The Need for Equivalence Testing in Economics

Jack Fitzgerald

Vrije Universiteit Amsterdam

September 13, 2024

"No Detectable Effects"

Bessone et al. (2021, QJE): Sleep improvement RCT with 400 people in Chennai, India

"No Detectable Effects"

Bessone et al. (2021, QJE): Sleep improvement RCT with 400 people in Chennai, India

- At baseline, avg. participant has sleep patterns mirroring clinical insomnia

"No Detectable Effects"

Bessone et al. (2021, QJE): Sleep improvement RCT with 400 people in Chennai, India

- At baseline, avg. participant has sleep patterns mirroring clinical insomnia
- The intervention is very effective (27 extra minutes of night sleep)

"No Detectable Effects"

Bessone et al. (2021, QJE): Sleep improvement RCT with 400 people in Chennai, India

- At baseline, avg. participant has sleep patterns mirroring clinical insomnia
- The intervention is very effective (27 extra minutes of night sleep)

However, per their abstract...

> *\Contrary to expert predictions and a large body of sleep research, increased nighttime sleep had* **no detectable e ects** *on cognition, productivity, decision making, or well being... "*

"No Detectable Effects"

Bessone et al. (2021, QJE): Sleep improvement RCT with 400 people in Chennai, India

- At baseline, avg. participant has sleep patterns mirroring clinical insomnia
- The intervention is very effective (27 extra minutes of night sleep)

However, per their abstract...

> *\Contrary to expert predictions and a large body of sleep research, increased nighttime sleep had* **no detectable e ects** *on cognition, productivity, decision making, or well being... "*

By their own admission, these findings contradict expert priors and large bodies of research

"No Detectable Effects"

Bessone et al. (2021, QJE): Sleep improvement RCT with 400 people in Chennai, India

- At baseline, avg. participant has sleep patterns mirroring clinical insomnia
- The intervention is very effective (27 extra minutes of night sleep)

However, per their abstract...

> \Contrary to expert predictions and a large body of sleep research, increased nighttime sleep had **no detectable e ects** on cognition, productivity, decision making, or well being... "

By their own admission, these findings contradict expert priors and large bodies of research

- So what do they mean by **`no detectable e ects?'**

Null Estimates in Bessone et al. (2021)

What they mean: Results are not stat. sig. di erent from zero

Null Estimates in Bessone et al. (2021)

What they mean: Results are not stat. sig. di erent from zero

- **They are not alone** in interpreting insigni cant results in this way

This Happens All the Time

From 2020-2023, 279 null claims made in abstracts of 158 articles in T5 journals are defended by statistically insigni cant results Detailed Results

This Happens All the Time

From 2020-2023, 279 null claims made in abstracts of 158 articles in T5 journals are defended by statistically insigni cant results Detailed Results

- > 72% of these null claims aren't quali ed by references to statistical signi cance, estimate magnitudes, or a lack of evidence

This Happens All the Time

From 2020-2023, 279 null claims made in abstracts of 158 articles in T5 journals are defended by statistically insigni cant results Detailed Results

- > 72% of these null claims aren't quali ed by references to statistical signi cance, estimate magnitudes, or a lack of evidence

Researchers and readers interpret such ndings as evidence of null/negligible relationships (McShane & Gal 2016, McShane & Gal 2017)

Why Is This a Problem?

Generally inferring that stat. insig. results are null results is known to be bad scienti c practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

- Statistical insigni cance may just re ect imprecision

Why Is This a Problem?

Generally inferring that stat. insig. results are null results is known to be bad scienti c practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

- Statistical insigni cance may just re ect imprecision

Under standard NHST, null results and imprecision are con ated. Credibility problems follow:

Why Is This a Problem?

Generally inferring that stat. insig. results are null results is known to be bad scienti c practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

- Statistical insigni cance may just re ect imprecision

Under standard NHST, null results and imprecision are con ated. Credibility problems follow:

- Null result penalty from beliefs of low quality and unpublishability (McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024)

Why Is This a Problem?

Generally inferring that stat. insig. results are null results is known to be bad scienti c practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

- Statistical insigni cance may just re ect imprecision

Under standard NHST, null results and imprecision are con ated. Credibility problems follow:

- Null result penalty from beliefs of low quality and unpublishability (McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024)
- Publication bias from non-publication of null results (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019)

Why Is This a Problem?

