.74% – 49.26%). Rather than specifying the null hypothesis as a difference of 0, the researcher would specify the null hypothesis as a difference of 0.015.

Alternatively, the researcher could decide that even if the true difference in the population is not exactly

Corresponding Author:
Daniël Lakens, Den Dolech 1, IPO 1.33, 5600 MB, Eindhoven, The Netherlands
E-mail: D.Lakens@tue.nl

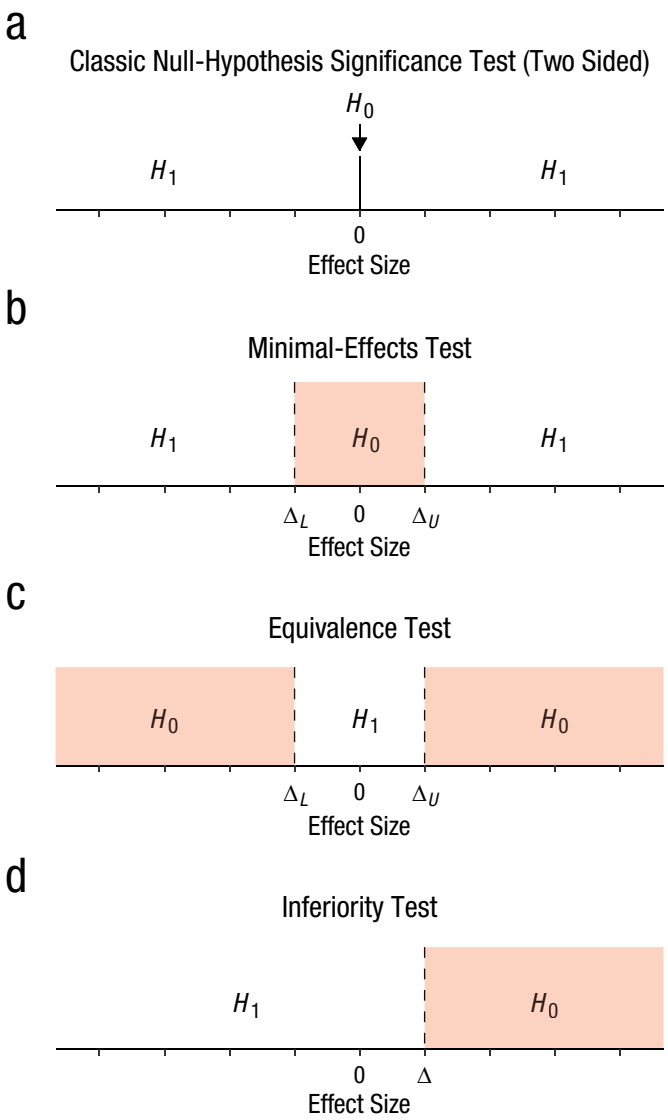

Fig. 1. Illustration of a null hypothesis (H_0) and alternative hypothesis (H_1) for each of four different types of significance tests. The null-hypothesis significance test (a) tests if the null hypothesis that an effect is equal to zero can be rejected. The minimal-effects test (b) tests if the null hypothesis that an effect falls between the lower equivalence bound, Δ_L, and the upper equivalence bound, Δ_U, can be rejected. The equivalence test (c) tests if the null hypothesis that an effect is at least as small as Δ_L or at least as large as Δ_U can be rejected. The inferiority test (d) tests if the null hypothesis that an effect is at least as large as Δ can be rejected.

0.015, the null hypothesis should consist of a range of values around 0.015 that can be considered trivially small. The researcher could, for example, test if the difference is smaller than −0.005 or larger than 0.035. This test against two bounds, with H_0 being a range rather than one value (see Fig. 1b), is known as a *minimal-effects test* (Murphy, Myors, & Wolach, 2014).

Equivalence tests can be seen as the opposite of minimal-effects tests: They examine whether the hypothesis that there are effects extreme enough to be considered meaningful can be rejected (see Fig. 1c). Note that rejecting the hypothesis of a meaningful effect does not imply that there is no effect at all. In this example, the researcher can perform an equivalence test to examine whether the gender difference in application rates is at least as extreme as the *smallest effect size of interest* (SESOI). After an extensive discussion with experts, the researcher decides that as long as the gender difference does not deviate from the population difference by more than .06, it is too small to care about. Given an expected true difference in the population of .015, the researcher will test if the observed difference falls outside the boundary values (or *equivalence bounds*) of −.055 and .075. If differences at least as extreme as these boundary values can be rejected in two one-sided tests (also known as one-tailed tests; i.e., the TOST procedure), the researcher will conclude that the application rates are statistically equivalent; the gender difference will be considered trivially small, and no money will be spent on addressing a gender difference in participation. When we refer to values as being "statistically equivalent" or to a "conclusion of statistical equivalence," we mean the difference between groups is smaller than what is considered meaningful and statistically falls within the interval indicated by the equivalence bounds.

In any one-sided test, for an alpha level of .05, one can reject H_0 when the 90% confidence interval (CI) around the observed estimate is in the predicted direction and does not contain the value the estimate is being tested against (e.g., 0). In the TOST procedure, the first one-sided test is used to test the estimate against values at least as extreme as the lower equivalence bound (Δ_L) and the second one-sided test is used to test the estimate against values at least as extreme as the upper equivalence bound (Δ_U). Even though the TOST procedure consists of two one-sided tests, it is not necessary to control for multiple comparisons because both tests need to be statistically significant for the researcher to draw a conclusion of statistical equivalence. Consequently, when reporting an equivalence test, it suffices to report the one-sided test with the smaller test statistic (e.g., t) and thus the larger p value. A conclusion of statistical equivalence is warranted when the larger of the two p values is smaller than alpha. If the observed effect is neither statistically different from zero nor statistically equivalent, there is insufficient data to draw conclusions. Further studies are needed, and they can be analyzed using a (small-scale) meta-analysis. The additional data will narrow the confidence interval around the observed effect, allowing the researcher to reject the null, reject effects at least as extreme as the SESOI, or both. In null-hypothesis significance tests, large sample sizes are

needed to have sufficient statistical power to detect small effects. Similarly, in equivalence tests, as the SESOI becomes smaller, the equivalence bounds become narrower (i.e., closer to zero), and a larger sample size is needed in order to obtain a sufficiently narrow confidence interval to conclude that that the observed estimate is statistically equivalent (i.e., that a meaningful effect is absent).

