Equivalence Testing: Applications in Economics and Econometrics

Jack Fitzgerald

Vrije Universiteit Amsterdam and Tinbergen Institute

January 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)

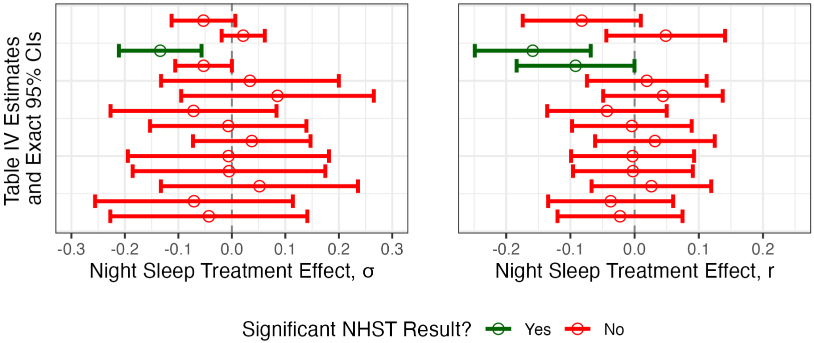
However, per their abstract...

> *“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?’**

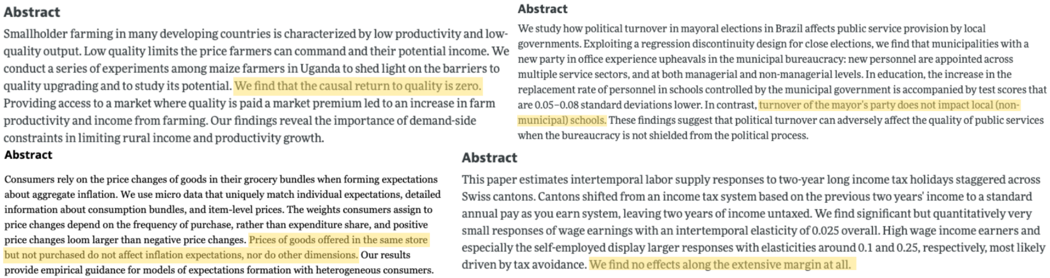
Null Estimates in Bessone et al. (2021)



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

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

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

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)

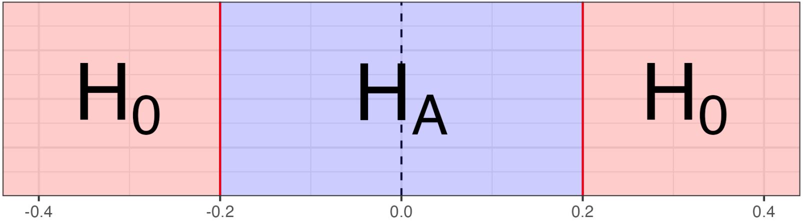
- Statistical insignificance may just reflect imprecision

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)

This Talk

Introducing equivalence testing and its necessity (JMP)

Fitzgerald, J. (2024). "The Need for Equivalence Testing in Economics." *Institute for Replication Discussion Paper Series* No. 125. https://hdl.handle.net/10419/296190.

Extending to practical significance testing (briefly)

- Isager, P. & Fitzgerald, J. (2024). "Three-Sided Testing to Establish Practical Significance: A Tutorial." *Tinbergen Institute Discussion Paper Series* No. 2024-077/III. https://papers.tinbergen.nl/24077.pdf.

Extending equivalence testing to econometric methodology

- Fitzgerald, J. (2024). "Manipulation Tests in Regression Discontinuity Design: The Need for Equivalence Testing. *Institute for Replication Discussion Paper Series* No. 136. https://www.econstor.eu/handle/10419/300277.

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

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 contraints. 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

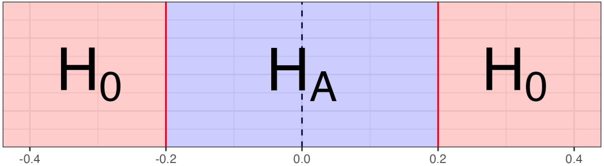
We begin by setting a range of values $[\epsilon_-, \epsilon_+]$, where $\epsilon_- < \epsilon_+$, called the *region of practical equivalence (ROPE)*

