

# Three-Sided Testing to Establish Practical Significance: A Tutorial

Peder Isager & Jack Fitzgerald

Vrije Universiteit Amsterdam

December 12, 2024



# The Fundamental Problem of Standard NHST

This whole equivalence testing workshop is motivated by one key problem of the **standard NHST framework**; i.e.,

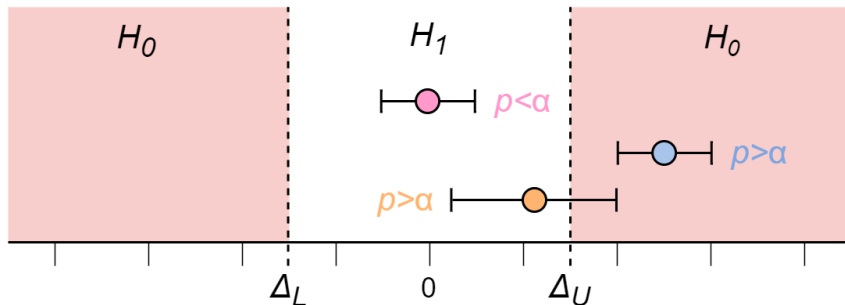
$$H_0 : \mu = 0$$

$$H_A : \mu \neq 0$$

In its simplest form, standard NHST procedures never let us accept  $H_0$

- I.e., without additional analyses, stat. insig. results under standard NHST do not let us conclude that  $\mu = 0$  or even that  $\mu \approx 0$ !

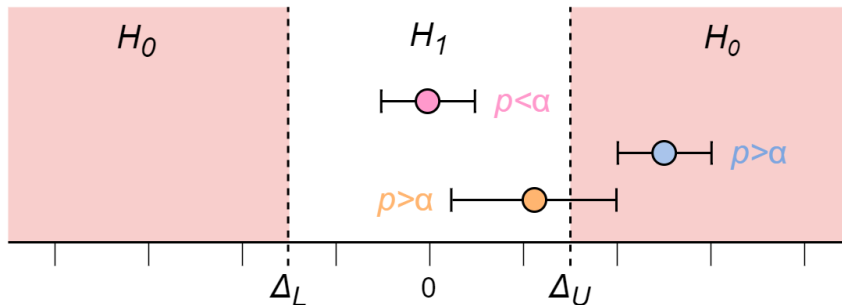
## Where Equivalence Testing Gets Us...



The upper pink estimate allows us to reject the null hypothesis that  $\mu$  is in one of the 'meaningfully large'  $H_0$  regions

- That allows us to conclude that the upper pink estimate  $\approx 0$ !

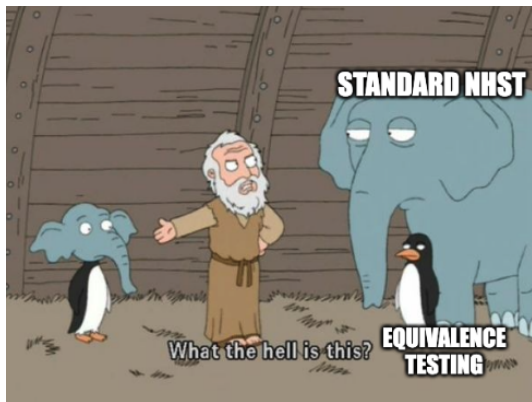
## ... And Where It Falls Short



... but just like standard NHST, equivalence testing never lets us 'accept the null' either

- Under equivalence testing, stat. insig. results do not let us conclude that  $\mu$  is meaningfully large, or even that  $\mu \neq 0$ !
- I.e., under equivalence testing, we'd make the exact same conclusions about the middle blue estimate and the lower orange estimate just because the equivalence testing  $p > \alpha$

# Stapler Solutions: Equivalence Testing + Standard NHST

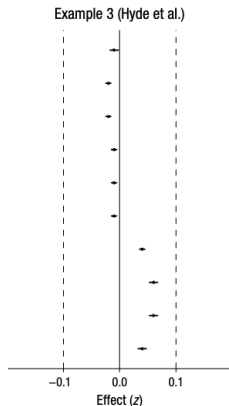


In practice, researchers deal with this by just stapling together equivalence testing with standard NHST

