

Equivalence Testing and Practical Significance Testing for Economics

Jack Fitzgerald

Vrije Universiteit Amsterdam and Tinbergen Institute

May 22, 2025



“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)

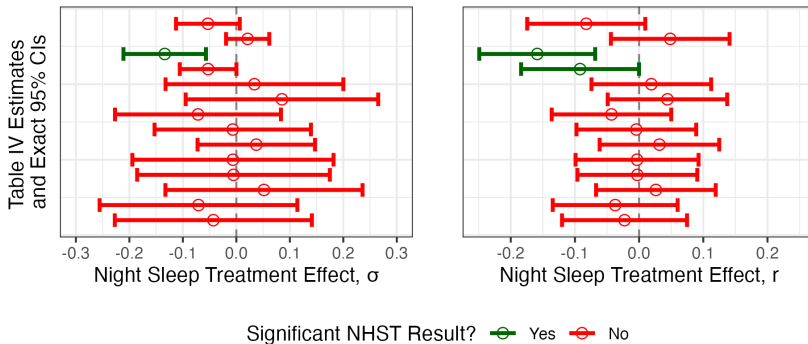
However, per their abstract...

*“Contrary to expert predictions and a large body of sleep research, increased nighttime sleep had **no detectable effects** 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 effects?**’

Null Estimates in Bessone et al. (2021)



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

- **They are not alone** in interpreting insignificant results in this way

This Happens All the Time

Abstract

Smallholder farming in many developing countries is characterized by low productivity and low-quality output. Low quality limits the price farmers can command and their potential income. We conduct a series of experiments among maize farmers in Uganda to shed light on the barriers to quality upgrading and to study its potential. **We find that the causal return to quality is zero.** Providing access to a market where quality is paid a market premium led to an increase in farm productivity and income from farming. Our findings reveal the importance of demand-side constraints in limiting rural income and productivity growth.

Abstract

Consumers rely on the price changes of goods in their grocery bundles when forming expectations about aggregate inflation. We use micro data that uniquely match individual expectations, detailed information about consumption bundles, and item-level prices. The weights consumers assign to price changes depend on the frequency of purchase, rather than expenditure share, and positive price changes loom larger than negative price changes. **Prices of goods offered in the same store but not purchased do not affect inflation expectations, nor do other dimensions.** Our results provide empirical guidance for models of expectations formation with heterogeneous consumers.

Abstract

We study how political turnover in mayoral elections in Brazil affects public service provision by local governments. Exploiting a regression discontinuity design for close elections, we find that municipalities with a new party in office experience upheavals in the municipal bureaucracy: new personnel are appointed across multiple service sectors, and at both managerial and non-managerial levels. In education, the increase in the replacement rate of personnel in schools controlled by the municipal government is accompanied by test scores that are 0.05–0.08 standard deviations lower. In contrast, **turnover of the mayor's party does not impact local (non-municipal) schools.** These findings suggest that political turnover can adversely affect the quality of public services when the bureaucracy is not shielded from the political process.

Abstract

This paper estimates intertemporal labor supply responses to two-year long income tax holidays staggered across Swiss cantons. Cantons shifted from an income tax system based on the previous two years' income to a standard annual pay as you earn system, leaving two years of income untaxed. We find significant but quantitatively very small responses of wage earnings with an intertemporal elasticity of 0.025 overall. High wage income earners and especially the self-employed display larger responses with elasticities around 0.1 and 0.25, respectively, most likely driven by tax avoidance. **We find no effects along the extensive margin at all.**

From 2020-2023, 279 null claims made in abstracts of 158 articles in T5 journals are defended by statistically insignificant results [Detailed Results](#)

- > 72% of these null claims aren't qualified by references to statistical significance, estimate magnitudes, or a lack of evidence

Researchers and readers interpret such findings 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 scientific practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

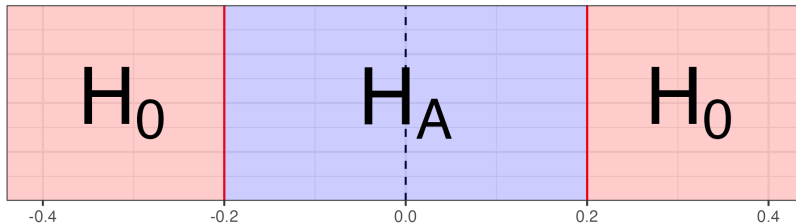
- ▶ Statistical insignificance may just reflect imprecision

Under standard NHST, null results and imprecision are conflated. 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 δ would be **practically equivalent** to zero (i.e., *economically insignificant*)
2. Use interval tests to assess if $\hat{\delta}$ 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)

The Need for Equivalence Testing in Economics

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
- ▶ I then show how we can extend this to general practical significance testing

This Talk

Introducing equivalence testing and its necessity (JMP)

Fitzgerald, J. (2025). "The Need for Equivalence Testing in Economics." *MetaArXiv*, https://doi.org/10.31222/osf.io/d7sqr_v1.