- 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

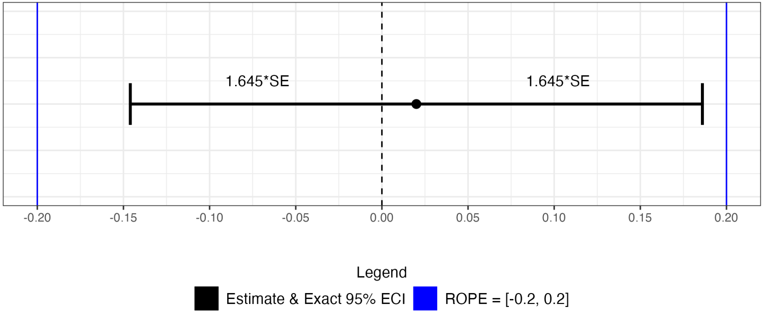
$$H_0 : \delta < \epsilon_- \text{ or } \delta > \epsilon_+$$
$$H_A : \delta \geq \epsilon_- \text{ and } \delta \leq \epsilon_+$$

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

$$H_0 : \delta < \epsilon_- \qquad\qquad H_0 : \delta > \epsilon_+$$
$$H_A : \delta \geq \epsilon_- \qquad\qquad H_A : \delta \leq \epsilon_+$$

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

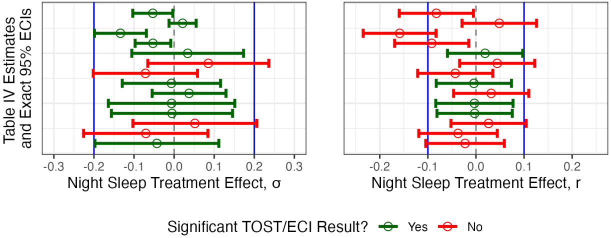
Equivalence Confidence Intervals (ECIs)



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

- 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

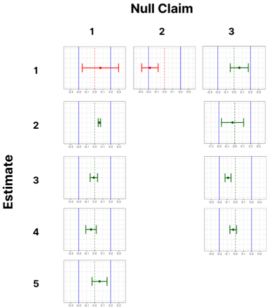
- However, **28%** of the 'null' estimates in Bessone et al. (2021) aren't significantly bounded beneath $|\sigma| = 0.2$
- **71%** aren't significantly bounded beneath $|r| = 0.1$

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



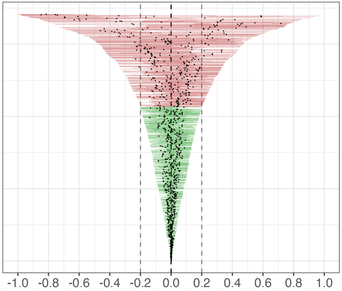
	1	2	3
# estimates not significantly bounded within the ROPE	1	1	0
Total # estimates defending the claim	5	1	4
Equivalence testing failure rate	**20%**	**100%**	**0%**
Average claim-level equivalence testing failure rate	(20% + 100% + 0%) / 3 = **40%**		

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

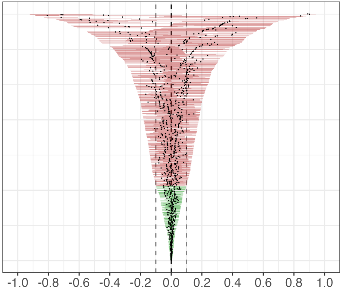
- **First ROPE**: $r \in [-0.1, 0.1]$
- $|r| = 0.1$ is larger than over 25% of published results in economics (Doucouliagos 2011) Effect Size Standardization
- **Second ROPE**: $\sigma \in [-0.2, 0.2]$
- $|\sigma| = 0.2$ is quite large for economic effect sizes Benchmarking Sample

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



Estimates and 95% ECIs, σ

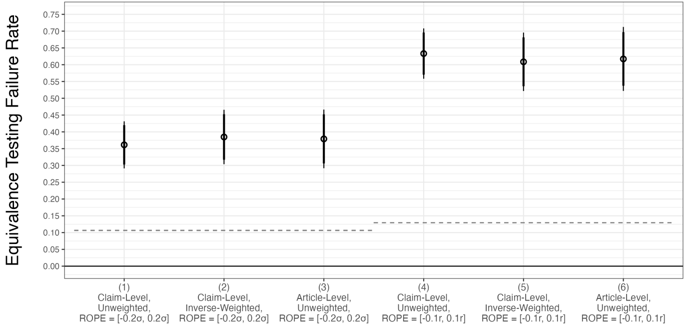


Estimates and 95% ECIs, r

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 $0.1r$

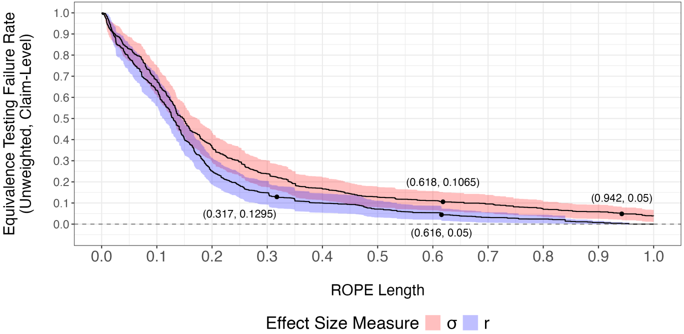
Equivalence Testing Failure Rates are Unacceptably High