- ▶ Campbell & Gustafson (2018) formalize this practice as 'conditional equivalence testing'
- ▶ First standard NHST, then conditional on statistical insignificance, equivalence testing (specifically TOST)

Several problems with this approach

# Two Kinds of 'Significant'



Estimates can be both stat. sig. diff. from zero and stat. sig. bounded between SESOI thresholds

- ▶ This situation shows up often in high-powered settings (e.g., large-scale RCTs, natural experiments, 'big data')
- ▶ Big issue in analyses with administrative data
- ▶ Fang & Fang (2024) find that  $> 80\%$  of published sociology studies using register data apply standard NHST, with only 13.5% offering any justification

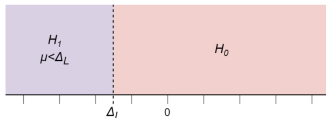
Standard NHST results are often preferred in these settings

- ▶ In fact, following conditional equivalence testing exactly, researchers would never set SESOLs or run equivalence tests in these settings

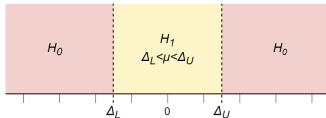
Source: Lakens, Scheel, & Isager (2018)

# Construct Validity

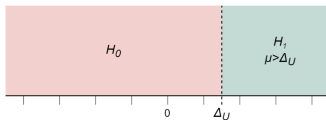
A) Inferiority



B) Equivalence (TOST)



C) Superiority



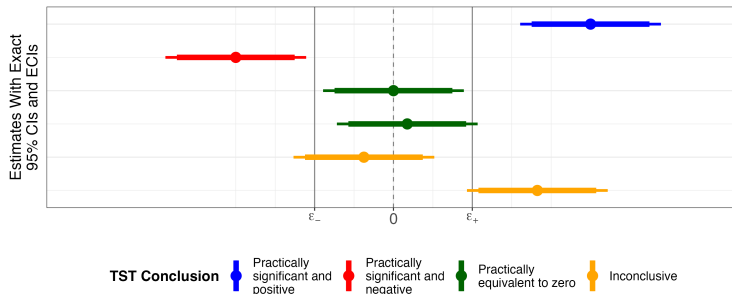
You learn that an estimate is stat. sig. diff. from zero. **So what?**

- ▶ Once we set SESOIs, it's clear that there are some nonzero values of  $\mu$  that are practically irrelevant
- ▶ Now rejecting the standard NHST null hypothesis that  $\mu = 0$  doesn't really tell us that  $\mu$  is 'significant'

If we want precise evidence that our estimate is practically significant, then we really want to show that the estimate is significantly larger than the SESOI

- ▶ This can be evidenced with **minimum effects tests** for inferiority or superiority (Murphy & Myers 1999)
- ▶ **What if we combine minimum effects tests with equivalence tests?**

# Three-Sided Testing (Goeman, Solari, & Stijnen 2010)



**Three-sided testing (TST)** is a comprehensive framework for assessing practical significance

- Under TST, we partition the parameter space into inferiority, equivalence, and superiority regions using the SESOI bounds
- We then see if there's stat. sig. evidence that the parameter is bounded in one of those regions

This is a direct test for the practical significance of an estimate



# Running Example

Congrats, you're hired! We work at a company that sells small, medium, and large service packages (increasing in price)

- ▶ Company leadership notices that in sales pitches, some branches attempt to anchor clients on the medium package by telling them it's the most popular purchase

There is debate about whether the anchoring tactic is a good idea, because it will...