In this article, we illustrate how the TOST procedure can be applied with free, easy-to-use software, by reanalyzing five previously published results. Note that we control the Type I error rate at 5% for all the examples, to mirror those original studies (but we recommend using substantive arguments to justify the Type I error rate in original research; Lakens et al., 2018). It may be easiest to think of an equivalence test as checking whether the entire 90% CI falls between the upper and lower equivalence bounds, but for any given study an equivalence test could also be conceptualized as determining whether an effect size or test statistic is closer to zero than to some critical value, or even whether the *p* value for a null-hypothesis significance test is larger than some *p*-value bound. Before presenting our examples, we discuss different approaches to determining the SESOI for psychological research, as this value determines the statistical question an equivalence test answers.

Disclosures

The code to reproduce the analyses reported in this article is available via the Open Science Framework, at https://doi.org/10.17605/OSF.IO/QAMC6. The project at this URL contains all the files necessary to reproduce the five examples in R, jamovi (jamovi project, 2017), and a spreadsheet. The manuscript, including the figures and statistical analyses, was created using RStudio (Version 1.1.383; RStudio Team, 2016) and R (Version 3.4.2; R Core Team, 2017) and the R packages bookdown (Version 0.5; Xie, 2016), ggplot2 (Version 2.2.1; Wickham, 2009), gridExtra (Version 2.3; Auguie, 2017), knitr (Version 1.17; Xie, 2015), lattice (Version 0.20.35; Sarkar, 2008), papaja (Version 0.1.0.9492; Aust & Barth, 2017), purrr (Version 0.2.4; Henry & Wickham, 2017), pwr (Version 1.2.1; Champely, 2017), rmarkdown (Version 1.7; Allaire et al., 2017), and TOSTER (Version 0.3; Lakens, 2017b).

Justifying the Smallest Effect Size of Interest

The TOST procedure is performed against equivalence bounds that are derived from the SESOI. The SESOI can sometimes be determined objectively (e.g., it can be based on just-noticeable differences, as we explain in the next paragraph). In lieu of objective justifications, the SESOI should ideally be based on a cost-benefit analysis. Because both costs and benefits are necessarily subjective, the SESOI will vary across researchers, fields, and time. The goal in setting a SESOI is to clearly justify conducting the study, that is, to explain why a study that has a high probability of rejecting effects at least as extreme as the specified equivalence bounds will contribute to the knowledge base. The SESOI is thus independent of the outcome of the study, and should be determined before looking at the data. A SESOI should be chosen such that inferences based on it answer meaningful questions. Although we use bounds that are symmetric around zero for all the equivalence tests in this Tutorial (e.g., $\Delta_L = -0.3$, $\Delta_U = 0.3$), it is also possible to use asymmetric bounds (e.g., $\Delta_L = -0.2$, $\Delta_U = 0.3$).

Objective justification of a smallest effect size of interest

An objectively determined SESOI should be based on quantifiable theoretical predictions, such as predictions derived from computational models. Sometimes, the only theoretical prediction is that an effect should be noticeable. In such circumstances, the SESOI can be based on the just-noticeable difference. For example, Burriss et al. (2015) examined whether women displayed increased redness in the face during the fertile phase of their ovulatory cycle. The hypothesis was that a slightly redder skin signals greater attractiveness and physical health, and that sending this signal to men yields an evolutionary advantage. This hypothesis presupposes that men can detect the increase in redness with the naked eye. Burriss et al. collected data from 22 women and showed that the redness of their facial skin indeed increased during their fertile period. However, this increase was not large enough for men to detect with the naked eye, so the hypothesis was falsified. Because the just-noticeable difference in redness of the skin can be measured, it was possible to establish an objective SESOI.

Another example of an objectively determined SESOI can be found in a study by Button et al. (2015). The minimal clinically important difference on the Beck Depression Inventory–II was determined by asking 1,039 patients when they subjectively felt an improvement in their depression and then relating these responses to the patients' difference scores on the depression inventory. Similarly, Norman, Sloan, and Wyrwich (2003) proposed that there is a surprisingly consistent minimally important difference of half a standard deviation, or $d = 0.5$, in health outcomes.

Subjective justification of a smallest effect size of interest

We distinguish three categories of subjective justifications of the SESOI. First, researchers can use benchmarks. For example, one might set the SESOI to a standardized effect size, such as $d = 0.5$, which would allow one to reject the hypothesis that the effect is at least as extreme as a medium-sized effect (Cohen, 1988). Similarly, one might set the SESOI at a Cohen's d of 0.1, which is sometimes considered a trivially small effect size (Maxwell, Lau, & Howard, 2015). Relying on a benchmark is the weakest possible justification of a SESOI and should be avoided. On the basis of a review of 112 meta-analyses, Weber and Popova (2012) concluded that setting a SESOI to a medium effect size ($r = .3$ or $d = 0.5$) would make it possible to reject only effects in the upper 25% of the distribution of effect sizes reported in communications research, and Hemphill (2003) suggested that a SESOI of $d = 0.5$ would imply rejecting only effects in the upper 33% of the distribution of effect sizes reported in the psychological literature.

