Equivalence Testing for Psychological Science

|  |
| --- |
| Daniel Lakens1, Anne M. Scheel1, & Peder Isager1 |
|  |
| 1 Eindhoven University of Technology |
|  |

# Author note

We would like to thank Courtney Soderbergh for creating the first version of the TOST function to test two independent proportions.

Correspondence concerning this article should be addressed to Daniel Lakens, Den Dolech 1, IPO 1.33, 5600 MB, Eindhoven, The Netherlands. E-mail: [D.Lakens@tue.nl](mailto:D.Lakens@tue.nl)

# Abstract

Psychologists need to be able to test for the absence of an effect. Using the Two-One-Sided Tests (TOST) procedure, researchers can easily test whether the observed effects are too small to be meaningful. By specifying a smallest effect size of interest (SESOI) researchers test whether observed effects are surprisingly closer to zero, assuming there was an effect the consider meaningful. 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.

*Keywords:* Equivalence Testing, NHST, power, TOST

Word count: X

# Equivalence Testing for Psychological Science

Psychologists should be able to falsify predictions. A common prediction in psychological research is that an effect exists in the population that differs from zero. For example, we might predict that priming American Asian women with their Asian identity will increase their performance on a math test compared to women who are primed with their female identity. To be able to falsify this hypothesis, and design a study that allows for strong inferences (Platt, 1964), it is important to specify which test result would *disprove* the hypothesis.

An equivalence test can be used to test whether the observed effect is surprisingly small, assuming a meaningful effect exists in the population. The test is a simple variation of the widely used null hypothesis significance tests (NHST). To understand the idea behind equivalence tests, it is useful to realize the null hypothesis we test against can be any numerical value. When we compare two groups, we often test whether the difference between these groups is zero, but we may sometimes want to test against other values 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. One could test whether the difference between boys and girls is zero. However, this would ignore the fact that the human sex ratio of newborn boys to girls is not exactly 1:1, and we should not expect 50% of participants to be boys. On average, 103 boys are born for every 100 girls (CIA, 2017), so approximately 50.74% of applicants should be boys, and 49.26% should be girls (with ). If boys and girls were exactly equally likely to be brought into the program by their parents, we would not expect a difference of zero, but 1.5% more boys. This can be used as a null hypothesis: Rather than testing against a difference in proportions of 0, the researcher tests against 0.015.

Even though 0.015 might be a more sensible value for a null hypothesis, our researcher knows that the true sex ratio in their population will differ slightly from that in the entire world. Rather than testing a point null hypothesis (setting to one specific value), the researcher decides to define a range of values around the difference in proportions of 0.015 that can be considered trivially small, even when the true ratio in the population is not exactly 0, or even 0.015. The researcher can for example test if the difference is smaller than -0.005, or larger than 0.035. This test against two bounds, with as a range rather than one value, is known as a *minimal effects test* (Murphy, Myors, & Wolach, 2014).

However, even if the researcher would consider any difference in percentages smaller than or larger than a rejection of the null hypothesis, not every difference is large enough to matter in practice. As long as gender differences in the participation rate are not too extreme, there is no need to spend money to address the gender imbalance in the voluntary participation rate, and all money can be spent on the core training program. To test whether the gender difference in participants is *not* large enough to matter, the researcher can perform an *equivalence test*. Equivalence tests examine whether values below and above a prespecified range can be rejected. Here, we would test if the presence of gender differences that are large enough to matter can be rejected. After an extensive discussion with experts, the researcher decides that as long as the difference in proportions is not larger than 6% the researcher can act as if the gender difference is too small to intervene. Given an expected true difference in the population of 0.015, the researcher will test if the observed difference in the data is smaller than -0.055, and larger than 0.075. If differences more extreme than both these boundary values can be rejected in two one-sided tests, the researcher will conclude statistical equivalence, and no money will be spent on addressing a gender difference in participation.