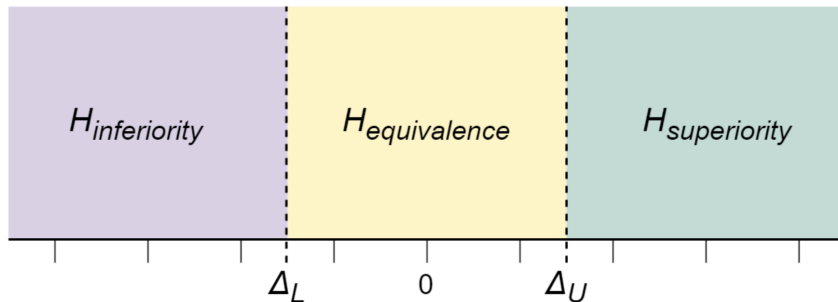
- ▶ Increase sales if medium packages substitute for small packages or no package at all, but...
- ▶ Decrease sales if medium packages substitute for large packages, or if anchoring discourages purchase of any package

We decide to run an experiment on the anchoring tactic's impact on sales

- ▶ Branches are randomized into a treatment group that uses the anchoring tactic in sales pitches and a control group that does not
- ▶ This experiment is effectively costless, and can in principle be run forever, so it's free to collect more observations (pretty common in private-sector A/B tests)

Our prestigious company's SESOI is \$10,000 in monthly branch sales

# The Hypothesis Framework



After partitioning  $\mu$ 's parameter space into  $H_{\text{Inf}}$ ,  $H_{\text{Eq}}$ , and  $H_{\text{Sup}}$  regions using the SESOI bounds  $\Delta_-$  and  $\Delta_+$ , TST assesses three sets of hypotheses at once:

$$H_0^{\{\text{Inf}\}} : \mu \geq \Delta_-$$

$$H_0^{\{\text{Eq}\}} : \mu < \Delta_- \text{ or } \mu > \Delta_+$$

$$H_0^{\{\text{Sup}\}} : \mu \leq \Delta_+$$

$$H_A^{\{\text{Inf}\}} : \mu < \Delta_-$$

$$H_A^{\{\text{Eq}\}} : \mu \geq \Delta_- \text{ and } \mu \leq \Delta_+$$

$$H_0^{\{\text{Sup}\}} : \mu > \Delta_+$$

# The Hypothesis Tests in Practice

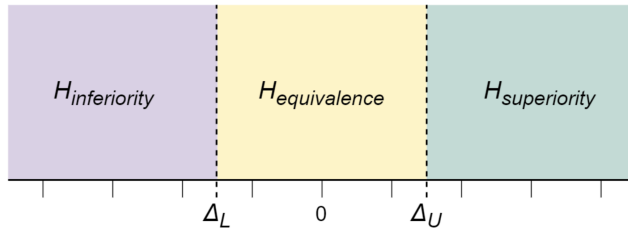
In other words, in TST, you run three tests at once:

1. An **inferiority test** assessing whether  $\mu < \Delta_-$
2. A **TOST procedure** assessing whether  $\mu \in [\Delta_-, \Delta_+]$
3. A **superiority test** assessing whether  $\mu > \Delta_+$

TST offers uniform error rate control over all three of these tests at significance level  $\alpha$

- ▶ Because TST begins by partitioning  $\mu$ 's parameter space into disjoint regions, TST benefits from the 'partitioning principle' (Finner & Straßburger 2002)
- ▶ Intuitively,  $\mu$  can only be in one of  $H_{\text{Inf}}$ ,  $H_{\text{Eq}}$ , or  $H_{\text{Sup}}$  at once
- ▶ Because it's impossible for  $\mu$  to be in  $H_{\text{Inf}}$  and  $H_{\text{Sup}}$  at the same time, we don't need to adjust for multiple hypothesis testing to control TST's error rates from these tests

## A Multiple Testing Caveat (1/2)



Suppose that  $\mu$  truly sits in  $H_{Eq}$ . Then it's possible to make *two* Type I errors in TST:

- ▶ I can conclude that  $\mu$  sits in  $H_{Inf}$  when it really sits in  $H_{Eq}$
- ▶ I can conclude that  $\mu$  sits in  $H_{Sup}$  when it really sits in  $H_{Eq}$

This issue doesn't show up for TST's equivalence test

- ▶ This is because in order to conclude that  $\mu$  sits in  $H_{Eq}$ , I already have to reject *both* that it sits in  $H_{Inf}$  *and* that it sits in  $H_{Sup}$