Second, the SESOI can be based on related studies in the literature. Ideally, researchers who publish novel research would always specify their SESOI, but this is not yet common practice. It is thus up to researchers who build on earlier work to decide which effect size is too small to be meaningful when they examine the same hypothesis. Simonsohn (2015) recently proposed setting the SESOI as the effect size that an earlier study would have had 33% power to detect. With this *small-telescopes* approach, the equivalence bounds are thus primarily based on the sample size in the original study. For example, consider a study in which 100 participants answered a question, and the results were analyzed with a one-sample t test. A two-sided test with an alpha of .05 would have had 33% power to detect an effect of $d = 0.15$. Another example of how previous research can be used to determine the SESOI can be found in Kordsmeyer and Penke (2017), who based the SESOI on the mean of effect sizes reported in the literature. Thus, in their replication study, they tested whether they could reject effects at least as extreme as the average reported in the literature. Given random variation and bias in the literature, a more conservative approach could be to use the lower end of a confidence interval around the meta-analytic estimate of the effect size (cf. Perugini, Gallucci, & Costantini, 2014).

Another justifiable option when choosing the SESOI on the basis of earlier work is to use the smallest observed effect size that could have been statistically significant in a previous study. In other words, the researcher decides that effects that could not have yielded a p less than α in an original study will not be considered meaningful in the replication study either, even if those effects are found to be statistically significant in the replication study. The assumption here is that the original authors were interested in observing a significant effect, and thus were not interested in observed effect sizes that could not have yielded a significant result. It might be likely that the original authors did not consider which effect sizes their study had good statistical power to detect, or that they were interested in smaller effects but gambled on observing an especially large effect in the sample purely as a result of random variation. Even then, when building on earlier research that does not specify a SESOI, a justifiable starting point might be to set the SESOI to the largest effect size that, when observed in the original study, would not have been statistically significant. Given only a study's alpha level and sample size, one can calculate the critical test value (e.g., t, F, z). This critical test value can be transformed to a standardized effect size (e.g., $d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$), which can thus be interpreted as the *critical effect size*.[1] Any observed effect size smaller than the critical effect size would not have been statistically significant in the original study, given the alpha and sample size of that study. With the SESOI set as the critical effect size, an equivalence test can reject all observed effect sizes that could have been detected in the earlier study.

The third approach to subjective justification of a SESOI is to base it on a resource question. Not all psychological theories make quantifiable predictions, for example, by specifying the size of a predicted effect (the lack of quantifiable predictions has been blamed on a strong reliance on null-hypothesis significance testing in which the alternative hypothesis is not a specific effect size but *any* observed difference; see Meehl, 1978). Researchers often have more precise ideas about the amount of data that they can afford to collect, or that other researchers in their field commonly collect, than about the effect size they predict. The amount of data that is collected limits the inferences one can make. Given the alpha level and the planned sample size for a study, researchers can calculate the smallest effect size that they have sufficient power to detect.[2]

An equivalence test based on this approach does not answer any theoretical question (after all, the equivalence bounds are not based on any theoretical prediction). When theories only allow for directional predictions, but do not predict effects of a specific size, the sample-size justification can be used to determine which hypothesized effects can be studied reliably, and

thus would be interesting to reject. For example, imagine a line of research in which a hypothesis has almost always been tested by performing a one-sample *t* test on a sample smaller than 100 observations. A one-sample *t* test on 100 observations, using an alpha of .05 (two sided), has 90% power to detect an effect of $d = 0.33$. In a new study, concluding equivalence in a test using equivalence bounds of $\Delta_L = -0.33$ and $\Delta_U = 0.33$ would suggest that effects at least as extreme as those the previous studies were sensitive enough to detect can be rejected. Such a study does not test a theoretical prediction, but it contributes to the literature by answering a resource question: It suggests that future studies need to collect data on samples larger than 100 observations to examine this hypothesis.

When there is no previous literature on a topic, researchers can justify their sample size on the basis of reasonable resource limitations, which are specific to scientific disciplines and research questions (and can be expected to change over time). Whereas 22 patients with a rare neurological disorder might be the largest sample a single researcher can recruit, a study on Amazon Mechanical Turk might easily collect data on hundreds or thousands of individuals. Thus, whether or not the resource question that is answered by an equivalence test is interesting must be evaluated by peers, preferably when the study design is submitted to an ethical review board or as a Registered Report.

Additional subjective justifications for a SESOI are possible. For example, the U.S. Food and Drug Administration has set equivalence bounds for bioequivalence studies, taking the decision out of the hands of individual researchers (Senn, 2007). We have described several approaches to help researchers justify equivalence bounds, but want to repeat the warning by Rogers et al. (1993): "The thoughtless application of 'cookbook' prescriptions is ill-advised, regardless of whether the goal is to establish a difference or to establish equivalency between treatments" (p. 564). By transparently reporting and adequately justifying their SESOI, researchers can communicate how their study contributes to the literature and provide a starting point for a discussion about what a reasonable SESOI may be.

Raw Versus Standardized Equivalence Bounds

The SESOI, and thus the equivalence bounds, can be set in terms of a standardized effect size (e.g., Cohen's *d* of 0.5) or as a raw mean difference (e.g., 0.5 points on a 7-point scale). The key difference is that equivalence bounds set as raw differences are independent of the standard deviation, whereas equivalence bounds set as standardized effects are dependent on the standard deviation (because they are calculated as $\frac{X_1 - X_2}{SD}$). The observed standard deviation varies randomly across samples. In practice, this means that when the equivalence bounds are based on standardized differences, the equivalence test depends on the standard deviation in the sample. In the case of a raw mean difference of 0.5, the standardized upper equivalence bound will be 0.5 when the standard deviation is 1, but it will be 1 when the standard deviation is 0.5.