Generally inferring that stat. insig. results are null results is known to be bad scienti c practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

- Statistical insigni cance may just re ect imprecision

Under standard NHST, null results and imprecision are con ated. Credibility problems follow:

- Null result penalty from beliefs of low quality and unpublishability (McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024)
- Publication bias from non-publication of null results (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019)
- High Type II error rates , given current practices and power levels (Ioannidis, Stanley, & Doucouliagos 2017; Askarov et al. 2023)

Why Is This a Problem?

Generally inferring that stat. insig. results are null results is known to be bad scienti c practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

- Statistical insigni cance may just re ect imprecision

Under standard NHST, null results and imprecision are con ated. Credibility problems follow:

- Null result penalty from beliefs of low quality and unpublishability (McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024)
- Publication bias from non-publication of null results (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019)
- High Type II error rates , given current practices and power levels (Ioannidis, Stanley, & Doucouliagos 2017; Askarov et al. 2023)

It doesn't have to be this way.

Equivalence Testing in a Nutshell

1. Set a region around zero wherein relationship of interest would be practically equivalent to zero (i.e., economically insignificant)

Equivalence Testing in a Nutshell

1. Set a region around zero wherein relationship of interest would be practically equivalent to zero (i.e., economically insignificant)
2. Use interval tests to assess if $\hat{\beta}$ is sig. bounded within this region

Equivalence Testing in a Nutshell

1. Set a region around zero wherein relationship of interest would be practically equivalent to zero (i.e., economically insignificant)
2. Use interval tests to assess if $\hat{\beta}$ is sig. bounded within this region

Common in medicine, political science, and psychology (see e.g., Piaggio et al. 2012; Hartman & Hidalgo 2018; Lakens, Scheel, & Isager 2018)

This Project

What is equivalence testing?

- I introduce simple frequentist equivalence testing techniques to economists

This Project

What is equivalence testing?

- I introduce simple frequentist equivalence testing techniques to economists

Why do we need to use it?

- 36-63% of estimates defending null claims in top economics journals fail lenient equivalence tests
- Type II error rates in economics are likely quite high

This Project

What is equivalence testing?

- I introduce simple frequentist equivalence testing techniques to economists

Why do we need to use it?

- 36-63% of estimates defending null claims in top economics journals fail lenient equivalence tests
- Type II error rates in economics are likely quite high

How do we perform equivalence testing credibly?

- I develop software commands and guidelines for credible and relatively easy implementation

The Wrong Hypotheses: NHST

Standard NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

The Wrong Hypotheses: NHST

Standard NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

When trying to show that $= 0$ using NHST, two key problems:

The Wrong Hypotheses: NHST

Standard NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

When trying to show that = 0 using NHST, two key problems:

1. The burden of proof is shifted: Researchers start by assuming they're right

The Wrong Hypotheses: NHST

Standard NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

When trying to show that = 0 using NHST, two key problems:

1. The burden of proof is shifted: Researchers start by assuming they're right
2. Imprecision is `good': Less precision = higher chance of stat. insig. results

The Wrong Hypotheses: NHST

Standard NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

When trying to show that = 0 using NHST, two key problems:

1. The burden of proof is shifted: Researchers start by assuming they're right
2. Imprecision is `good': Less precision ⇒ higher chance of stat. insig. results

It's thus a logical fallacy to generally infer that stat. insig. results are null results (appeal to ignorance)

The Right Hypotheses: Equivalence Testing

We'll fix these problems by 1) flipping the hypotheses and 2) relaxing the contraints. As a reminder, NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

And now equivalence testing hypotheses:

$$H_0: \quad \geq 0$$
$$H_A: \quad < 0$$

The Right Hypotheses: Equivalence Testing

We'll x these problems by 1) ipping the hypotheses and 2) relaxing the contraints. As a reminder, NHST hypotheses:

$$H_0: \quad = 0$$
$$H_A: \quad \neq 0$$

And now equivalence testing hypotheses:

$$H_0: \quad \geq 0$$
$$H_A: \quad < 0$$

If we can set a range of values $[\ ;\ _+]$ wherein $\quad 0$, then we can nd stat. sig. evidence for H_A with a simple interval test

The Equivalence Testing Framework

We begin by setting a range of values [$_{+}$], where $<$ $_{+}$, called theregion of practical equivalence (ROPE)

