The Need for Equivalence Testing in Economics

Jack Fitzgerald, Vrije Universiteit Amsterdam
and Tinbergen Institute*

December 8, 2025

Most Recent Version

Abstract

Equivalence testing is an interval testing approach that can provide statistically significant evidence that economic relationships are bounded beneath practically negligible effect sizes. I demonstrate its necessity in a large-scale robustness replication of estimates defending 135 null claims made in 81 recent articles from top economics journals. 36-63% of estimates defending the average null claim fail lenient equivalence tests. In a prediction platform survey, researchers accurately predict that equivalence testing failure rates will significantly exceed levels which they deem acceptable. Obtaining equivalence testing failure rates that these researchers deem acceptable requires arguing that nearly 75% of published estimates in economics are practically equal to zero. These results imply that Type II error rates are unacceptably high throughout economics, and that many null findings in economics reflect low power rather than truly negligible relationships. I provide economists with guidelines and commands in Stata and R for conducting credible equivalence testing and practical significance testing in future research.

*Email: j.f.fitzgerald@vu.nl. I am grateful to Sanchayan Banerjee, Abel Brodeur, Katharina Brütt, Eve Ernst, Jelle Goeman, Yi He, Florian Heine, Guido Imbens, Peder Isager, Nick Koning, Stan Koobs, Andre Lucas, Derek Mikola, Devin Pope, David Romer, Jonathan Roth, Martin Schumann, and Arjen van Witteloostuijn for valuable input on this paper, as well as several authors who answered my questions about their research and replication data. I also thank conference and seminar participants from the European Association of Young Economists Annual Meeting, European Commission CC-ME COMPIE Conference, International Behavioural Public Policy Conference, KVS New Paper Sessions, Leibniz Open Science Day, Lindau Nobel Laureate Meetings, MAER-Net Colloquium, Metascience Conference, Nederlandse Economendag, Netherlands Reproducibility Network/Platform for Young Meta-Scientists Symposium, ODISSEI Conference for Social Sciences, PhD-EVS Seminar, RWI Essen, Technische Universiteit Eindhoven, Tinbergen Institute, University of Innsbruck, and Vrije Universiteit Amsterdam for comments and feedback. All errors are my own. I am grateful to the Amsterdam Law and Behavior Institute for financial support. This paper was initially developed when I was serving a 12-month term as a member of the Superforecaster Panel for the Social Science Prediction Platform (SSPP; see DellaVigna, Pope, & Vivalt 2019). The views expressed in this paper do not necessarily represent the views of the SSPP, nor of the researchers who created and/or operate the SSPP. This research has Ethical Review Board approval from the School of Business and Economics at Vrije Universiteit Amsterdam. The Online Appendix to this paper is available here.

1 Introduction

An economist runs a regression to estimate the relationship between two variables. As it turns out, the regression estimate is not statistically significantly different from zero. Assuming that this result is not 'shoved in the file drawer', how would most economists report this finding? I show that nearly 70% of article abstracts in top economics journals report such results by claiming that the two variables have no meaningful relationship at all. Readers and researchers also interpret such findings in this way (McShane & Gal 2016; McShane & Gal 2017). However, inferring that statistically insignificant results are evidence of null relationships is widely-recognized as bad scientific practice, because under standard null hypothesis significance testing, a statistically insignificant estimate may reflect a large relationship whose estimate is simply noisy and imprecise (see Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016).

This paper introduces economists to a testing framework that is more appropriate for evidencing null relationships, known as 'equivalence testing'. In this framework, the researcher first sets a 'region of practical equivalence (ROPE)' around zero, representing the range of values for the relationship of interest that are 'practically equal to zero', or in economic parlance, 'economically insignificant'. The equivalence testing framework assumes in the null hypothesis that the estimate is bounded *outside* of the ROPE. If the estimate is significantly bounded *inside* the ROPE, then there is credible evidence that the relationship of interest is practically equal to zero. Equivalence testing is commonly used in medicine, and is being rapidly adopted in psychology and political science (see Piaggio et al. 2012; Hartman & Hidalgo 2018; Lakens, Scheel, & Isager 2018). This paper shows empirically that economists must also adopt equivalence testing to ensure that null findings are robust, and demonstrates how to credibly apply this testing framework.

In a large-scale re-analysis, I show that the standard testing procedures that economists use to make and defend null claims likely tolerate high Type II error

rates. Using publicly available replication data, I re-estimate and standardize the estimates defending 135 null claims made in the abstracts of 81 articles published in Top 5 economics journals from 2020-2023. I also survey 62 researchers on the Social Science Prediction Platform to obtain their judgments and predictions on equivalence testing results in my replication sample (see DellaVigna, Pope, & Vivalt 2019).

To assess how estimates in my replication sample perform under equivalence testing, I set ROPEs with boundaries based on Cohen's (1988) widely-used small effect size benchmarks. These benchmarks are larger than many published estimates in economics, so these ROPEs are quite lenient (see Doucouliagos 2011). One should expect that estimates defending null claims in top economics journals are significantly bounded within these ROPEs, and thus 'pass' lenient equivalence tests. Across the 135 null claims in my sample, I estimate 'equivalence testing failure rates (ETFRs)' by computing the average proportion of estimates defending each claim that 'fail' these lenient tests.

Even for these lenient ROPEs, 36-63% of estimates defending the average null claim fail equivalence tests. To obtain ETFRs that my prediction platform sample considers acceptable, one must claim that nearly 75% of all published effect sizes in economics are practically equal to zero. These results imply that null claims in top economics journals exhibit unacceptably high error rates.

My prediction platform data shows that researchers actually *expect* ETFRs to be unacceptably high. The median researcher considers ETFRs of 10.65-12.95% to be acceptable, but predicts ETFRs from 35.1-38.35%. These predictions are quite accurate, roughly aligning with the lower bound of my ETFR estimates. On average, researchers expect ETFRs to exceed acceptable levels by around 23 percentage points. Though researchers distrust many null results achieved using standard testing procedures, this mistrust appears to be relatively well-placed. These results imply that equivalence testing is a necessary addition to empirical practice.

Given the clear need for equivalence testing in economics, I provide guidelines for

conducting credible equivalence testing in economic research. I offer several credible methods for setting ROPEs, such as pre-registered cost-benefit analyses and surveys that elicit judgments on effect sizes from experts or relevant stakeholders. I also discuss the 'three-sided testing (TST)' procedure, a general framework for testing an estimate's practical significance (Goeman, Solari, & Stijnen 2010).

An estimate may be too imprecise to be reliably classified as either significant *or* practically equal to zero. In such cases, the testing frameworks that I discuss in this paper require researchers to concede that their results are 'inconclusive'. This ensures that imprecise estimates are not misinterpreted as evidence of null relationships.

I also provide the `tsti` command in Stata and the `tst` command in the `eqtesting` R package, which compute immediate testing results under the TST framework for a given estimate, standard error, and ROPE. Because standard equivalence testing procedures are nested in the TST framework, both `tsti` and `tst` can in principle be used exclusively for equivalence testing. The `tsti` command is available on SSC, and the `eqtesting` package is available on CRAN.

Equivalence testing is new to economists, which can create easily-addressable misunderstandings. In Online Appendix A, I thus answer 'frequently asked questions' about equivalence testing, null results in economics, and this paper's empirics.

Section 2 of this paper details the replication data underlying my empirical analysis. In Section 3, I use this data to document problems with current practice for evidencing null claims in economics. Section 4 introduces equivalence testing procedures that address these issues. Section 5 provides methodological details for my empirical analysis, and Section 6 details my results. Section 7 provides guidelines and extensions for credible equivalence testing and practical significance testing in future research. Section 8 concludes.

2 Data

I obtain a systematically-selected sample of 2346 estimates defending 279 null claims made in the abstracts of 158 articles published from 2020-2023 in Top 5 economics journals (i.e., *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*).[1] The systematic selection procedure is detailed in Online Appendix B. All null claims selected for this sample are likely to be interpreted by readers as claims of negligible or nonexistent relationships or phenomena (see McShane & Gal 2016; McShane & Gal 2017). I refer to this first sample of articles, claims, and estimates as the 'intermediate sample'.

The 'final sample' contains all conformable estimates in the intermediate sample that can be re-estimated using publicly available data.[2] The final sample comprises 876 estimates that defend 135 null claims made in the abstracts of 81 articles. Claims are typically defended by more than one estimate due to different model specifications or robustness checks. The final sample stores the standardized regression coefficient σ and standard error s for each estimate, which are standardized such that point estimates and standard errors can be interpreted on the scale of 'standard deviation effects' (see Section 5.1 for details). For each estimate, I also store sample size N, residual degrees of freedom df,[3] reproducibility status, conformability status, outcome and exposure variables (with dummies indicating if each is binary), and the initial p-value under the standard null hypothesis significance testing (NHST) framework (without conformability changes, if applicable). In Online Appendix C, I provide the articles represented in the final sample, alongside additional data repositories attached to these articles (when applicable). In Online Appendix D, I provide the articles in

[1] This includes articles that were not yet published in print, but were digitally published as corrected proofs, before the search date; see Online Appendix B for further details.

[2] For the purposes of this paper, 'publicly available' data includes data stored in repositories of the Inter-university Consortium for Political Science Research (ICPSR), where data is freely available to most researchers who create an ICPSR account.

[3] When df is not directly provided by software output, I impute $df = N - b$, where b is the number of covariates plus one (for a constant term). This imputation is conservative for the purposes of this paper, if anything deflating ETFRs for partial correlation coefficients (see Sections 5.1 and 5.2).

	Min	P10	P25	P50	P75	P90	Max	Mean	SD	N
Panel A: Article-Level										
# of Claims, Intermediate Sample	1	1	1	1	2	3	11	1.766	1.369	158
# of Estimates, Intermediate Sample	1	1	3	6	14	28.3	288	14.848	32.197	158
# of Claims, Final Sample	1	1	1	1	2	3	5	1.667	1.025	81
# of Estimates, Final Sample	1	1	3	6	14	24	82	10.815	13.145	81
Panel B: Claim-Level										
# of Estimates, Intermediate Sample	1	1	2	4	8	16	288	8.409	22.372	279
# of Estimates, Final Sample	1	1	2	4	7.5	14.6	55	6.489	8.128	135
Panel C: Estimate-Level										
σ	-1.671	-0.12	-0.026	0.004	0.044	0.118	1.817	0.001	0.201	876
$\lvert\sigma\rvert$	0	0.004	0.013	0.036	0.102	0.244	1.817	0.096	0.176	876
s	0	0.012	0.027	0.068	0.13	0.208	5.783	0.107	0.259	876
Initial NHST p-value	0	0.054	0.231	0.484	0.739	0.899	1	0.482	0.302	876
N	12	171	616	3558	14606	197768	12353303	92508.845	629132.708	876
df	10	36.5	91	180	1045	11104	1076398	6356.906	51866.319	876
Power to detect $\lvert\sigma\rvert = 0.2$	0.031	0.157	0.33	0.829	1	1	1	0.685	0.341	876

Note: This table reports summary statistics aggregated at each clustering level of the data. All data at the estimate level arises from the final sample.