Extending to practical significance testing

- ▶ Isager, P. & Fitzgerald, J. (2024). "Three-Sided Testing to Establish Practical Significance: A Tutorial." *PsyArXiv*, <https://doi.org/10.31234/osf.io/8y925>.

Applying equivalence testing in econometric methodology (briefly)

- ▶ Fitzgerald, J. (2025). "Manipulation Tests in Regression Discontinuity Design: The Need for Equivalence Testing." *MetaArXiv*, https://doi.org/10.31222/osf.io/2dgrp_v1.

The Wrong Hypotheses: NHST

Standard NHST hypotheses:

$$H_0 : \delta = 0$$

$$H_A : \delta \neq 0$$

When trying to show that $\delta = 0$ 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 \rightarrow 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 constraints.

As a reminder, **NHST hypotheses**:

$$H_0 : \delta = 0$$

$$H_A : \delta \neq 0$$

And now **equivalence testing hypotheses**:

$$H_0 : \delta \not\approx 0$$

$$H_A : \delta \approx 0$$

If we can set a range of values $[\epsilon_-, \epsilon_+]$ wherein $\delta \approx 0$, then we can find stat. sig. evidence for H_A with a simple interval test

The Equivalence Testing Framework

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

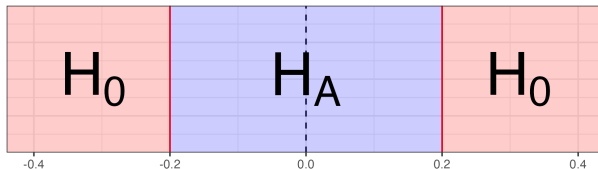
- ▶ The ROPE is the range of δ values we'd call *economically insignificant*
- ▶ This is a subjective judgment call that will differ for different relationships of interest
- ▶ I show how to credibly aggregate ROPEs later in this talk Credible ROPE-Setting

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

$$H_0 : \delta \notin [\epsilon_-, \epsilon_+]$$

$$H_A : \delta \in [\epsilon_-, \epsilon_+]$$

Two One-Sided Tests (TOST)



We can identically write the equivalence testing hypotheses as

$$H_0 : \delta < \epsilon_- \text{ or } \delta > \epsilon_+$$

$$H_A : \delta \geq \epsilon_- \text{ and } \delta \leq \epsilon_+$$

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

$$H_0 : \delta < \epsilon_-$$

$$H_A : \delta \geq \epsilon_-$$

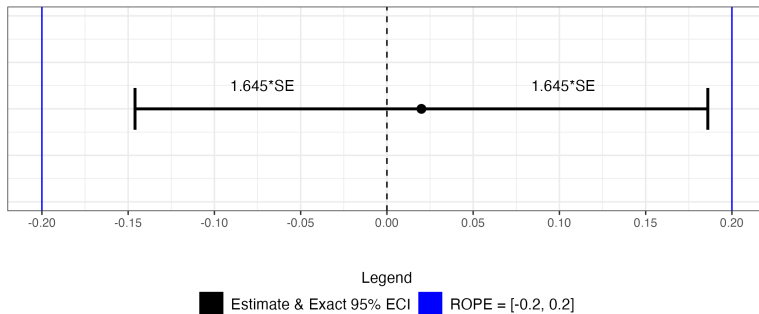
$$H_0 : \delta > \epsilon_+$$

$$H_A : \delta \leq \epsilon_+$$

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

[Procedural Details](#)[Visualization](#)

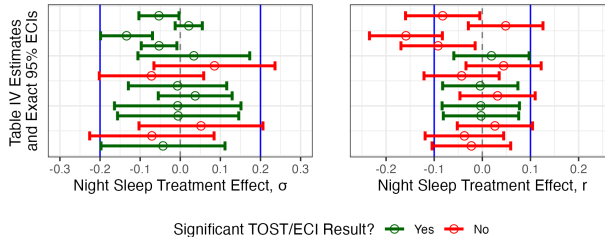
Equivalence Confidence Intervals (ECIs)



$\hat{\delta}$'s $(1 - \alpha)$ **equivalence confidence interval (ECI)** is just its $(1 - 2\alpha)$ CI Mechanisms

- If $\hat{\delta}$'s $(1 - \alpha)$ ECI is entirely bounded in the ROPE, then we have size- α evidence under the TOST procedure that $\delta \approx 0$ (Berger & Hsu 1996)

Revisiting Bessone et al. (2021)



Estimates defending null claims should be significantly bounded within reasonably wide ROPEs

- ▶ However, **28%** of the 'null' estimates in Bessone et al. (2021) aren't significantly bounded beneath $|\sigma| = 0.2$
- ▶ **71%** aren't significantly 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

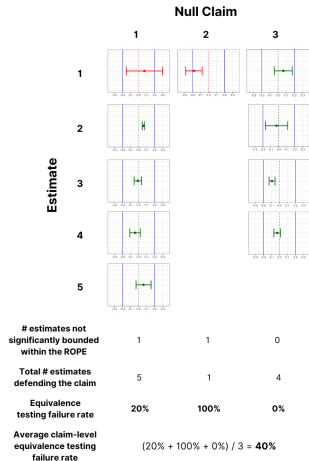
1. Systematically-selected replication sample