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

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

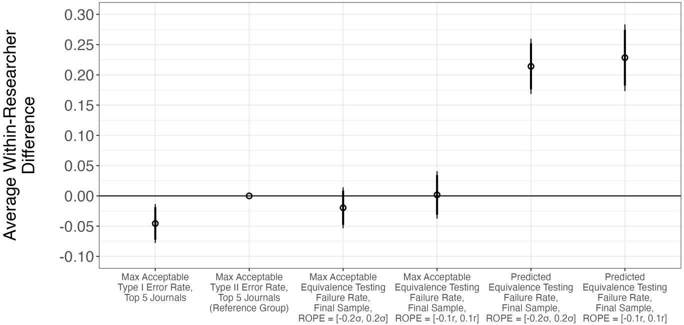
Failure Curves



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)

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

Practical Matters

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

Given this ROPE, an estimate, and an SE, you can directly obtain equivalence testing results using my `tsti` Stata command or the `tst` command in my `eqtesting` R package

- Both can be found at github.com/jack-fitzgerald

The Next Step: Practical Significance Testing

Natural to want to combine equivalence testing with tests for δ's practical significance

- Can be done using the **three-sided testing (TST) framework** (Goeman, Solari, & Stijnen 2010)

Given ROPE $[\epsilon_-, \epsilon_+]$, the idea is to assess δ's practical significance using three tests:

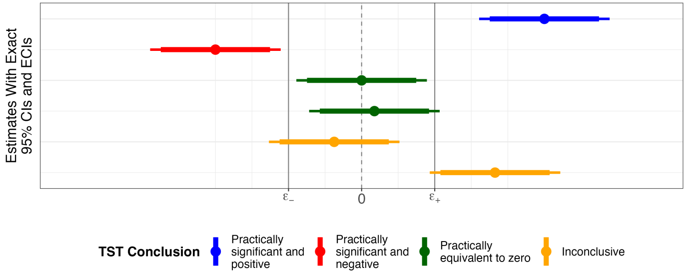
1. Two-sided test: Is $\delta < \epsilon_-$?
2. TOST procedure: Is $\delta \in [\epsilon_-, \epsilon_+]$?
3. Two-sided test: Is $\delta > \epsilon_+$?

Significance conclusions can be derived from the smallest of these three *p*-values

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

Peder Isager and I have written a tutorial on this method, available at osf.io/preprints/psyarxiv/8y925

The Three-Sided Testing Framework Visualized

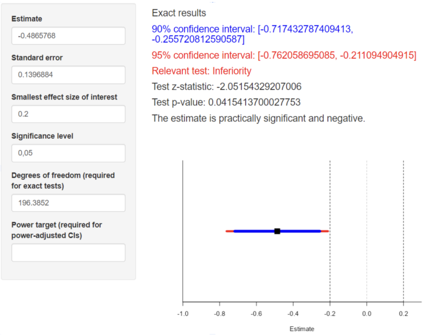


Under TST, given their 95% ECIs and CIs, these estimates are respectively:

- Practically significant and above the ROPE
- Practically significant and below the ROPE
- Practically equivalent to zero
- Inconclusive

Main Results

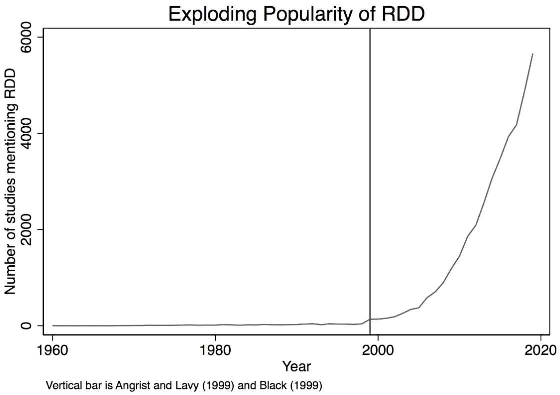
ShinyTST App



Peder Isager & I have built the ShinyTST app, available at jack-fitzgerald.shinyapps.io/shinyTST/

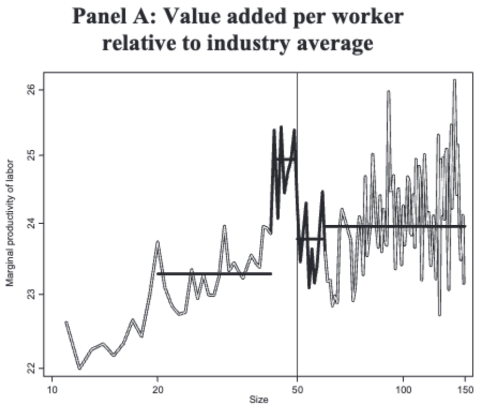
ShinyTST App

Regression Discontinuity Design (RDD)