Table 1: Summary Statistics

the intermediate sample that are excluded from the final sample.

Table 1 displays summary statistics. The majority of articles in my data make only one null claim, and over 90% make between one and three null claims. The median null claim in my data is defended by four estimates. Effect sizes throughout the final sample are quite small, with the median standardized coefficient magnitude at 0.036σ. The median estimate in the final sample arises from a model with $N = 3558$ and $df = 180$.[4] At a 5% significance level, there is at least 80% power to detect an effect size of 0.2σ under standard NHST for the majority of estimates in the final sample. However, there is a concentrated group of underpowered estimates. For 32% of estimates in the final sample, there is not even 50% power to detect a 0.2σ effect under standard NHST. Over 90% of estimates in the final sample are statistically insignificant under the standard NHST framework at a 5% significance level. The 10% of estimates that are initially statistically significant nearly always arise alongside other statistically insignificant estimates that collectively defend their null claim.[5]

I omit summary statistics on several binary variables from Table 1. 8.3% of es-

[4]This large difference between N and df arises largely due to clustering. When standard errors are clustered, df is constrained by the number of clusters rather than the number of observations.

[5]One claim – the only null claim in its article – is defended with a single statistically significant result (Fuster, Kaplan, & Zafar 2021).

timates in the final sample are not fully reproducible; i.e., my best attempts to reproduce the article's findings using its replication repository do not yield the exact same results as those published in the article. Further, 7.9% of estimates in the final sample arise from models that are adjusted with conformability modifications for my analysis; i.e., the model used to obtain the estimate in the final sample differs from the model used to produce the estimate in the published article.[6] 22.9% of estimates in the final sample correspond to outcome and exposure variables that are both continuous, whereas 25.5% correspond to outcome and exposure variables that are both binary. The most common type of estimate corresponds to a continuous outcome variable and a binary exposure variable, representing 35.7% of estimates in the final sample.

3 Null Claims in Economics: Theory and Practice

In practice, economists usually estimate relationships using linear models of the form $Y = \delta D + X\phi + \mu$, where Y is the outcome variable of interest, D is the exposure variable of interest, and X is a matrix of b other covariates, which typically includes a constant term. The parameter of interest is δ, the linear association between Y and D. Point estimate $\hat{\delta}$ and standard error $s > 0$ can be estimated in a regression model whose residual μ exhibits df degrees of freedom. When economists are interested in testing whether there is a relationship between Y and D, they predominantly apply a two-tailed test to $\hat{\delta}$ under the standard NHST framework (Imbens 2021).

Definition 3.1 (The Standard Null Hypothesis Significance Testing Framework)**.** *The researcher wants to assess whether* $\delta \neq 0$ *using a test with Type I error rate*

[6]For example, average marginal effects must be estimated for a probit or logit estimate to be appropriately interpreted as a linear relationship with the outcome variable.

$\alpha \in (0, 1)$. *They formulate null and alternative hypotheses as*

$$H_0 : \delta = 0 \qquad\qquad H_A : \delta \neq 0 \tag{1}$$

and compute test statistic $t_{NHST} = \hat{\delta}/s$. *Let* $F(t,\ df)$ *be the cumulative density function of the t-distribution with df degrees of freedom. The exact critical value is*

$$t^*_{\alpha/2,\ df} = F^{-1}\left(1 - \frac{\alpha}{2},\ df\right). \tag{2}$$

The researcher rejects H_0 *and concludes that* $\delta \neq 0$ *if and only if* $\hat{\delta}$ *is statistically significant, where* $\hat{\delta}$ *is statistically significant if and only if* $|t_{NHST}| \geq t^*_{\alpha/2,\ df}$.

Economists using the standard NHST framework typically conclude that there is a relationship between Y and D if H_0 is rejected, and that there is no relationship between Y and D if H_0 is not rejected (Romer 2020; Imbens 2021). Table 2 details how economists make null claims when H_0 is not rejected. Specifically, I use a slightly modified version of the categorization from Gates & Ealing's (2019) survey of null claims in medical journals to classify all null claims in my intermediate sample.[7] I additionally divide claims into four categories: main, mechanism, robustness, and subgroup. First, main claims are among the primary 'headline' conclusions of the paper. Second, null mechanism claims rule out one or more potential mechanisms for the main claim(s). Third, null robustness claims highlight robustness checks such as placebo tests which provide evidence against the prospect that confounding drives the main result(s). Fourth and finally, null subgroup claims emphasize that meaningful statistical relationships are not observed in certain subsets of the data.

Table 2 shows that most null claims being made in articles from top economics journals are main claims of the article, rather than secondary claims. 162 of the 279 claims in my intermediate sample (58%) are main claims. Mechanism claims,

[7]No claim in the intermediate sample would fall into categories 9 or 10 in Gates & Ealing (2019); categories 9 and 10 in Table 2 serve as replacements. I also adjust the wording of claim types.

Claim Category	Claim Type	Claim Example	# Total Claims	% of Total Claims	# Main Claims	% of Main Claims	# Mechanism Claims	% of Mechanism Claims	# Robustness Claims	% of Robustness Claims	# Subgroup Claims	% of Subgroup Claims
1	Claim that a relationship/phenomenon does not exist or is negligible	D has no effect on Y.	105	37.6%	55	34%	22	51.2%	18	50%	10	26.3%
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.	35	12.5%	24	14.8%	4	9.3%	2	5.6%	5	13.2%
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.	29	10.4%	21	13%	4	9.3%	3	8.3%	1	2.6%
4	Claim that a relationship/phenomenon does not (meaningfully) hold in a given direction	D has no positive effect on Y.	50	17.9%	33	20.4%	0	0%	7	19.4%	10	26.3%
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%	2	1.2%	1	2.3%	1	2.8%	0	0%
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.	7	2.5%	4	2.5%	2	4.7%	0	0%	1	2.6%
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%	7	4.3%	1	2.3%	0	0%	2	5.3%
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%	3	1.9%	1	2.3%	3	8.3%	0	0%
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%	3	1.9%	8	18.6%	2	5.6%	9	23.7%
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%	10	6.2%	0	0%	0	0%	0	0%
	Unqualified null claim	Categories 1, 4, or 8-10	194	69.5%	104	64.2%	31	72.1%	30	83.3%	29	76.3%
	Qualified null claim	Categories 2-3 or 5-7	85	30.5%	58	35.8%	12	27.9%	6	16.7%	9	23.7%
	Total		279	100%	162	100%	43	100%	36	100%	38	100%

Note: Data is based on the 158 articles and 279 null claims in the intermediate sample (see Section 2).

Table 2: Types of Null Claims in the Economics Literature

robustness claims, and subgroup claims each represent just 13-15% of the intermediate sample. This implies that if there are issues with null claims in economics, then this will often implicate the primary claims of influential articles.[8]

Table 2 provides both good news and bad news concerning economists' reporting of null claims. The good news is that over 30% of null claims in the intermediate sample are qualified with references to statistical significance, estimate magnitude, or a lack of evidence. Thus in recent years, some null claims in economics are being reported with appropriate attention paid to effect sizes and power. The bad news is that these well-nuanced null claims are in the minority. Nearly 70% of null claims in the intermediate sample make no reference to statistical significance, effect sizes, or a lack of evidence, and are in this sense 'unqualified'. These unqualified null claims are unambiguous assertions that the relationship of interest is negligible or nonexistent. Unqualified null claims are only slightly less frequent amongst main claims, representing 64% of main claims. Accordingly, the proportion of null claims that are unqualified is higher when those claims are secondary. 72% of null claims about mechanisms, 76% of those about subgroup analyses, and 83% of those about robustness checks are unqualified.

Of course, if $\hat{\delta}$ is statistically insignificant, this does not necessarily imply that δ is negligibly small. A statistically insignificant result could simply reflect imprecision due to low power. As s grows arbitrarily large, any arbitrarily large $\hat{\delta}$ may be 'insignificant' under the standard NHST framework. Therefore, generally inferring null results from statistically insignificant estimates can often result in erroneously concluding that genuinely meaningful relationships do not exist, among other negative consequences.

To formalize these intuitions, the standard NHST framework can produce Type I and Type II errors. Type I errors occur when one incorrectly rejects the null hypothesis that $\delta = 0$, whereas Type II errors occur when one fails to reject that hypothesis when one should. Type I error rates are largely controlled by the significance level α, which is conventionally set at 0.05. Type II error rate $\beta \in (0, 1)$ relates to the power $(1 - \beta)$

[8]Online Appendix Table A8 provides direct empirical evidence of this.

with which a relationship of magnitude $\epsilon > 0$ can be detected using standard NHST. As the complement of the standard NHST Type II error rate for effect size ϵ, $(1 - \beta)$ represents the probability that $\hat{\delta}$ is statistically significant under the standard NHST framework if $\left|\hat{\delta}\right| \geq \epsilon$, which depends on the precision of $\hat{\delta}$ (Ioannidis, Stanley, & Doucouliagos 2017). In principle, if published estimates in economics are sufficiently powered to detect reasonably small ϵ values, then statistically insignificant results in the economics literature usually reflect true nulls, and there is no need to change current testing practices in economics.

Unfortunately, the risk of Type II errors is quite high in economics, as power is usually remarkably low throughout the economics literature. As discussed in Section 2, there is not even 50% power to detect a 0.2σ effect under standard NHST for 32% of estimates in my final sample. Ioannidis, Stanley, & Doucouliagos (2017) estimate median power to detect true effects in the economics literature at 18% or less. Askarov et al. (2023) obtain median power estimates of 7% in leading economics journals.

These low power levels are not necessarily due to poor research practices, and can naturally arise from inherent constraints of economic research. Answering important economic questions frequently requires researchers to work with pre-existing datasets. Economists are thus frequently 'at the mercy' of existing sample sizes, and often cannot summon new data at will to improve power.