For an alpha level of 5%, we can conclude statistical equivalence whenever a 90% confidence interval around the observed estimate does not contain the smallest effect size of interest (SESOI). A 90% confidence interval is used because one-sided tests are performed, and it is not necessary to control for multiple comparisons when performing two one-sided tests because both tests need to be statistically significant to conclude equivalence. Note that although we will use an alpha of 5% in all examples because this alpha level was used in the papers we reanalyze, researchers should carefully decide upon a desired Type 1 error rate, and justify their choice (Lakens et al., 2017). Figure 1 illustrates the possible conclusions we can draw from null hypothesis, minimal effect, and equivalence tests. Although thinking about equivalence tests in terms of the values that can be rejected based on a confidence interval might be the easiest conceptualization, for any given study, you can also think of an equivalence test as determining whether effect sizes or test statistics are closer to zero than some critical, or even whether the *p*-value for a NHST is larger than some *p*-value bound.

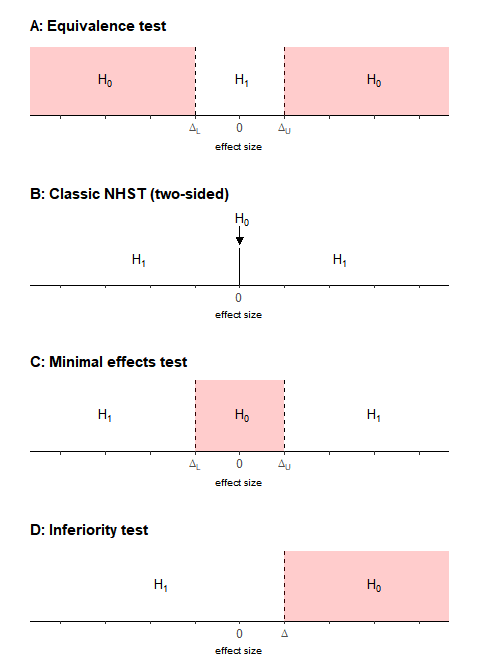


Figure caption: Figure 1. Illustration of null hypotheses () and alternative hypotheses () for different types of significance tests. **A)** Equivalence test: Tests the hypothesis () that an effect is smaller than *or* larger than . **B)** NHST: Tests the hypothesis () that an effect is equal to 0. **C)** Minimal effects test: Tests the hypothesis () that an effect is larger than *and* smaller than . **D)** Inferiority test: Tests the hypothesis () that an effect is larger than .

Even though equivalence tests are just a small variation of traditional frequentist null-hypothesis significance tests (and test against the SESOI, rather than against 0), their use in psychology was limited, perhaps due to the lack of user-friendly software to perform the calculations in the past (Lakens, 2017). In this article, we provide several examples of equivalence tests which illustrate the procedure in free software (R, a spreadsheet, and in Jamovi). Furthermore, we discuss different approaches to determining the SESOI for psychological research, and provide detailed reproducible examples of how to perform power analyses when designing equivalence tests, and statistical re-analyses of published psychology experiments.

# Justifying the Smallest Effect Size of Interest

Equivalence tests are performed against a value that is considered the smallest effect size of interest (SESOI). Although we use symmetrical bounds around 0 in the examples we will discuss below (e.g., , ), it is possible to use asymmetric bounds, for example based on an explicit cost-benefit analysis (e.g., , ). The SESOI can sometimes be determined objectively, for example based on just noticeable differences . Most often, however, it is a subjective decision that varies across researchers, fields, and time. In lieu of objective justifications, the SESOI should ideally be based on a cost-benefit analysis. Since both costs and benefits are necessarily subjective, the SESOI will depend on the researcher who designs the study. The goal of setting a SESOI is to clearly justify why designing a study that has a high probability of rejecting effects larger than a specified value contributes to our knowledge base. A SESOI should be chosen such that inferences based on it answer a meaningful question.