## A Multiple Testing Caveat (2/2)

We therefore need multiple hypothesis corrections for TST's inferiority and superiority tests, but not for its equivalence test

- ▶ Goeman, Solari, & Stijnen (2010) deal with this using a simple **Bonferroni correction** on the critical values for TST's inferiority and superiority tests
- ▶ For  $k$  tests, the Bonferroni correction controls the family-wise error rate across all  $k$  tests at level  $\alpha$  by setting an effective significance level of  $\alpha/k$
- ▶ This is a special case where  $k = 2$ , so this correction just cuts effective significance levels in half

We can control the error rate across all three tests in TST at  $\alpha$  (e.g., 5%) by using two effective significance levels:

- ▶ The equivalence test is conducted with significance level  $\alpha$  (e.g., 5%), just as usual
- ▶ The inferiority and superiority tests are conducted with significance level  $\alpha/2$  (e.g., 2.5%)
- ▶ Convenient coincidence – because the inferiority/superiority tests are one-sided tests, we can identically just do two-sided tests for inferiority/superiority at significance level  $\alpha$  (e.g., 5%)!

TST thus gives researchers already using equivalence testing a 'free lunch' – we can add minimum effects tests without any loss to power/error rate control for the equivalence test!

# TST: The $p$ -value Approach

Get the  $p$ -values from three tests:

1.  $p_{\text{Inf}}$ : One-sided test of whether  $\mu < \Delta_L$
2.  $p_{\text{Eq}}$ : Two one-sided tests of whether  $\Delta_L \leq \mu$  and  $\mu \leq \Delta_U$  (take the *larger* of these two  $p$ -values)
3.  $p_{\text{Sup}}$ : One-sided test of whether  $\mu > \Delta_U$

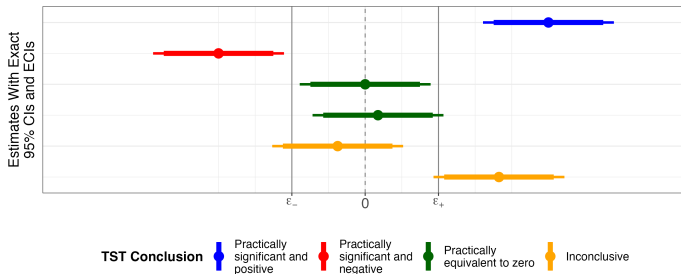
Reach conclusions as follows (only one can be true at a time):

- ▶ If  $p_{\text{Inf}} < \alpha/2$ , then conclude that  $\mu$  is practically significant and negative
- ▶ If  $p_{\text{Eq}} < \alpha$ , then conclude that  $\mu$  is practically equivalent to zero
- ▶ If  $p_{\text{Sup}} < \alpha/2$ , then conclude that  $\mu$  is practically significant and positive
- ▶ If none of these conditions hold, conclude that your results are inconclusive

# Double-Banded Confidence Intervals

Working with two effective significance levels can be clunky

- ▶ However, remember the convenient property from before – a one-sided critical value at significance level  $\alpha/2$  is equal to the two-sided critical value at significance level  $\alpha$
- ▶ We can thus visualize TST using a combination of  $(1 - \alpha)$  and  $(1 - 2\alpha)$  confidence intervals (CIs)



A **double-banded CI** displays an estimate's  $(1 - 2\alpha)$  CI (e.g., 90% CI) in thicker bands and its  $(1 - \alpha)$  CI (e.g., 95% CI) in thinner bands

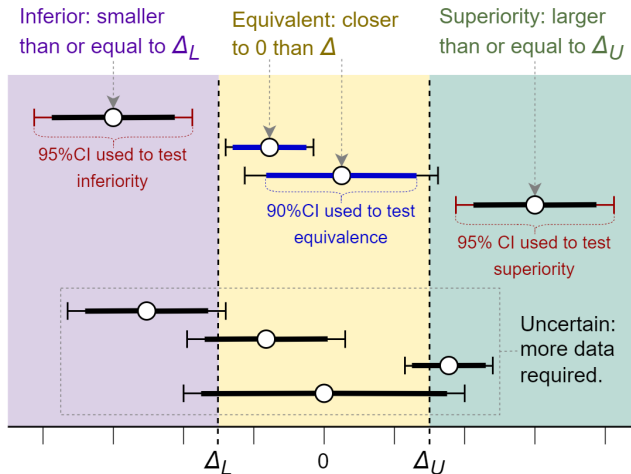
## TST: The CI Approach (1/2)