Cunningham (2021) documents 5600 RDD papers published in 2019 alone

RDD's 'Experimental Appeal'



Source: Garicano, Lelarge, & van Reenen (2016)

In principle, when an agent's running variable (RV) crosses the assignment cutoff, the agent should be effectively randomized into or out of treatment

RV Manipulation at the Cutoff

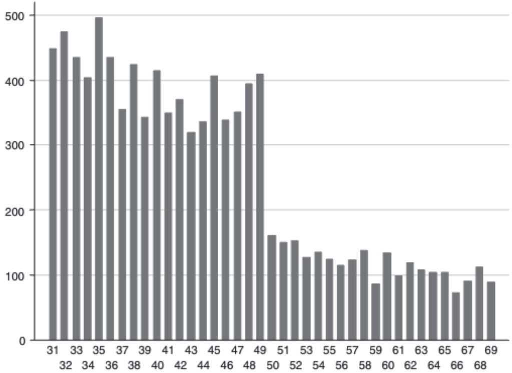


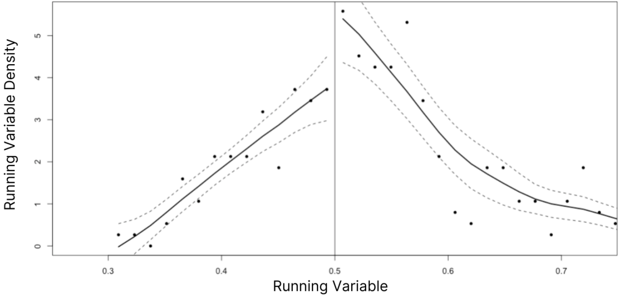
FIGURE 2. NUMBER OF FIRMS BY EMPLOYMENT SIZE IN FRANCE

Source: Garicano, Lelarge, & van Reenen (2016)

Endogenous manipulation of RV values near the cutoff induces selection biases

- Agents can often effectively select into/out of treatment

RV Manipulation Tests



RV manipulation tests estimate and assess discontinuities in the RV's density at the cutoff

- Well-known versions include `DCdensity` and `rddensity` (McCrary 2008; Cattaneo, Jansson, & Ma 2018; Cattaneo, Jansson, & Ma 2020)
- Per Web of Science, these tests have over 2100 citations between them

... and How They're Misused

R Code   Stata Code

```
* If necessary, findit rddensity and install the rddensity package
causaldata gov_transfers_density.dta, use clear download

* Limit to the bandwidth ourselves
keep if abs(income_centered) < .02
* Run the discontinuity check
rddensity income_centered, c(0)
```

As expected, we find no statistically significant break in the distribution of income at the cutoff. Hooray!

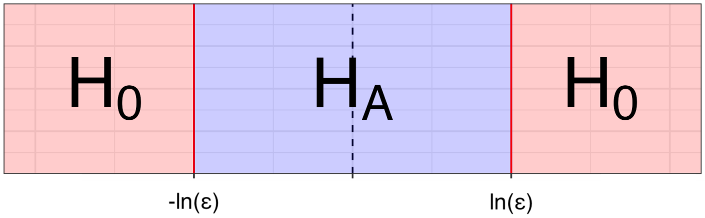
Source: Huntington-Klein (2022)

Unfortunately, researchers (mis)interpret *stat. insig.* manipulation as evidence of *negligible* manipulation

- This is a well-known fallacy (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016)

Meaningful manipulation may go undetected if these tests are underpowered

An Alternative Testing Framework



Ideal: Stat. sig. evidence that RV manipulation ≈ 0. We can get this using equivalence testing:

1. Define the smallest practically/economically significant RV density discontinuities at the cutoff for our given research setting
2. Use interval tests to assess whether the RV density discontinuity at the cutoff is bounded beneath this effect size

This Project

Novel equivalence testing procedure for RV manipulation tests

- Can provide sig. evidence that RV manipulation ≈ 0, which is what applied researchers usually want to show
- Also augments standard RV manipulation tests with bootstrap algorithms for finite-sample (cluster-)robust inference

Empirical evidence of its necessity in applied RDD research

- Replicating 36 published RDD papers shows that $> 44\%$ of RV density discontinuity magnitudes can't be stat. sig. bounded beneath a 50% upward jump

Guidelines and statistical software commands for credible implementation

- `lddtest` command in Stata and in the `eqtesting` R package

Setup (1/2)