## Objective Justifications of a SESOI

An objectively determined SESOI should be based on quantifiable theoretical predictions, such as computational models. Sometimes, the only theoretical prediction is that an effect should be noticeable. In such circumstances, the SESOI can be set based on just noticeable differences. For example, Burriss and colleagues (2015) examined whether women displayed an increase in 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 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 and colleagues collected data from 22 women and showed that there was indeed an increase in redness of the facial skin during their fertile period. However, this increase was not large enough for men to detect with the naked eye, thus falsifying their hypothesis. Because the just noticeable difference in redness of the skin can be measured, it is possible to objectively establish the SESOI.

Another example of an objectively determined SESOI can be found in Button et al. (2015) where the minimal clinically important difference on the Beck Depression Inventory - II was determined by asking 1039 patients when they subjectively felt less depressed (i.e., when they personally noticed an improvement) and relating this to the corresponding difference score on the depression inventory.

Norman, Sloan, and Wyrwich (2003) used a wide range of methods to determine the minimally important (or detectable) difference in measures for the quality of life in chronic diseases, and observe a surprisingly consistent minimally important difference of half a standard deviation, or d = 0.5. They make the argument that this bound of approximately half a standard deviation could be related to the limits of human discriminability. Therefore, even though a SESOI of d = 0.5 might seem quite large, it could be a plausible equivalence bound when examining effects that should be noticeable by single individuals.

## Subjective justifications of a SESOI

We distinguish between three categories of subjective justifications for the SESOI. First, researchers can use benchmarks. For example, one might set the SESOI to a standardized effect size of d = 0.5, which would allow one to reject effect as large or larger than a 'medium' effect size (Cohen, 1988). Similarly, effect sizes smaller than a Cohen's d of 0.1 are sometimes considered trivially small (Maxwell, Lau, & Howard, 2015). Relying on a benchmark is the weakest possible justification of a SESOI, and should be avoided. Based on a review of 112 meta-analyses, Weber and Popova (2012) conclude that setting a SESOI in communication research to a medium effect size (*r* = 0.3, or d = 0.5) corresponds to rejecting the highest 25% of effect sizes reported in communications research, and Hemphill (2003) suggests that in psychology a SESOI of d = 0.5 would imply rejecting effects as large as the upper 33% of effect sizes reported in the psychological literature.

Second, researchers can determine the SESOI based on previous work in the literature. Ideally, researchers would specify their SESOI in studies they report, 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, given an earlier study. Simonsohn (2015) recently proposed to set the SESOI to the effect size that an earlier study on the same effect would have had 33% power to detect. With this so-called 'small telescopes' approach, the equivalence bounds are thus determined 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. For a two-sided test with an alpha of 0.05, this test 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 defined the SESOI based on the mean of effect sizes reported in the literature. Thus, they examined whether their replication study could reject effects as large or larger than had on average been reported in the literature. Given random variation and bias in the literature, a more conservative approach could be to use a lower bound of a confidence interval around the meta-analytic effect size estimate (cf. Perugini, Gallucci, & Constantini, 2014).

Another justifiable choice for the SESOI is to use the smallest observed effect size that could have been statistically significant in a previous study. In other words, we decide that effects that could not have yielded a in an original study will not be considered meaningful in the replication study either, even if they are found to be statistically significant. Based only on the alpha level and the sample size, we 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., ), which can thus be interpreted as a *critical effect size*[[1]](#footnote-29). All observed effect sizes 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. By setting the SESOI to the critical effect size, an equivalence test can reject all observed effect sizes that could have been detected in an earlier study.