Let  $\hat{\mu}$  be an estimate of  $\mu$ . Reach conclusions as follows (only one can be true at a time):

- ▶ If  $\hat{\mu}$ 's  $(1 - \alpha)$  CI (e.g., 95% CI) is entirely bounded below  $\Delta_L$ , then conclude that  $\mu$  is practically significant and negative
- ▶ If  $\hat{\mu}$ 's  $(1 - 2\alpha)$  CI (e.g., 90% CI) is entirely bounded between  $\Delta_L$  and  $\Delta_U$ , then conclude that  $\mu$  is practically equal to zero
- ▶ If  $\hat{\mu}$ 's  $(1 - \alpha)$  CI (e.g., 95% CI) is entirely bounded above  $\Delta_U$ , then conclude that  $\mu$  is practically significant and positive
- ▶ If none of these conditions hold, conclude that your results are inconclusive



# TST: The CI Approach (2/2)



# ShinyTST App



Link:

<https://jack-fitzgerald.shinyapps.io/shinyTST/>

In the paper, we provide three tools for researchers to easily apply TST

1. The ShinyTST app, a Shiny app
2. The `tst` command in my `eqtesting` R package (<https://github.com/jack-fitzgerald/eqtesting/>)
3. Jamovi code using Caldwell's (2022) `TOSTER` module (<https://doi.org/10.31234/osf.io/ty8de>)

In this tutorial, we focus on ShinyTST

## Example 1 (1/2)

Suppose our anchoring experiment shows that treated branches make \$18,000 more per month than control branches ( $SE = \$3000$ )

- ▶ **Which test is relevant: inferiority, equivalence, or superiority?** Superiority! \$18,000 is positive, and  $|\$18,000|$  is larger than our SESOI of \$10,000

Now we just input our estimate, SE, and SESOI into ShinyTST

- ▶ We can also specify degrees of freedom to get exact tests, rather than asymptotically approximate tests

## Example 1 (2/2)

**Estimate**

**Standard error**

**Smallest effect size of interest**

**Significance level**

**Degrees of freedom (required for exact tests)**

**Power target (required for power-adjusted CIs)**

Asymptotically approximate results

90% confidence interval: [13065.4391191456, 22934.5608808544]

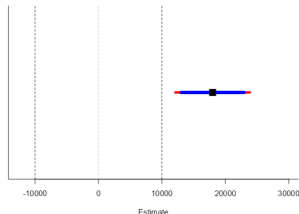
95% confidence interval: [12120.1080463798, 23879.8919536202]

Relevant test: Superiority

Test z-statistic: 2.66666666666667

Test p-value: 0.00766076113517947

The estimate is practically significant and positive.



As expected, ShinyTST tells us that the superiority test is relevant, so the **red**  $(1 - \alpha)$  CI should guide our significance conclusions

- ▶ The superiority test statistic is 2.67,  $>$  the two-sided  $\alpha = 5\%$  critical value of 1.96
- ▶ Accordingly, the **95% CI** is entirely bounded above  $\Delta_U$ , and the TST  $p$ -value is  $< 5\%$

Here we'd conclude that sales pitches with the anchoring tactic are superior to sales pitches without the anchoring tactic

- ▶ As a company, we then might mandate use of the tactic in all sales pitches

## Example 2

**Estimate**

**Standard error**

**Smallest effect size of interest**

**Significance level**

**Degrees of freedom (required for exact tests)**

**Power target (required for power-adjusted CIs)**

Asymptotically approximate results

90% confidence interval: [-22934.5608808544, -13065.4391191456]

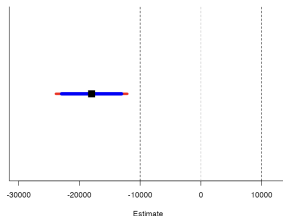
95% confidence interval: [-23879.8919536202, -12120.1080463798]

Relevant test: Inferiority

Test z-statistic: -2.66666666666667

Test p-value: 0.00766076113517947

The estimate is practically significant and negative.