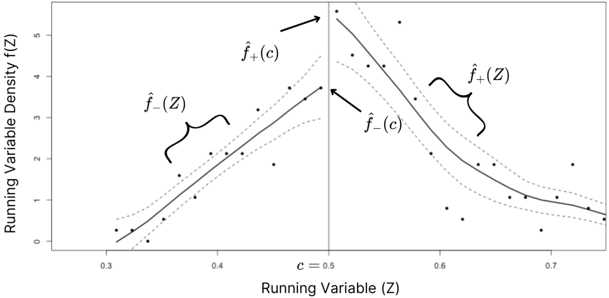
Standard cross-sectional RDD setup (cluster/panel setup possible via bootstrap)

- Agents i have some running variable Z_i
- Agents are assigned to treatment if Z_i crosses cutoff c:

$$D_i = \begin{cases} 1 \text{ if } Z_i \geq c \\ 0 \text{ if } Z_i < c \end{cases} \quad \text{or} \quad D_i = \begin{cases} 1 \text{ if } Z_i \leq c \\ 0 \text{ if } Z_i > c \end{cases}$$

- Z_i exhibits probability density function $f(Z_i)$

Setup (2/2)



We'll test for RV manipulation by testing a continuity assumption: $\lim_{Z_i \to c^-} f(Z_i) = \lim_{Z_i \to c^+} f(Z_i)$

- RV manipulation tests estimate density functions on each side of the cutoff, $\hat{f}_-(Z_i)$ and $\hat{f}_+(Z_i)$
- Our estimates of the LHS and RHS density limits are respectively $\hat{f}_-(c)$ and $\hat{f}_+(c)$

The Wrong Hypotheses: Standard NHST

Standard RV manipulation tests effectively assess the hypotheses

$$H_0 : \lim_{Z_i \to c^-} f(Z_i) = \lim_{Z_i \to c^+} f(Z_i)$$
$$H_A : \lim_{Z_i \to c^-} f(Z_i) \neq \lim_{Z_i \to c^+} f(Z_i).$$

There are many problems with this standard NHST approach

- **No burden of proof**: Researchers assume in the null hypotheses that what they want to show is true
- For most researchers, **imprecision is 'good'**
- **Negligible manipulation can be 'significant'** in high-powered research settings

Creates perverse incentives for **'reverse *p*-hacking'** by setting restrictive bandwidths or not reporting RV manipulation tests (see Dreber, Johanneson, & Yang 2024)

The Right Hypotheses: Equivalence Testing

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

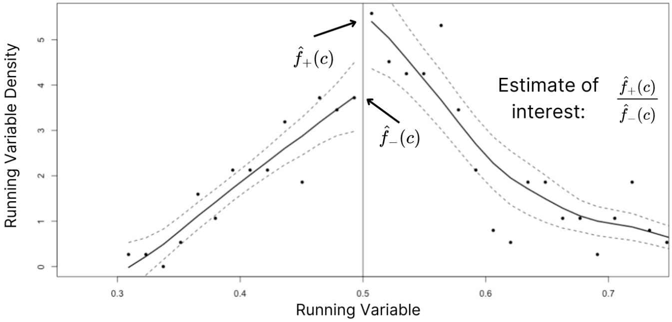
$$H_0 : \lim_{Z_i \to c^-} f(Z_i) = \lim_{Z_i \to c^+} f(Z_i)$$
$$H_A : \lim_{Z_i \to c^-} f(Z_i) \neq \lim_{Z_i \to c^+} f(Z_i).$$

And now **equivalence testing hypotheses**:

$$H_0 : \lim_{Z_i \to c^-} f(Z_i) \not\approx \lim_{Z_i \to c^+} f(Z_i)$$
$$H_A : \lim_{Z_i \to c^-} f(Z_i) \approx \lim_{Z_i \to c^+} f(Z_i).$$

If we can set a range of values wherein the RV's density jump at the cutoff ≈ 0, then we can get stat sig. evidence for H_A with a simple interval test

Step 1: Set the Effect Size Threshold



Set largest practically/economically insignificant RTL density ratio $\epsilon > 1$ for our research setting

- RTL density ratios are useful effect sizes because they are always comparable across datasets
- This threshold can be credibly set by surveying other researchers for their judgments

Step 2: Estimate the Logarithmic Density Discontinuity

McCrary's (2008) `DCdensity` procedure estimates **logarithmic density discontinuities**:

