Three-Sided Testing to Establish Practical Significance: A Tutorial

Peder Isager & Jack Fitzgerald*

December 11, 2024

Abstract

Researchers may want to know whether an observed statistical relationship is either meaningfully negative, meaningfully positive, or small enough to be considered practically equivalent to zero. Such a question can not be addressed with standard null hypothesis significance testing, nor with standard equivalence testing. Three-sided testing (TST) is a procedure to address such questions, by simultaneously testing whether an estimated relationship is significantly below, within, or above predetermined smallest effect sizes of interest. TST is a natural extension of the standard two one-sided test for equivalence (TOST). TST offers a more comprehensive decision framework than TOST with no penalty to error rates or statistical power. In this paper, we give a non-technical introduction to TST, provide commands for conducting TST in both R and Jamovi, and provide a Shiny app for easy implementation. Whenever a meaningful smallest effect size of interest can be specified, TST should be combined with null hypothesis significance testing as the default frequentist testing procedure.

Keywords: Three-sided testing, equivalence test, hypothesis test, TST, NHST, TOST, SESOI, tutorial, R, Jamovi

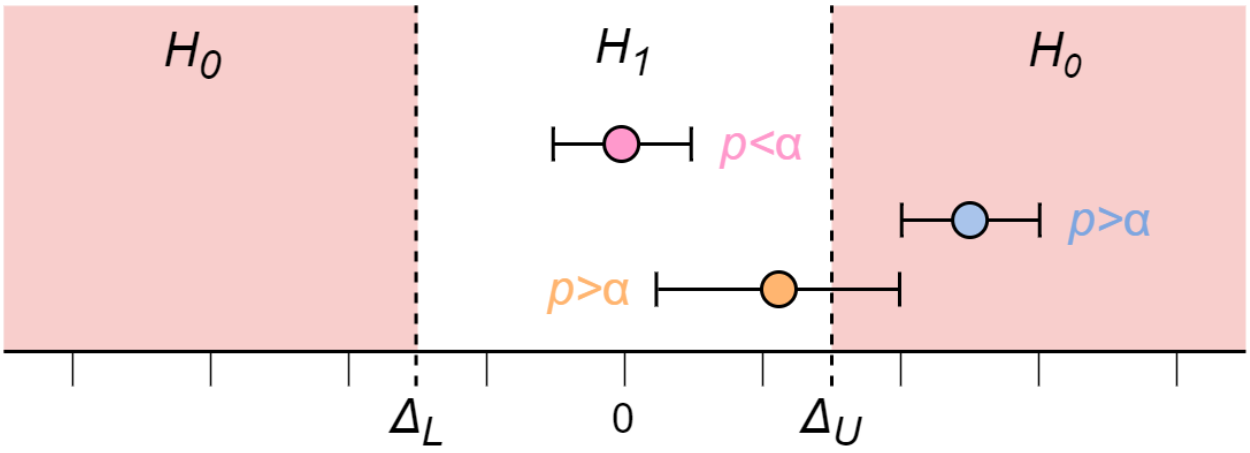
*Isager: Oslo New University College, Peder.Isager@oslonh.no. Fitzgerald: Vrije Universiteit Amsterdam, j.f.fitzgerald@vu.nl. We thank Jelle Goeman for helpful comments and feedback.

1 Introduction

In standard null hypothesis significance testing (NHST), we test whether a given statistical relationship exists by assessing whether we can reject the null-hypothesis (H_0) that this relationship does not exist. If H_0 is rejected, we conclude that the relationship of interest is not zero. A problem with standard NHST is that we can only ever *reject* H_0; we can never *accept* it. Put differently, standard NHST never lets us conclude that the relationship of interest *is* zero. This is because statistically insignificant results are ambiguous in standard NHST. Large p-values can signal precisely measured null relationships or large relationships that are imprecisely measured. Standard NHST makes no distinction between these cases.

This limitation of standard NHST leads to two substantial problems. First, researchers commonly misinterpret statistically insignificant results as evidence that the relationship of interest is zero, even when the chance of a meaningfully large, noisily estimated relationship is high (Greenland et al. 2016; Aczel et al. 2018; Gates & Ealing 2019; Fitzgerald 2024). Second, researchers who are only familiar with NHST have no method to formally test the hypothesis that a relationship is null, which they may often want to. To provide just a few examples, we may want to test whether a drug has *no* negative side effects, whether a new treatment works *as well* as existing treatments, whether a replication result is *equivalent* to the original study result, whether two populations are *similar* on some attribute, etc. None of these predictions can be tested with standard NHST.

A remedy to this situation is the frequentist two one-sided tests (TOST) procedure for equivalence testing (see Lakens, Scheel, & Isager 2018). The TOST procedure allows researchers to test whether an estimate of a relationship is smaller than the 'smallest effect size(s) of interest (SESOI)' for that relationship. In this procedure, a relationship is considered to be 'practically equal to zero' if it can be significantly bounded both (1) above Δ_L and (2) below Δ_U using one-sided tests, where Δ_L and Δ_U are the lower and upper SESOI bounds (respectively). For example, under the TOST procedure, we would conclude that the upper pink estimate in Figure 1 is practically equal to zero, whereas we would not reach this conclusion for the middle blue or bottom orange estimates. The TOST procedure gives researchers a method to distinguish imprecise estimates of potentially large relationships



Note: Estimates are presented alongside $100 \times (1 - 2\alpha)\%$ confidence intervals. Δ_L and Δ_U represent the lower and upper SESOI bounds (respectively). The red H_0 regions more extreme than the SESOI, whereas the H_1 region is closer to zero than the SESOI.

Figure 1: Examples of the Two One-Sided Test of Equivalence

from precise estimates of trivially small relationships.

However, the TOST procedure has its own limitation: in TOST, it is not possible to accept the hypothesis that the relationship of interest *is* large enough to care about – i.e., that it is more extreme than the SESOI. Put differently, the TOST procedure does not distinguish between estimates that significantly exceed the SESOI (e.g., the middle blue estimate in Figure 1) and estimates whose relationship with the SESOI is uncertain (e.g., the lower orange point estimate in Figure 1). Combining NHST and TOST is not sufficient to address this issue. For example, the bottom orange estimate in Figure 1 is significantly different from zero and not significantly smaller than the SESOI, but it obviously does not indicate strong evidence for a relationship larger than the SESOI.

A common but flawed solution to this problem is to combine standard NHST with the TOST procedure (e.g., see Campbell & Gustafson 2018). This allows us to test both whether the relationship of interest is different from zero *and* whether it is smaller than the SESOI. However, the procedure does not actually test if the relationship is large enough to care about. Even a combination of standard NHST and equivalence testing approaches cannot tell us whether there is significant evidence that a relationship is larger than its SESOI. Consequently, it is not safe to assume that a significant standard NHST result plus a statistically insignificant TOST result indicates evidence for a practically large relationship. Therefore, even if we combine standard NHST with TOST, these frameworks do not allow

us to formally conclude that a relationship is larger than the SESOI.

To obtain significant evidence that a relationship is larger than its SESOI, we need minimum-effects tests (see Murphy & Myors 1999). Such procedures can test for inferiority, assessing whether a relationship is at least as negative as its lower SESOI bound (see Figure 2A). They can also test for superiority, assessing whether a relationship is at least as positive as its upper SESOI bound (see Figure 2C).