Things would be similar if the observed effect of the anchoring tactic was -\$18,000 per month instead of +\$18,000 per month

- Now the inferiority test is relevant instead of the superiority test
- The inferiority test statistic is -2.67,  $<$  the two-sided  $\alpha = 5\%$  critical value of -1.96
- Accordingly, the **95% CI** is entirely bounded below  $\Delta_L$ , and the TST  $p$ -value is  $< 5\%$

Here we'd conclude that sales pitches with the anchoring tactic are inferior to sales pitches without the anchoring tactic

- As a company, we then might ban use of the tactic in all sales pitches

## Example 3 (1/2)

Estimate

Standard error

Smallest effect size of interest

Significance level

Degrees of freedom (required for exact tests)

Power target (required for power-adjusted CIs)

Asymptotically approximate results

90% confidence interval: [-434.560880854415, 9434.56088085441]

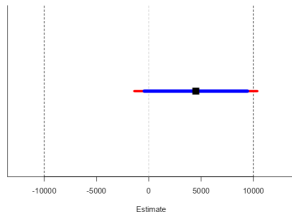
95% confidence interval: [-1379.89195362016, 10379.8919536202]

Relevant test: Equivalence

Test z-statistic: -1.83333333333333

Test p-value: 0.0333765075848173

The estimate is practically equal to zero.



Now suppose that holding all else constant, we observe that treated branches make \$4500 per month more than control branches

- Because  $|\$4500| < \$10,000$ , ShinyTST tells us that the equivalence test is relevant

Now the blue  $(1 - 2\alpha)$  CI guides our significance conclusions

## Example 3 (2/2)

Estimate

Standard error

Smallest effect size of interest

Significance level

Degrees of freedom (required for exact tests)

Power target (required for power-adjusted CIs)

Asymptotically approximate results

90% confidence interval: [-434.560880854415, 9434.56088085441]

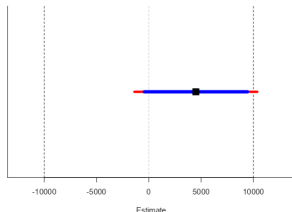
95% confidence interval: [-1379.89195362016, 10379.8919536202]

Relevant test: Equivalence

Test z-statistic: -1.83333333333333

Test p-value: 0.0333765075848173

The estimate is practically equal to zero.



The equivalence test statistic is 1.83,  $>$  the one-sided  $\alpha = 5\%$  critical value of 1.96

- The **90% CI** is thus entirely bounded between  $\Delta_L$  and  $\Delta_U$ , and the TST  $p$ -value is  $< 5\%$
- For the equivalence test, it doesn't matter that the **95% CI** crosses an SESOI bound; we only care about the **90% CI** here

Here we'd conclude that sales pitches with the anchoring tactic produce practically equivalent revenue compared to sales pitches without the anchoring tactic

- As a company, we then might allow branches to choose whether they use the anchoring tactic at will

## Example 4

**Estimate**

**Standard error**

**Smallest effect size of interest**

**Significance level**

**Degrees of freedom (required for exact tests)**

**Power target (required for power-adjusted CIs)**

Asymptotically approximate results

90% confidence interval: [-20434.5608808544, -10565.4391191456]

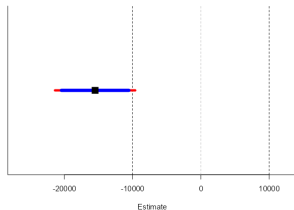
95% confidence interval: [-21379.8919536202, -9620.10804637984]

Relevant test: Inferiority

Test z-statistic: -1.83333333333333

Test p-value: 0.0667530151696345

The practical significance of the estimate is inconclusive.



What if the estimated effect of the anchoring tactic was -\$15,500?