The Equivalence Testing Framework

We begin by setting a range of values $[\ _{+}]$, where $<\ _{+}$, called the region of practical equivalence (ROPE)

- The ROPE is the range of values we'd call economically insigni cant

The Equivalence Testing Framework

We begin by setting a range of values $[\ _{+}]$, where $\ < \ _{+}$, called the region of practical equivalence (ROPE)

- The ROPE is the range of values we'd call economically insigni cant
- This is a subjective judgment call that will di er for di erent relationships of interest
- I show how to credibly aggregate ROPEs later in this t

Credible ROPE-Setting

The Equivalence Testing Framework

We begin by setting a range of values $[\ _{-}\ _{+}]$, where $\ _{-} < \ _{+}$, called the *region of practical equivalence* (ROPE)

- The ROPE is the range of values we'd call *economically insignificant*
- This is a subjective judgment call that will differ for different relationships of interest
- I show how to credibly aggregate ROPEs later in this talk (Credible ROPE-Setting)

Once we have a ROPE, we can set up the equivalence testing hypotheses:

$$H_0: \ \not\in [\ _{-};\ _{+}]$$
$$H_A: \ \in [\ _{-};\ _{+}]$$

Two One-Sided Tests (TOST)

We can identically write the equivalence testing hypotheses as

H_0 : $<$ or $>$ $+$

H_A : and $+$

Two One-Sided Tests (TOST)

We can identically write the equivalence testing hypotheses as

$$H_0: \quad < \quad \text{or} \quad > \quad +$$
$$H_A: \quad \quad \text{and} \quad +$$

Further, we can assess the joint H_A using two one-sided tests:

$$H_0: \quad < \qquad\qquad H_0: \quad > \quad +$$
$$H_A: \qquad\qquad H_A: \quad +$$

Two One-Sided Tests (TOST)

We can identically write the equivalence testing hypotheses as

$$H_0: \quad < \quad \text{or} \quad > \quad +$$
$$H_A: \quad \quad \text{and} \quad \quad +$$

Further, we can assess the joint H_A using two one-sided tests:

$$H_0: \quad < \qquad\qquad H_0: \quad > \quad +$$
$$H_A: \qquad\qquad H_A: \quad \quad +$$

Stat. sig. evidence for both H_A statements using one-sided tests is stat. sig. evidence that 0 (Schuirmann 1987; Berger & Hsu 1996)

Procedural Details Visualization

Equivalence Con dence Intervals (ECIs)

^'s (1) equivalence con dence interval (ECI) is just its (1 2) CI

- If ^'s (1) ECI is entirely bounded in the ROPE, then we have size- evidence under the TOST procedure that 0 (Berger & Hsu 1996)

Revisiting Bessone et al. (2021)

Estimates defending null claims should be signi cantly bounded within reasonably wide ROPEs

Revisiting Bessone et al. (2021)

Estimates defending null claims should be signi cantly bounded within reasonably wide ROPEs

- However, 28% of the `null' estimates in Bessone et al. (2021) aren't signi cantly bounded beneath $|\delta| = 0.2$
- 71% aren't signi cantly bounded beneath $|\delta| = 0.1$

Revisiting Bessone et al. (2021)

Estimates defending null claims should be signi cantly bounded within reasonably wide ROPEs

- However, 28% of the `null' estimates in Bessone et al. (2021) aren't signi cantly bounded beneath $|\delta| = 0.2$
- 71% aren't signi cantly bounded beneath $|r| = 0.1$

Takeaway: Bessone et al. (2021) cannot guarantee precise nulls for a large proportion of their `null' estimates, which `fail' lenient equivalence tests

Data

1. Systematically-selected replication sample
   - 876 estimates defending 135 null claims in abstracts of 81 articles in T5 economics journals published from 2020-202 Claim Example
   - Estimates defending these null claims are reproducible with publicly-available data

Data

1. Systematically-selected replication sample
   - 876 estimates defending 135 null claims in abstracts of 81 articles in T5 economics journals published from 2020-202 Claim Example
   - Estimates defending these null claims are reproducible with publicly-available data
2. Prediction platform data
   - I survey 62 researchers on the Social Science Prediction Platform for predictions and judgments on equivalence testing results in my sample