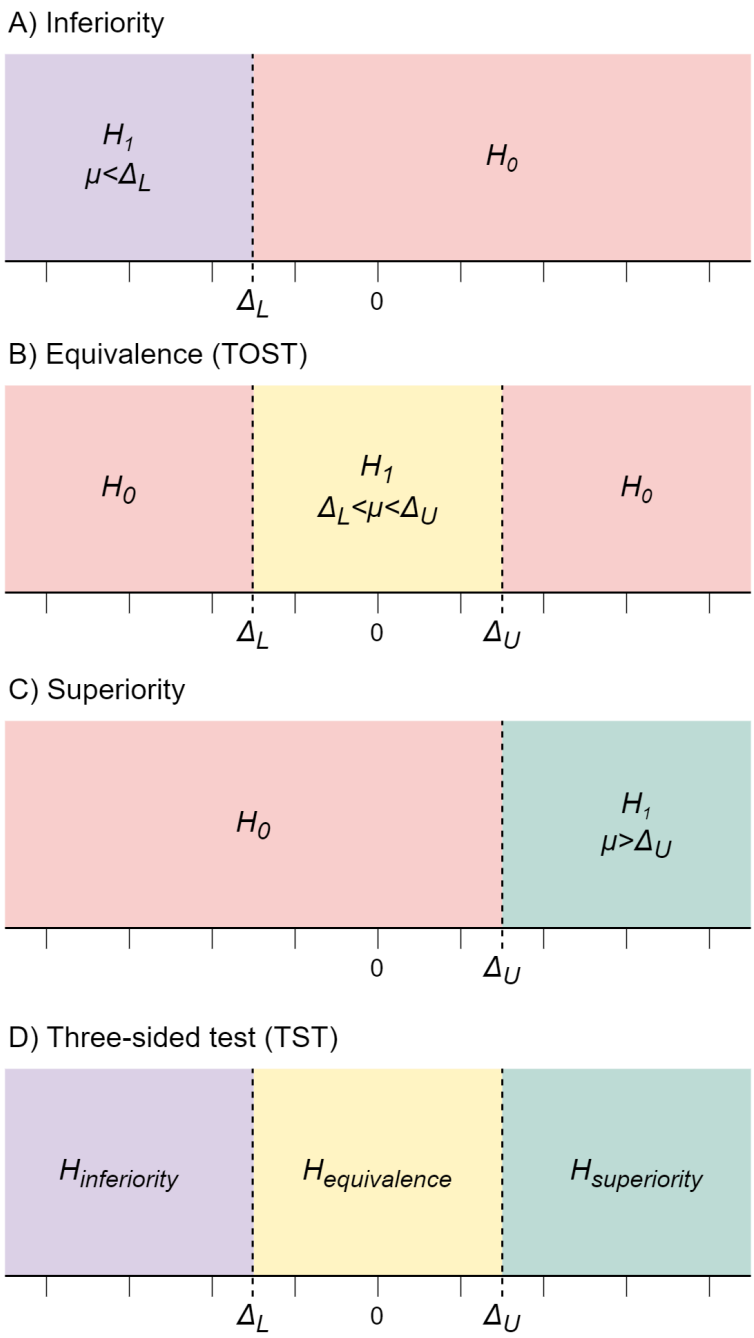
In practice, we may often want to test for superiority, inferiority, and equivalence all at the same time. Fortunately, there is a formal procedure that lets us conduct all three tests simultaneously while controlling false positive rates across all tests, known as the three-sided testing (TST) procedure. TST is simply a combination of TOST and minimum-effects tests. It was initially developed in the field of medical statistics (Goeman, Solari, & Stijnen 2010), and was recently proposed as a default testing approach for economic relationships (Fitzgerald 2024), but the relevance of TST is not restricted to these fields. TST is a uniform improvement over the TOST procedure in any setting where researchers wish to test statistical equivalence, permitting the same power for equivalence testing while simultaneously allowing researchers to test for significant evidence that relationships are bounded outside of their SESOIs. TST thus gives researchers a comprehensive testing framework to assess the practical significance of their estimates.

2 What Is the Three-Sided Testing Procedure?

The TST procedure tests three mutually exclusive hypotheses, all of which are related to the same relationship of interest, and each of which should lead to a meaningfully different conclusion if accepted:

1. H_{Inf}: Relationship of interest $< \Delta_L$ (inferiority)
2. H_{Eq}: $\Delta_L \leq$ Relationship of interest $\leq \Delta_U$ (equivalence)
3. H_{Sup}: $< \Delta_U$ Relationship of interest (superiority)

Δ_L and Δ_U refer to the lower and upper SESOI bounds, respectively. In words, we are testing whether the relationship is smaller than, larger than, or bounded by the SESOI.



Note: Δ_L and Δ_U represent the lower and upper SESOI bounds (respectively). Panel A indicates hypotheses for an inferiority test. Panel B indicates hypotheses for an equivalence test, such as the TOST procedure. Panel C indicates hypotheses for a superiority test. Panel D indicates hypotheses for the three-sided testing procedure.

Figure 2: Hypotheses for (Constitutent Tests of) the Three-Sided Testing Approach

These hypotheses can be tested separately using a two-sided test against the lower SESOI bound to test inferiority (Figure 2A), a TOST procedure to test for equivalence (Figure 1),

and a two-sided test against the upper SESOI bound to test for superiority (Figure 2C). Though the inferiority and superiority hypotheses are technically one-sided, two-sided tests on these hypotheses are required to control error rates from simultaneous tests on all three hypotheses (Goeman, Solari, & Stijnen 2010).

The three-sided testing procedure simply entails running all the tests listed above at once, and then using the combined result of all tests to make an overall inference about the practical significance about the effect (see Figure 2D). For a more technical introduction to TST, see Goeman, Solari, & Stijnen (2010). In this paper, we present an identical formulation of the original TST procedure, combining a TOST procedure, an inferiority test, and a superiority test to make the procedure more easily understandable for researchers who are already familiar with the TOST procedure for equivalence testing.

2.1 Example: Behavioral Research on Sales Tactics

Suppose we do research for a services company. The company sells service packages to clients, offering small, medium, or large packages (increasing in price). Company leadership has noticed that some branches attempt to anchor clients on purchasing the medium package by telling clients that the medium package is the most popular purchase. There is debate amongst company leadership as to whether this sales tactic is a good idea. Anchoring clients on the medium package will increase sales if this tactic draws clients away from purchasing the small package or no package at all, but will decrease sales if the medium package instead substitutes for the large package, or if the anchoring tactic discourages purchases of any package at all.

The company runs an experiment to assess the effects of the anchoring tactic in sales pitches. Company branches are randomized into a treatment group that uses the anchoring tactic and a control group that does not. The company decides that any difference in branch-level monthly sales less than $10,000 is practically equal to zero, and consequently that any difference in such sales exceeding $10,000 is practically meaningful.

The company's official stance on the anchoring tactic will depend on the practical significance of the tactic's impact on sales. If the tactic leads to a loss in revenue larger than -$10,000 per month, then the company will conclude that sales pitches with anchoring are

inferior to pitches without anchoring, and will ban its use in sales pitches across all branches. If the effect is significantly bounded between -$10,000 and $10,000 per month, then the company will conclude that the tactic's effect is practically equal to zero, and will allow branches to freely choose whether or not to use the tactic. If the tactic's effect is significantly greater than $10,000 per month, then the company will conclude that the sales tactic is superior, and will mandate its use in sales pitches throughout the company. If none of these conclusions can be reached with statistical significance, then the company will continue the experiment, collecting data until significant conclusions can be reached.

3 How to Conduct and Interpret Three-Sided Testing

In what follows, we will first present two identical methods for conducting TST. The methods can be applied to any statistical estimate for which (1) a standard error/confidence interval can be computed and (2) test statistics can be reasonably assumed to be Student t- or normally-distributed. We will then give examples of how to conduct TST in practice using the ShinyTST app (https://jack-fitzgerald.shinyapps.io/shinyTST/). In Online Appendices A and B, we also show how to conduct TST in R and in Jamovi (respectively). These supplementary files can easily be adapted to fit different datasets and specifications.

3.1 Specifying a Smallest Effect Size of Interest

The first step in TST is to specify a meaningful smallest effect size of interest (SESOI). Just as in equivalence testing, specifying a meaningful SESOI is the most challenging aspect of TST. A thorough guide to SESOI specification is outside the scope of this article (see Fitzgerald 2024 for further discussion), but we will briefly offer some advice and relevant references to the interested reader.

First, there is a growing consensus that using Cohen's benchmark values to justify the smallest effect size of interest is not advisable (Funder & Ozer 2019; Bakker et al. 2019). Instead, if benchmark values must be used, choose values that are based on the actual distribution of effect sizes in your own research field (e.g., see Richard, Bond, & Stokes-Zoota 2003; Gignac Szodorai 2016; Brydges 2019; Lovakov Agadullina 2021; Nordahl-Hansen

et al. 2024). However, even empirically informed benchmarks are a weak justification for the SESOI. Within fields, there can be large variations in what effect sizes are considered interesting, depending on the specific research topic, the variables in question, the research design used, etc.

A better starting point would be to consult the existing literature on SESOI justification and effect size interpretation. Lakens, Scheel, & Isager (2018) provides a good general introduction to objective and subjective SESOI justification strategies. Funder & Ozer (2019) provides a good introduction to effect size interpretation in general, though we recommend disregarding their proposed benchmark values unless you can justify how they relate to effect sizes in your own field. Beyond these, several insightful discussions of SESOI justification are worth consulting, such as the use of anchor-based methods for justification (Anvari & Lakens 2021), justifications based on elicited stakeholder judgements (Fitzgerald 2024), guidelines for effect size interpretation in intervention research (Kraft 2020; Peetz et al. 2024), and mechanisms that amplify and counteract the practical importance of a relationship (Anvari et al. 2023).

Finally, it may be helpful to study practical examples of how other researchers have justified their SESOI in equivalence testing applications. Equivalence testing primers such as Lakens (2017) and Lakens, Scheel & Isager (2018) have been cited in hundreds of published papers that provide practical examples of how SESOIs are set in applied research settings. We recommend interested readers to search the citing articles for examples of how SESOIs are set in practice.

3.2 Three-Sided Testing Using One-Sided Tests

The first way to conduct TST is by obtaining p-values from the four tests involved in a TST, and then interpreting the joint test result. In the statistical software of your choice, for any relationship you wish to test, run all four of the one-sided tests that make up the inferiority, superiority, and equivalence tests specified in Figure 2A-C. Let μ represent the relationship of interest. Then the four relevant one-sided tests are as follows:

1. **Inferiority.** $H_0: \ \mu \geq \Delta_L$. $H_A: \ \mu < \Delta_L$.

2. **Lower-bound equivalence test.** $H_0: \ \mu < \Delta_L$. $H_A: \ \mu \geq \Delta_L$.

3. **Upper-bound equivalence test.** $H_0: \ \mu > \Delta_U$. $H_A: \ \mu \leq \Delta_U$.