- The inferiority test is relevant, so we care about the **95% CI**

The **95% CI** crosses an SESOI bound, and accordingly:

- The inferiority test statistic (-1.83) is smaller than the two-sided 5% critical value (-1.96)
- The TST  $p$ -value  $> 5\%$

The TST results on this estimate are inconclusive



## Example 5

Estimate

Standard error

Smallest effect size of interest

Significance level

Degrees of freedom (required for exact tests)

Power target (required for power-adjusted CIs)

Asymptotically approximate results

90% confidence interval: [-12434.5608808544, -2565.43911914559]

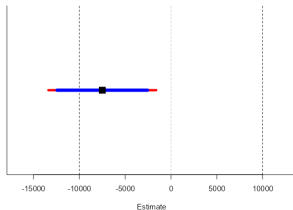
95% confidence interval: [-13379.8919536202, -1620.10804637984]

Relevant test: Equivalence

Test z-statistic: 0.833333333333333

Test p-value: 0.202328380963643

The practical significance of the estimate is inconclusive.



Similar results arise is the estimated effect of the anchoring tactic is -7500

- Now the equivalence test is relevant, so we care about the 90% CI

Here the 90% CI crosses an SESOI bound, and accordingly:

- The equivalence test statistic (-0.83) is smaller than the one-sided 5% critical value (-1.645)
- The TST  $p$ -value  $> 5\%$

Because results are inconclusive, as a company, we could continue running the experiment until sufficiently precise results can be achieved

# What to Mention When Reporting TST Results

Which test is relevant: inferiority, equivalence, or superiority?

- ▶ This will depend on where your estimate is in relation to the SESOI bounds

What is the TST  $p$ -value?

- ▶ This will depend on which test is relevant, the estimate, its SE, and the SESOI

What are the  $(1 - \alpha)$  and  $(1 - 2\alpha)$  confidence intervals?

- ▶ This depends only on your estimate and SE

How should we interpret your TST results?

- ▶ Estimates can be (1) practically significant and negative, (2) practically equal to zero, (3) practically significant and positive, or (4) inconclusive
- ▶ Important to verbally interpret the results, and tell us what we can actionably conclude!

# Power Analysis

In TST, you commit to conducting three tests, which makes power analysis tricky

- ▶ To have sufficient *a priori* power for TST, you need sufficient *a priori* power for *all three tests*
- ▶ However, in practice, researchers usually only anticipate seeing one estimate

What if you plan power for an estimate greater than the SESOI, but then the estimate ends up being smaller than the SESOI?

- ▶ Then you've powered your experiment for the wrong test!

To deal with this issue, we recommend that researchers use a **safeguard power** approach (Perugini, Galluci, & Costantini 2014)

- ▶ This basically involves powering your experiment to two different effect sizes, one that you actually anticipate and one 'just in case'
- ▶ E.g., if you expect your estimate to be smaller than the SESOI, maybe consider a 'worst case scenario' for how close to the SESOI the estimate could be if it were beyond the SESOI bounds
- ▶ We leave specific details to the preprint (last slide)

# Return of the Stapler: TST + Standard NHST

What if our SESOI can change over time?

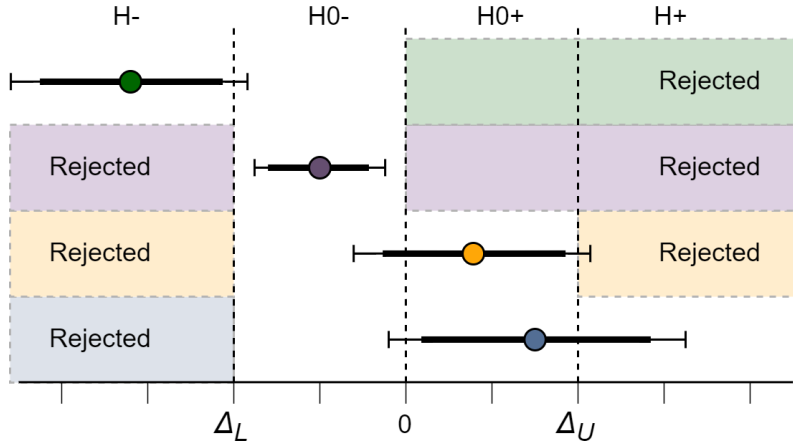
- ▶ E.g., the SESOI for monthly branch sales in our company may increase if the company grows or decrease in times of financial distress