Both raw equivalence bounds and standardized equivalence bounds have specific benefits and limitations (for a discussion, see Baguley, 2009). When raw mean differences are meaningful and of theoretical interest, it makes sense to set equivalence bounds based on raw effect sizes. When the raw mean difference is of less theoretical importance, or different measures are used across lines of research, it is often easier to base equivalence bounds on standardized differences. Researchers should realize that an equivalence test with bounds based on raw differences asks a slightly different question than an equivalence test with bounds based on standardized differences and should justify their choice between these options. Ideally, when equivalence bounds are based on earlier research, such as in replication studies, they would yield the same result whether they are based on raw or standardized differences, and large differences in standard deviations between studies are as important to interpret as are large differences in means.

In the remainder of this article, we provide five detailed examples of equivalence tests, reanalyzing data from published studies. These concrete and easy-to-follow examples illustrate all the types of approaches to setting equivalence bounds that we have discussed (except benchmarks, which we advise against using) and demonstrate how to perform and report equivalence tests.

Example 1: Not Statistically Equivalent and Not Statistically Different

Moon and Roeder (2014), replicating work by Shih, Pittinsky, and Ambady (1999), conducted a study to investigate whether Asian American women would perform better than a control group on a math test when primed with their Asian identity. In contrast to Shih et al., they found that the Asian primed group ($n = 53$, $M = 0.46$, $SD = 0.17$) performed worse than the control group ($n = 48$, $M = 0.50$, $SD = 0.18$), but the difference was not significant, $d = -0.21$, $t(97.77) = -1.08$, $p = .284$ (two sided).[3] The nonsignificant null-hypothesis test does not necessarily mean that there was no meaningful effect, however; the design may not have had sufficient

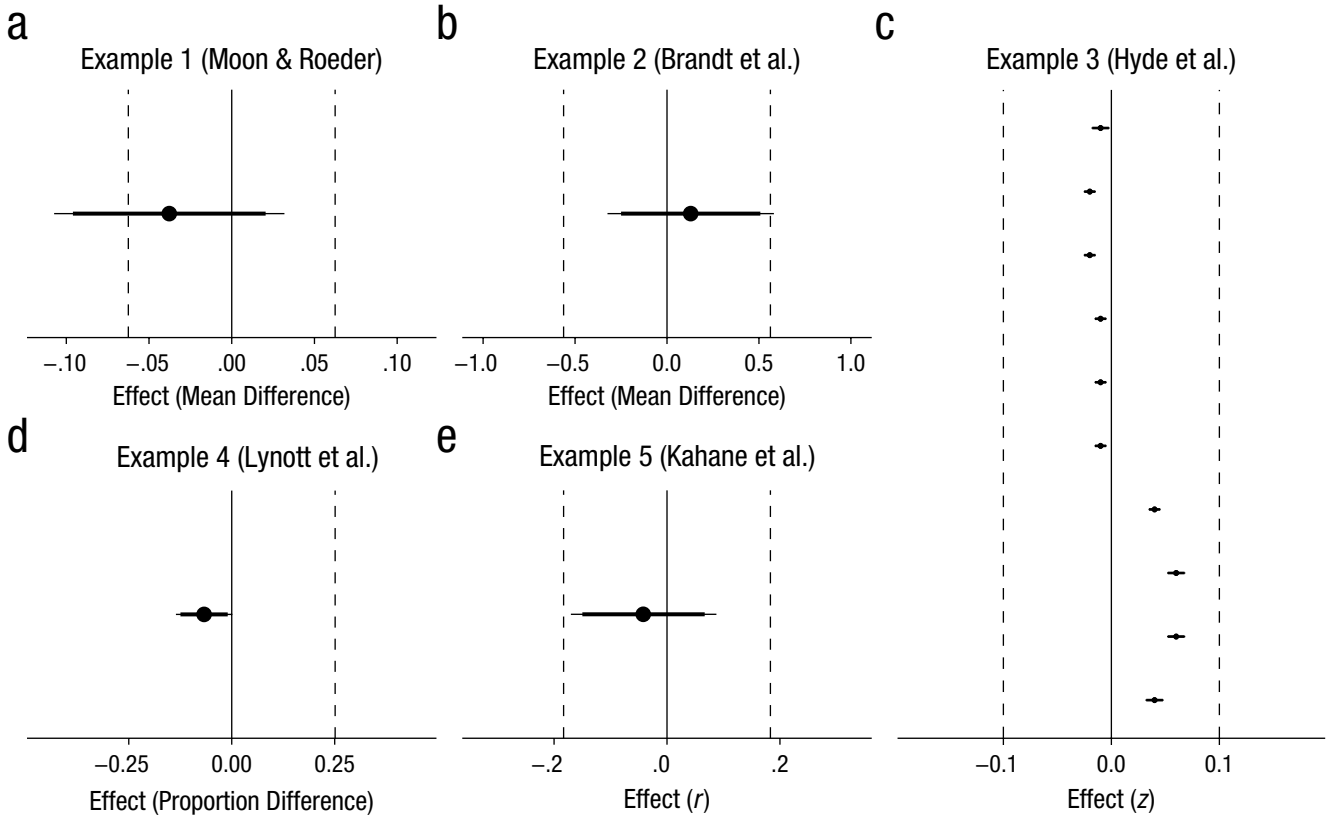

Fig. 2. Plots of the five examples: (a) difference between means in Moon and Roeder (2014; Example 1); (b) difference between means in Brandt, IJzerman, and Blanken (2014; Example 2); (c) differences between meta-analytic effect sizes in Hyde, Lindberg, Linn, Ellis, and Williams (2008; Example 3; separate results for each grade); (d) difference between proportions in Lynott et al. (2014; Example 4); and (e) difference between Pearson correlations in Kahane, Everett, Earp, Farias, and Savulescu (2015; Example 5). In the plots, the thick horizontal lines indicate the 90% confidence intervals from the two one-sided tests procedure, the thin horizontal lines indicate the 95% confidence intervals from null-hypothesis significance tests, the solid vertical lines indicate the null hypothesis, and the dashed vertical lines indicate the equivalence bounds (in raw scores).

statistical power to show whether a meaningful effect was present or absent.