- ▶ 876 estimates defending 135 null claims in abstracts of 81 articles in T5 economics journals published from 2020-2023 [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

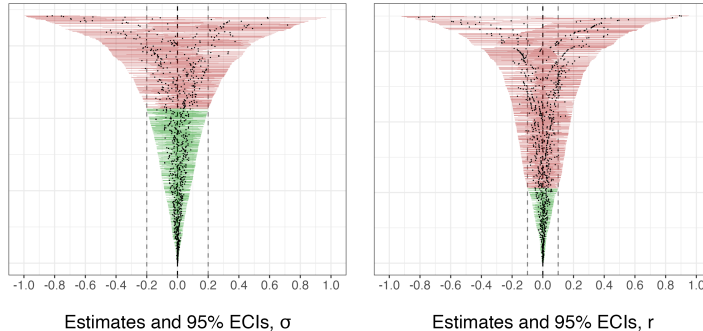


I compute avg. **equivalence testing failure rates** in the replication sample

- **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
- **Second ROPE:** $\sigma \in [-0.2, 0.2]$
- $|\sigma| = 0.2$ is quite large for economic effect sizes
Benchmarking Sample

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

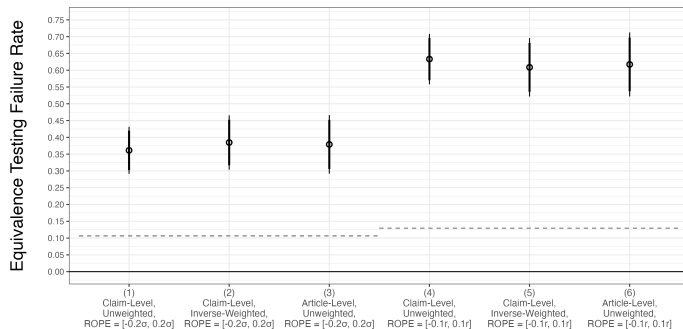
Many 'Null' Estimates Fail Lenient Equivalence Tests



Over 39% of the 'null' estimates in my sample can't be significantly bounded beneath 0.2σ

- Over 69% can't be significantly bounded beneath $0.1r$

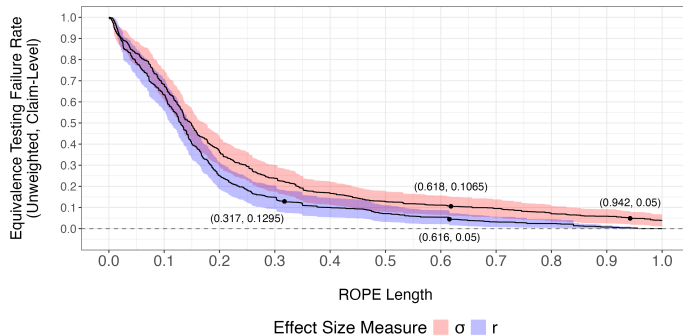
Equivalence Testing Failure Rates are Unacceptably High



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

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

Failure Curves



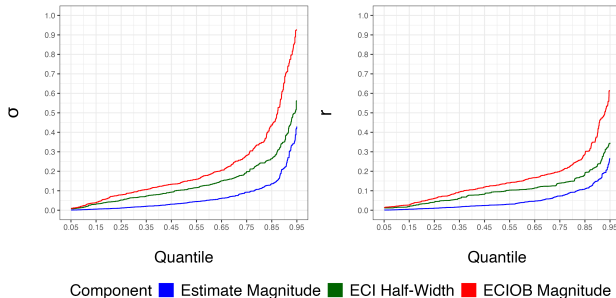
Equivalence testing failure rates stay unacceptably high even as ROPEs become ridiculously large

- ▶ To obtain acceptable failure rates, you'd need to argue that $|0.317r|$ is practically equal to zero
- ▶ $|0.317r|$ is larger than nearly 75% of published effects in economics (Doucouliagos 2011)

Mechanisms: Large Estimates or Low Power?

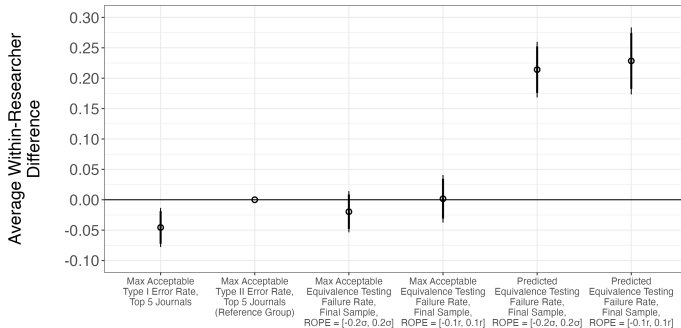
Estimates can fail equivalence tests either because they're **large** or because they're **imprecise**

- We can assess the relative contribution of **effect sizes** and **power** to ETFRs by decomposing ECI outer bounds into **estimate magnitudes** and **ECI half-widths** (ECIs)



Imprecision stochastically dominates **effect size** throughout the distribution of estimates