This low power challenges the credibility of null claims in economics. When a researcher trying to show that $\delta = 0$ uses the standard NHST framework in Definition 3.1, they begin by assuming in the null hypothesis that what they want to show is true – that $\delta = 0$ – and only conclude otherwise if the estimate is statistically significant. This shifts the burden of proof off of the researcher. Thus for researchers trying to show that $\delta = 0$, imprecision is 'good', as the probability of finding a statistically insignificant result is inversely related to statistical precision. This dynamic drives 'reverse p-hacking', a common practice in placebo tests where null results are desirable (Dreber, Johannesson, & Yang 2024).

Because researchers using the standard NHST framework to show that $\delta = 0$ face no effective burden of proof, generally concluding that statistically insignificant results are null results is a logical fallacy (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016). Formally, researchers who make this inference engage in 'appeals to ignorance', which arise when one infers that a claim is correct simply because no one has yet produced significant evidence against the claim. Though null relationships can sometimes be inferred from statistically insignificant results, this inference is only valid for sufficiently well-powered results. Generally inferring null relationships from statistically insignificant estimates without any regard to the Type II error control implied by those estimates' statistical power can result in researchers unwittingly tolerating high Type II error rates. The low power documented in both my replication data and reviews of the economic literature, combined with the high frequency of unqualified null claims documented in Table 2, therefore imply that economists often effectively tolerate large Type II error rates.

Because estimates may be statistically insignificant either because they are negligible or because they are noisy, standard testing procedures offer researchers no framework through which null results can be distinguished from imprecise results, which contributes to many problems in the economics literature. Researchers associate null results so strongly with imprecision that even when precision measures are held constant, researchers perceive null results as being more imprecisely-estimated than statistically significant results (Chopra et al. 2024). Researchers also assign strong 'null result penalties', viewing null results as low-quality and unpublishable (see McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024). Partly due to this uncertainty around null results' credibility, even though statistically insignificant results can often be more informative (i.e., prior-shifting) than statistically significant results (Abadie 2020), null findings remain far less likely to be published in economics journals (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019). Even amongst published null findings, the high Type II error rates effectively toler-

ated by current economic practice imply that many null findings in economics are likely 'false negatives' that wrongfully declare meaningful economic relationships to be nonexistent (Ioannidis, Stanley, & Doucouliagos 2017; Askarov et al. 2024).

Testing frameworks that provide better error control for null results can mitigate these problems. If researchers understand these dynamics in the current research landscape, then biases against null results may partially arise from researchers associating null results with imprecise results (see Chopra et al. 2024). Therefore, testing frameworks that can credibly distinguish precise nulls from imprecise estimates may yield the added benefit of mitigating null result penalties, and could in turn reduce publication bias against null results.

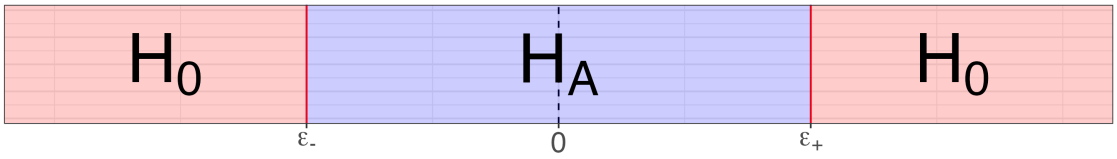
4 Equivalence Testing

One can construct a credible framework for testing whether relationships are practically null by adjusting the standard NHST framework in two ways. First, flipping the null and alternative hypotheses in Equation 1 restores the burden of proof for researchers trying to show that $\delta = 0$. Second, to make the test feasible, the constraints in Equation 1 can be relaxed. Rather than assessing whether $\delta = 0$ strictly, one can instead assess whether $\delta \approx 0$. The hypotheses then take the form

$$H_0 : \delta \not\approx 0 \qquad\qquad H_A : \delta \approx 0. \tag{3}$$

This is a feasible hypothesis test if one can define a range of values within which $\delta \approx 0$, as one can test whether $\hat{\delta}$ is significantly bounded within that range using a simple interval test. This is the core idea of equivalence testing.

Definition 4.1 (The Equivalence Testing Framework)**.** *The researcher wants to test whether* $\delta \approx 0$*. Let* $[\epsilon_-, \epsilon_+]$ *be a range where* $\epsilon_- < \epsilon_+$*, where* $0 \in [\epsilon_-, \epsilon_+]$*, and where* $\delta \approx$



Note: $[\epsilon_-, \epsilon_+]$ is the ROPE for these hypotheses.

Figure 1: Visualization of the Equivalence Testing Hypothesis Framework

0 *when* $\delta \in [\epsilon_-, \epsilon_+]$. *The researcher thus formulates null and alternative hypotheses:*

$$H_0 : \delta \notin [\epsilon_-, \epsilon_+] \qquad\qquad H_A : \delta \in [\epsilon_-, \epsilon_+]\,. \tag{4}$$

The researcher rejects H_0*, concluding that* $\delta \approx 0$*, if and only if* $\hat{\delta}$ *is statistically significantly bounded within* $[\epsilon_-, \epsilon_+]$.

Figure 1 visualizes the equivalence testing hypothesis framework. $[\epsilon_-, \epsilon_+]$ is the 'region of practical equivalence (ROPE)', which is the range of δ values that are 'practically equal to zero' or 'economically insignificant'. As I detail in Section 7.1, these boundaries can be credibly set in multiple ways, such as through surveying other researchers for independent effect size judgments or through pre-registered cost-benefit analysis. ROPEs are often set symmetrically around zero such that $\epsilon_- = -\epsilon_+$, but ROPEs can also be asymmetric. Symmetric ROPEs can be described as having a 'length' of $\epsilon > 0$ and written as $[-\epsilon, \epsilon]$. I discuss several tests that assess whether $\hat{\delta}$ is statistically significantly bounded within the ROPE throughout the rest of this section.

4.1 The Two One-Sided Tests Procedure

The hypotheses in Equation 4 can be rewritten as

$$H_0 : \delta < \epsilon_- \quad \text{or} \quad \delta > \epsilon_+ \qquad\qquad H_A : \delta \geq \epsilon_- \quad \text{and} \quad \delta \leq \epsilon_+. \tag{5}$$

These joint hypotheses can be assessed using two one-sided tests:

$$\begin{aligned} H_0 &: \delta < \epsilon_- & \qquad H_0 &: \delta > \epsilon_+ \\ H_A &: \delta \geq \epsilon_- & \qquad H_A &: \delta \leq \epsilon_+. \end{aligned} \tag{6}$$

Under Definition 4.1, significant evidence that $\delta \approx 0$ can be obtained by showing significant evidence against both H_0 statements in Equation 6. This is the idea behind the 'two one-sided tests (TOST)' procedure.

Definition 4.2 (The Two One-Sided Tests Procedure)**.** *The researcher wants to test the hypotheses in Definition 4.1 using a size-α test. They thus formulate test statistics*

$$t_- = \frac{\hat{\delta} - \epsilon_-}{s} \qquad\qquad t_+ = \frac{\hat{\delta} - \epsilon_+}{s} \tag{7}$$

and compute

$$t_{TOST} = \underset{t \in \{t_-, t_+\}}{\arg\min} \{|t|\}. \tag{8}$$

Let t^*_α be defined as in Equation 2. If $t_{TOST} = t_-$, then the researcher concludes that $\hat{\delta}$ is statistically significantly bounded within $[\epsilon_-, \epsilon_+]$ if and only if $t_{TOST} \geq t^*_{\alpha,\ df}$. If $t_{TOST} = t_+$, then the researcher concludes that $\hat{\delta}$ is statistically significantly bounded within $[\epsilon_-, \epsilon_+]$ if and only if $t_{TOST} \leq -t^*_{\alpha,\ df}$.

Put simply, at a 5% significance level, the TOST procedure (asymptotically) deems $\hat{\delta}$ to be significantly bounded within a ROPE if it is both 1.645 standard errors *above* the ROPE's *lower* bound and 1.645 standard errors *below* the ROPE's *upper* bound. The TOST procedure's size is preserved at nominal level α despite the use of two simultaneous tests because the relevant test statistic is the smaller of its two t-statistics. The TOST procedure is thus an intersection-union test of two level-α tests (Berger & Hsu 1996; Lakens, Scheel, & Isager 2018). This procedure is established by Schuirmann (1987), who shows that the TOST procedure has more power to evidence

null relationships than traditional 'power approaches'; for further details on this latter point, see Section 4.3.

4.2 Equivalence Confidence Intervals

At a significance level of α, the TOST procedure can be inverted using an identical confidence interval-based approach that makes use of the symmetric $(1 - 2\alpha)$ confidence interval (Berger & Hsu 1996). Following Hartman & Hidalgo (2018), I refer to this interval as the 'equivalence confidence interval (ECI)'.

Definition 4.3 (The Equivalence Confidence Interval Approach)**.** *A researcher wants to test the hypotheses in Definition 4.1 using a size-α test. They formulate real interval*

$$\left[\Delta_-, \Delta_+\right] = \left[\hat{\delta} - \left(s \times t^*_{\alpha,\ df}\right), \hat{\delta} + \left(s \times t^*_{\alpha,\ df}\right)\right], \tag{9}$$

where $t^*_{\alpha,\ df}$ is defined as in Equation 2. The researcher concludes that $\hat{\delta}$ is statistically significantly bounded within $[\epsilon_-, \epsilon_+]$ if and only if $[\Delta_-, \Delta_+] \subset [\epsilon_-, \epsilon_+]$.

Because the $(1 - \alpha)$ ECI is the $(1 - 2\alpha)$ confidence interval, computing ECIs is simple. For example, the 95% ECI is just the 90% confidence interval. However, ECIs and confidence intervals are used to judge statistical significance in different ways. In the standard NHST framework, statistical significance judgments can be made based on the confidence interval's relationship with zero. In contrast, significance judgments in equivalence testing can be made based on the ECI's relationship with the ROPE. An estimate is significantly bounded within the ROPE at significance level α if and only if the $(1 - \alpha)$ ECI of that estimate is entirely bounded within the ROPE. This decision rule yields identical conclusions to the TOST procedure.