4. **Superiority.** $H_0: \ \mu \leq \Delta_U$. $H_A: \ \mu > \Delta_U$.

Running these four tests will yield four one-sided p-values. We combine all four results to form one overall conclusion about the relationship of interest based on the TST procedure. Suppose that we set a significance level of $\alpha = 5\%$. The procedure will yield one of the following four conclusions concerning the relationship of interest:

1. **Practically significant and negative.** If the one-sided inferiority test statistic is significant at level $\alpha/2$ (i.e., $2p < 0.05$ or $p < 0.025$), then treat the relationship as inferior to the lower SESOI bound. All other tests will be non-significant.

2. **Practically equal to zero.** If both of the two one-sided tests for equivalence are significant at level α (i.e., $p < 0.05$), then treat the relationship as if it is no further from zero than the SESOI. The superiority and inferiority tests will be non-significant.

3. **Practically significant and negative.** If the one-sided superiority test statistic is significant at level $\alpha/2$ (i.e., $2p < 0.05$ or $p < 0.025$), then treat the relationship as superior to the upper SESOI bound. All other tests will be non-significant.

4. **Inconclusive**. In all other cases, remain uncertain about the practical significance of the relationship.

Notice that even though we have set a significance level of 5%, the α threshold that we consider for the inferiority and superiority tests is half that, set at 2.5%. This is necessary to adjust for multiple hypothesis testing. If our relationship of interest is truly bounded between Δ_L and Δ_U, then it is possible to make two Type I errors: one where we mistakenly conclude that the relationship is bounded above Δ_U, and one where we mistakenly conclude that it is bounded below Δ_L. Because the TST procedure involves simultaneous tests both for inferiority and for superiority, a simple Bonferroni correction across these two tests – which halves the effective α for these tests – effectively controls the size of the test at our nominal significance level (Goeman, Solari, & Stijnen 2010). An equivalent method is to double the

one-sided p-values for the inferiority and superiority tests[1], which is the method commonly used in statistical software.

No multiple comparison correction are required for the lower- and upper-bound equivalence tests, because TOST already requires that *both* one-sided tests are significant at level α before estimates are deemed practically equal to zero (see Berger & Hsu 1996). Even though the inferiority and superiority tests in TST must be conducted at significance level $\alpha/2$ to control TST's error rates, we can still safely conduct TST's equivalence test at significance level α. This is a useful property of TST, as it implies that TST can allow researchers to augment TOST with minimum effects testing for inferiority and superiority without sacrificing any power or error rate control for the equivalence test.

3.3 Three-Sided Testing Using Confidence Intervals

An identical procedure for conducting TST is to compute two confidence intervals (CIs) for the relationship of interest – one at $100(1-2\alpha)\%$ confidence and one at $100(1-\alpha)\%$ confidence. We can then inspect where these intervals fall relative to the upper and lower SESOI bounds. Assuming $\alpha = 5\%$, the wider 95% CI is used to evaluate whether the inferiority and superiority tests yield significant results, and the 90% CI is used to assess whether the equivalence test yields significant results.

In the statistical software of your choice, compute the 95% CI and 90% CI for an estimate of your relationship of interest interest. Then compare the upper and lower CI bounds with the upper and lower SESOI bounds. As in Section 3.2, this CI-based procedure will yield one of the following four conclusions about the relationship of interest (assuming $\alpha = 5\%$).

1. **Practically significant and negative.** If the $100(1-2\alpha)\%$ CI (i.e., the 95% CI) falls entirely below the lower SESOI bound, then treat the relationship as inferior to the lower SESOI bound.

2. **Practically equal to zero.** If the $100(1-\alpha)\%$ CI (i.e., the 90% CI) falls entirely within the SESOI bounds, then treat the relationship as if it is no further from zero

[1]Of course, for inferiority/superiority tests that produce $p > 0.5$, this adjusted p-value is mechanically set to $p = 1$.

than the SESOI.

3. **Practically significant and positive.** If the $100(1-2\alpha)\%$ CI (i.e., the 95% CI) falls entirely above the upper SESOI bound, then treat the relationship as superior to the upper SESOI bound.

4. **Inconclusive.** If none of the above conditions hold, then remain uncertain about the practical significance of the relationship.

Figure 3 illustrates this CI procedure. For the equivalence test, what matters is that the entire *90%* CI falls within the H_{Eq} region. Note for example that the third estimate from the top is equivalent even though the *95%* CI overlaps with the SESOI bounds. In contrast, inferiority and superiority tests are only significant if the entire 95% CI falls outside the SESOI bounds. For instance, looking to the fifth estimate in Figure 3, its 95% CI crosses the lower SESOI threshold, meaning that we cannot conclude at a 5% significance level that this estimate is inferior to Δ_L.

3.4 Three-Sided Testing Using the ShinyTST Application

To enable as many readers as possible to apply TST in their own data analyses, we have developed a stand-alone application that can produce TST results for most statistical estimates. The ShinyTST app requires four inputs and has two optional inputs. Required inputs include an estimate (e.g., a mean difference or a correlation coefficient), the standard error of that estimate, the smallest effect size of interest (in the same units as the estimate), and the significance level α. Optional inputs include the degrees of freedom of the test (which are required for exact tests, rather than asymptotically approximate tests) and a power target (which is used to set power-adjusted CIs).[2] Based on this input, the calculator computes 90% CIs, 95% CIs, TST test statistics and p-values, and provides an appropriate conclusion based on the test results. ShinyTST is available online at https://jack-fitzgerald.shinyapps.io/shinyTST/.

[2]Whereas the half-width of an ordinary $100(1-\alpha)\%$ CI is given by the standard error times the two-sided critical value at significance level α, the half-width of a $100(1-\alpha)\%$ power-adjusted CI at $100(1-\beta)\%$ power is given by the minimal detectable effect size at $100(1-\beta)\%$ power given the estimate's standard error (see Bloom 1995).

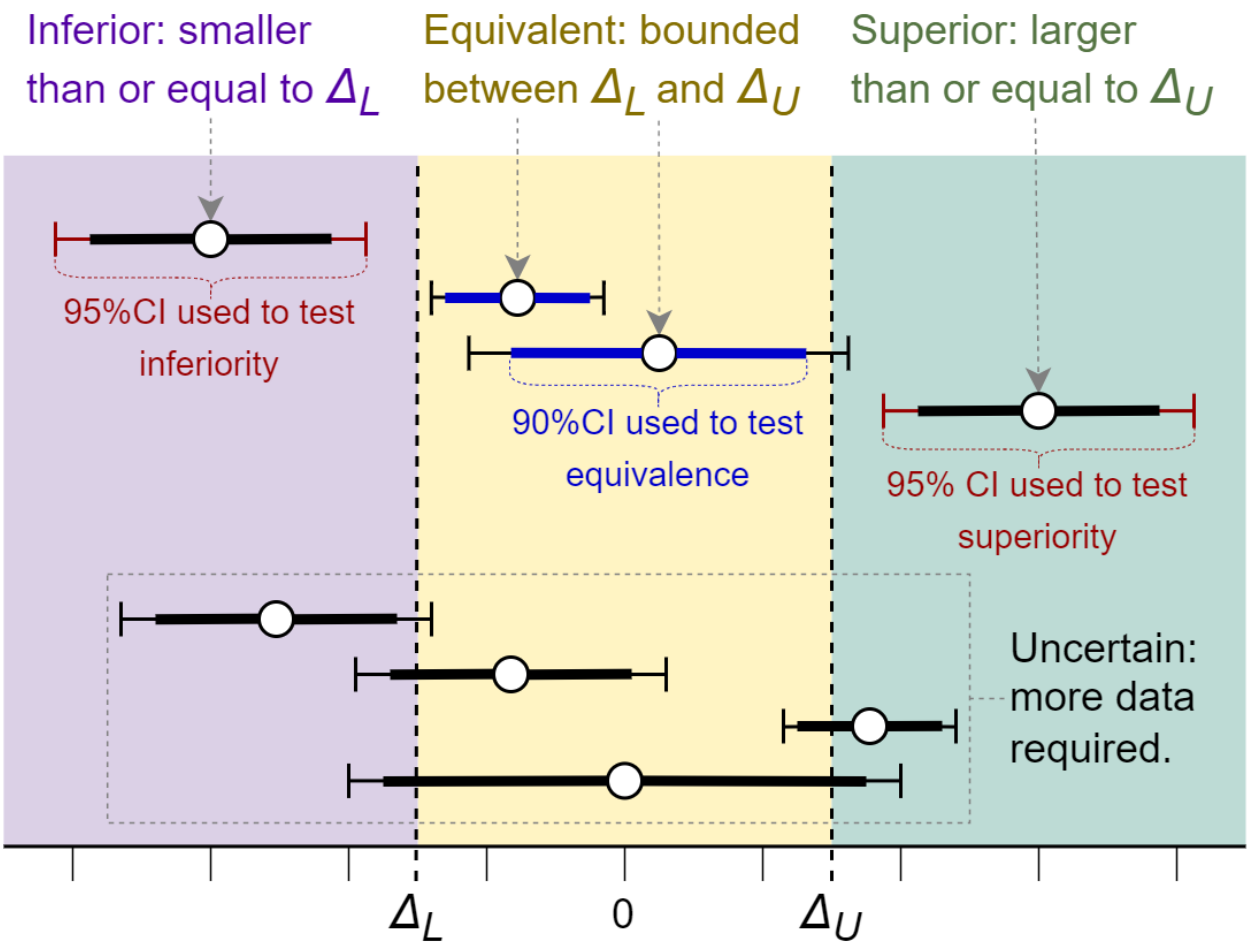


Figure 3: Illustration of how different test results should be interpreted in TST at a significance level of $\alpha = 5\%$. Estimates are plotted alongside 90% CIs (thicker bands) and 95% CIs (thinner bands).