In order to distinguish between these possibilities, one can define what a meaningful effect would be and use that value to set the bounds for an equivalence test (remember that these bounds should normally be specified before looking at the data). One option would be to think in terms of letter grades set in increments of 6.25 percentage points (F = 0%–6.25% correct, . . . , A+ = 93.75%–100% correct) and decide that only test-score differences corresponding to an increase or decrease of at least 1 grade are of interest. Thus, the SESOI is a difference in raw scores of 6.25 percentage points, or .0625. One can then perform an equivalence test for a two-sample Welch's t test, using equivalence bounds of ±.0625. The TOST procedure consists of two one-sided tests, and yields a nonsignificant result for the test against Δ_L, $t(97.77) = 0.71$, $p = .241$, and a significant result for the test against Δ_U, $t(97.77) = -2.86$, $p = .003$. Although the t test against Δ_U indicates that one can reject differences at least as large as .0625, the test against Δ_L shows that one cannot reject effects at least as extreme as −.0625. The equivalence test is therefore nonsignificant, which means one cannot reject the hypothesis that the true effect is at least as extreme as 6.25 percentage points (Fig. 2a). The result would be reported as $t(97.77) = 0.71$, $p = .241$, because typically only the one-sided test yielding the higher p value is reported in the Results section.[4]

Example 2: Statistically Equivalent and Not Statistically Different

Banerjee, Chatterjee, and Sinha (2012) reported that 40 participants who had been asked to describe an unethical deed from their past judged the room they were in to be darker than did participants who had been asked to describe an ethical deed, $t(38) = 2.03$, $p = .049$, $d = 0.65$. In a close replication with 100 participants, Brandt, IJzerman, and Blanken (2014) found no significant

Box 1. Calculating an Equivalence Test in R

Equivalence tests can be performed with summary statistics (e.g., means, standard deviations, and sample sizes) using the TOSTER package in the open-source programming language R. Using TOSTER in R requires no more programming experience than it takes to reproduce three lines of code.

The following code, which can be typed into R or RStudio, will reproduce the result of Example 2:

```
install.packages("TOSTER")

library(TOSTER)

TOSTtwo(m1 = 4.7857, m2 = 4.6569, sd1 = 1.0897, sd2 = 1.1895, n1 = 49, n2 = 51,
low_eqbound_d = -0.4929019, high_eqbound_d = 0.4929019, alpha = 0.05,
var.equal = FALSE)
```

The parameters of the test are defined inside the parentheses in the last line of code. To perform the test on your own data, simply copy these lines of code, replace the values with the corresponding values from your own study, and run the code. Results and a plot will be printed automatically. Running the code "help("TOSTtwo")" provides a help file with more detailed information.

effect (unethical condition: $M = 4.79$, $SD = 1.09$; ethical condition: $M = 4.66$, $SD = 1.19$), $t(97.78) = 0.57$, $p = .573$, $d = 0.11$. Before looking at the data, one can choose the small-telescopes approach and set the SESOI for the equivalence test as $d = 0.49$ (the effect size the original study had 33% power to detect). The TOST procedure for Welch's t test for independent samples, with equivalence bounds of $\Delta_L = -0.49$ and $\Delta_U = 0.49$, reveals that the effect observed in the replication study is statistically equivalent, because the larger of the two p values is less than .05, $t(97.78) = -1.90$, $p = .030$ (Fig. 2b; see Box 1 for an example of how to reproduce this test using R). According to the Neyman-Pearson approach, this means that one can reject the hypothesis that the true effect is smaller than $d = -0.49$ or larger than $d = 0.49$ and act as if the effect size falls within the equivalence bounds, with the understanding that (given the chosen alpha level) one will not be wrong more than 5% of the time.

Example 3: Statistically Equivalent and Statistically Different

Hyde, Lindberg, Linn, Ellis, and Williams (2008) reported a meta-analysis of effect sizes for gender differences in performance on mathematics tests across 7 million students in the United States. They concluded that the gender differences in Grades 2 through 11 were trivial, according to their definition of *trivial* as d smaller than 0.1. For example, in Grade 3, the d value was 0.04, with a standard error of 0.002. Equivalence tests on the meta-analytic effect sizes of the difference in math performance, using an alpha level of .005 to correct for multiple comparisons and using equivalence bounds of $\Delta_L = -0.1$ and $\Delta_U = 0.1$, show that the estimates of the effect size are statistically equivalent. That is, performance was measured with such high precision that the gender differences can be considered trivially small according to the authors' definition (Fig. 2c; e.g., for Grade 3, $z = 70.00$, $p < .001$). However, note that all of the effects were also statistically different from zero, as one might expect when there is no random assignment to conditions and samples sizes are very large (e.g., for Grade 3, $z = 20.00$, $p < .001$). This example shows how equivalence tests allow researchers to distinguish between *statistical* significance and *practical* significance, and thus how equivalence tests can improve hypothesis-testing procedures in psychological research.

Example 4: Statistically Inferior and Not Statistically Different

Lynott et al. (2014) conducted a study to investigate the effect of being exposed to physically warm or cold stimuli on subsequent judgments related to interpersonal warmth and prosocial behavior (a replication of Williams & Bargh, 2008). They observed that 50.74% of participants who received a cold pack ($n = 404$) opted to receive a reward for themselves, and 57.46% of participants who received a warm pack ($n = 409$) did the same. A z test indicated that the difference of 6.71 percentage points was not statistically significant ($z = -1.93$, $p = .054$). Had the authors planned to perform both a null-hypothesis significance test and an equivalence test, the latter would have allowed them to distinguish between an inconclusive outcome and a statistically equivalent result.