Third, researchers can set the SESOI based on the sample size they are planning to collect. The amount of data you can collect limits the inferences you can make. Given the alpha level and the sample size, researchers can calculate the smallest effect size that they have sufficient power to detect[[2]](#footnote-30). This approach can be used when there are no quantitative predictions made by a theory, and no previous studies have been performed, but researchers *can* justify their sample size. In essence, this approach allows researchers to reject the presence of effects that can be reliably detected given a specific sample size. For example, a researcher who plans to perform a (two-sided) one-sample *t*-test using an alpha of 5%, based on data from 100 observations, has 90% power to detect an effect of d = 0.364. Using equivalence bounds of and in an equivalence test allows one to reject effects the study had high statistical power to detect. An equivalence test based on this approach does not answer any theoretical question (after all, the equivalence bounds are not based on any theoretical predictions) but it answers a *resource* question: Can effects be rejected that could reliably be detected given a specific sample size? In this case, concluding equivalence would suggest that effects which could reasonably be detected with a sample size of 100 can be rejected, and that future studies probably need larger sample sizes to examine a specific question. Whether or not the answer such an equivalence test provides is interesting depends on the sample sizes used in the study, and in part on how typical the chosen sample size is for the research area. An equivalence test based on 15 observations testing against a SESOI of d = 1 cannot be expected to substantially increase our knowledge, given that most effects in psychology are substantially smaller than these bounds and require larger sample sizes to be detected. By transparently reporting the effects one can detect and reject, based on the study design, researchers can communicate the information their study contributes, and provide a starting point for a discussion about what a reasonable SESOI is.

# Raw versus standardized equivalence bounds

Equivalence bounds can be set in terms of standardized effect sizes, or raw differences. For example, the SESOI might be a raw mean difference of 0.5 points on a 7-point scale, or a Cohen’s d of 0.5. The main difference is that equivalence bounds set in raw differences are independent of the standard deviation, while equivalence bounds set as standardized effects are dependent on the standard deviation (since they are calculated as [FORMULA (X1-X2)/SD]. The observed standard deviation randomly varies across samples. In practice, this means that when you use standardized differences as bounds (e.g., d = 0.35) the equivalence test depends on the standard deviation in your the sample. The equivalence bound for a raw mean difference of 0.5 equals a standardized equivalence bound of d = 0.5 when the standard deviation is 1, but a standardized equivalence bound of d = 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 research lines, it is often easier to set equivalence bounds based on standardized differences. Researchers should realize that equivalence bounds based on raw differences or standardized differences ask slightly different questions, and justify their choice for an equivalence bound. When setting equivalence bounds based on earlier research, such as in replication studies, equivalence bounds based on raw or standardized differences would ideally give the same result, and large differences in standard deviations between studies are as important to interpret as large differences in means.

In the remainder of this article we will provide five detailed examples of equivalence tests performed on published studies. These concrete and easy to follow examples will illustrate all approaches to setting equivalence bounds discussed above, and demonstrate how to perform and report equivalence tests. The code to reproduce these analyses is available at XXXX.XXXX.

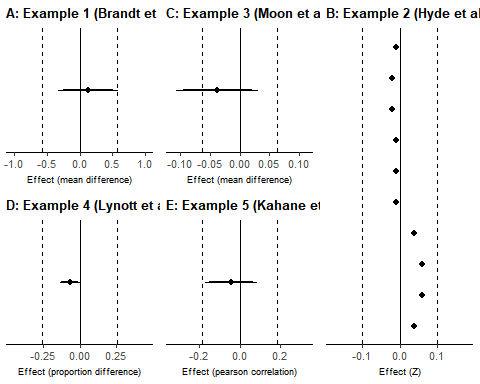


Figure caption: Figure 2. Effects with TOST CIs (thick lines) and NHST CIs (thin lines), NHST H0 (solid vertical line) and TOST H0 (dashed vertical lines) displayed. **A)** Example 1 - The mean difference. **B)** Example 2 - The effects included in the meta-analysis. **C)** Example 3 - The mean difference. **D)** - The proportion difference.

## Example 1: Statistically Equivalent and Not Statistically Different

Banerjee, Chatterjee, & Sinha (2012) reported that 40 participants who had been asked to describe an unethical deed from their past judged the room to be darker than participants who had been asked to describe an ethical deed (, , , , , , ). A close replication by Brandt, IJzerman, & Blanken (2014) with 100 participants found no significant effect (, , ). Following the small telescopes approach, we can calculate the effect size the original study had power to detect --- --- and use this as our SESOI. If we run a TOST with Welch's *t*-test for independent samples and equivalence bounds of and , we indeed find that the effect reported by the replication study is statistically equivalent, , (Figure 2A).