In this section we will demonstrate how to use and interpret TST results in ShinyTST through four applied examples.

Consider again our experiment on anchoring tactics in sales pitches. Based on the observed results in our experiment, we want to decide if sales pitches using the anchoring tactic yield inferior, equivalent, or superior revenue when compared to sales pitches that do not use the tactic. We can conduct TST in the ShinyTST app to help us make this decision.

3.4.1 Example 1: A Practically Significant Estimate

Think back to the services company example from section 2.1. Suppose that in our experiment, we observe that branches employing the anchoring tactic make $18,000 more in

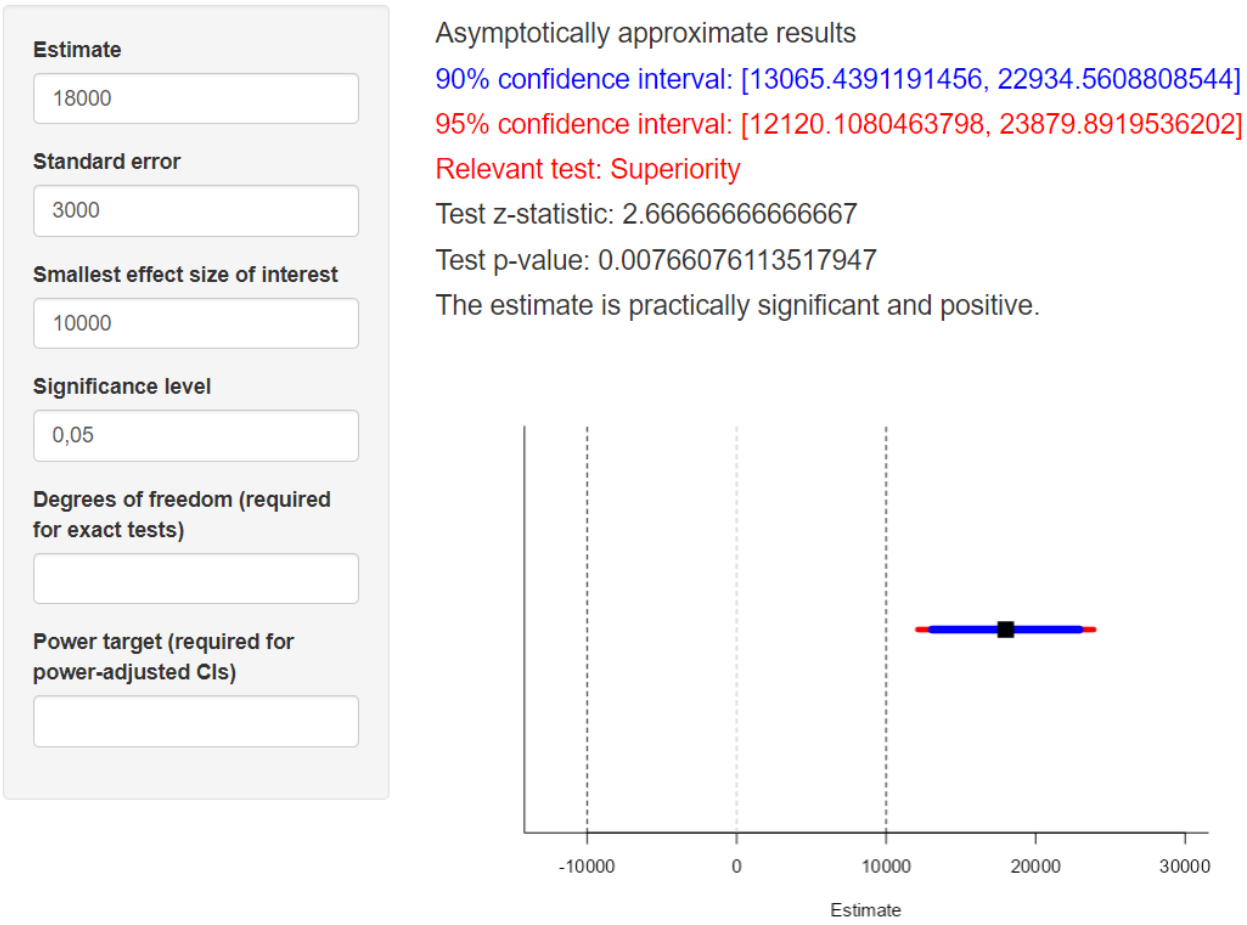


Figure 4: ShinyTST example for an estimate of 18,000, standard error of 3000, and SESOI of 10,000.

monthly sales than branches in the control group, with a standard error of \$3000. Our significance level is 5%, and as aforementioned, our SESOI bounds are $\pm\$10,000$. Figure 4 displays where to input these parameters in ShinyTST, as well as the resulting output.

Following the TST procedure, we would conclude that this estimate is superior to our SESOI. The "Relevant test" output tells us that the relevant test to look at in this case is the superiority test, as the observed estimate is more positive than the upper SESOI bound. The color of the "Relevant test" output signals which reported confidence interval is relevant for drawing a conclusion. In this case, the red-colored 95% CI is the relevant one, and we can see both in the numeric output and the accompanying graph that the lower 95% CI bound is greater than the upper SESOI bound of \$10,000. I.e., the entire 95% CI falls above the smallest effect size of interest. Accordingly, the test p-value falls below α. Note that

ShinyTST has already doubled the superiority test's one-sided p-value, so we do not need to multiply p or divide α by two in this case; this same functionality is provided in this paper's companion `tst` command in R. As reported in the output, these results indicate that "The estimate is practically significant and positive." In other words, in this example, sales pitches using the anchoring tactic appear to be superior to pitches that do not use the anchoring tactic to a practically significant degree.

3.4.2 Example 2: An Estimate Practically Equal to Zero

Now suppose instead that branches employing the anchoring tactic make \$4500 more in monthly sales than branches in the control group, keeping everything else as before. This is a small estimate (in the scale of our company), but is it precise enough that we can confidently rule out practically meaningful differences in monthly sales? Figure 5 displays the TST results for this estimate.

In this case, we would conclude that the estimate is practically equal to zero. I.e., the estimated effect is significantly bounded within $\pm\$10,000$. The "Relevant test" output indicates that the blue-colored 90% CI is the relevant CI for this test (because the point estimate is within the SESOI bounds). This 90% CI falls entirely within the SESOI bounds. The p-value output now displays the p-value of the equivalence test, which is accordingly beneath 5%. As reported in the output, this indicates that "The estimate is practically equal to zero." In other words, sales pitches using the anchoring tactic appear to yield practically equivalent revenue to sales pitches that do not use the anchoring tactic.

3.4.3 Example 3: Inconclusive Results

Now suppose we observe that treated branches make \$15,500 less in sales than control branches. This estimate is more negative than our lower SESOI bound, but is it estimated precisely enough that we should ban usage of the anchoring tactic? Figure 6 displays the TST results for this estimate.

In this case, we arrive at an inconclusive result. As in Figure 4, the "Relevant test" output indicates that the red-colored 95% CI is relevant for this conclusion, as the estimate is more negative than the lower SESOI bound. The p-value output now displays the (adjusted) p-