Because this was a replication study, one could decide in advance to test whether the observed effect size was smaller than the smallest effect size that the original study could have detected. The critical z value (~1.96 in a two-sided test with an alpha of .05) would then be used to set the equivalence bounds ($\Delta = \pm 1.96$). To calculate the difference corresponding to the critical z value in the original study, one would multiply that z value by the standard error (1.96×0.13) and find that the original study could have observed a significant effect if the difference had been at least 25%. Because the original study had a clear directional hypothesis, however, the replication study was aimed at determining whether receiving a warm pack would increase the proportion of people who chose a gift for a friend, and thus a test for inferiority is appropriate (see Fig. 1d and note 4). In such a test, one concludes that a meaningful effect is absent if the observed effect size is reliably lower than the SESOI. The inferiority test on the data from Lynott et al. reveals that one can reject effects larger than $\Delta = 0.25$, $z = -9.12$, $p < .001$ (see Fig. 2d). Thus, the statistically nonsignificant effect in the replication study is also statistically smaller than the SESOI.

Example 5: Statistically Equivalent and Not Statistically Different

Kahane, Everett, Earp, Farias, and Savulescu (2015) investigated how responses to moral dilemmas in which participants have to decide whether or not they would sacrifice one person's life to save several other lives relate to other indicators of moral orientation. Traditionally, greater endorsement for sacrificing one life to save several others has been interpreted as a more "utilitarian" moral orientation (i.e., a stronger concern for the greater good). Kahane et al. contested this interpretation in a number of studies. In their Study 4, they compared the traditionally used dilemmas with a set of new dilemmas in which the sacrifice for the greater good was not another person's life, but something the participant would have a partial interest in (e.g., one dilemma concerned the choice between buying a new mobile phone and donating the money to save lives in a distant country). The authors found no significant correlation between responses to the two sets of dilemmas, $r(229) = -.04$, $p = .525$, $N = 231$.[5] They concluded that "a robust association between 'utilitarian' judgment and genuine concern for the greater good seems extremely unlikely" (p. 206), given the statistical power their study had to detect meaningful effect sizes.

This interpretation can be formalized by performing an equivalence test for correlations, with equivalence bounds set to an effect size the study had reasonable power to detect (as determined before looking at the data). With 231 participants, the study had 80% power to detect a correlation of .18. Given $\Delta_L = -.18$ and $\Delta_U = .18$, the TOST procedure indicates that the outcome was statistically equivalent, $r(229) = -.04$, $p = .015$. This means that r values at least as extreme as ±.18 can be rejected at an alpha level of .05 (see Fig. 2e). If other researchers are interested in examining the presence of a smaller effect, they can design studies with a larger sample size.

Discussion

Equivalence testing is a simple statistical technique for determining whether one should reject the presence of an effect at least as extreme as the SESOI. As long as it is possible to calculate a confidence interval around a parameter estimate, one can compare that estimate with the SESOI. The result of an equivalence test can be obtained by mere visual inspection of the confidence interval (Seaman & Serlin, 1998; Tryon, 2001) or by performing two one-sided tests (i.e., the TOST procedure).

As is the case with any statistical test, the usefulness of the result of an equivalence test depends on the specific question asked. That question manifests itself in the bounds that are specified: Is the observed effect surprisingly close to zero, assuming that there is a true effect at least as extreme as the SESOI? If one tests against very wide bounds, finding that the observed effect is statistically equivalent can hardly be considered surprising, given that most effects in psychology are small to medium (Hemphill, 2003). Examining publications citing Lakens (2017a), we found that some researchers state, but do not justify, the SESOI used in their equivalence tests (e.g., Brown, Rodriguez, Gretak, & Berry, 2017; Schumann, Klein, Douglas, & Hewstone, 2017). An equivalence test using a SESOI of $d = 0.5$ might very well answer a question the researchers are interested in (for one possible justification based on minimally important differences, see Norman et al., 2003), but researchers should always explain why they chose a particular SESOI. It makes little sense to report a statistical test without explaining why you are interested in the question it answers.

Equivalence bounds should be specified before results are known, ideally as part of a preregistration (cf. Piaggio et al., 2006). In the most extreme case, a researcher can always first look at the data and then choose equivalence bounds wide enough for the test to yield a statistically equivalent result (for a discussion, see Weber & Popova, 2012). However, fixed error rates are no longer valid when bounds are determined on the basis of the observed data. The value of an equivalence test is determined by the strength of the justification of

the equivalence bounds. If the bounds chosen are based on the observed data, an equivalence test becomes meaningless. We have proposed several ways to justify equivalence bounds, but in the end these discussions must happen among peers, and the best occasion for these discussions is during peer review of proposals for Registered Reports.

As is the case with null-hypothesis significance tests, equivalence tests interpreted from a Neyman-Pearson perspective on statistical inferences are used to guide the actions of researchers while controlling error rates. Research sometimes requires dichotomous choices. For example, a researcher might want to decide to discontinue a line of research if an observed effect is too small to be considered meaningful. Equivalence tests based on carefully chosen equivalence bounds and error rates can be used to govern researchers' behavior (Neyman & Pearson, 1933). An equivalence test and a null-hypothesis significance test examine two different hypotheses and can therefore be used in concert (Weber & Popova, 2012). We recommend that researchers by default perform both a null-hypothesis significance test and an equivalence test on their data, as long as they can justify a SESOI, in order to improve the falsifiability of predictions in psychological science.

The biggest challenge for researchers will be to specify a SESOI. Psychological theories are usually too vague for deriving precise predictions, and if there are no theoretical reference points, natural constraints, or prior studies a researcher can use to define the SESOI for a hypothesis test, any choice will be somewhat arbitrary. In some lines of research, researchers might use equivalence tests to simply reject consecutively smaller effect sizes by performing studies with increasingly large sample sizes while controlling error rates, until no one is willing to invest the time and resources needed to explore the possibility of even smaller effects. Nevertheless, it is important to realize that not specifying a SESOI for research questions at all will severely hinder theoretical progress (Morey & Lakens, 2016). Incorporating equivalence tests in your statistical toolbox will in time allow you to contribute to better—and more falsifiable—theories.