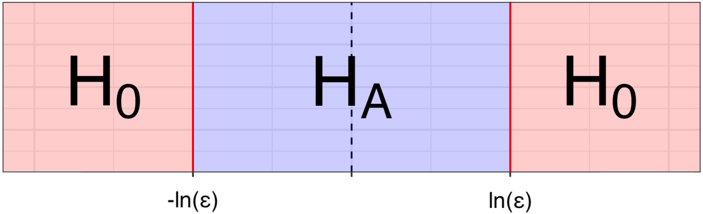
$$\hat{\theta} \equiv \ln\left(\hat{f}_+(c)\right) - \ln\left(\hat{f}_-(c)\right) = \ln\left(\frac{\hat{f}_+(c)}{\hat{f}_-(c)}\right)$$

McCrary (2008) also shows that $\hat{\theta}$ is consistent and asymptotically normal

- We can thus use $\hat{\theta}$ and $\mathrm{SE}\left(\hat{\theta}\right)$ from `DCdensity` for standard Gaussian inference

I also develop (cluster) bootstrap procedures for finite-sample (cluster-)robust inference

Step 3: Equivalence Testing

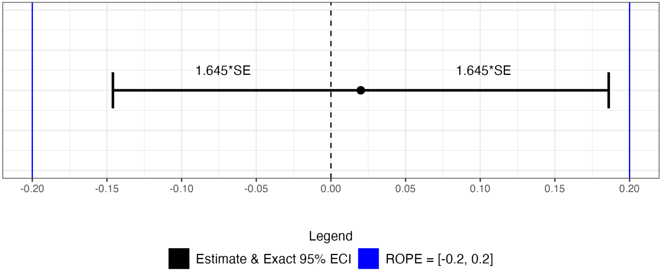


We'll test whether $\hat{\theta}$ is stat. sig. bounded between $-\ln(\epsilon)$ and $\ln(\epsilon)$ w/ two one-sided tests of the form

$$H_0 : \theta < -\ln(\epsilon) \qquad H_0 : \theta > \ln(\epsilon)$$
$$H_A : \theta \geq -\ln(\epsilon) \qquad H_A : \theta \leq \ln(\epsilon)$$

If both tests are stat. sig. at level α, then there's size-α stat. sig. evidence that RV manipulation at the cutoff is practically equal to zero (see Schuirmann 1987; Berger & Hsu 1996) Visualization

Equivalence Confidence Interval (ECI) Approach



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

- If $\hat{\theta}$'s $(1-\alpha)$ ECI is entirely bounded in $[-\ln(\epsilon), \ln(\epsilon)]$, then we have size-$\alpha$ evidence under the TOST procedure that RV manipulation at the cutoff ≈ 0 (Berger & Hsu 1996)

We can use this for (percentile) bootstrap inference by constructing $(1-\alpha)$ bootstrap ECIs

Replication Data

I leverage replication data from Stommes, Aronow, & Sävje (2023a; 2023b), who run robustness checks on 36 published RDD papers in *AJPS*, *APSR*, and *JOP* from 2009-2018

- Some papers use multiple datasets; I run RV manipulation tests in each dataset (45 in total)

Designs in this dataset include close election designs, spatial discontinuities, and age discontinuities

- Per Lee & Lemieux (2010), 42% of published RDD papers in economics use one of these RV classes

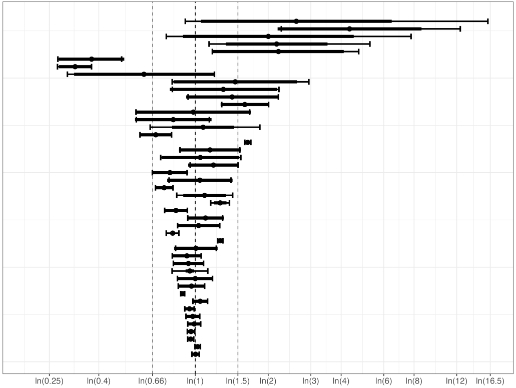
Equivalence Testing Performance

I re-examine these papers with my equivalence-based RV manipulation test, using a lenient threshold of $\epsilon = 1.5$ Why?

- I.e., each test asks: *Can we significantly bound RV manipulation at the cutoff beneath a 50% upward jump/33.3% downward jump?*
- Given the caliber of journals, these RVs should 'pass' this lenient equivalence test

I then compute **equivalence testing failure rates** – the proportion of these equivalence tests that are *not* significant at a 5% level