Figure 2 shows an example of an exact 95% ECI and its uses. In this example, $\hat{\delta} = 0.02$, $s = 0.1$, and $df = 100$. The 95% ECI of this estimate can be roughly written as $[-0.146, 0.186]$, because $t^*_{0.05,100} \approx 1.66$. As in Section 4.1, if the ROPE is set as $[-0.2, 0.2]$, then $\hat{\delta}$ is significantly bounded within the ROPE at a 5% significance

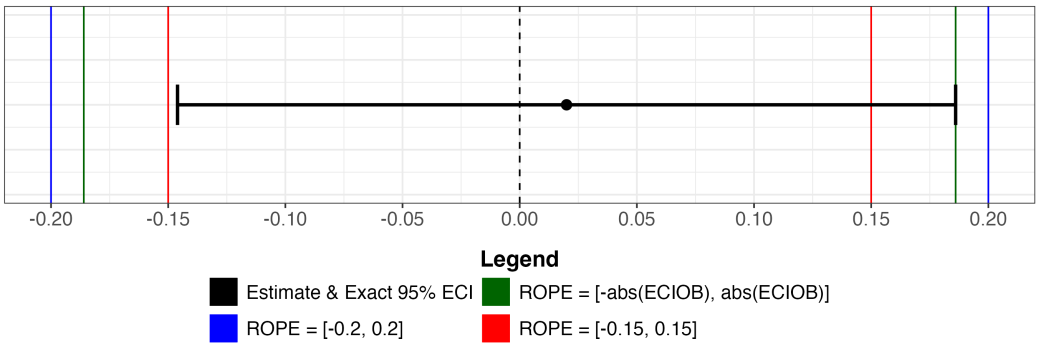
level. This is because the entire 95% ECI is bounded within this ROPE. However, the same conclusion cannot be reached if the ROPE is instead specified as $[-0.15, 0.15]$. $\hat{\delta}$'s $(1-\alpha)$ ECI is the smallest ROPE wherein one can significantly bound δ at a significance level of α.

The 'ECI outer bound (ECIOB)' of an estimate is the bound of that estimate's ECI that is furthest from zero. For example, the ECIOB of the estimate in Figure 2 is 0.186, because the upper bound of this estimate's ECI is further from zero than the lower bound. The magnitude of the ECIOB is the length of the smallest symmetric ROPE around zero wherein there is statistically significant evidence that $\delta \approx 0$. Therefore, the ECIOB's magnitude serves as a measure of how closely to zero an estimate can be significantly bounded.

ECIOB magnitudes are interesting for many applied economists, as the ECIOB magnitude is the smallest effect size for an estimate that can be 'ruled out' with statistically significant evidence using equivalence testing. At present, when economists wish to make similar judgments, they typically point to the outer bound of an estimate's $(1-\alpha)$ confidence interval, arguing that larger effect sizes can be 'ruled out' with reasonable certainty. However, an estimate's $(1-\alpha)$ ECIOB magnitude is always smaller than the magnitude of its $(1-\alpha)$ confidence interval's outer bound. This means that focusing on ECIOB magnitudes allows a researcher to credibly report that an estimate is bounded closer to zero than would be possible if that researcher instead focused on the outer bound of the estimate's confidence interval. Using equivalence testing thus offers power improvements over typical inference approaches for evidencing null relationships using confidence intervals.

4.3 Relationship to Statistical Power

A key credibility advantage of equivalence testing over standard NHST is that in equivalence testing, low power/precision is always 'bad' – one can never get closer to conclusive null results by obtaining more underpowered and imprecise estimates. As



Note: The estimate in this figure exhibits $\hat{\delta} = 0.02$, $\text{SE}\left(\hat{\delta}\right) = 0.1$, and $df = 100$, implying that ECIOB $= 0.186$.

Figure 2: An ECI Example

an estimator becomes more well-powered to detect a given effect size, its standard error shrinks, and so too does its ECI. Therefore, under equivalence testing, more power and precision enables researchers to more tightly bound their estimates within a given ROPE. I demonstrate empirical evidence of this property in Section 6.4.

Equivalence testing and 'power approaches' for evidencing null relationships are naturally related through this mechanism. Researchers typically presume that if an estimator $\hat{\delta}$ could detect an effect size ϵ with 80% power, and if $\hat{\delta}$ is not statistically significantly different from zero, then there is sufficient error control to conclude that the estimate is bounded beneath ϵ (Nikiforakis & Slonim 2015; Ioannidis, Stanley, & Doucouliagos 2017). Researchers with statistically insignificant results will thus often point to the minimum detectable effect size for $\hat{\delta}$ at 80% power and a 5% significance level, assert that this minimum detectable effect size is sufficiently small (effectively arguing that it is bounded within δ's ROPE), and consequently conclude that δ is practically equal to zero. As in equivalence testing, under a power approach, estimates with more power to detect a given effect size can be more tightly bounded to zero than estimates with less power.

Though equivalence testing and such power approaches are related, statistically insignificant estimates with 'sufficient' power to detect a given effect size cannot

always be significantly bounded beneath that effect size under equivalence testing. Let the standard error (SE) of $\hat{\delta}$ be written as $\text{SE}\left(\hat{\delta}\right)$, and let ϵ be the minimum detectable effect size of $\hat{\delta}$ with 80% power at a 5% significance level, implying that $\epsilon \approx 2.801 \times \text{SE}\left(\hat{\delta}\right)$ (Bloom 1985). Under TOST, as $df \to \infty$, $\hat{\delta}$ is significantly bounded beneath ϵ if and only if $\left|\hat{\delta}\right| \leq (2.801 - 1.645) \times \text{SE}\left(\hat{\delta}\right) = 1.156 \times \text{SE}\left(\hat{\delta}\right)$. The estimate will not be statistically significantly different from zero if $\left|\hat{\delta}\right| < 1.96$. Thus in this scenario, if $\left|\hat{\delta}\right| \in \left(1.156 \times \text{SE}\left(\hat{\delta}\right), 1.96 \times \text{SE}\left(\hat{\delta}\right)\right)$, then $\hat{\delta}$ will be statistically insignificant with 80% power to detect an effect size of ϵ, but will not be able to be significantly bounded beneath ϵ under equivalence testing.

One reason for this discrepancy is that traditional disciplinary thresholds for Type I and Type II error rates tolerate much more Type II error than Type I error. A statistical significance threshold of 5% and a sufficient power threshold of 80% jointly imply that researchers tolerate Type II errors four times as often as they tolerate Type I errors (Ioannidis, Stanley, & Doucouliagos 2017). By employing traditional statistical significance thresholds, equivalence testing sets the error rate tolerances for null claims and positive claims equal to one another.

Though one can avert this discrepancy by raising the power target to ensure that statistically insignificant results are considered precise nulls under both equivalence testing and the power approach, doing so actually renders the power approach less well-powered than equivalence testing to detect practically negligible relationships. Let $G(t, \mu, df)$ be the cumulative density function of the noncentral t-distribution with noncentrality parameter μ and df degrees of freedom. The minimum detectable effect size at significance level α and power target $(1 - \beta)$ is then $G^{-1}\left(1 - \beta, t^*_{\alpha/2,\ df}, df\right) \times \text{SE}\left(\hat{\theta}\right)$ (see Bloom 1995). To obtain a power level at which statistically insignificant results would always be considered precise nulls under both equivalence testing and the power approach, one would need to change the power target by adjusting β, identifying some target Type II error rate value β^* at which $G^{-1}\left(1 - \beta^*, t^*_{\alpha/2,\ df}, df\right) = t^*_{\alpha/2,\ df} + t^*_{\alpha,\ df}$. This is because the minimum detectable

effect size at power target $(1-\beta^*)$ would need to be large enough that any point estimate whose $(1-\alpha)$ confidence interval intersects zero would also have a $(1-\alpha)$ ECI that is entirely beneath the minimum detectable effect size; $t^*_{\alpha/2,\ df}$ covers the half-width of the confidence interval while $t^*_{\alpha,\ df}$ covers the half-width of the ECI. In contrast, the equivalence testing approach can significantly bound an estimate beneath any $\epsilon \geq t^*_{\alpha,\ df}$ (see Section 4.2). Thus because $t^*_{\alpha,\ df} < t^*_{\alpha/2,\ df} + t^*_{\alpha,\ df}$, any power approach that sufficiently implies that a statistically insignificant relationship is practically negligible will always have less power to detect practically negligible relationships than equivalence testing. Similar properties were noted by Schuirmann (1987) when establishing the TOST procedure.

The combination of a statistically insignificant result and 'sufficient' power to detect a given effect size under standard NHST is therefore neither necessary nor sufficient for demonstrating that an estimate is significantly bounded beneath that effect size under equivalence testing. Ironically, these results imply that compared to equivalence testing, the power approach is underpowered to detect precise nulls.

5 Methods

5.1 Standardization and Effect Sizes

I standardize all regression results obtained in the final sample into two effect size measures. The first is the 'standardized coefficient' σ, calculated along with its standard error s as

$$\sigma = \begin{cases} \frac{\hat{\delta}}{\sigma_Y} \text{ if } D \text{ is binary} \\ \frac{\hat{\delta}\sigma_D}{\sigma_Y} \text{ otherwise} \end{cases} \qquad s = \begin{cases} \frac{\text{SE}(\hat{\delta})}{\sigma_Y} \text{ if } D \text{ is binary} \\ \frac{\text{SE}(\hat{\delta})\sigma_D}{\sigma_Y} \text{ otherwise} \end{cases} . \tag{10}$$

σ_D and σ_Y respectively represent the standard deviations of the exposure and outcome variables of interest within the estimation sample, and $\hat{\delta}$ is the estimated linear

association between Y and D. Standardized coefficients can be interpreted as 'standard deviation effects', and closely relate to the widely-used Cohen's d effect size metric when exposure variables are binary (see Cohen 1988, pg. 20). This standardization ensures that Y has a standard deviation of one. Thus when D is binary, σ is the number of standard deviations of Y associated with a switch of D from zero to one. When D is not binary, σ is the number of standard deviations of Y associated with a one-standard deviation increase in D. The scale of σ does not depend on the size of estimate $\hat{\delta}$, and only depends on the standard deviation of outcome Y (and, if exposure D is non-binary, the standard deviation of D).

The second effect size measure I use is the 'partial correlation coefficient' r, a widely-used effect size measure in meta-analyses. Per Stanley & Doucouliagos (2012), regression coefficients can be sequentially converted first into partial correlations and then into corresponding standard errors as

$$r = \frac{t_{\text{NHST}}}{\sqrt{t^2_{\text{NHST}} + df}} \qquad \text{SE}(r) = \frac{1 - r^2}{\sqrt{df}}. \tag{11}$$

t_{NHST} is the standard NHST t-statistic described in Definition 3.1, where $\hat{\delta} = \sigma$ and s is the standard error of σ.[9] This measure is akin to the 'correlation per df' measure employed in other large-scale re-analyses of replication data (e.g., Open Science Collaboration 2015; Camerer et al. 2016; Camerer et al. 2018).