Equivalence Testing Failure Rates

I compute avg. equivalence testing failure rates in the replication sample

Equivalence Testing Failure Rates

I compute avg. equivalence testing failure rates in the replication sample

- First ROPE: $r \in [-0.1, 0.1]$
- $|r| = 0.1$ is larger than over 25% of published results in economics (Doucouliagos 2011)

Effect Size Standardization

Equivalence Testing Failure Rates

I compute avg. equivalence testing failure rates in the replication sample

- First ROPE: $r \in [-0.1, 0.1]$
- $|r| = 0.1$ is larger than over 25% of published results in economics (Doucouliagos 2011)
  E ect Size Standardization
- Second ROPE: $\in [-0.2, 0.2]$
- $|\ | = 0.2$ is quite large for economic e ect sizes Benchmarking Sample

Equivalence Testing Failure Rates

I compute avg. equivalence testing failure rates in the replication sample

- First ROPE: r 2 [0:1; 0:1]
- jrj = 0:1 is larger than over 25% of published results in economics (Doucouliagos 2011) E ect Size Standardization
- Second ROPE: 2 [0:2; 0:2]
- j j = 0:2 is quite large for economic e ect sizes Benchmarking Sample

Models defending null claims in T5 journals should have no trouble signi cantly bounding estimates within ROPEs this wide

Equivalence Testing Failure Rates are Unacceptably High

Equivalence testing failure rates range from 36-63%

Robustness Checks TST Framework Mechanisms

Equivalence Testing Failure Rates are Unacceptably High

Equivalence testing failure rates range from 36-63% Robustness Checks TST Framework Mechanisms

- Interpretation : 62% of estimates defending the average null claim can't signi cantly bound their estimates beneath $|r_j| = 0.1$ (see Model 4)

Failure Curves

Equivalence testing failure rates stay unacceptably high even as ROPEs become ridiculously large

Failure Curves

Equivalence testing failure rates stay unacceptably high even as ROPEs become ridiculously large

- To obtain acceptable failure rates, you'd need to argue that $|0.317 - r_j|$ is practically equal to zero

Failure Curves

Equivalence testing failure rates stay unacceptably high even as ROPEs become ridiculously large

- To obtain acceptable failure rates, you'd need to argue that $|\hat{\beta}| < 0.317\sigma_{\hat{\beta}}$ is practically equal to zero
- $|\hat{\beta}| < 0.317\sigma_{\hat{\beta}}$ is larger than nearly 75% of published effects in economics (Doucouliagos 2011)

Researchers Anticipate Unacceptably High Failure Rates

The median researcher nds failure rates from 11-13% acceptable, but (pretty accurately) predicts failure rates from 35-38%.Takeaways:

Researchers Anticipate Unacceptably High Failure Rates

The median researcher nds failure rates from 11-13% acceptable, but (pretty accurately) predicts failure rates from 35-38%.Takeaways:

1. Researchers don't trust null results under standard NHST, but this mistrust is well-placed

Researchers Anticipate Unacceptably High Failure Rates

The median researcher nds failure rates from 11-13% acceptable, but (pretty accurately) predicts failure rates from 35-38%.Takeaways:

1. Researchers don't trust null results under standard NHST, but this mistrust is well-placed
2. More credible testing frameworks are necessary to restore trust

Credible ROPE-Setting

ROPEs need to be set independently to be credible (Lange & Freitag 2005; Ofori et al. 2023)

- `ROPE-hacking' is a key concern

Credible ROPE-Setting

ROPEs need to be set independently to be credible (Lange & Freitag 2005; Ofori et al. 2023)

- `ROPE-hacking' is a key concern
- To maintain independence & credibility, you shouldn't set your ROPEs { you should get other people to set them for you

Credible ROPE-Setting

ROPEs need to be set independently to be credible (Lange & Freitag 2005; Ofori et al. 2023)

- `ROPE-hacking' is a key concern
- To maintain independence & credibility, you shouldn't set your ROPEs { you should get other people to set them for you

Solution: Survey independent experts/stakeholders for their judgments

Credible ROPE-Setting

ROPEs need to be set independently to be credible (Lange & Freitag 2005; Ofori et al. 2023)

- `ROPE-hacking' is a key concern
- To maintain independence & credibility, you shouldn't set your ROPEs { you should get other people to set them for you