## Example 2: Statistically Equivalent and Statistically Different

Hyde, Lindberg, Linn, Ellis, and Williams (2008) report that effect sizes for gender differences in mathematics tests across the 7 million students in the US represent trivial differences, where the authors specify a trivial difference as an effect size smaller than d = 0.1. For example, in grade 3 the difference is d = 0.04, with a standard error of 0.002. When we perform equivalence tests on the meta-analytic effect sizes of IQ difference for grades 2 to 11 (using an alpha level of 0.005 to correct for multiple comparisons), using equivalence bounds of d = -0.1 and d = 0.1, we see that indeed, all effect size estimates are measured with such high precision, we can conclude that they fall within the equivalence bound of d = -0.1 and d = 0.1, and can be considered trivially small, according to the bounds set by the authors (figure 2B). However, note that all of the effects are also statistically different from zero, as one might expect when there is no random assignment to conditions, and samples sizes are huge. This is one way in which the use of equivalence tests improve hypothesis testing procedures in science, by distinguishing between statistical significance and practical significance.

## Example 3: Not Statistically Equivalent and Not Statistically Different

Moon & Roeder (2014), replicating Shih, Pittinsky, and Ambady (1999), conducted a study to investigate whether Asian-American women would perform better on a maths test when primed with their Asian identity. They observed a slightly reversed difference between the Asian primed group (*n* = 53, *M* = 0.46, *SD* = 0.17) and the control (*n* = 48, *M* = 0.50, *SD* - 0.18) which was not significant, d = -0.23, *t*(96.62) = 1.15, *p* = 0.26, two-sided). A non-significant null-hypothesis test leaves two possible conclusions:

1. The data suggest the absence of a meaningful effect, and we can reject effects large enough to be meaningful.
2. The data are not sensitive enough to tell us whether a meaningful effect is present or absent.

Let us imagine that we consult a teacher about the test, and discover that that grades for this test are set at every 6.25% increase in correct answers (F = 0% to 6.25% ... A+ = 93.75% to 100%). We can decide that we are only interested in test differences so far as they correspond to at least a 1 grade point increase or decrease. Thus, our SESOI becomes a difference in raw scores of 6.25%, or 0.0625.

We then calculate the TOST for a two-sample Welch’s t-test, assuming an alpha of 0.05 (two-sided), and using +/- the SESOI of 0.0625 as our equivalence bounds ().

We find that the TOST is non-significant, t(97.77) = 0.71, p = 0.24 (figure 2C). Thus, we cannot reject the possibility that the true effect really is larger than our SESOI.[[3]](#footnote-36) When results are neither statistically different from zero nor statistically equivalent, there is insufficient data to draw conclusions. Further studies are needed, which can be analyzed using a (small-scale) meta-analysis. The additional data will narrow the confidence interval around the observed effect, allowing us to reject the null, reject the SESOI, or both. Analogous to the large sample sizes needed to detect small effects, a lot of data is needed to reject the SESOI when very narrow bounds are used for the equivalence test (e.g., = -0.1 and = 0.1).

## 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 (replicating Williams and Bargh, 2008). They observed that 51% of participants who received a cold pack () opted to receive a reward for themselves, while 57.00% of participants who received a warm pack () did the same. In a z-test for the difference between the proportions, this effect is not statistically significant (%, , ). Having obtained a non-significant result, we are left with the same two alternatives as in example 3.