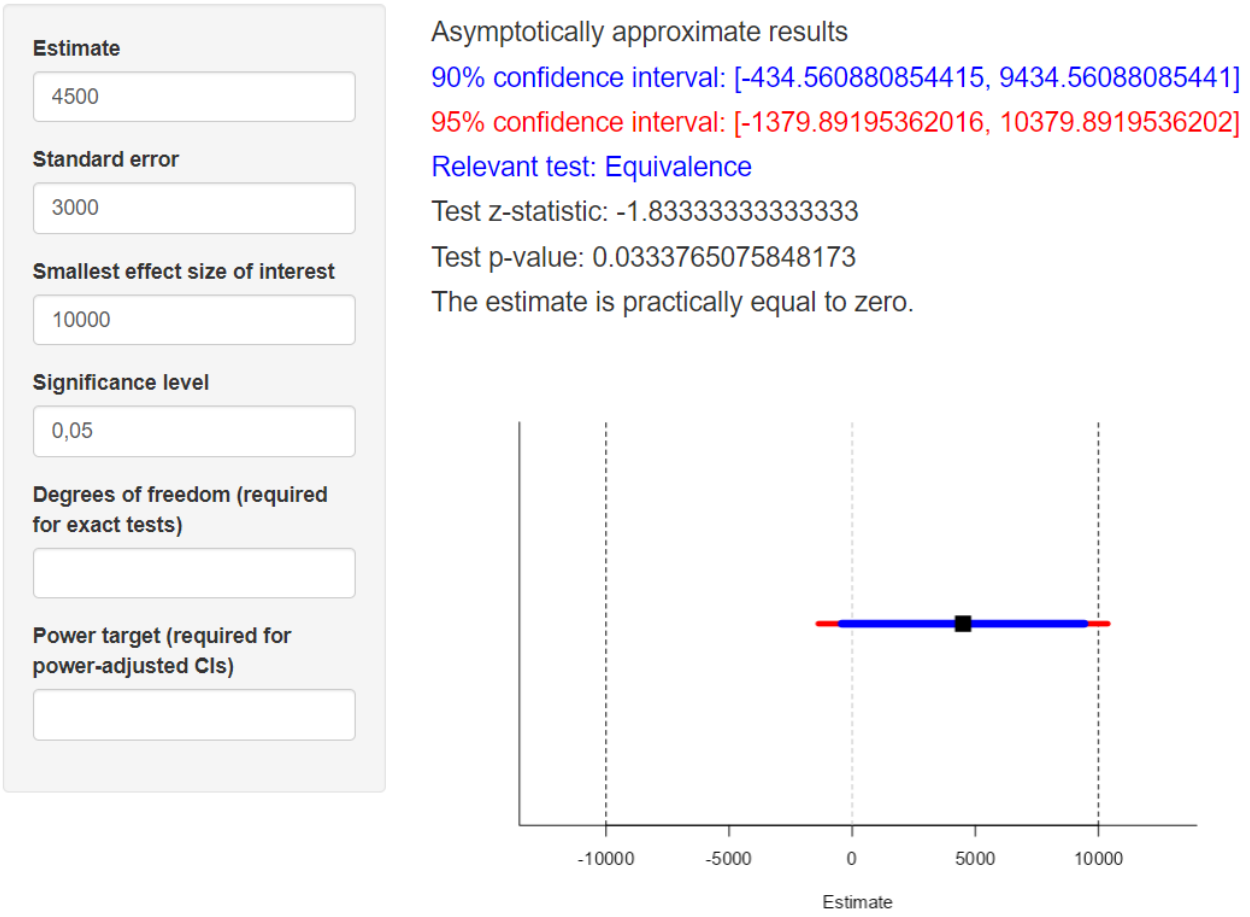


Figure 5: ShinyTST example for an estimate of 4500, standard error of 3000, and SESOI of 10,000.

value of the inferiority test, which exceeds 5%. Accordingly, the 95% CI intersects one of the SESOI bounds (specifically, the lower bound). As reported in the output, given a significance threshold of $\alpha = 5\%$, these results indicate that "The practical significance of the estimate is inconclusive."

Finally, suppose instead that branches using the anchoring tactic make \$7500 less in monthly sales than branches in the control group. As in Example 2, this estimate is smaller in magnitude than the SESOI. But is it precisely bounded enough that we should give branches free reign to use the anchoring tactic at will? Figure 7 displays TST results for this estimate.

This case also yields an inconclusive result. As in Figure 5, the "Relevant test" output indicates that the blue-colored 90% CI is relevant for this conclusion because the estimate

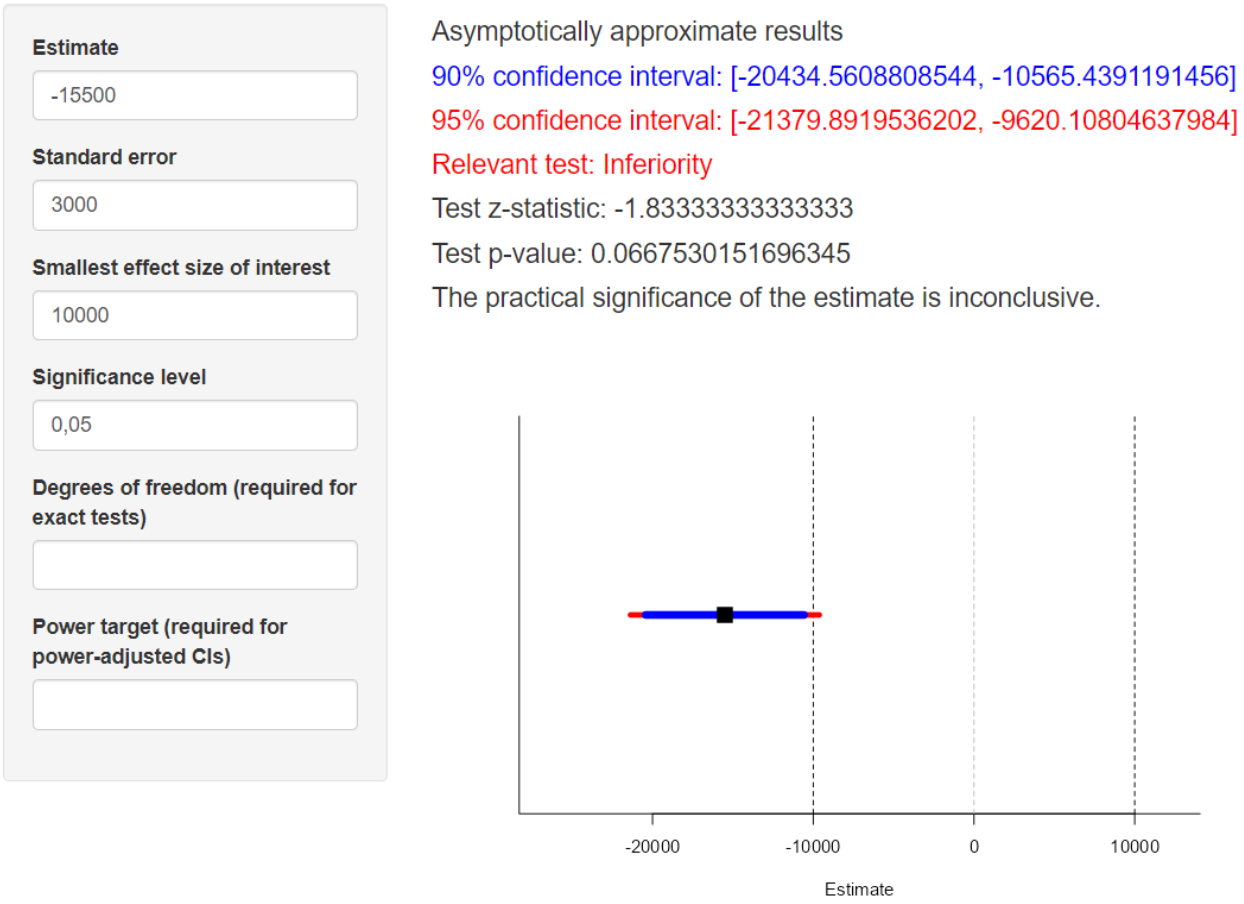


Figure 6: ShinyTST example for an estimate of -15,500, standard error of 3000, and SESOI of 10,000.

is between the SESOI bounds. This estimate's 90% CI intersects one of the SESOI bounds, and the TST p-value accordingly exceeds 5%. In both this case and the case displayed in Figure 5, results are inconclusive. In both cases, this implies that we should continue the experiment and collect more data until sufficiently certain conclusions can be reached about the relative efficacy of the anchoring tactic.

3.5 Three-Sided Testing in R and Jamovi

Some researchers prefer statistical software to point-and-click applications. We therefore supplement the ShinyTST app and the tutorials in this paper with examples of how to conduct TST in both R and Jamovi in Online Appendices A and B (respectively). These examples work with data from a case similar to that discussed in Section 3.4.1, and respectively utilize