Main Equivalence Testing Failure Rate Estimates

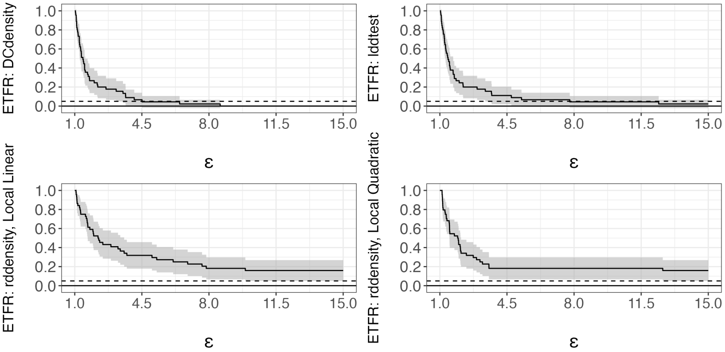


Logarithmic RV Density Discontinuity at the Cutoff

Failure rates for my equivalence-based RV manipulation test range from 44-75%

- **Interpretation:** Over 44% of RV density discontinuity magnitudes at the cutoff can't be significantly bounded beneath a 50% upward jump

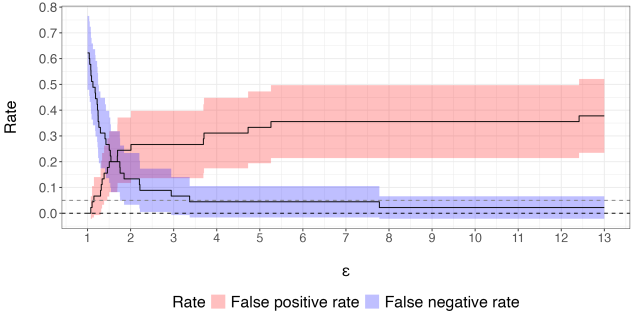
Failure Curves



To obtain equivalence testing failure rates beneath 5%, we'd have to be willing to argue that a 350% upward density jump is practically equal to zero

- **Takeaway:** Meaningful RV manipulation at the cutoff is still a serious problem in RDD research

Confusion Curves



17.7% of LDD estimates at cutoff are **false positives**: Stat. sig., but sig. bounded within $\epsilon \in [2/3, 3/2]$

- Likewise, 26.6% of LDD estimates at the cutoff are **false negatives**: Not stat. sig., but not sig. bounded within $\epsilon \in [2/3, 3/2]$

Takeaway: Standard NHST often misclassifies the practical significance of RV manipulation at cutoff

RDD: Practical Considerations

I recommend setting ϵ by **surveying other researchers for their judgments** of the smallest practically/economically significant RV density jump at the cutoff

- Practical using online resources such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019)
- Data from these researcher surveys can be useful for reasons beyond this test

If this is not feasible (or an RV fails my manipulation test), consider the `rdbounds` partial identification procedure (Gerard, Rokkanen, & Rothe 2020)

- But for most projects, the procedures I'm proposing will be feasible

With the ϵ threshold in hand, you can use my `lddtest` Stata command or the `lddtest` command in my `eqtesting` R package

- Both of my packages are available at github.com/jack-fitzgerald
- Leo Stimpfle has also created a Python version, available at github.com/leostimpfle/lddtest

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

Cattaneo, M. D., M. Jansson, and X. Ma (2018, Mar).
Manipulation testing based on density discontinuity.
The Stata Journal 18(1), 234–261.

Cattaneo, M. D., M. Jansson, and X. Ma (2020, Sep).
Simple local polynomial density estimators.
Journal of the American Statistical Association 115(531), 1449–1455.

Chen, H., P. Cohen, and S. Chen (2010, Apr).
How big is a big odds ratio? interpreting the magnitudes of odds ratios in epidemiological studies.
Communications in Statistics - Simulation and Computation 39(4), 860–864.

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.

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

References IV

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.

Garicano, L., C. Lelarge, and J. Van Reenen (2016).
Firm size distortions and the productivity distribution: Evidence from france.
American Economic Review 106(11), 3439–3479.

References V

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.

Hartman, E. (2021, Oct).
Equivalence testing for regression discontinuity designs.
Political Analysis 29(4), 505–521.

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.

Huntington-Klein, N. (2022, Aug).
The Effect: An introduction to research design and causality (1 ed.).
CRC Press.

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).

References VII

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.

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.

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

McCrary, J. (2008, Feb).
Manipulation of the running variable in the regression discontinuity design: A density test.
Journal of Econometrics 142(2), 698–714.

References VIII

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.

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.