For my main results, I assess the equivalence testing performance of estimates in my final sample by testing whether these estimates can be significantly bounded beneath Cohen's (1988) small effect size benchmarks. For each estimate, I separately test whether $\sigma \in [-0.2, 0.2]$ and whether $r \in [-0.1, 0.1]$. These ROPEs are quite lenient. $|r| = 0.1$ is larger than over 25% of all published estimates in economics (Doucouliagos 2011). Online Appendix E shows that both $|r| = 0.1$ and $|\sigma| = 0.2$ are large effect sizes even amongst a benchmark sample of plausibly large economic

[9]Per Equation 10, the value of t_{NHST} derived using σ and s after my standardization procedure is identical to that which would be derived from the original regression results before standardization.

effects. Thus when an article in a top economics journal claims that a relationship is null or negligible, it should be easy to show that the estimates defending that claim are significantly bounded beneath $|\sigma| = 0.2$ or $|r| = 0.1$, as these are lenient thresholds.

5.2 Measuring Equivalence Testing Failure

I define the 'equivalence testing failure rate (ETFR)' as the average partition-level proportion of estimates that fail to be significantly bounded within a given ROPE at a 5% significance level. E.g., consider a toy dataset of estimates defending three null claims. Suppose that for 20% of estimates defending the first claim, σ cannot be significantly bounded within a ROPE of $[-0.2, 0.2]$ at a 5% significance level. Suppose that the same is true of all estimates defending the second claim and no estimates defending the third claim. The average claim-level ETFR for standardized coefficients within a ROPE of $[-0.2, 0.2]$ would thus be $(20\% + 100\% + 0\%)/3 = 40\%$. Online Appendix Figure A1 provides a visualization of these ETFR calculations.

I compute claim-level and article-level ETFRs, as well as an inverse-weighted claim-level ETFR that ensures all articles receive the same weight in the sample. I estimate precision using the standard error of the mean partition-level failure rate. Online Appendix H provides precise computational details for partition-level ETFRs and their standard errors.

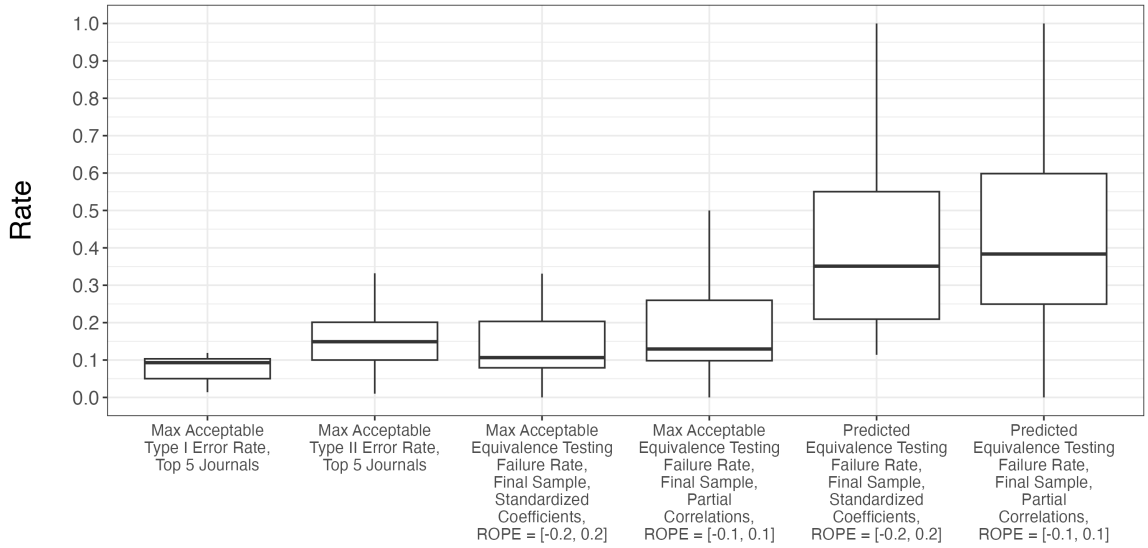
An estimate $\hat{\delta}$ 'failing' an equivalence test does not necessarily imply that the underlying relationship δ is actually large – it simply means that there is not precise enough evidence to say with high confidence that the relationship is practically equal to zero. When an estimate is not statistically significantly different from zero nor significantly bounded within its ROPE, this yields an *inconclusive* result, as there is not enough power or precision to say whether or not the underlying relationship is meaningfully large. I elaborate further on inconclusive results in Section 7.2. ETFRs thus primarily measure how frequently estimates defending null claims in economics

yield inconclusive results, and therefore serve as a measure of the robustness of null claims in economics to equivalence testing.

5.3 Prediction Platform Survey

In addition to my main replication data, I obtain data from a Qualtrics-based survey conducted on the Social Science Prediction Platform (SSPP) from 30 March to 30 April 2024 (see DellaVigna, Pope, & Vivalt 2019).[10] The SSPP survey asks social science researchers to provide their predictions and judgments concerning equivalence testing results in my final sample. Specifically, considering a 5% significance level, I ask respondents to predict ETFRs in the final sample for standardized coefficients σ when the ROPE is $[-0.2, 0.2]$. Thereafter, I ask respondents to provide the largest ETFRs that they would consider to be acceptable for that ROPE. To minimize confusion, I then ask each respondent whether they anticipate that these ETFRs will differ for partial correlation coefficients r when the ROPE is $[-0.1, 0.1]$. If they answer 'yes', then they are asked to provide these same predictions and judgments of ETFRs for partial correlation coefficients within a ROPE of $[-0.1, 0.1]$. If they answer 'no', then they are not shown these new questions, and their predictions and judgments of ETFRs within a partial correlation ROPE of $[-0.1, 0.1]$ are imputed based on their responses for the standardized coefficient ROPE of $[-0.2, 0.2]$. I also ask respondents to provide judgments on acceptable Type I and Type II error rates in Top 5 economics journals. After screening out respondents who report familiarity with my analysis' results or give incomplete responses, I possess a sample of judgments and predictions from 62 researchers. Further details about this sample can be found in Online Appendix G.

[10]The survey and the original Qualtrics file can be found at https://socialscienceprediction.org/s/602202.



Note: Each box plot displays the 25th, 50th, and 75th percentile of its respective rate in the SSPP sample, along with whiskers that extend to the largest (smallest) point that lies within 1.5 interquartile ranges above (below) the box.

Figure 3: Distributions of SSPP Predictions and Judgments

6 Results

6.1 Predictions and Judgments

Figure 3 presents box plots of the SSPP sample's predictions and judgments. The first two box plots show judgments of acceptable Type I and Type II error rates in Top 5 economics journals. The final four box plots show predictions and judgments of equivalence testing failure rates in the final sample.

Interestingly, the SSPP sample's error rate tolerance for Top 5 economics journals does not conform to disciplinary standards. The median SSPP respondent is willing to tolerate Type I error rates of 9.3% (quite above the classical 5% prescription) and Type II error rates of 14.9% (quite below the classical 20% prescription; see Cohen 1988). Respondents' median tolerance for ETFRs is somewhere between their median tolerances for Type I and Type II errors. The median respondent deems ETFRs up to 10.65% to be acceptable for a standardized coefficient ROPE of $[-0.2, 0.2]$. This median ETFR tolerance increases to 12.95% for a partial correlation ROPE of

$[-0.1, 0.1]$. However, respondents predict that ETFRs will substantially exceed these thresholds. Median predictions for ETFRs are 35.1% for a standardized coefficient ROPE of $[-0.2, 0.2]$ and 38.35% for a partial correlation ROPE of $[-0.1, 0.1]$. Section 6.2 shows that these predictions are fairly accurate, although the median ETFR prediction for a partial correlation ROPE of $[-0.1, 0.1]$ is an underestimate.

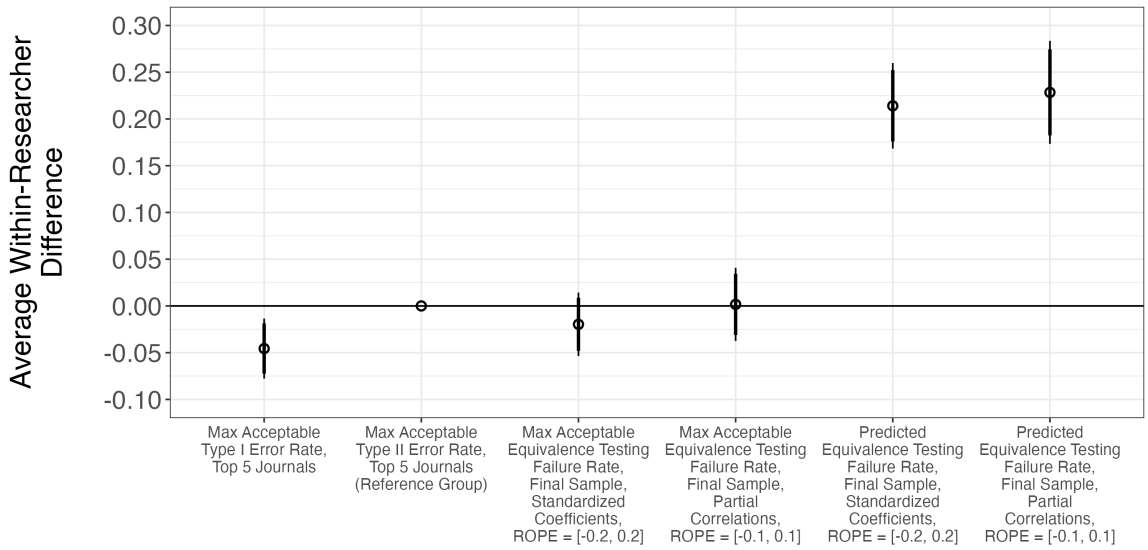
To formally test how survey participants' ETFR predictions differ from the ETFRs they would judge to be acceptable, I leverage the within-subject design of my survey to construct a respondent-rate panel dataset. This panel dataset allows me to obtain within-respondent differences between rates using a panel data regression model that controls for respondent fixed effects. Let i index the respondent and r index one of the six rates displayed in Figure 3. I estimate the model

$$\text{Rate}_{i,r} = \theta + \gamma_r + \lambda_i + \mu_{i,r}. \tag{12}$$

Figure 4 displays these within-respondent estimates of differences between rates. Specifically, Figure 4 shows γ_r estimates from a model of Equation 12 that treats judgments on Type II error rates as the reference group.[11] The average respondent reports that for results in Top 5 economics journals, their tolerance for Type I error rates is 4.561 percentage points lower than their tolerance for Type II error rates. This is consistent with higher disciplinary error rate thresholds for Type II errors than for Type I errors (see Section 4.3), and with prior evidence showing that researchers find null claims to be less important than claims that statistical relationships are significant (see Chopra et al. 2024).