In this case, it is not really clear what an objective criterion for the smallest effect size of interest could be. However, since this is a replication, we could decide that we do not care about an effect if the difference is smaller than the smallest effect that the original study could detect. This essentially means that we use the critical z value (~1.96 in a two-tailed test with an alpha of 0.05) as our bounds (). To figure out what difference corresponds to a critical z in the original study, we multiply critical z with the standard error:

Having determined the SESOI, we may also decide that since this is a replication, we are only interested in whether the effect of receiving a cold pack is significantly *lower* than the smallest detectable effect of the original study. In this case, we can still use a TOST to analyse the data, but we only need to consider the test against because an effect with CIs crossing can still be *lower* than the original effect.

We find that the test against is significant, -9.12, < 0.001 (see figure 2D). Thus, we can conclude that the statistically non-significant effect is also statistically inferior to our SESOI.

## Example 5: Statistically Equivalent and not Statistically Different

In a paper published in 2015, Kahane and colleagues investigate moral dilemma vignettes in which participants have to decide whether or not they would sacrifice one person's life to save several other lives, or judge how morally admissible such actions are. Traditionally, greater endorsement for sacrificing a life to save others has been interpreted as a more "utilitarian" moral orientation, i.e. a stronger concern for the greater good (in total, fewer people lose their lives). In a number of studies, Kahane et al. contest this interpretation, for example by showing that greater endorsement for such "utilitarian" choices correlates with sub-clinical psychopathy, and, crucially, by comparing the traditionally used vignettes with a set of new ones that pit partial motivations against an impartial concern for the greater good (e.g. buying a new mobile phone vs. donating the money to save lives in a distant country, study 4).

Kahane et al. find no significant correlation between the perceived wrongness of a utilitarian choice in the classical sacrificial dilemmas and the new "greater good" dilemmas, , ([[4]](#footnote-39)). They conclude that the classical vignettes fail to capture "true"" utilitarianism, but also grant that this conclusion is dependent on the power of their study and that better-powered future studies might overturn their verdict: "Thus, while we cannot rule out the possibility that such an association could emerge in future studies using an even larger number of subjects or different measures, we submit that, in light of the present results, a robust association between ‘utilitarian’ judgment and genuine concern for the greater good seems extremely unlikely." (p. 206). This inference --- that the result would be surprising if there was a true effect of a size the study could have detected --- can be formalized with an equivalence test for correlations and bounds set to an effect size the study had reasonable power to detect: With 231 participants, the study had power to detect effects of . Given bounds of and , we find that the result is indeed statistically equivalent, , . This means that values smaller than and larger than can be rejected at an alpha level of and used as an initial benchmark which future studies could challenge using larger samples and more narrow equivalence bounds.

## Box 1: Calculating an equivalence test in R

If you want to calculate an equivalence test based on summary statistics, you can use the "TOSTER" package in the open-source programming language R. Using TOSTER requires very little experience with R, or programming in general. Below are the few simple steps required to calculate an equivalence test with TOSTER.

1. Download, install, and open R.
2. Run install.packages("TOSTER") in the console.
3. Run library(TOSTER) in the console.
4. The example below reproduces the result of example 1 in R. The parameters of the test are defined inside the parentheses, and you can run ?TOSTtwo in the console to get information about them. Simply copy the example to the console, replace the values with the corresponding values of your own study, and run the code. Results and a plot will automatically be printed.

TOSTtwo(m1 = 4.7857, m2 = 4.6569, sd1 = 1.0897, sd2 = 1.1895,   
 n1 = 49, n2 = 51, low\_eqbound\_d = -0.6401696, high\_eqbound\_d = 0.6401696,   
 alpha = 0.05, var.equal = FALSE)

1. The example above uses the TOSTtwo() function in the TOSTER package which calculates an equivalence test for an independent samples t-test. If you need to calculate TOST for a different type of test (e.g. a correlation), you can find documentation for TOSTr() and other functions of the TOSTER package by going to <https://cran.r-project.org/web/packages/TOSTER/TOSTER.pdf> in your web browser.

# Discussion

Equivalence testing is a relatively simple and flexible technique to falsify hypotheses. As long as we can calculate a confidence interval around a parameter estimate of interest, we can compare it to a smallest effect size of interest. The result of an equivalence test can be obtained by mere visual inspection of the CI (Seaman & Serlin, 1998; Tryon, 2001), or by performing two one-sided tests.