Solution: Survey independent experts/stakeholders for their judgments

- Practically feasible using online platforms such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019)

Credible ROPE-Setting

ROPEs need to be set independently to be credible (Lange & Freitag 2005; Ofori et al. 2023)

- `ROPE-hacking' is a key concern
- To maintain independence & credibility, you shouldn't set your ROPEs { you should get other people to set them for you

Solution: Survey independent experts/stakeholders for their judgments

- Practically feasible using online platforms such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019)
- Example from this project : Alongside predictions of failure rates, I elicit what failure rates researchers deem acceptable

The Equivalence Testing Framework

Software Commands & More Information

tsti Stata command (QR: Github)	equivtest R package, containing tst command (QR: Github)	Working paper (QR: PDF link, personal website)

Website: https://jack- tzgerald.github.io
Email: j.f. tzgerald@vu.nl

References I

Altman, D. G. and J. M. Bland (1995).
Statistics notes: Absence of evidence is not evidence of absence.
BMJ 311(7003), 485{485.

Andrews, I. and M. Kasy (2019).
Identi cation of and correction for publication bias.
American Economic Review 109(8), 2766{2794.

Askarov, Z., A. Doucouliagos, H. Doucouliagos, and T. D. Stanley (2023).
Selective and (mis)leading economics journals: Meta-research evidence.
Journal of Economic Surveys, Forthcoming

Berger, R. L. and J. C. Hsu (1996).
Bioequivalence trials, intersection-union tests and equivalence con dence sets.
Statistical Science 11(4).

References II

Chopra, F., I. Haaland, C. Roth, and A. Stegmann (2024).
The null result penalty.
The Economic Journal 134(657), 193{219.

Cohen, J. (1988).
Statistical power analysis for the behavioral sciences (2 ed.).
L. Erlbaum Associates.

DellaVigna, S., D. Pope, and E. Vivalt (2019).
Predict science to improve science.
Science 366(6464), 428{429.

Doucouliagos, H. (2011).
How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics.
Working Paper SWP 2011/5, Deakin University, Geelong, Australia.

References III

Dreber, A., M. Johannesson, and Y. Yang (2024).
Selective reporting of placebo tests in top economics journals.
Economic Inquiry.

Fanelli, D. (2012).
Negative results are disappearing from most disciplines and countries.
Scientometrics 90(3), 891{904.

Finner, H. and K. Strassburger (2002).
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).
Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority.
Statistics in Medicine 29(20), 2117{2125.

References IV

Hartman, E. and F. D. Hidalgo (2018).
An equivalence approach to balance and placebo tests.
American Journal of Political Science 62(4), 1000–1013.

Imai, K., G. King, and E. A. Stuart (2008).
Misunderstandings between experimentalists and observationalists about causal inference.
Journal of the Royal Statistical Society Series A: Statistics in Society 171(2), 481–502.

Ioannidis, J. P., T. D. Stanley, and H. Doucouliagos (2017).
The power of bias in economics research.
The Economic Journal 127(605).

Lakens, D., A. M. Scheel, and P. M. Isager (2018).
Equivalence testing for psychological research: A tutorial.
Advances in Methods and Practices in Psychological Science 1(2), 259–269.

References V

Lange, S. and G. Freitag (2005).
Choice of delta: Requirements and reality { results of a systematic review.
Biometrical Journal 47(1), 12{27.

McShane, B. B. and D. Gal (2016).
Blinding us to the obvious? The e ect of statistical training on the evaluation of evidence.
Management Science 62(6), 1707{1718.

McShane, B. B. and D. Gal (2017).
Statistical signi cance and the dichotomization of evidence.
Journal of the American Statistical Association 112(519), 885{895.

References VI

Ofori, S., T. Cafaro, P. Devereaux, M. Marcucci, L. Mbuagbaw, L. Thabane, and G. Guyatt (2023).

Noninferiority margins exceed superiority e ect estimates for mortality in cardiovascular trials in high-impact journals.

Journal of Clinical Epidemiology 16,120{27.

Piaggio, G., D. R. Elbourne, S. J. Pocock, S. J. Evans, and D. G. Altman (2012).

Reporting of noninferiority and equivalence randomized trials.

JAMA 308(24), 2594{2604.

Schuirmann, D. J. (1987).

A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability.

Journal of Pharmacokinetics and Biopharmaceutics 15(6), 657{680.