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	D has no effect on Y.	111	39.8%
2	Claim that a relationship/phenomenon does not exist or is negligible, qualified by reference to statistical significance	D has no significant effect on Y.	33	11.8%
3	Claim that a relationship/phenomenon does not exist or is negligible, qualified by reference to something other than statistical significance	D has no meaningful effect on Y.	24	8.6%
4	Claim that a relationship/phenomenon does not (meaningfully) hold in a given direction	D has no positive effect on Y.	53	19%
5	Claim that a relationship/phenomenon does not (meaningfully) hold in a given direction, qualified by reference to statistical significance	D has no significant positive effect on Y.	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	D has no meaningful positive effect on Y.	5	1.8%
7	Claim that there is a lack of evidence for a (meaningful) relationship/phenomenon	There is no evidence that D has an effect on Y.	10	3.6%
8	Claim that a variable holds similar values regardless of the values of another variable	Y 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 D on Y is concentrated in older respondents.	22	7.9%
10	Claim that a relationship/phenomenon stabilizes for some values of another variable	D has a short term effect on Y that dissipates after Z 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%

This Happens All the Time

The TOST Procedure

First, compute test statistics

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

The relevant test statistic is the smaller of the two:

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

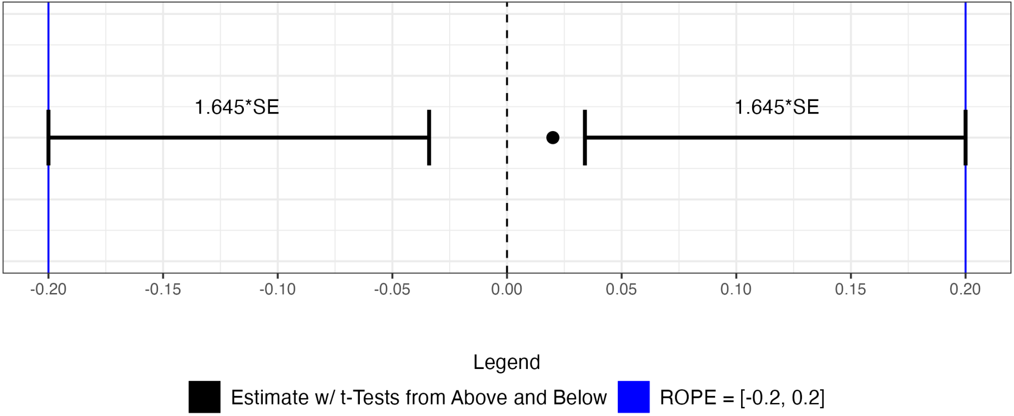
The critical value for a size-α TOST procedure is the **one-sided** critical value t^*_α

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



TOST Concept

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**:

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

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)**:

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

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

- PCCs are widely-used in economic meta-analyses

Failure Rates Introduction

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 & Shaprio (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 Rates

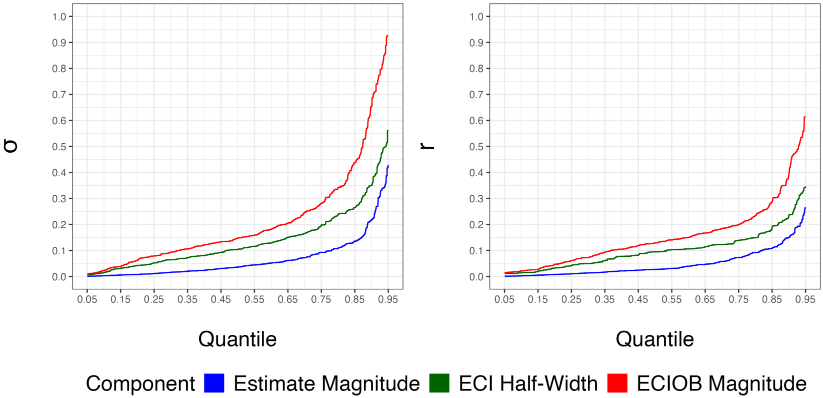
Failure Rate Robustness

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

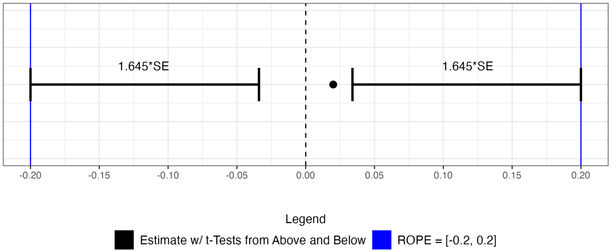
Mechanisms



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

Main Results

Two One-Sided Tests (TOST) Procedure



In other words, we have stat. sig. evidence at the 5% level that $\theta \approx 0$ if

1. $\hat{\theta}$ is 1.645 SEs above $-\ln(\epsilon)$, **and**
2. $\hat{\theta}$ is 1.645 SEs below $\ln(\epsilon)$

Step 3

Why $\epsilon = 1.5$?

- Chen, Cohen, & Chen (2010) show that an odds ratio of 1.5 corresponds closely w/ a Cohen's (1988) $d = 0.2$, the classic small effect size benchmark
- Same effect size proposed by Hartman (2021)
- Practically large in many research-relevant RDD settings (e.g., elections)

Back