The estimates in Figure 4 show that ETFR tolerance is quantitatively close to Type II error rate tolerance. The average respondent's tolerance for Type II errors is two percentage points higher than their tolerance for ETFRs within a standardized coefficient ROPE of $[-0.2, 0.2]$, and is 0.2 percentage points lower than their tolerance for ETFRs within a partial correlation ROPE of $[-0.1, 0.1]$. Though one could use

[11] A table version of these within-respondent estimates is provided in Online Appendix Table A2.

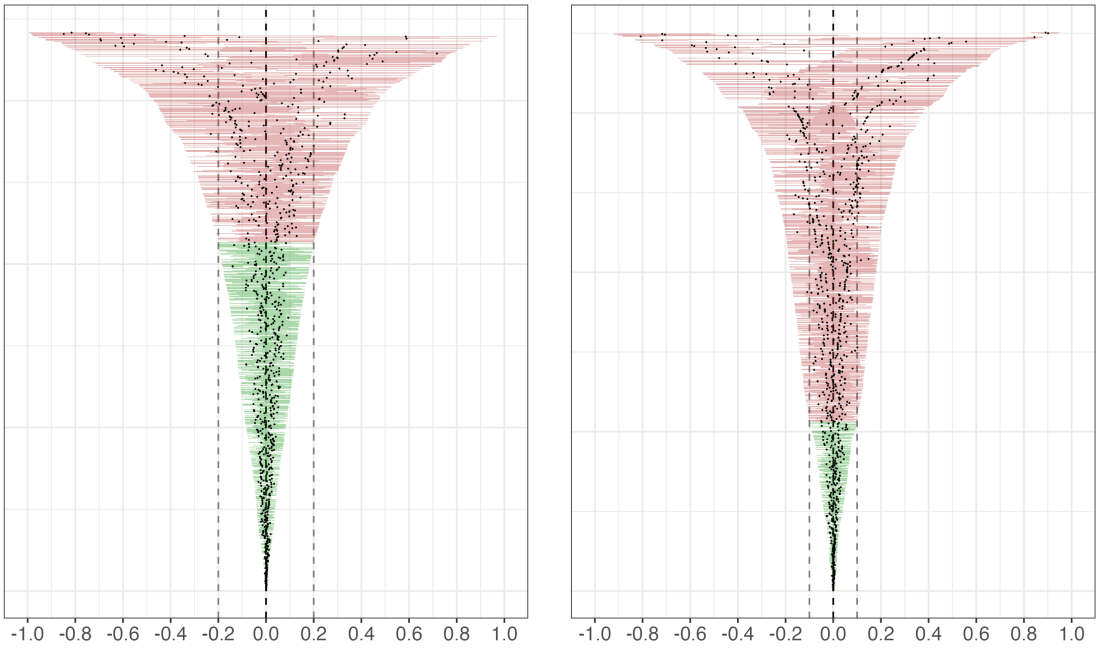


Note: γ_r estimates from Equation 12 are provided along with 95% ECIs (thicker bands) and confidence intervals (thinner bands). Standard errors are clustered at the respondent level using a CR3 cluster-robust variance estimator.

Figure 4: Within-Respondent Estimates of Differences in Predictions/Judgments

equivalence testing to significantly bound these two estimates within a five percentage point difference of Type II error rate tolerance, it is not clear that such a five percentage point difference is practically equal to zero in this context. There is thus insufficient power to say that ETFR tolerances are practically equal to Type II error rate tolerance.

However, researchers predict that ETFRs in my final sample will far exceed any of these acceptability thresholds. The average respondent predicts that ETFRs will exceed their personal Type II error rate tolerance by 21.4 percentage points within a standardized coefficient ROPE of $[-0.2, 0.2]$, and by 22.8 percentage points within a partial correlation ROPE of $[-0.1, 0.1]$. After accounting for the differences between Type II error rate tolerance and ETFR tolerances, these estimates imply that the average respondent predicts that ETFRs will exceed their personal acceptability thresholds by around 23 percentage points. This is evidence that researchers believe that current testing practices in top economics journals produce null claims that exhibit unacceptably high error rates. My ETFR estimates in the rest of this section



Note: Estimates in the final sample are plotted along with 95% ECIs. Estimates are sorted from top to bottom by ECIOB magnitude; estimates with larger ECIOB magnitudes are closer to the top. Green (red) estimates in the left-hand graph can (not) be significantly bounded in a standardized coefficient ROPE of $[-0.2, 0.2]$. Green (red) estimates in the right-hand graph can (not) be significantly bounded in a partial correlation ROPE of $[-0.1, 0.1]$. For visibility, I remove outlier estimates with standardized coefficient ECIOB magnitudes > 1 when producing the left-hand graph. This change drops 2.5% of the estimates in the final sample.

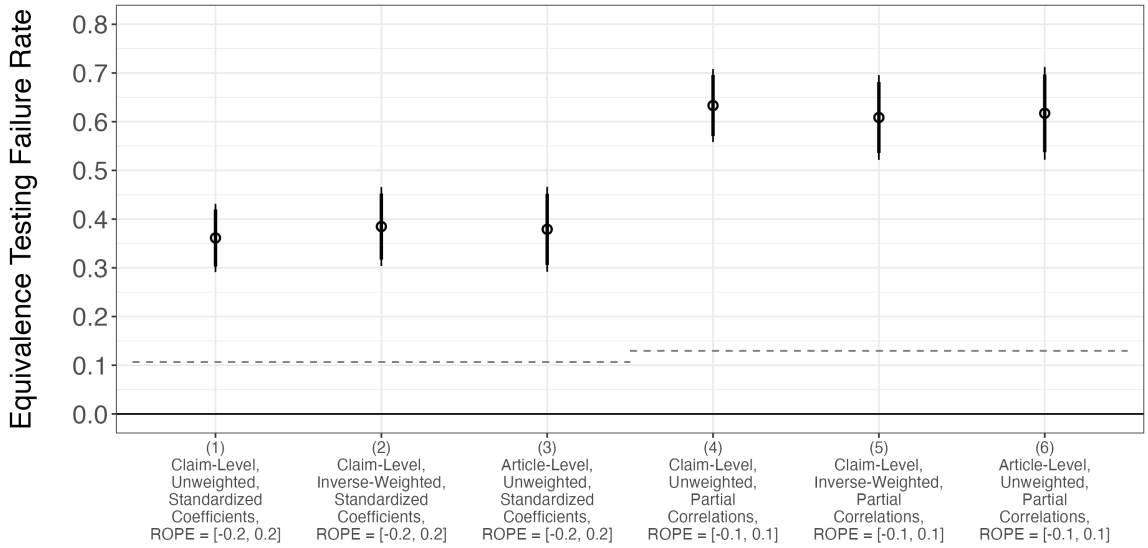
Figure 5: Estimates and ECIs in the Final Sample

show that this prediction is quite accurate.

6.2 Main Results

Many estimates defending null claims in top economics journals fail lenient equivalence tests. Figure 5 plots the estimates in the final sample. Under equivalence testing, over 39% of these estimates cannot be significantly bounded beneath a 0.2σ effect size, and over 69% cannot be significantly bounded beneath a $0.1r$ effect size.

This poor equivalence testing performance is not explained by a small group of papers with disproportionately many estimates – such performance is poor across



Note: ETFRs are provided along with 95% ECIs (thicker bands) and confidence intervals (thinner bands). These intervals are based on the standard error of the mean for unweighted ETFRs and the standard error of the weighted mean for weighted ETFRs (see Online Appendix H). Dashed lines represent the median SSPP respondent's maximum acceptable claim-level ETFR for the given ROPE at a 5% significance level (see Section 6.1).

Figure 6: Main Equivalence Testing Failure Rate Estimates

all claims and articles in the final sample. Figure 6 displays the main ETFR estimates.[12] The dashed lines represent the median SSPP respondent's thresholds for acceptable ETFRs (see Section 6.1). ETFRs are significantly greater than both zero and these thresholds. For a standardized coefficient ROPE of $[-0.2, 0.2]$, ETFRs range from 36.1-38.5%. These ETFRs are even higher for a partial correlation ROPE of $[-0.1, 0.1]$, ranging from 60.9-63.3%. For null claims in top economics journals, ETFRs within lenient ROPEs thus range from 36-63%.

The significance of these ETFRs is robust to many checks. First, Figure 6 shows ETFRs for different effect size measures and aggregation procedures. Second, Online Appendix Table A4 shows ETFRs after I remove estimates from the sample that are initially statistically significant under standard NHST. Third, Online Appendix Table A5 displays ETFRs after I employ a leave-one-out approach where subsamples of regressor type combinations are removed from the sample. Fourth, Online

[12] A table version of these ETFR estimates is provided in Online Appendix Table A3.

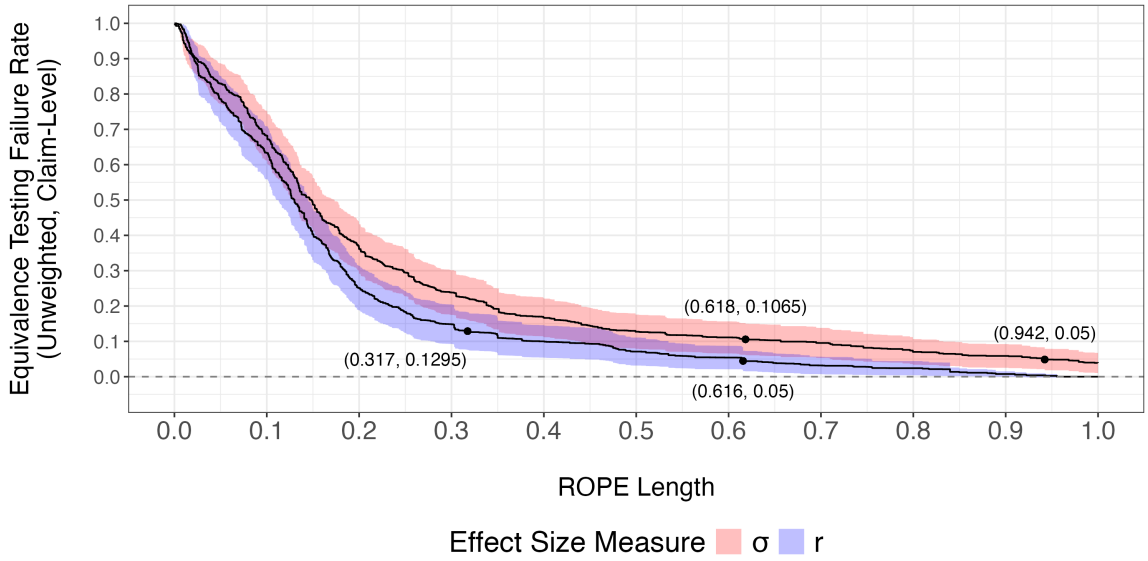
Appendix Table A6 uses the same leave-one-out approach to check ETFR robustness to coding choices, first removing estimates that are not fully reproducible, and thereafter removing estimates from models that require conformability modifications. Fifth, Online Appendix Table A7 splits the sample into estimates defending 'qualified' and 'unqualified' null claims, depending on whether those claims are qualified by references to effect sizes, statistical significance, or imprecision (see Section 3). Sixth and finally, Online Appendix Table A8 splits the sample into estimates defending main claims and secondary claims (where the latter includes null claims about mechanisms, robustness checks, and subgroup analyses; see Section 3). In all cases, ETFRs always remain significantly bounded above median SSPP acceptability thresholds.