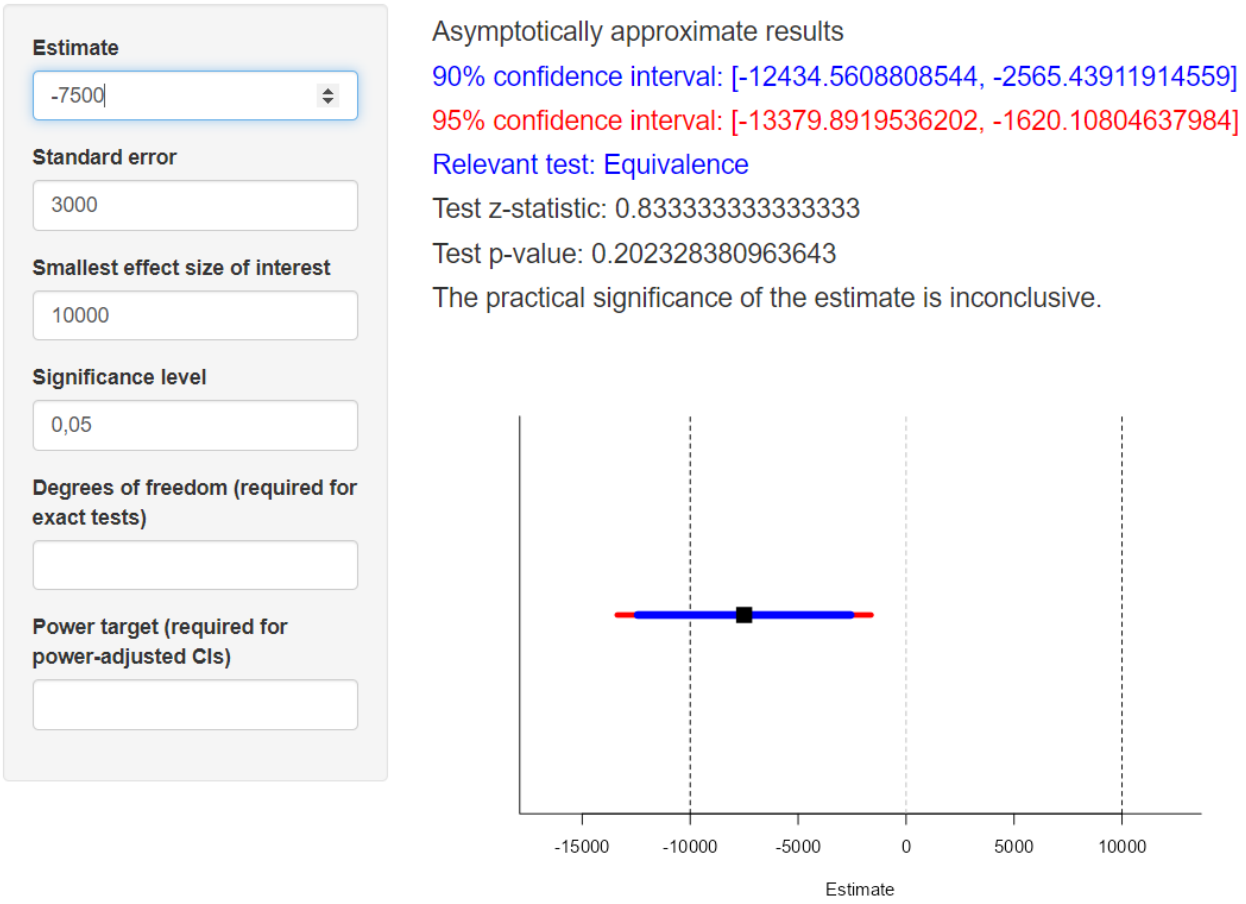


Figure 7: ShinyTST example for an estimate of -7500, standard error of 3000, and SESOI of 10,000.

the `tst` command in the `eqtesting` R package (Fitzgerald 2024) and the `TOSTER` command in Jamovi (Lakens 2017).[3]

4 Reporting Three-Sided Testing Results

When reporting TST, include the following information:

1. **Which test is relevant: inferiority, equivalence, or superiority?** This information is automatically provided by the ShinyTST app and the `tst` command in R, but this can also be inferred directly by examining the point estimate's relationship with the SESOI bounds.

[3]The `eqtesting` R package can be downloaded from GitHub at https://github.com/jack-fitzgerald/eqtesting. Aaron Caldwell deserves credit for developing the `TOSTER` command in R and Jamovi.

2. **What is the TST p-value?** If the superiority (inferiority) test is the relevant test, then this is two times the one-sided p-value of the superiority (inferiority) test; p-values reported by the ShinyTST app and the `tst` R command are already adjusted and do not need to be doubled. If the equivalence test is the relevant test, then the TST p-value is the larger of the two one-sided p-values for the equivalence tests; the p-value reported by the ShinyTST app and the `tst` R command is the larger of these two equivalence testing p-values.

3. **What are the bounds of the $100(1-2\alpha)\%$ and $100(1-\alpha)\%$ confidence intervals?** This can be textually reported with two sets of CI bounds, and visually reported using double-banded confidence intervals such as those produced by the ShinyTST app.

4. **What is the conclusion of the test?** Interpret whether the estimate is practically significant and negative, practically equivalent to zero, practically significant and positive, or whether the TST results are inconclusive.

In addition to visualizing double-banded confidence intervals of the form in Figures 4-7, the TST results in Section 3.4.1 could be reported as follows:

> We conduct three-sided testing to assess whether there is significant evidence that the anchoring tactic's effect on branch-level sales is more extreme than $\pm\$10,000$ per month, which is the smallest effect size of interest in this setting. Our experimental results imply that the anchoring tactic increases branch-level sales by \$18,000 per month (90% CI: [\$13,065.44, \$22,934.56], 95% CI: [\$12,120.11, \$23,879.89]). The superiority test is the relevant test under TST, which yields an adjusted p-value of 0.008 (see Isager & Fitzgerald 2024). The anchoring sales tactic therefore yields significantly more than \$10,000 in additional revenue for company branches, implying that the anchoring tactic increases sales to a practically significant degree.

5 Power Analysis and Sample Size Planning for Three-Sided Testing

Best practices for computing necessary sample sizes for sufficient power differ between standard NHST and TST. Unlike standard NHST, TST procedures test several hypotheses at once. Consequently, having sufficient *a priori* power for conclusive TST results requires having sufficient power for *all tests*. Our recommendations for computing sample sizes for sufficient power thus differ depending on whether you expect to obtain an estimate larger or smaller in magnitude than the SESOI. Online Appendix C provides interested readers with a comprehensive guide.

6 Combining TST With NHST

If we expect the SESOI to change in the future, then it may be worth combining TST with standard NHST. Returning to our experiment on anchoring sales tactics, the company's SESOI for branch-level sales may increase over time as the company grows in size. Likewise, the smallest practically meaningful change in branch-level sales may decline in periods when the company experiences financial distress. It thus may be useful to record whether the anchoring tactic has some nonzero effect on sales to inform future experiments and policies.

To combine TST and standard NHST, simply add a two-sided test against zero to the TST procedure outlined in this paper. Figure 8 shows how this alteration augments the TST procedure. This effectively splits the H_{Eq} region in Figure 2D into two parts: H_{0-} (effect sizes greater than Δ_L but less than zero) and H_{0+} (effects greater than zero but less than Δ_U). As a result, combining TST and NHST does not require additional control for multiple comparisons, and can thus be performed with no loss to statistical power (see Goeman, Solari, & Stijnen 2010). Whether the additional nuance afforded by NHST is of any interest must be for the individual scientist to decide. If the justification for SESOI is strong, then it is not as interesting to know if the effect is different from exactly zero. In such cases, TST should replace standard NHST as the default test procedure. Even if TST and NHST are reported together, the results of the TST should take precedence in the results and discussion.

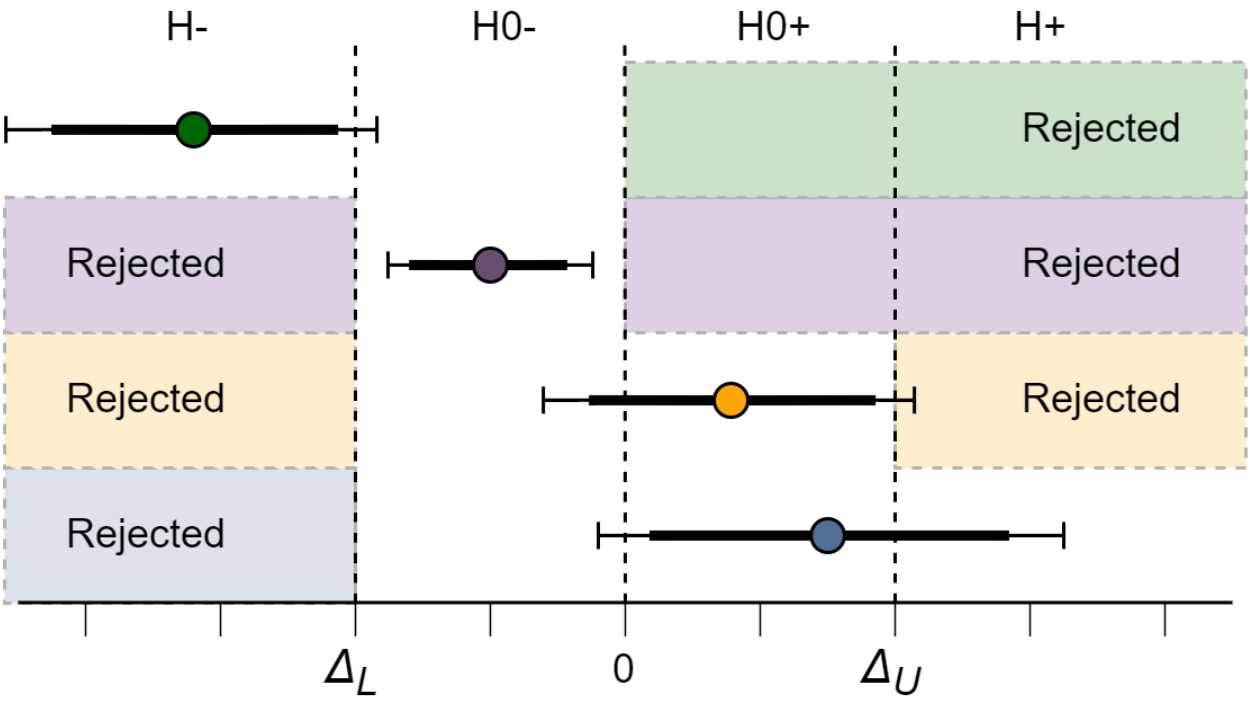


Figure 8: A combined TST and NHST test procedure. Four meaningful test regions can be rejected. H- and H+ are both rejected if the 90% CI is bounded between Δ_L and Δ_U. If the 95% CI is entirely bounded below H_{0+} (above H_{0-}), then H_{0+} (H_{0-}) is rejected. In cases like the green-colored estimate, we may not be able to conclude whether the effect is smaller or larger than Δ, but we can establish the direction of the effect (negative in this case). The colored horizontal bars in the plot illustrate which test regions can be rejected for each estimate in the plot, based on the 90% and 95% CIs of the estimate.