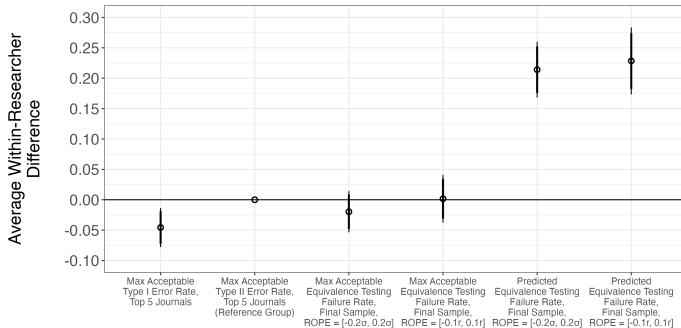
Researchers Anticipate Unacceptably High Failure Rates



The median researcher finds 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

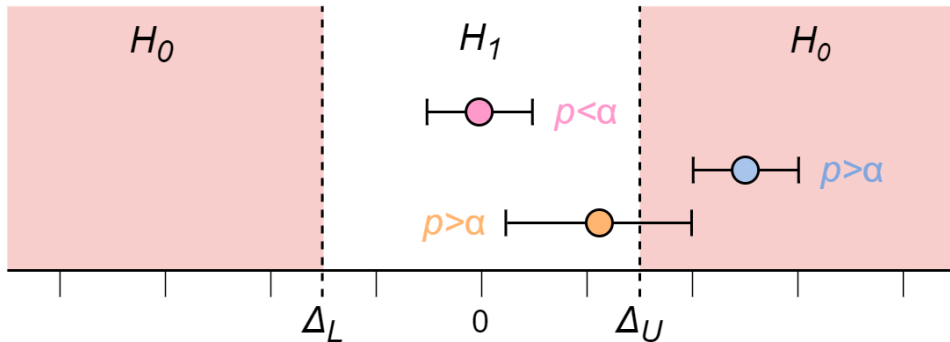
Researchers Anticipate Unacceptably High Failure Rates



The median researcher finds 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

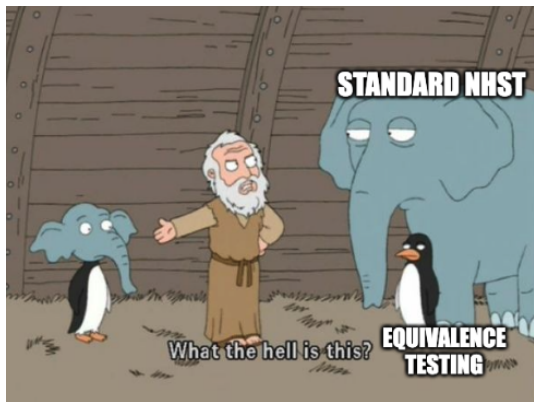
Where Equivalence Testing Falls Short



Under equivalence testing, stat. insign. results do not let us conclude that δ is meaningfully large, or even that $\delta \neq 0$!

- 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

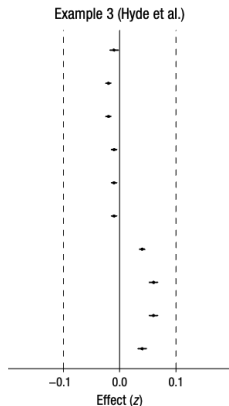


In practice, researchers deal with this by just stapling together equivalence testing with standard NHST

- ▶ 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 in the ROPE

- ▶ 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

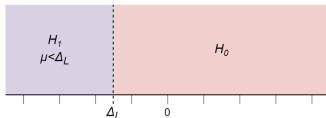
Standard NHST results are often preferred in these settings

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

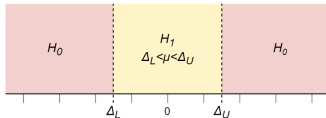
Source: Lakens, Scheel, & Isager (2018)

Construct Validity

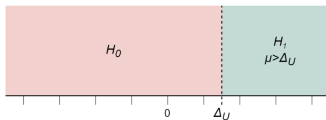
A) Inferiority



B) Equivalence (TOST)



C) Superiority



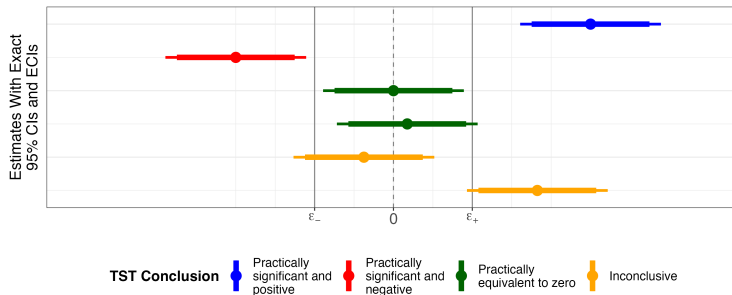
You learn that an estimate is stat. sig. diff. from zero. **So what?**

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

If we want precise evidence that our estimate is practically significant, then we really want to show that the estimate is significantly outside the ROPE

- ▶ 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)