References VII

Sha er, J. P. (1986).
Modi ed sequentially rejective multiple test procedures.
Journal of the American Statistical Association 8(395), 826{831.

Wasserstein, R. L. and N. A. Lazar (2016).
The ASA statement onp-values: Context, process, and purpose.
The American Statistician 70(2), 129{133.

Null Claim Classi cation

This Happens All the Time

The TOST Procedure

First, compute test statistics

$$t = \frac{\hat{}}{s} \qquad t_+ = \frac{\hat{} \; +}{s}$$

The relevant test statistic is the smaller of the two:

$$t_{TOST} = \arg\min_{t \; 2 \; f t \; ; t_+ g} f j \, t j g$$

The critical value for a size- TOST procedure is the one-sided critical value t

1. If $t_{TOST} = t$, then there is stat. sig. evidence that 2 [; +] i t t
2. If $t_{TOST} = t_+$, then there is stat. sig. evidence that 2 [; +] i t_+ t

A single TOST procedure maintains size even without multiple hypothesis corrections (Berger & Hsu 1996)

TOST Concept

TOST Example

Claim Example

The bolded text represents the two null claims made by this abstract:

> \This article estimates peer e ects originating from the ability composition of tutorial groups for undergraduate students in economics. We manipulated the composition of groups to achieve a wide range of support, and assigned students-conditional on their prior ability-randomly to these groups. The data support a speci cation in which the impact of group composition on achievement is captured by the mean and standard deviation of peers' prior ability, their interaction, and interactions with students' own prior ability. When we assess the aggregate implications of these peer e ects regressions for group assignment, we nd that low-and medium-ability students gain on an average 0.19 SD units of achievement by switching from ability mixing to three-way tracking. Their dropout rate is reduced by 12 percentage points (relative to a mean of 0.6). High-ability students are una ected . Analysis of survey data indicates that in tracked groups, low-ability students have more positive interactions with other students, and are more involved. We nd no evidence that teachers adjust their teaching to the composition of groups. "

Data

Standardized E ect Sizes

I aggregate all regression results into two e ect size measures

1. Standardized coe cients :

$$= \begin{cases} \frac{}{Y} & \text{if D is binary} \\ \frac{D}{Y} & \text{otherwise} \end{cases} \qquad s = \begin{cases} \frac{SE(\ )}{Y} & \text{if D is binary} \\ \frac{SE(\ )\ D}{Y} & \text{otherwise} \end{cases}$$

$_Y$ and $_D$ are respectively within-sample SDs of Y and D

- is closely related to the classical Cohen's d e ect size

2. Partial correlation coe cients (PCCs) :

$$r = p\frac{t_{NHST}}{t_{NHST}^2 + df} \qquad SE(r) = \frac{1}{p}\frac{r^2}{df}:$$

t_{NHST} is the usual t-statistic and df is degrees of freedom

- PCCs are widely-used in economic meta-analyses

Failure Rates Introduction

Benchmarking Sample

The Three-Sided Testing Framework

The three-sided testing (TST) framework (Goeman, Solari, & Stijnen 2010) uses ROPE [; $_{+}$] to assess 's practical signi cance using three tests:

The Three-Sided Testing Framework

The three-sided testing (TST) framework (Goeman, Solari, & Stijnen 2010) uses ROPE [; $_{+}$] to assess 's practical signi cance using three tests:

1. Two-sided test: Is < ?
2. TOST procedure: Is 2 [; $_{+}$]?
3. Two-sided test: Is > $_{+}$?

The Three-Sided Testing Framework

The three-sided testing (TST) framework (Goeman, Solari, & Stijnen 2010) uses ROPE [; +] to assess 's practical signi cance using three tests:

1. Two-sided test: Is < ?
2. TOST procedure: Is 2 [; +]?
3. Two-sided test: Is > + ?

Signi cance conclusions can be derived from the smallest of these three values (Sha er 1986; Finner & Strassburger 2002)

The Three-Sided Testing Framework

The three-sided testing (TST) framework (Goeman, Solari, & Stijnen 2010) uses ROPE [; +] to assess 's practical signi cance using three tests:

1. Two-sided test: Is < ?
2. TOST procedure: Is 2 [; +]?
3. Two-sided test: Is > + ?

Signi cance conclusions can be derived from the smallest of these three p-values (Sha er 1986; Finner & Strassburger 2002)