7 Conclusion

Whenever we can specify a meaningful SESOI, TST is a superior testing procedure compared with both standard NHST and TOST. TST allows us to detect significant evidence that a relationship is practically significant or or practically null, yielding a more unambiguous interpretation of non-significant results, and allows us to test more meaningful predictions than the standard NHST test of whether relationships are different from zero (see also Meehl 1967; Lakens 2022). Researchers should therefore strongly consider TST as their default frequentist test procedure if they can specify a meaningful SESOI.

8 Disclosures

8.1 Author Contributions

Conceptualization: P. Isager and J. Fitzgerald; Software: P. Isager and J. Fitzgerald; Validation: P. Isager; Resources: P. Isager and J. Fitzgerald; Writing – Original Draft: P. Isager; Writing – Review & Editing: P. Isager and J. Fitzgerald; Visualization: P. Isager and J. Fitzgerald; Project Administration: P. Isager and J. Fitzgerald; Funding Acquisition: P. Isager and J. Fitzgerald.

8.2 Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

8.3 Acknowledgments

We thank Jelle Goeman for helpful comments and feedback. All errors are our own.

8.4 Funding

J. Fitzgerald gratefully acknowledges PhD funding support from the Amsterdam Law and Behavior Institute. Early drafts of this manuscript were written when J. Fitzgerald was a member of the Superforecaster Panel for the Social Science Prediction Platform. The views expressed in this paper do not necessarily represent the views of the SSPP, nor of the researchers who created and/or operate the SSPP.

References

Aczel, Balazs et al. (2018). "Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation". *Advances in Methods and Practices in Psychological Science* 1.3, pp. 357–366. DOI: 10.1177/2515245918773742.

Anvari, Farid, Rogier Kievit, et al. (Mar. 2023). "Not All Effects Are Indispensable: Psychological Science Requires Verifiable Lines of Reasoning for Whether an Effect Matters". *Perspectives on Psychological Science* 18.2, pp. 503–507. ISSN: 1745-6916. DOI: 10.1177/17456916221091565. (Visited on 06/12/2024).

Anvari, Farid and Daniël Lakens (Sept. 2021). "Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest". *Journal of Experimental Social Psychology* 96, p. 104159. ISSN: 0022-1031. DOI: 10.1016/j.jesp.2021.104159. (Visited on 06/12/2024).

Bakker, Arthur et al. (Sept. 2019). "Beyond Small, Medium, or Large: Points of Consideration When Interpreting Effect Sizes". *Educational Studies in Mathematics* 102.1, pp. 1–8. ISSN: 1573-0816. DOI: 10.1007/s10649-019-09908-4. (Visited on 06/12/2024).

Berger, Roger L. and Jason C. Hsu (1996). "Bioequivalence Trials, Intersection-Union Tests and Equivalence Confidence Sets". *Statistical Science* 11.4. DOI: 10.1214/ss/1032280304.

Bloom, Howard S. (Oct. 1995). "Minimum Detectable Effects: A Simple Way to Report the Statistical Power of Experimental Designs". *Evaluation Review* 19.5, pp. 547–556. ISSN: 0193-841X, 1552-3926. DOI: 10.1177/0193841X9501900504. (Visited on 09/27/2024).

Brydges, Christopher R (Aug. 2019). "Effect Size Guidelines, Sample Size Calculations, and Statistical Power in Gerontology". *Innovation in Aging* 3.4, igz036. ISSN: 2399-5300. DOI: 10.1093/geroni/igz036. (Visited on 06/12/2024).

Cohen, Jacob (July 1988). *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. New York: Routledge. ISBN: 978-0-203-77158-7. DOI: 10.4324/9780203771587.

— (1992). "A Power Primer." *Psychological Bulletin* 112.1, pp. 155–159. ISSN: 1939-1455, 0033-2909. DOI: 10.1037/0033-2909.112.1.155. (Visited on 06/13/2024).

Fitzgerald, Jack (2024). *The need for equivalence testing in economics*. Institute for Replication Discussion Paper Series No. 125. URL: https://www.econstor.eu/handle/10419/296190.

Funder, David C. and Daniel J. Ozer (June 2019). "Evaluating Effect Size in Psychological Research: Sense and Nonsense". *Advances in Methods and Practices in Psychological Science* 2.2, pp. 156–168. ISSN: 2515-2459, 2515-2467. DOI: 10.1177/2515245919847202. (Visited on 06/17/2024).

Gates, Simon and Elizabeth Ealing (2019). "Reporting and interpretation of results from clinical trials that did not claim a treatment difference: Survey of four general medical journals". *BMJ Open* 9.9. DOI: 10.1136/bmjopen-2018-024785.

Gignac, Gilles E. and Eva T. Szodorai (2016). "Effect Size Guidelines for Individual Differences Researchers". *Personality and Individual Differences* 102, pp. 74–78. ISSN: 1873-3549. DOI: 10.1016/j.paid.2016.06.069.

Goeman, Jelle J., Aldo Solari, and Theo Stijnen (Sept. 2010). "Three-Sided Hypothesis Testing: Simultaneous Testing of Superiority, Equivalence and Inferiority". *Statistics in Medicine* 29.20, pp. 2117–2125. ISSN: 02776715. DOI: 10.1002/sim.4002. (Visited on 02/23/2024).

Greenland, Sander et al. (Apr. 2016). "Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations". *European Journal of Epidemiology* 31.4, pp. 337–350. ISSN: 1573-7284. DOI: 10.1007/s10654-016-0149-3. (Visited on 05/24/2024).

Kraft, Matthew A. (May 2020). "Interpreting Effect Sizes of Education Interventions". *Educational Researcher* 49.4, pp. 241–253. ISSN: 0013-189X. DOI: 10.3102/0013189X20912798. (Visited on 06/12/2024).

Lakens, Daniël (2017). "Equivalence Tests: A Practical Primer for t-Tests, Correlations, and Meta-Analyses". *Social Psychological and Personality Science* 8.4, pp. 355–362. DOI: 10.1177/1948550617697177.

— (2022). "Chapter 5. Asking Statistical Questions". *Improving Your Statistical Inferences*.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager (June 2018). "Equivalence Testing for Psychological Research: A Tutorial". *Advances in Methods and Practices in Psychological Science* 1.2, pp. 259–269. ISSN: 2515-2459, 2515-2467. DOI: 10.1177/2515245918770963. (Visited on 08/16/2020).

Lovakov, Andrey and Elena R. Agadullina (Apr. 2021). "Empirically Derived Guidelines for Effect Size Interpretation in Social Psychology". *European Journal of Social Psychology* 51.3, pp. 485–504. ISSN: 0046-2772, 1099-0992. DOI: 10.1002/ejsp.2752. (Visited on 06/13/2024).

Meehl, Paul E. (June 1967). "Theory-Testing in Psychology and Physics: A Methodological Paradox". *Philosophy of Science* 34.2, pp. 103–115. ISSN: 0031-8248, 1539-767X. DOI: 10.1086/288135. (Visited on 08/25/2019).

Murphy, Kevin R. and Brett Myors (1999). "Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model." *Journal of Applied Psychology* 84.2, pp. 234–248. DOI: 10.1037/0021-9010.84.2.234.

Nordahl-Hansen, Anders et al. (Jan. 2024). "Redefining Effect Size Interpretations for Psychotherapy RCTs in Depression". *Journal of Psychiatric Research* 169, pp. 38–41. ISSN: 0022-3956. DOI: 10.1016/j.jpsychires.2023.11.009. (Visited on 06/24/2024).

Peetz, Hannah Katharina et al. (Jan. 2024). *Evaluating Interventions: A Practical Primer for Specifying the Smallest Effect Size of Interest*. DOI: 10.31234/osf.io/3qmj4. (Visited on 07/31/2024).

Richard, F. D., Charles F. Bond, and Juli J. Stokes-Zoota (Dec. 2003). "One Hundred Years of Social Psychology Quantitatively Described". *Review of General Psychology* 7.4, pp. 331–363. ISSN: 1089-2680. DOI: 10.1037/1089-2680.7.4.331. (Visited on 06/17/2024).

Online Appendix

A Three-Sided Testing in R

Supplementary material SM1 (https://osf.io/rfyas) contains a script demonstrates three ways to conduct TST in R. First, one can conduct four one-sided tests in base R. Second, one can compare the 90% and 95% CIs of an estimate with its SESOI bounds. Third and finally, one can use the `tst` function in the `eqtesting` R package (Fitzgerald 2024). This is an R-based version of the ShinyTST app.