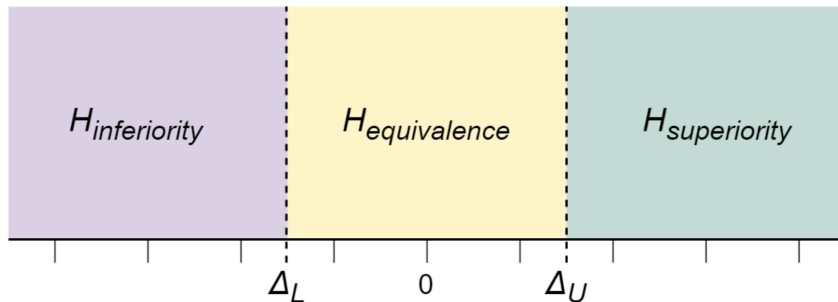


Three-sided testing (TST) is a comprehensive framework for assessing practical significance

- ▶ Under TST, we partition the parameter space into inferiority, equivalence, and superiority regions using the ROPE 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

The Hypothesis Framework



After partitioning δ 's parameter space into H_{Inf} , H_{Eq} , and H_{Sup} regions using the ROPE bounds ϵ_- and ϵ_+ , TST assesses three sets of hypotheses at once:

$$H_0^{\{Inf\}} : \delta \geq \epsilon_-$$

$$H_0^{\{Eq\}} : \delta < \epsilon_- \text{ or } \delta > \epsilon_+$$

$$H_0^{\{Sup\}} : \delta \leq \epsilon_+$$

$$H_A^{\{Inf\}} : \delta < \epsilon_-$$

$$H_A^{\{Eq\}} : \delta \geq \epsilon_- \text{ and } \delta \leq \epsilon_+$$

$$H_A^{\{Sup\}} : \delta > \epsilon_+$$

The Hypothesis Tests in Practice

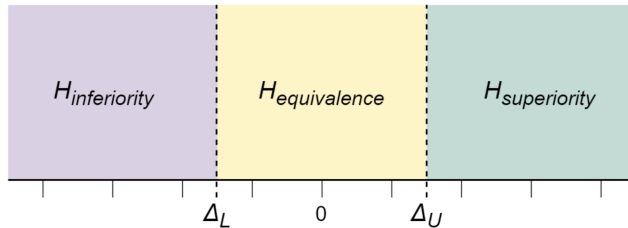
In other words, in TST, you run three tests at once:

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

Because TST begins by partitioning δ 's parameter space into disjoint regions, TST benefits from the 'partitioning principle' (Finner & Straßburger 2002)

- Intuitively, δ can only be in one of H_{Inf} , H_{Eq} , or H_{Sup} at once

A Multiple Testing Caveat (1/2)



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

- ▶ I can conclude that δ sits in H_{Inf} when it really sits in H_{Eq}
- ▶ I can conclude that δ 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 δ 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

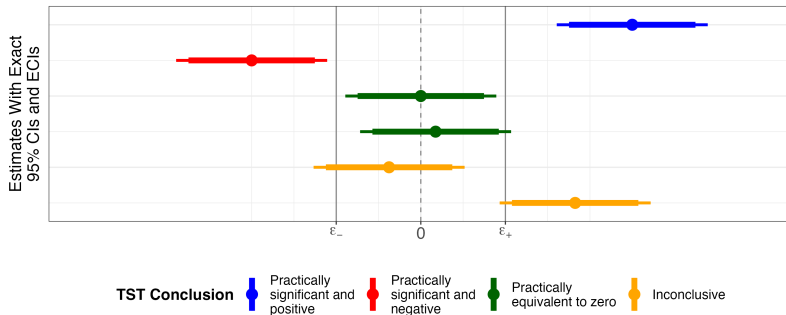
- ▶ Goeman, Solari, & Stijnen (2010) deal with this using a simple Bonferroni correction on the critical values for TST's inferiority and superiority tests
- ▶ This is a special case with two simultaneous tests, so this correction just cuts effective significance levels in half

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

- ▶ The equivalence test is conducted with significance level α (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 α (e.g., 5%)!

Double-Banded Confidence Intervals

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

Return of the Stapler: TST + Standard NHST

What if our ROPE can change over time?

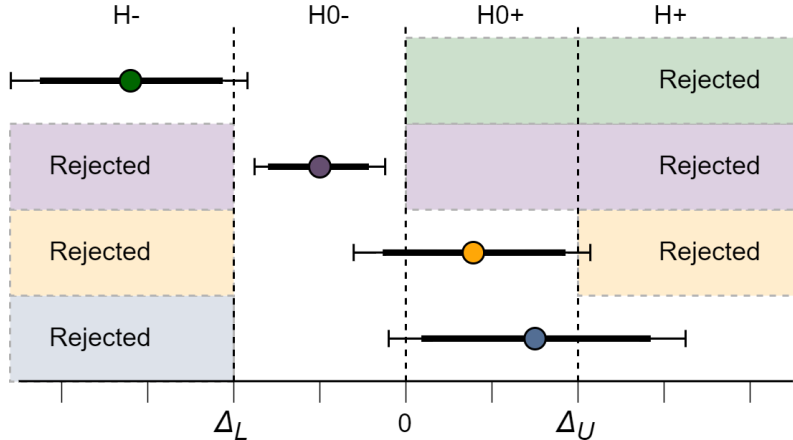
- ▶ E.g., what seems like a small financial gain to a healthy organization may become game-changing in times of financial distress