Action Editor

Daniel J. Simons served as action editor for this article.

Author Contributions

D. Lakens conceptualized the article, programmed the TOSTER package and spreadsheet (https://osf.io/qzjaj/), and was the main contributor to the introduction and discussion sections of this article. A. M. Scheel and P. M. Isager were the main contributors to the presentation of the examples and figures. All the authors contributed to the section on justifying the smallest effect size of interest. All the authors edited and revised the manuscript, reviewed it for submission, and contributed to additional development of the TOSTER package. A. M. Scheel and P. M. Isager contributed equally to the manuscript, and the order of their authorship was determined by executing the following commands in R:

```
set.seed(7998976/5271)

x <- sample(c("Anne", "Peder"), 1)

print(paste("The winner is", x, "!"))
```

ORCID iDs

Daniël Lakens https://orcid.org/0000-0002-0247-239X
Peder M. Isager https://orcid.org/0000-0002-6922-3590
Anne M. Scheel https://orcid.org/0000-0002-6627-0746

Acknowledgments

We would like to thank Courtney Soderberg for creating the first version of the two one-sided tests function for testing two independent proportions. We would also like to thank Dermot Lynott and Katie Corker for their helpful responses to our inquiries about Lynott et al. (2014); Jim Everett and Brian Earp for their helpful responses regarding Study 4 in Kahane, Everett, Earp, Farias, and Savulescu (2015) and for providing the raw data for that study; and Neil McLatchie and Jan Vanhove for their valuable comments on an earlier version of this manuscript.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Open Practices

The code to reproduce the analyses reported in this article has been made publicly available via the Open Science Framework and can be accessed at https://osf.io/qamc6/. The complete Open Practices Disclosure for this article can be found at http://journals.sagepub.com/doi/suppl/10.1177/2515245918770963. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

Prior Versions

A preprint version of this manuscript prior to peer review is available on the Open Science Framework, at https://osf.io/qmgtn/?show=revision (Version 1). For the accepted version

of this manuscript, we updated our recommendations on how to justify the smallest effect size of interest on the basis of available resources. The updated preprint version (Version 2) is available at https://doi.org/10.17605/OSF.IO/V3ZKT.

Notes

1. This will typically, although not always, correspond to the effect size the study had 50% power to detect (see Lenth, 2007). This procedure will result in a wider interval between the equivalence bounds than when one follows the small-telescopes approach, which uses the effect size a study had 33% power to detect.
2. This approach is conceptually very similar to the *power approach*, in which the effect size a study had 95% power to detect is calculated, and the presence of effects at least as extreme as this value is rejected if a traditional null-hypothesis significance test yields a p value larger than .05. However, Meyners (2012) explained that the power approach is not recommended (even though it is common) because it ignores the possibility that effects are both significant and equivalent, and error rates are not controlled accurately.
3. Fractional degrees of freedom in t tests are a result of using Welch's t test instead of Student's t test, the former of which is the recommended default when the groups being compared have unequal sample sizes (Delacre, Lakens, & Leys, 2017). In the TOSTER package, Welch's t test can be used by setting "var.equal = FALSE."
4. Because this is a replication study, it also would be reasonable to focus on rejecting effects in the same direction as the effect in the original study. After all, although effects that are much smaller than those observed in the original study may indicate nonreplication, so too do large effects in the opposite direction. To test if one can reject effects in the same direction as the originally observed effect, one would perform an inferiority test (see Fig. 1d) against Δ. As Figure 2a shows, the 90% CI for the effect Moon and Roeder observed does not overlap with Δ_U, so one can reject effects as large or larger than .0625. However, the choice between an equivalence or inferiority test must be made before running the test to prevent inflated error rates.
5. The final sample size was 232, but because of missing data in one case, the correlation reported was based on a sample of 231.

References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., . . . Chang, W. (2017). rmarkdown: Dynamic documents for R (Version 1.7) [Computer software]. Retrieved from https://CRAN.R-project.org/package=rmarkdown

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics (Version 2.3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F., & Barth, M. (2017). papaja: Create APA manuscripts with R Markdown (Version 0.1.0.9492) [Computer software]. Retrieved from https://github.com/crsh/papaja

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology*, *100*, 603–617. doi:10.1348/000712608X377117

Banerjee, P., Chatterjee, P., & Sinha, J. (2012). Is it light or dark? Recalling moral behavior changes perception of brightness. *Psychological Science*, *23*, 407–409. doi:10.1177/0956797611432497

Brandt, M. J., IJzerman, H., & Blanken, I. (2014). Does recalling moral behavior change the perception of brightness? A replication and meta-analysis of Banerjee, Chatterjee, and Sinha (2012). *Social Psychology*, *45*, 246–252. doi:10.1027/1864-9335/a000191

Brown, M., Rodriguez, D. N., Gretak, A. P., & Berry, M. A. (2017). Preliminary evidence for how the behavioral immune system predicts juror decision-making. *Evolutionary Psychological Science*, *3*, 325–334. doi:10.1007/s40806-017-0102-z

Burriss, R. P., Troscianko, J., Lovell, P. G., Fulford, A. J. C., Stevens, M., Quigley, R., . . . Rowland, H. M. (2015). Changes in women's facial skin color over the ovulatory cycle are not detectable by the human visual system. *PLOS ONE*, *10*(7), Article e0130093. doi:10.1371/journal.pone.0130093

Button, K. S., Kounali, D., Thomas, L., Wiles, N. J., Peters, T. J., Welton, N. J., . . . Lewis, G. (2015). Minimal clinically important difference on the Beck Depression Inventory - II according to the patient's perspective. *Psychological Medicine*, *45*, 3269–3279. doi:10.1017/S0033291715001270

Central Intelligence Agency. (2016). *The CIA World Factbook 2017*. New York, NY: Skyhorse.