- If no p-value < , then results areinconclusive the researcher must stay agnostic about the practical signi cance of

The Three-Sided Testing Framework

The **three-sided testing (TST) framework** (Goeman, Solari, & Stijnen 2010) uses ROPE [; $_{+}$] to assess 's practical signi cance using three tests:

1. Two-sided test: Is < ?
2. TOST procedure: Is 2 [; $_{+}$]?
3. Two-sided test: Is > $_{+}$?

Signi cance conclusions can be derived from the smallest of these three *p*-values (Sha er 1986; Finner & Strassburger 2002)

- If no *p*-value < , then results are *inconclusive*: the researcher must stay agnostic about the practical signi cance of
- Embracing this uncertainty may be uncomfortable/limiting, but my results show that standard practice tolerates high error rates

The Three-Sided Testing Framework

The **three-sided testing (TST) framework** (Goeman, Solari, & Stijnen 2010) uses ROPE [; $_{+}$] to assess 's practical signi cance using three tests:

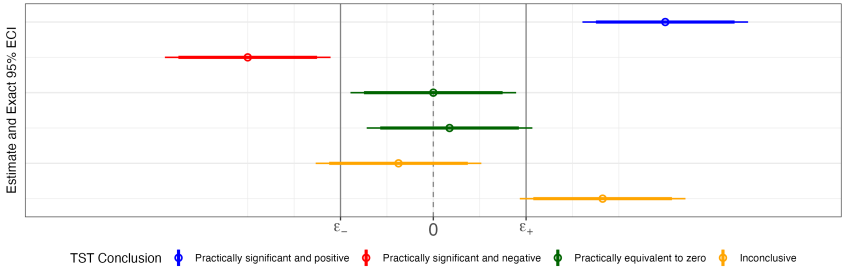
1. Two-sided test: Is < ?
2. TOST procedure: Is 2 [; $_{+}$]?
3. Two-sided test: Is > $_{+}$?

Signi cance conclusions can be derived from the smallest of these three p-values (Sha er 1986; Finner & Strassburger 2002)

- If no p-value < , then results are *inconclusive*: the researcher must stay agnostic about the practical signi cance of
- Embracing this uncertainty may be uncomfortable/limiting, but my results show that standard practice tolerates high error rates

Example from this project: I show that my failure rates are signi cantly bounded above the median failure rates that researchers deem acceptable Main Results

The TST Framework Visualized



Under TST, given their 95% ECIs and CIs, these estimates are respectively:

- Practically signi cant and above the ROPE
- Practically signi cant and below the ROPE
- Practically equivalent to zero
- Inconclusive

Main Results

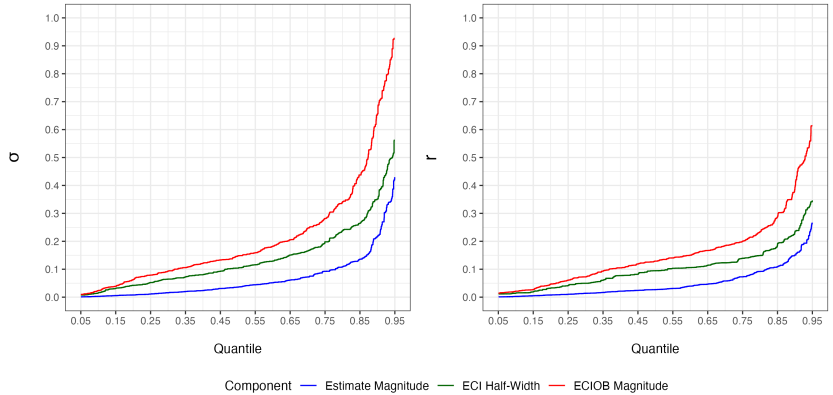
Failure Rate Robustness

These failure rates remain large and significant when...

- Switching from to r
- Switching from exact to asymptotically approximate tests
- Switching aggregation procedures
- Removing initially stat. sig. estimates
- Separating models by regressor type combination (i.e., binary vs. non-binary)
- Removing non-replicable estimates from the sample
- Removing models that require conformability modifications from the sample (e.g., logit/probit models put through `margins, dydx()`)

Main Results

Mechanisms



Power is a greater driver of equivalence testing failure rates than effect size

Main Results