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 a ROPE bound and/or imprecise, it may not be possible to find stat. sig. evidence that δ 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 δ 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
- ▶ This requires scientists 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 ROPEs, 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 ROPE
- ▶ 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, α 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

Software



Link:

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

We provide several tools for researchers to easily apply TST

1. The ShinyTST app, a Shiny app
2. The `tst` function in my `eqtesting` R package (<https://github.com/jack-fitzgerald/eqtesting/>), forthcoming at CRAN
3. The `tsti` command in Stata, now downloadable from SSC (`ssc install tsti`)

Credible ROPE-Setting

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

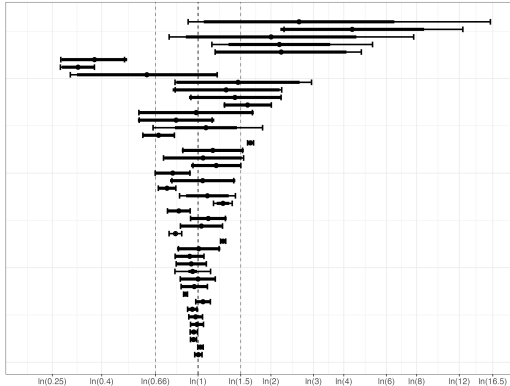
- ▶ ‘ROPE-hacking’ is a key concern
- ▶ To maintain independence & credibility, *you* shouldn’t set your ROPEs – you should get other people to set them for you

Solution: Survey independent experts/stakeholders for their judgments

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

The Equivalence Testing Framework

Extension: Manipulation Testing in Regression Discontinuity



Logarithmic RV Density Discontinuity at the Cutoff

In RDD, it's common to test whether there are significant jumps in running variable (RV) density at treatment cutoffs

- ▶ Like many robustness checks, null results are desirable (Dreber, Johannesson, & Yang 2024)

I develop equivalence testing procedures for these RV manipulation tests

- ▶ In a sample of 36 published RDD papers, I find similar robustness failures
- ▶ 44% of RV density discontinuities at treatment cutoffs can't be significantly bounded beneath a 50% upward jump

Equivalence testing → entire classes of new methods

Replication-Based Methods Research

Common attitude towards research methods: “If it ain’t broke, don’t fix it”

- ▶ Many historical success stories in advancing econometric methods have come after showing empirically that there are big problems with existing methods
- ▶ E.g., LaLonde (1986); Bound, Jaeger, & Baker (1995); de Chaisemartin & D’Haultfœuille (2020)

Replication-based methods research can quickly establish the importance of adopting methodological improvements

- ▶ Immediately guarantees your method’s applicability in the literature
- ▶ Systematic searches can guard against cherry-picking
- ▶ Credibility advantages over simulations, where DGPs can be hacked for favorable conditions

This work is time-intensive, but worthwhile and better in teams (contact me!)

- ▶ **What research practices annoy you?**

General Takeaways

Social scientists need to start testing hypotheses with effect sizes in mind

- ▶ For many relationships, there's a meaningful smallest effect size of interest
- ▶ We should leverage this information to test the *practical significance* of estimates

There is great opportunity to develop better methods with equivalence testing

- ▶ Researchers often want to test whether relationships are practically equal to zero
- ▶ The RDD paper is a proof-of-concept; more is coming

Many debates on methods can be resolved with replication-based methods research

- ▶ JMP originated because I was told (repeatedly) that in top journals, $p > 0.05$ is a good indicator of null relationships; 81 replications later, clearly that's not true
- ▶ Growing subfield (see Hainmueller, Mummolo, & Xu 2019; Muralidharan, Romero, & Wutrich 2023; Stommes, Aronow, & Sävje 2023; Chiu et al. 2024; Lal et al. 2024)

Thank You For Your Attention!



These Slides

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

Email: j.f.fitzgerald@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).

Identification 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 confidence sets.
Statistical Science 11(4).

References II

-  Bessone, P., G. Rao, F. Schilbach, H. Schofield, and M. Toma (2021).
The economic consequences of increasing sleep among the urban poor.
The Quarterly Journal of Economics 136(3), 1887–1941.
-  Bound, J., D. A. Jaeger, and R. M. Baker (1995).
Problems with instrumental variables estimation when the correlation between the instruments
and the endogenous explanatory variable is weak.
Journal of the American Statistical Association 90(430), 443–450.
-  Campbell, H. and P. Gustafson (2018).
Conditional equivalence testing: An alternative remedy for publication bias.
PLOS ONE 13(4).
-  Chiu, A., X. Lan, Z. Liu, and Y. Xu (2024).
Causal panel analysis under parallel trends: Lessons from a large reanalysis study.

References III



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.



Cunningham, S. (2021, Aug).

Causal inference: The mixtape (1 ed.).