6.3 Failure Curves

Perhaps the most important sensitivity check concerns the choice of ROPE. Figure 7 plots 'failure curves', which show how claim-level ETFRs vary with the choice of ROPE length ϵ. The downward-sloping shape of the failure curves reflects the intuition that ETFRs decline when one is willing to tolerate larger ROPEs. Figure 7 shows that ETFRs remain significantly above nominal and acceptable levels even as ROPE lengths grow quite large. These findings hold for both effect size measures (i.e., both σ and r).

The failure curves are also useful for a thought experiment on the credibility of standard testing practices. Suppose one wanted to assert that current testing practices for null claims in economics are sufficiently credible, and that ETFRs are beneath acceptable levels for reasonably-sized ROPEs. How large is the smallest ROPE that one would need to tolerate in order to make such a claim?

Figure 7's annotated points show that one must tolerate enormous ROPEs to obtain acceptable ETFRs. As discussed in Section 6.1, the median SSPP respondent's maximum ETFR tolerance for a partial correlation ROPE of $[-0.1, 0.1]$ is 12.95%. To obtain claim-level ETFRs beneath 12.95%, one must set a partial correlation



Note: Failure curves are annotated by points indicating the ROPEs that must be tolerated to bound ETFRs beneath 1) 5% and 2) the median SSPP respondent's maximum tolerance for claim-level ETFRs within the benchmark ROPEs tested when producing Figure 6's estimates. Uncertainty bands represent 95% confidence intervals based on the unweighted claim-level ETFR's standard error of the mean (see Online Appendix H).

Figure 7: Failure Curves

ROPE of $[-0.317, 0.317]$. A researcher who sets this ROPE is effectively arguing that $|r| = 0.317$ is practically negligible. This is a very large effect size. Consider Doucouliagos (2011), who maps the distribution of partial correlation coefficients in the economics literature using roughly 22,000 estimates from published papers used in economic meta-analyses. Based on this distribution, $|r| = 0.317$ is larger than nearly 75% of published results in economics. To obtain claim-level ETFRs beneath 5%, one must be willing to claim that $|r| = 0.616$ is practically negligible. This effect size is extremely large.

The ROPEs one must tolerate to obtain acceptable ETFRs are large regardless of the effect size measure considered. As aforementioned in Section 6.1, the median SSPP respondent's maximum ETFR tolerance for a standardized coefficient ROPE of $[-0.2, 0.2]$ is 10.65%. One must be willing to tolerate a ROPE length of 0.618σ to obtain claim-level ETFRs beneath 10.65%, and a ROPE length of 0.942σ to achieve claim-level ETFRs beneath 5%. Although the distribution of standardized coefficient

magnitudes throughout the economics literature is not yet known, Online Appendix E shows that these magnitudes are quite large even in a sample of plausibly large economic effects.

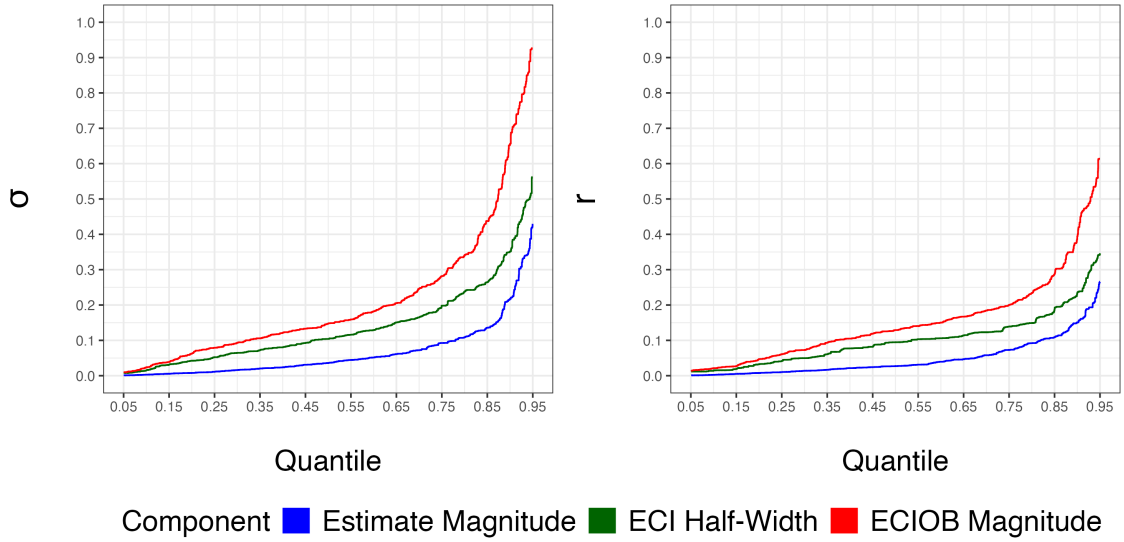
Effect sizes this large are clearly not practically equal to zero. Given this, a more sensible alternative conclusion arises: the current testing paradigm economists use to make and defend null claims tolerates unacceptably high error rates. Many meaningful economic relationships are thus likely erroneously dismissed as negligible or nonexistent under standard NHST.

6.4 Mechanisms

Are these high ETFRs caused more by large effect sizes or by imprecision? Section 4.2 establishes that the magnitude of the ECI outer bound (ECIOB) is the length of the smallest symmetric ROPE around zero wherein one can significantly bound $\hat{\delta}$. ECIOB magnitudes thus directly determine ETFRs for a given symmetric ROPE. Therefore, the mechanisms of ETFRs can be examined by decomposing the exact 95% ECIOB magnitude into its two constituent parts: the estimate's magnitude $\left|\hat{\delta}\right|$, which measures effect size, and the estimate's 95% ECI half-width $s \times t^*_{0.05,df}$, which measures imprecision (see Definition 4.3).

Figure 8 plots the distributions of exact 95% ECIOB magnitudes and their components. If effect sizes were the only driver of ETFRs, then one would expect the distribution of ECI half-widths to be a flat horizontal line, and the distribution of estimate magnitudes would run parallel to the distribution of ECIOB magnitudes. However, ECI half-widths stochastically dominate estimate magnitudes throughout the distribution. Though both large effect sizes and low precision contribute to high ETFRs, low precision is the dominant driver.

Though larger relationships may naturally have larger ECIs, Table 3 provides further empirical evidence of imprecision's dominance in determining ETFRs which is not susceptible to this scaling issue. Specifically, Table 3 displays constant elasticity



Note: The figure shows the central 90% of the inverse cumulative density functions for each component of the ECIOB magnitude and the ECIOB magnitude itself. Cumulative density functions arise from a weighted inverse density that ensures each claim receives the same weight in the data.

Figure 8: Inverse Cumulative Density Functions of ECIOB Magnitudes and Their Components

estimates of the relationships between 95% ECIOB magnitudes and both effect sizes and 95% ECI half-widths. Both effect sizes and ECI half-widths are significantly positively associated with ECIOB magnitudes, which is intuitive. However, ECI half-widths display noticeably stronger relationships with ECIOB magnitudes than effect sizes. This provides additional evidence that though large effect sizes are an important factor for explaining high ETFRs, imprecision is the dominant determinant.

Table 3 also provides encouraging evidence on equivalence testing's empirical properties. Section 3 notes that when the researcher is trying to show that $\delta = 0$ using the standard NHST framework, imprecision is 'good' because there is an inverse relationship between precision and the probability of obtaining null results. However, the second and fourth columns in Table 3 show that when using equivalence testing, more precise estimates can be bounded significantly closer to zero.

This demonstrates that equivalence testing restores the proportional relationship between precision and the probability of reaching a null conclusion. This property

	Effect Size	ECI Half-Width	Effect Size	ECI Half-Width
Elasticity w/ |ECIOB|	0.575 (0.127)	0.668 (0.051)	0.422 (0.031)	0.958 (0.059)
N	876	876	876	876
Adj. R^2	0.604	0.936	0.767	0.76
Effect Size Measure	σ	σ	r	r

Note: Each column's elasticity is calculated using a weighted univariate linear regression where the dependent variable is the 95% ECIOB magnitude in units specified by the column, the independent variable is specified by the column, and observations are weighted by an inverse density that ensures all claims receive the same weight in the data. The linear regression estimates are transformed into elasticities using the `marginaleffects` post-estimation suite in R (Arel-Bundock, Greifer, & Heiss 2024). The adjusted R^2 is that for the original weighted linear regression model. Standard errors are clustered by claim and reported in parentheses.

Table 3: Mechanisms of ECIOB Magnitudes

eliminates the perverse incentive for researchers to limit the power and precision of their estimates when a null result is desirable (e.g., in robustness checks; see Dreber, Johannesson, & Yang 2024). More broadly, this property makes null findings evidenced by equivalence testing quite credible, because in contrast to standard NHST, equivalence testing eliminates the possibility that null results arise simply due to low power and imprecision.

7 Applying Equivalence Testing Methods

Section 6 uses equivalence testing to show that economists' current practices for making and defending null claims likely tolerate unacceptably high error rates. This implies that many null findings in the economics literature are likely false negatives. Fortunately, the tool used to demonstrate this problem is also the problem's solution. By eliminating the conflation between imprecision and null results inherent to the standard NHST framework, equivalence testing restores researchers' ability to credibly make null claims with reasonable error rate coverage. Equivalence testing is a first-order robustness check for null findings. Because virtually any relationship may be practically equal to zero, every researcher should be prepared to perform equivalence

testing on their estimates of interest. The rest of this section is dedicated to showing researchers how they can credibly apply equivalence testing in future research.