In this case, it may be useful to record standard NHST results

- ▶ Knowing an intervention has some nonzero effect may be useful to inform future experiments/policies

You can staple standard NHST onto TST without any loss in power/error rate control

# Why Stapling NHST Onto TST Works



# Inconclusive Results

If your estimate is close to the SESOI and/or imprecise, it may not be possible to find stat. sig. evidence that  $\mu$  is bounded in any TST region

- ▶ The conclusion to reach in this setting is that your results are **inconclusive**
- ▶ I.e., you don't have enough power to say the TST region in which  $\mu$  lies with sufficient certainty

This is an uncomfortable finding, but it is an important admission

- ▶ If a doctor doesn't know whether you have a disease, you don't want them to tell you 'yes, you have it' or 'no, you don't' – you want them to get more data and run more tests!
- ▶ Distinguishing between imprecise results and precise nulls requires us to be able to name when an estimate is imprecise
- ▶ For this to work, scientists have to refrain from making conclusive statements about statistical relationships when they don't have enough power/precision to make reasonably certain conclusions

# What Statistical Significance Means in TST

Statistical significance in standard NHST is a certainty guarantee over error rates when you argue a nonzero relationship exists

- ▶ If all goes well, we won't be wrong more than  $100\alpha\%$  of the time when we say a stat. sig. relationship is nonzero

Without SESOs, researchers lean on these certainty guarantees to get a sense of whether relationships are 'significant'

- ▶ This abuses the word 'significant' a bit

## **TST decouples 'certainty' and 'significance'**

- ▶ In TST, 'significance' is determined from an estimate's relationship with the SESOI
- ▶ Statistical testing is then just about making sure we have enough precision to say certainly, with error rate control, which TST region a relationship lies in

In this context,  $\alpha$  significance thresholds specify error rates on practical significance conclusions

- ▶ I.e., if all goes well, we won't be wrong more than  $100\alpha\%$  of the time when we make conclusive claims about the practical significance of an estimate

## More Information

For more information, visit our  
PsyArXiv preprint (left QR)

- Includes links to the  
ShinyTST app and the  
eqtesting R package

Alternatively, save these slides  
(right QR) and revisit this page  
for the PsyArXiv link



**PsyArXiv Preprint**

[https://doi.org/  
10.31234/osf.io/8y925](https://doi.org/10.31234/osf.io/8y925)



**These Slides**





[https://jack-fitzgerald.github.io/  
files/TST\\_Tutorial.pdf](https://jack-fitzgerald.github.io/files/TST_Tutorial.pdf)

**Website:** <https://jack-fitzgerald.github.io>



**Email:** [j.f.fitzgerald@vu.nl](mailto:j.f.fitzgerald@vu.nl)



# References I

-  Campbell, H. and P. Gustafson (2018).  
Conditional equivalence testing: An alternative remedy for publication bias.  
*PLOS ONE* 13(4).
-  Finner, H. and K. Strassburger (2002, August).  
The partitioning principle: A powerful tool in multiple decision theory.  
*The Annals of Statistics* 30(4), 1194–1213.
-  Goeman, J. J., A. Solari, and T. Stijnen (2010, September).  
Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority.  
*Statistics in Medicine* 29(20), 2117–2125.
-  Lakens, D., A. M. Scheel, and P. M. Isager (2018, June).  
Equivalence Testing for Psychological Research: A Tutorial.  
*Advances in Methods and Practices in Psychological Science* 1(2), 259–269.

## References II

-  Murphy, K. R. and B. Myers (1999).  
Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model.  
*Journal of Applied Psychology* 84(2), 234–248.
-  Perugini, M., M. Gallucci, and G. Costantini (2014).  
Safeguard power as a protection against imprecise power estimates.  
*Perspectives on Psychological Science* 9(3), 319–332.