Yale University Press.



de Chaisemartin, C. and X. D'Haultfœuille (2020).





Two-way fixed effects estimators with heterogeneous treatment effects.

American Economic Review 110(9), 2964–2996.





References IV

-  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.
-  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.





References V

-  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.
-  Franco, A., N. Malhotra, and G. Simonovits (2014).
Publication bias in the social sciences: Unlocking the file drawer.
Science 345(6203), 1502–1505.
-  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.
-  Hainmueller, J., J. Mummolo, and Y. Xu (2019).
How much should we trust estimates from multiplicative interaction models? simple tools to improve empirical practice.
Political Analysis 27(2), 163–192.




References VI

-  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 VII

-  Lal, A., M. Lockhart, Y. Xu, and Z. Zu (2024).
How much should we trust instrumental variable estimates in political science? practical advice based on 67 replicated studies.
Political Analysis, 1–20.
-  LaLonde, R. J. (1986).
Evaluating the econometric evaluations of training programs with experimental data.
American Economic Review 76(4), 604–620.
-  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 effect of statistical training on the evaluation of evidence.
Management Science 62(6), 1707–1718.

References VIII

-  McShane, B. B. and D. Gal (2017).
Statistical significance and the dichotomization of evidence.
Journal of the American Statistical Association 112(519), 885–895.
-  Muralidharan, K., M. Romero, and K. Wüthrich (2023).
Factorial designs, model selection, and (incorrect) inference in randomized experiments.
The Review of Economics and Statistics, 1–44.
-  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.

References IX

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

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

Journal of Clinical Epidemiology 161, 20–27.

-  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 X

-  Stommes, D., P. M. Aronow, and F. Sävje (2023a, Apr).
On the reliability of published findings using the regression discontinuity design in political science.
Research & Politics 10(2), 205316802311664.
-  Stommes, D., P. M. Aronow, and F. Sävje (2023b, Mar).
Replication data for: On the reliability of published findings using the regression discontinuity design in political science.
Dataset V1, Harvard Dataverse, Cambridge, MA, U.S.A.
-  Wasserstein, R. L. and N. A. Lazar (2016).
The ASA statement on p -values: Context, process, and purpose.
The American Statistician 70(2), 129–133.

Null Claim Classification

Category	Claim Type	Example	# Claims	% of Claims
1	Claim that a relationship/phenomenon does not exist or is negligible	<i>D</i> has no effect on <i>Y</i> .	111	39.8%
2	Claim that a relationship/phenomenon does not exist or is negligible, qualified by reference to statistical significance	<i>D</i> has no significant effect on <i>Y</i> .	33	11.8%
3	Claim that a relationship/phenomenon does not exist or is negligible, qualified by reference to something other than statistical significance	<i>D</i> has no meaningful effect on <i>Y</i> .	24	8.6%
4	Claim that a relationship/phenomenon does not (meaningfully) hold in a given direction	<i>D</i> has no positive effect on <i>Y</i> .	53	19%
5	Claim that a relationship/phenomenon does not (meaningfully) hold in a given direction, qualified by reference to statistical significance	<i>D</i> has no significant positive effect on <i>Y</i> .	4	1.4%
6	Claim that a relationship/phenomenon does not (meaningfully) hold in a given direction, qualified by reference to something other than statistical significance	<i>D</i> has no meaningful positive effect on <i>Y</i> .	5	1.8%
7	Claim that there is a lack of evidence for a (meaningful) relationship/phenomenon	There is no evidence that <i>D</i> has an effect on <i>Y</i> .	10	3.6%
8	Claim that a variable holds similar values regardless of the values of another variable	<i>Y</i> is similar for those in the treatment group and the control group.	7	2.5%
9	Claim that a relationship/phenomenon holds only or primarily in a subset of the data	The effect of <i>D</i> on <i>Y</i> is concentrated in older respondents.	22	7.9%
10	Claim that a relationship/phenomenon stabilizes for some values of another variable	<i>D</i> has a short term effect on <i>Y</i> that dissipates after <i>Z</i> months.	10	3.6%
Unqualified null claim		Categories 1, 4, or 8-10	203	72.8%
Qualified null claim		Categories 2-3 or 5-7	76	27.2%

The TOST Procedure

First, compute test statistics

$$t_- = \frac{\hat{\delta} - \epsilon_-}{s}$$

$$t_+ = \frac{\hat{\delta} - \epsilon_+}{s}$$

The relevant test statistic is the smaller of the two:

$$t_{\text{TOST}} = \arg \min_{t \in \{t_-, t_+\}} \{|t|\}$$

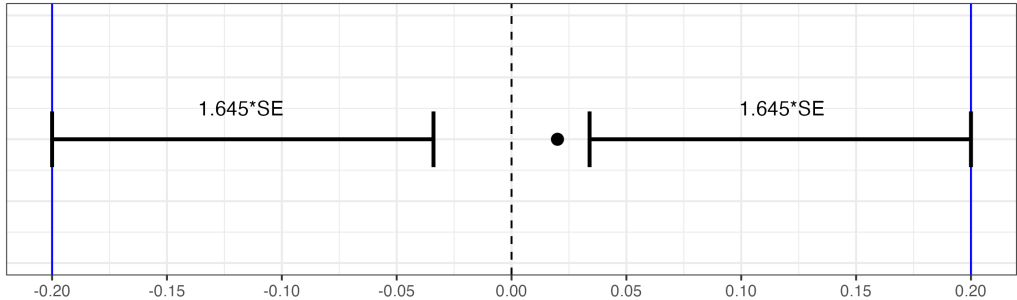
The critical value for a size- α TOST procedure is the **one-sided** critical value t_α^*