Champely, S. (2017). pwr: Basic functions for power analysis (Version 1.2.1) [Computer software]. Retrieved from https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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*, 92–101. doi:10.5334/irsp.82

Goertzen, J. R., & Cribbie, R. A. (2010). Detecting a lack of association: An equivalence testing approach. *British Journal of Mathematical and Statistical Psychology*, *63*, 527–537. doi:10.1348/000711009X475853

Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist*, *58*, 78–80.

Henry, L., & Wickham, H. (2017). purrr: Functional programming tools (Version 0.2.4) [Computer software]. Retrieved from https://CRAN.R-project.org/package=purrr

Hyde, J. S., Lindberg, S. M., Linn, M. C., Ellis, A. B., & Williams, C. C. (2008). Gender similarities characterize math performance. *Science*, *321*, 494–495. doi:10.1126/science.1160364

jamovi project. (2017). Jamovi (Version 0.8) [Computer software]. Retrieved from https://www.jamovi.org/download.html

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). "Utilitarian" judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition*, *134*, 193–209. doi:10.1016/j.cognition.2014.10.005

Kordsmeyer, T. L., & Penke, L. (2017). The association of three indicators of developmental instability with mating success in humans. *Evolution & Human Behavior*, *38*, 704–713. doi:10.1016/j.evolhumbehav.2017.08.002

Lakens, D. (2017a). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological*

& Personality Science, *8*, 355–362. doi:10.1177/1948550617697177

Lakens, D. (2017b). TOSTER: Two one-sided tests (TOST) equivalence testing (Version 0.3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=TOSTER

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., . . . Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behavior*, *2*, 168–171. doi:10.1038/s41562-018-0311-x

Lenth, R. V. (2007). *Post hoc power: Tables and commentary*. Iowa City: University of Iowa, Department of Statistics and Actuarial Science.

Lynott, D., Corker, K. S., Wortman, J., Connell, L., Donnellan, M. B., Lucas, R. E., & O'Brien, K. (2014). Replication of "Experiencing physical warmth promotes interpersonal warmth" by Williams and Bargh (2008). *Social Psychology*, *45*, 216–222. doi:10.1027/1864-9335/a000187

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. doi:10.1037/a0039400

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. doi:10.1037/0022-006X.46.4.806

Meyners, M. (2012). Equivalence tests – a review. *Food Quality and Preference*, *26*, 231–245. doi:10.1016/j.foodqual.2012.05.003

Moon, A., & Roeder, S. S. (2014). A secondary replication attempt of stereotype susceptibility (Shih, Pittinsky, & Ambady, 1999). *Social Psychology*, *45*, 199–201. doi:10.1027/1864-9335/a000193

Morey, R., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. doi:10.5281/zenodo.838685

Murphy, K. R., Myors, B., & Wolach, A. (2014). *Statistical power analysis: A simple and general model for traditional and modern hypothesis tests* (4th ed.). New York, NY: Routledge.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, *231*, 289–337. doi:10.1098/rsta.1933.0009

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of changes in health-related quality of life: The remarkable universality of half a standard deviation. *Medical Care*, *41*, 582–592.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, *9*, 319–332. doi:10.1177/1745691614528519

Piaggio, G., Elbourne, D. R., Altman, D. G., Pocock, S. J., Evans, S. J. W. & the CONSORT Group. (2006). Reporting of noninferiority and equivalence randomized trials: An extension of the CONSORT Statement. *Journal of the American Medical Association*, *295*, 1152–1160. doi:10.1001/jama.295.10.1152

Platt, J. R. (1964). Strong inference. *Science*, *146*, 347–353.

Quertemont, E. (2011). How to statistically show the absence of an effect. *Psychologica Belgica*, *51*, 109–127. doi:10.5334/pb-51-2-109

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.2) [Computer software]. Retrieved from https://www.R-project.org/

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin*, *113*, 553–565.

RStudio Team. (2016). RStudio: Integrated development environment for R (Version 1.1.383) [Computer software]. Boston, MA: RStudio.

Sarkar, D. (2008). Lattice: Multivariate data visualization with R (Version 0.20.35) [Computer software]. Retrieved from http://lmdvr.r-forge.r-project.org

Schumann, S., Klein, O., Douglas, K., & Hewstone, M. (2017). When is computer-mediated intergroup contact most promising? Examining the effect of out-group members' anonymity on prejudice. *Computers in Human Behavior*, *77*, 198–210. doi:10.1016/j.chb.2017.08.006

Seaman, M. A., & Serlin, R. C. (1998). Equivalence confidence intervals for two-group comparisons of means. *Psychological Methods*, *3*, 403–411. doi:10.1037/1082-989X.3.4.403

Senn, S. (2007). *Statistical issues in drug development* (2nd ed.). Chichester, England: John Wiley & Sons.

Shih, M., Pittinsky, T. L., & Ambady, N. (1999). Stereotype susceptibility: Identity salience and shifts in quantitative performance. *Psychological Science*, *10*, 80–83. doi:10.1111/1467-9280.00111

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559–569. doi:10.1177/0956797614567341

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, *6*, 371–386.

Weber, R., & Popova, L. (2012). Testing equivalence in communication research: Theory and application. *Communication Methods and Measures*, *6*, 190–213. doi:10.1080/19312458.2012.703834

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer-Verlag.

Williams, L. E., & Bargh, J. A. (2008). Experiencing physical warmth promotes interpersonal warmth. *Science*, *322*, 606–607. doi:10.1126/science.1162548

Xie, Y. (2015). *knitr: Elegant, flexible, and fast dynamic report generation with R*. Retrieved from https://yihui.name/knitr/

Xie, Y. (2016). *bookdown: Authoring books and technical documents with R Markdown*. Retrieved from https://github.com/rstudio/bookdown