B Three-Sided Testing in Jamovi

Supplementary material SM2 (https://osf.io/ncxvg) provides a Jamovi file that contains the same simulated data as the R script, and a TST analysis conducted using Jamovi's TOSTER module (version 1.4.1). In the Jamovi file, the "Results" pane contains two identical TOST independent sample *t*-test analyses. In the first analysis, the "Hypothesis" parameter is set to "Equivalence Test", which will run the TOST procedure for the difference between groups. The 90% CI is given in the "Effect Sizes" table, and can be visualized with the "Plot Effect Sizes" parameter. In the second analysis, the "Hypothesis" parameter is set to "Minimal Effects Test", which will run an inferiority test ("TOST Lower" in the "TOST Results" table) and a superiority test ("TOST Upper" in the "TOST Results" table) for the difference between groups. In this analysis the "Alpha level" parameter is set to 0.025 so that 95% CIs are computed in the table and plots.

C Examples of Power Analysis for TST

In what follows, we assume that you have a good *a priori* expectation of the effect size that you will find in your main study. Such an expectation can be formed from previous studies, pilot experiments, or expert predictions, among other resources.

C.1 Case 1: The Expected Effect Size Is Larger Than the Smallest Effect Size of Interest

Returning to our hypothetical anchoring experiment, suppose that you expect to find a positive effect of $15,000 per month. Because this effect is larger in magnitude than our upper SESOI bound of $10,000, you should be prepared to detect a practically significant effect size of at least $15,000 with sufficient power. Suppose we want to have at least 80% power to detect such an effect at a 5% significance level. To calculate the required sample size for a superiority test in R, use the base `power.t.test` function with the `delta` parameter set to the effect size you want to power for minus the upper SESOI bound:

```
sesoi = 10000
effect = 15000
min_size_ps = power.t.test(delta = effect - sesoi,
                           sig.level = 0.05,
                           power = 0.80,
                           type = 'two.sample',
                           alternative = 'two.sided')$n
```

A similar procedure can be used if you expect the relationship to be negative. If you believe the anchoring tactic's effect will be less than the lower SESOI bound – say, -$15,000 per month – then calculate the required sample size for the inferiority test. In R, use the `power.t.test` function with `delta` set to `abs(effect - sesoi)`, as `power.t.test` cannot handle negative effect size values:

```
sesoi = -10000
effect = -15000
min_size_ps = power.t.test(delta = abs(effect - sesoi),
                           sig.level = 0.05,
                           power = 0.80,
                           type = 'two.sample',
                           alternative = 'two.sided')$n
```

Though these procedures give you the necessary sample size to have 80% power to detect practically significant effects, this sample size is not necessarily large enough to guarantee sufficient power for the equivalence test if the relationship you end up estimating happens to be smaller in magnitude than the SESOI. This is important because if you commit to TST, then you commit to performing an equivalence test if the estimated relationship is smaller in magnitude than the SESOI, and so you should be prepared to perform this equivalence test with sufficient power. At the very least, you should have enough observations to conclude that an estimate of zero is practically equivalent to zero. To calculate the required sample size for such an equivalence test in R, use the `power_t_TOST` function in the `TOSTER` package (see Lakens & Caldwell 2024):

```
sesoi = 10000
effect = 0
library(TOSTER)
min_size_eq = power_t_TOST(delta = effect,
                           low_eqbound = -sesoi,
                           high_eqbound = sesoi,
                           alpha = 0.05,
                           power = 0.80,
                           type = 'two.sample')$n
```

However, in practice, this approach may leave you with insufficient power for the equivalence testing component of TST. This is because estimated relationships are virtually never exactly zero. Because you need more power to obtain significant equivalence test results when estimated relationships are further away from zero, having exactly 80% power to conclude that an estimate of zero is practically equivalent to zero necessarily means having less than 80% power to conclude that a nonzero estimate is practically equivalent to zero.

An alternative approach is to compute (sufficient sample sizes for) 'safeguard power' (see Perugini, Gallucci, & Costantini 2014). This approach obtains enough power to conduct equivalence testing in the event that your expected *a priori* effect size is an outlier, and the true relationship of interest is in fact a significant distance away from this *a priori* expectation

in the direction of zero. To make this concrete, suppose that your pilot experiment yields an estimated *a priori* effect size expectation of $15,000 per month, but that the lower bound of the pilot estimate's 95% confidence interval is $5000 per month. In the same vein, you could imagine that the median expert prediction of the anchoring tactic's impact on sales is $15,000 per month, but that the fifth percentile of predictions for the effect is $5000 per month. In such a case, you can obtain the 'safeguard sample size' that would be necessary to conclude that such an effect size is practically equivalent to zero by changing the `effect` parameter in the previous code from 0 to 5000.

This 'safeguard power' approach is only feasible if the 'safeguard effect size' is smaller in magnitude than the SESOI. This is because under TST, it is impossible to conclude that a relationship that is larger in magnitude than the SESOI is practically equivalent to zero. Obtaining sufficient sample size to attain such safeguard power always requires at least as many observations as a sample size determination approach that does not attain safeguard power. Regardless of which method is used to compute the necessary sample size for a sufficiently-powered equivalence test, the necessary sample size needed to ensure sufficient power for TST when the expected relationship is larger in magnitude than the SESOI is simply the larger of the two necessary sample sizes for the equivalence test and the superiority/inferiority test.

```
min_size = max(c(min_size_ps, min_size_eq))
min_size
```

C.2 Case 2: The Expected Effect Size Is Smaller Than the Smallest Effect Size of Interest

For this case, suppose that we expect to obtain an estimate of $4000 per month. Because this relationship is smaller in magnitude than our SESOI of $10,000 per month, we begin our power calculations by ensuring that we have 80% power to conclude that our expected estimate of $4000 per month is significantly bounded beneath a size of $10,000 per month:

```
sesoi = 10000
```

```
effect = 4000
library(TOSTER)
min_size_eq = power_t_TOST(delta = effect,
                           low_eqbound = -sesoi,
                           high_eqbound = sesoi,
                           alpha = 0.05,
                           power = 0.80,
                           type = 'two.sample')$n
```

However, it could be the case that our estimated relationship ends up being larger in magnitude than the SESOI. Committing to TST requires us to perform superiority/inferiority tests if this occurs. We should thus be prepared to conduct such a test with sufficient power if necessary.

As in Case 1, we can compute necessary sample sizes to obtain sufficient power for superiority/inferiority tests in this case by computing sufficient sample sizes for safeguard power. In this setting, this insures against the prospect that your expected effect size is unrepresentatively small, and the true relationship is significantly further away from zero. For example, suppose that a pilot study yields an estimate of $4000 per month, but that the outer bound of that pilot estimate's 95% confidence interval is $14,000 per month. We can similarly imagine that the median expert prediction of the relationship is $4000 per month, but that the 95th percentile of expert predictions for the relationship is $14,000 per month. In either setting, we can assess necessary sample size for the superiority test as follows:

```
sesoi = 10000
effect = 14000
min_size_ps = power.t.test(delta = effect - sesoi,
                           sig.level = 0.05,
                           power = 0.80,
                           type = 'two.sample',
                           alternative = 'two.sided')$n
```

We can similarly obtain sample sizes if the worst tail risk is a negative effect size. In such a case, make the same adjustment in Case 1 for an inferiority test to ensure that 'power.t.test' can handle the negative values:

```
sesoi = -10000
effect = -14000
min_size_ps = power.t.test(delta = abs(effect - sesoi),
                           sig.level = 0.05,
                           power = 0.80,
                           type = 'two.sample',
                           alternative = 'two.sided')$n
```

In cases where researchers have extremely strong prior expectations that the relationship of interest will be smaller in magnitude than the SESOI, researchers may have reasonable justification to compute minimal required sample sizes for sufficient equivalence testing power without regard for sufficient superiority/inferiority testing power. This is because similarly to Case 1, the 'safeguard power' augmentation to sufficient sample size determination under TST when the expected relationship is smaller in magnitude than the SESOI is only feasible if the 'safeguard effect size' is larger in magnitude than the SESOI. Under TST, it is impossible to conclude that an effect size smaller in magnitude than the SESOI is superior/inferior to the SESOI, so if no 'safeguard effect size' larger in magnitude than the SESOI can be specified, then there is no benchmark effect size for which to assess the power of superiority/inferiority tests.

This divergence in practice between Case 1 and Case 2 emerges because zero is a 'special number' for the TST framework discussed in this paper. By construction, zero benchmarks all thresholds used for TST, and TST always involves an equivalence test that assesses whether the estimate is practically equivalent to zero. However, there is no similarly-privileged nonzero number for TST. Therefore, zero can always be used as a benchmark effect size for computing power for equivalence testing, but there is no similar 'fallback' effect size for assessing power in superiority/inferiority testing.

We thus recommend the following decision rule for determining minimal required sample

sizes for TST when the expected relationship is smaller in magnitude than the SESOI. If you can establish some 'safeguard effect size' larger in magnitude than the SESOI, then compute the necessary sample size to detect the tail effect size with sufficient power using superiority/inferiority testing, compute the necessary sample size to conclude that the expected effect size is practically equal to zero with sufficient power, and then collect the larger of these two necessary sample sizes. If you are extremely certain that the estimated relationship will be smaller in magnitude than the SESOI, then simply collect the necessary sample size to conclude that your expected relationship is practically equivalent to zero with sufficient power. In R, this recommendation can be implemented using the following code.

```
if (exists( min_size_ps )) {
        min_size = max(c(min_size_ps, min_size_eq))
}
else {
        min_size = min_size_eq
}
```