Equivalence bounds should be specified before results are known, ideally as part of a pre-registration (cf. CONSORT guidelines, Piaggio et al., 2006). Researchers might be tempted to set equivalence bounds *after* looking at their data, but this "boundary hacking" will increase error rates.

As with any statistical test, the usefulness of the result of an equivalence test depends on the question we ask. The question manifests itself in the bounds we set: Is our data surprising if the true effect lies outside of the bounds? If we test against very wide bounds, observing statistical significance can hardly be considered surprising, given that most effects in psychology are small to medium (Hemphill, 2003). Examining the papers citing Lakens (2017), we see some researchers state a SESOI without providing a justification for it (Brown and colleagues, 2017; Schumann and colleagues, 2017)[[5]](#footnote-43). 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 provide a detailed justification of their chosen SESOI, given that this is in essence a justification for why they ask a specific question from their data.

For many researchers, the biggest challenge in designing studies where they will perform both NHST and equivalence tests by default will be the need to specify the smallest effect size of interest. Psychological theories are usually too vague to derive 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 new hypothesis, any choice will be arbitrary to some extent. In some research lines researchers might simply use equivalence tests to reject consecutively smaller effect sizes by performing studies with increasingly larger sample sizes while controlling error rates, until no one is willing to invest the time and resources needed to examine the presence of even smaller effects. Nevertheless, it is important to realize that not specifying a SESOI in our research questions will severely hinder theoretical progress. While our initial choices may be arbitrary and subjective, our goal should be to use them as starting points to develop better and more falsifiable theories over time, and ultimately be able to make more precise predictions than we can today.

# References

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(15), 3269-3279. <https://doi.org/10.1017/S0033291715001270>

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), e0130093. <https://doi.org/10.1371/journal.pone.0130093>

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Social Psychological and Personality Science, 8(4), 355–362.

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), e0130093. doi:[10.1371/journal.pone.0130093](https://doi.org/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*(15), 3269–3279. doi:[10.1017/S0033291715001270](https://doi.org/10.1017/S0033291715001270)

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

1. This will typically, although not always, correspond to the effect size the study had power to detect (see Lenth, 2007). This procedure will thus result in equivalence bounds that are substantially wider than the ones obtained using the small telescopes approach, which gives the effect size a study had power to detect. [↑](#footnote-ref-29)
2. This approach is conceptually very similar to the *power approach*, where the effect size you had 95% power to detect is calculated, and the presence of effects more extreme than this value is rejected after observing a *p*-value *larger than* 0.05 in a traditional NHST. However, Meyners (2012) explains this approach is not recommended (even though it is common) because it ignores the possibility that effects are significant *and* equivalent, and the approach does not accurately control error rates. [↑](#footnote-ref-30)
3. In this example, the bounds of the TOST was set based on a raw effect of 6.25%. We could have converted this raw effect into a standardized effect (Cohen's d) for the current sample, and we would have gotten the exact same result, only then in Cohen's d (. Since this is a replication of a previous study, we could also have chosen to calculate the Cohen's d that corresponds to the raw SESOI in the *original* study, and used this as the bounds for the replication. Now we would no longer be answering a question about a raw effect independent of the samples. Instead, we are asking whether we in the replication can reject standardized effects larger than a certain standardized effect in the original study. Since the standard deviations of the original study were only slightly higher than in the replication in this case, Cohen's d (and thus the bounds) would also be only slightly smaller (. As standard deviations of the original study compared to the replication become larger, the bounds become more narrow. [↑](#footnote-ref-36)
4. Study 4 in Kahane et al. (2015) had a final sample size of , but due to missing data in 1 case, the correlation reported here is based on a sample of only . [↑](#footnote-ref-39)
5. In these two cases, one could perhaps guess that benchmark values for a Cohen’s d of “medium” size was chosen. [↑](#footnote-ref-43)