1. If $t_{\text{TOST}} = t_-$, then there is stat. sig. evidence that $\delta \in [\epsilon_-, \epsilon_+]$ iff $t_- \geq t_\alpha^*$
2. If $t_{\text{TOST}} = t_+$, then there is stat. sig. evidence that $\delta \in [\epsilon_-, \epsilon_+]$ iff $t_+ \leq -t_\alpha^*$

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

TOST Concept

TOST Example



Legend



Estimate w/ t-Tests from Above and Below



ROPE = [-0.2, 0.2]

Claim Example

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

*“This article estimates peer effects 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 specification 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 effects regressions for group assignment, we find 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 unaffected.** Analysis of survey data indicates that in tracked groups, low-ability students have more positive interactions with other students, and are more involved. **We find no evidence that teachers adjust their teaching to the composition of groups.**”*

Data

Standardized Effect Sizes

I aggregate all regression results into two effect size measures

1. Standardized coefficients:

$$\sigma = \begin{cases} \frac{\delta}{\sigma_Y} & \text{if } D \text{ is binary} \\ \frac{\delta\sigma_D}{\sigma_Y} & \text{otherwise} \end{cases} \quad s = \begin{cases} \frac{SE(\delta)}{\sigma_Y} & \text{if } D \text{ is binary} \\ \frac{SE(\delta)\sigma_D}{\sigma_Y} & \text{otherwise} \end{cases}$$

σ_Y and σ_D are respectively within-sample SDs of Y and D

► σ is closely related to the classical Cohen's d effect size

2. Partial correlation coefficients (PCCs):

$$r = \frac{t_{\text{NHST}}}{\sqrt{t_{\text{NHST}}^2 + df}} \quad SE(r) = \frac{1 - r^2}{\sqrt{df}}.$$

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

► PCCs are widely-used in economic meta-analyses

Benchmarking Sample

Article	Setting	Outcome Variable	Exposure Variable	Initial p-Value	σ	r	Location
Acemoglu & Restrepo (2020)	Difference-in-differences analysis of U.S. commuting zones, 1990-2007	Employment rates (continuous)	Industrial robot exposure (continuous)	0.000	-0.206	-0.16	Table 7, Panel A, US exposure to robots, Model 3
Acemoglu et al. (2019)	Difference-in-differences analysis of countries, 1960-2010	Short-run log GDP levels (continuous)	Democratization (binary)	0.001	0.005	0.255	Table 2, Democracy, Model 3
Berman et al. (2017)	African 0.5×0.5 longitude-latitude cells with mineral mines, 1997-2010	Conflict incidence (binary)	Log price of main mineral (continuous)	0.012	0.521	0.007	Table 2, ln price x mines > 0, Model 1
Deschênes, Greenstone, & Shapiro (2017)	Difference-in-differences analysis of U.S. counties, 2001-2007	Nitrogen dioxide emissions (continuous)	Nitrogen dioxide cap-and-trade participation (binary)	0.000	-0.134	-0.468	Table 2, Panel A, NOx, Model 3
Haushofer & Shapiro (2016)	Experiment with low-income Kenyan households, 2011-2013	Non-durable consumption (continuous)	Unconditional cash transfer (binary)	0.000	0.376	0.195	Table V, Non-durable expenditure, Model 1
Benhassine et al. (2015)	Experiment with families of Moroccan primary school-aged students, 2008-2010	School attendance (binary)	Educational cash transfer to fathers (binary)	0.000	0.18	0.252	Table 5, Panel A, Attending school by end of year 2, among those 6-15 at baseline, Impact of LCT to fathers
Bloom et al. (2015)	Field experiment with Chinese workers, 2010-2011	Attrition (binary)	Voluntarily working from home (binary)	0.002	-0.397	-0.196	Table VIII, Treatment, Model 1
Duflo, Dupas, & Kremer (2015)	Experiment with Kenyan primary school-aged girls, 2003-2010	Reaching eighth grade (binary)	Education subsidy (binary)	0.023	0.1	0.125	Table 3, Panel A, Stand-alone education subsidy, Model 1
Hanushek et al. (2015)	OECD adult workers, 2011-2012	Log hourly wages (continuous)	Numeracy skills (continuous)	0.000	0.091	0.316	Table 5, Numeracy, Model 1
Oswald, Proto, & Sgroi (2015)	UK students, piece-rate laboratory task	Productivity (continuous)	Happiness (continuous)	0.018	0.753	0.244	Table 2, Change in happiness, Model 4

Failure Rate Robustness

These failure rates remain large and significant when...

- ▶ Switching from σ 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