7.1 ROPE Selection

What should the ROPE be for a given estimate? There is no one-size-fits-all answer to this question. Benchmark effect sizes can be useful for analyses that assess an entire literature, particularly when estimates from that literature are comprised of estimates from diverse regressor types, variable units, and models. However, benchmark effect sizes are not generally valid ROPEs for individual research questions (Lakens, Scheel, & Isager 2018). The true ROPEs for two different relationships will seldom be exactly the same. Therefore, a literature-wide effect size benchmark will rarely (if ever) be a useful boundary for an individual estimate's ROPE.

In practice, researchers need to assign different ROPEs for each relationship of interest, but this generates substantial researcher degrees of freedom. A key concern is 'ROPE-hacking', whereby researchers interested in showing that $\delta \approx 0$ adjust ROPEs *ad hoc* so that their estimates are significantly bounded within those ROPEs. There is already strong evidence of such ROPE-hacking in the medical literature (see Ofori et al. 2023). Given the prevalence of reverse p-hacking for placebo tests in top economics journals, it is not difficult to imagine that ROPE-hacking could similarly emerge in economic applications of equivalence testing (see Dreber, Johannesson, & Yang 2024). This is a problem that pre-registration alone cannot fix, as researchers interested in obtaining evidence of null findings can simply pre-register excessively wide ROPEs. Unsurprisingly, this practice can inflate error rates in equivalence testing (Campbell & Gustafson 2021). Below, I offer researchers two credible ROPE-setting methods.

7.1.1 Surveys

A flexible approach to control researcher degrees of freedom and ensure that ROPEs are set credibly and independently is by surveying judgments on minimal meaningful

effect sizes from independent parties, such as experts or relevant stakeholders. Such judgments are practical to elicit using recent research-centric survey platforms, such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019). Though the SSPP is primarily a prediction platform, and thus requires that researchers ask respondents to make predictions regarding some outcome, it is seamless to incorporate questions regarding the effect sizes that respondents would deem practically equal to zero. It is easy to follow the question "What do you predict the effect of this intervention will be?" with the question "What is the smallest effect that you would consider practically meaningful?" Users of the SSPP are a particularly useful group to pose such questions to, as they are typically experienced at leveraging theory and prior prediction experience to make judgments on effect sizes in social science research.

This paper provides an example of how to implement such a survey. In addition to asking respondents what equivalence testing failure rates they would predict, I also asked the largest ETFRs that they would find acceptable. This is the relevant measure of practical significance for the purposes of this paper.

Researchers can set ROPEs based on respondents' median responses to such questions. Further, even if researchers administer such surveys with the primary goal of eliciting ROPEs, the extra prediction data will still be useful to help inform posterior beliefs and evidence the novelty of research findings (DellaVigna, Pope, & Vivalt 2019). Of course, other survey platforms can be appropriate for such belief elicitation.

A key advantage of this ROPE-setting strategy is that even *post hoc*, non-pre-registered surveys can provide credible ROPEs. Many guides on equivalence testing stress the importance of pre-registering ROPEs to maintain credibility and avoid ROPE-hacking (Piaggio et al. 2012; Lakens, Scheel, & Isager 2018; Campbell & Gustafson 2021). The main advantage of pre-registration is that it makes the researcher's methodological choices independent of the data, preventing data-mining and p-hacking (Olken 2015; Campbell & Gustafson 2021). The same aim can be achieved by setting ROPEs based on independent survey data because other people

are effectively selecting the ROPE for the researcher. This renders ROPE selection independent of the main data even if the survey is conducted after the researcher has already seen and analyzed the data. Thus even if a pre-registered empirical analysis has already been completed, a peer-reviewer or colleague can still recommend equivalence testing to a researcher, and the researcher can still make the equivalence testing results credible by eliciting their ROPEs from independent parties.

7.1.2 Cost-Benefit Analysis

At points in the research process when researchers are sequestered enough from their data that statistical analyses can be credibly pre-registered, another economically intuitive method for setting ROPEs is cost-benefit analysis. Suppose that a one-unit increase in Y and a one-unit increase in D can respectively be monetarily valued at $V_Y > 0$ and $V_D > 0$. Then because a one-unit increase in D costs V_D and is associated with a monetary gain of $V_Y\delta$, one can test whether the monetary benefit of a unit increase in exposure is significantly smaller than the monetary cost of that unit increase by testing whether δ is significantly bounded within the ROPE of $[-V_D/V_Y, V_D/V_Y]$. If Y is a 'bad' rather than a 'good' outcome (i.e., $V_Y < 0$), or if D is an exposure that would be costly to abate (i.e., $V_D < 0$), then the ROPE becomes $[V_D/V_Y, -V_D/V_Y]$; if both are true, then the ROPE remains $[-V_D/V_Y, V_D/V_Y]$.

I emphasize the importance of pre-registration for this ROPE-setting approach because due to the fact that the researcher themselves typically specifies V_Y and V_D, these two valuations represent researcher degrees of freedom that can be used to ROPE-hack. That said, there are resources that can make these valuations more credible. E.g., the Center for Effective Global Action provides resources for systematic costing, such as cost reporting tables and detailed pre-analysis planning templates for costing estimates.[13] Boardman et al. (2018) provide a comprehensive review of methods for cost-benefit analysis, which includes numerous methods for defining both

[13]See https://cega.berkeley.edu/resources-systematic-cost-effectiveness-analysis/, accessed 4 December 2025.

V_Y and V_D. Though pre-registration alone does not guarantee credible ROPEs, pre-registration combined with robust methods for determining cost and benefit valuations can yield credible ROPEs for equivalence tests that assess whether interventions' benefits are significantly outweighed by their costs.

7.2 ROPEs and Research Conclusions

How should equivalence testing coexist with current frameworks that test whether relationships are significantly different from zero? Even when applied, equivalence testing is often treated as an afterthought, utilized only when statistically significant evidence cannot be obtained under the standard NHST framework (Campbell & Gustafson 2021). For example, medical trials with nominal aims of testing for equivalence seldom report a pre-specified ROPE (Piaggio et al. 2012). This implies that such trials first test an estimate using the standard NHST framework and move to equivalence testing only when the standard NHST framework does not yield statistically significant evidence. Even if not named explicitly, this common practice is functionally identical to the 'conditional equivalence testing (CET)' procedure formulated by Campbell & Gustafson (2018), who show empirically that the procedure produces similar results to inference with Bayes factors under inverse variance priors. Because over 90% of the estimates selected for my final sample are not statistically significantly different from zero, the equivalence testing I conduct in my main empirical analysis is primarily an application of the CET procedure.

Definition 7.1 (The Conditional Equivalence Testing Procedure)**.** *The researcher begins by testing $\hat{\delta}$ using the standard NHST framework in Definition 3.1. If the researcher rejects H_0 under the standard NHST framework, then the researcher concludes that $\delta \neq 0$. Otherwise, the researcher then tests $\hat{\delta}$ using the equivalence testing framework in Definition 4.1. If the researcher then rejects H_0 under the equivalence testing framework, then the researcher concludes that $\delta \approx 0$. Otherwise, the researcher concludes that the relationship between δ and zero is inconclusive.*

There are drawbacks to the CET procedure; e.g., in highly-powered research settings, $\hat{\delta}$ can be both significantly different from zero and significantly bounded within a ROPE (Lakens, Scheel, & Isager 2018). If the CET procedure is followed exactly, then researchers may reach misleading conclusions in this setting. The CET framework would deem $\hat{\delta}$ to be significantly different from zero in the first step, but then equivalence testing would never be performed. Readers (and potentially also the researcher) would therefore never learn that $\hat{\delta}$ is significantly bounded within its ROPE.

Further, the CET procedure begins with applying the standard NHST framework, which is not always construct-valid to employ once a ROPE is set. The knowledge that some non-zero values of δ are practically equal to zero implies that if the researcher wants to show that δ is practically significant, then it is not sufficient to provide significant evidence that $\delta \neq 0$. Rather, the researcher must demonstrate significant evidence that δ is bounded outside of the ROPE to conclude with certainty that the estimate is practically significant. This is not required by the CET procedure.

However, one useful feature of the CET procedure is that it can yield inconclusive results. The standard NHST framework currently results in a dichotomization of research findings – either a relationship is significant or it is not (McShane & Gal 2017). However, if an estimate is imprecise enough, it may neither be possible to find significant evidence that the estimate is different from zero nor to find significant evidence that the estimate is practically equal to zero. In such settings, researchers cannot make a claim about the estimate's significance with reasonable certainty, and thus the researcher's conclusions about the estimate should remain agnostic. This paper provides an example of such conclusions. In Section 6.1, I note that though the within-researcher point estimates of tolerances for ETFRs and Type II error rates may look quantitatively similar, there is ultimately insufficient power and precision to conclude whether these tolerances differ with reasonable error rate coverage.

Though embracing this uncertainty is likely uncomfortable and limiting to researchers who are used to being able to dichotomize research findings as 'significant'

or 'insignificant', the empirical results of this paper show that reaching research conclusions in this way is a dangerous practice that results in high error rates. This is likely a key contributor to the low faith that researchers have in the quality and publishability of null conclusions reached using the standard NHST framework (McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024). Researchers should thus be willing to admit when they do not have sufficient power to make reasonably certain conclusions regarding statistical relationships, and therefore should use testing frameworks that make it possible to reach inconclusive findings.

I advocate for researchers to test statistical relationships with a framework that retains the capacity to produce inconclusive findings while also addressing the CET procedure's flaws. Specifically, I advocate for using the 'three-sided testing (TST)' framework designed by Goeman, Solari, & Stijnen (2010).

Definition 7.2 (The Three-Sided Testing Framework)**.** *The researcher wishes to assess the practical significance of δ with a size-α test. The researcher thus sets a ROPE $[\epsilon_-, \epsilon_+]$ as in Definition 4.1 and establishes hypotheses*

$$\begin{array}{lll} H_0^{\{N\}} : \delta \geq \epsilon_- & H_0^{\{TOST\}} : \ \delta < \epsilon_- \ \textit{or} \ \delta > \epsilon_+ & H_0^{\{P\}} : \delta \leq \epsilon_+ \\ H_A^{\{N\}} : \delta < \epsilon_- & H_A^{\{TOST\}} : \ \delta \geq \epsilon_- \ \textit{and} \ \delta \leq \epsilon_+ & H_A^{\{P\}} : \delta > \epsilon_+. \end{array} \tag{13}$$

Test statistic t_{TOST} is computed as in Definition 4.2 along with test statistics

$$t_N = \frac{\hat{\delta} - \epsilon_-}{s} \qquad\qquad t_P = \frac{\hat{\delta} - \epsilon_+}{s}. \tag{14}$$

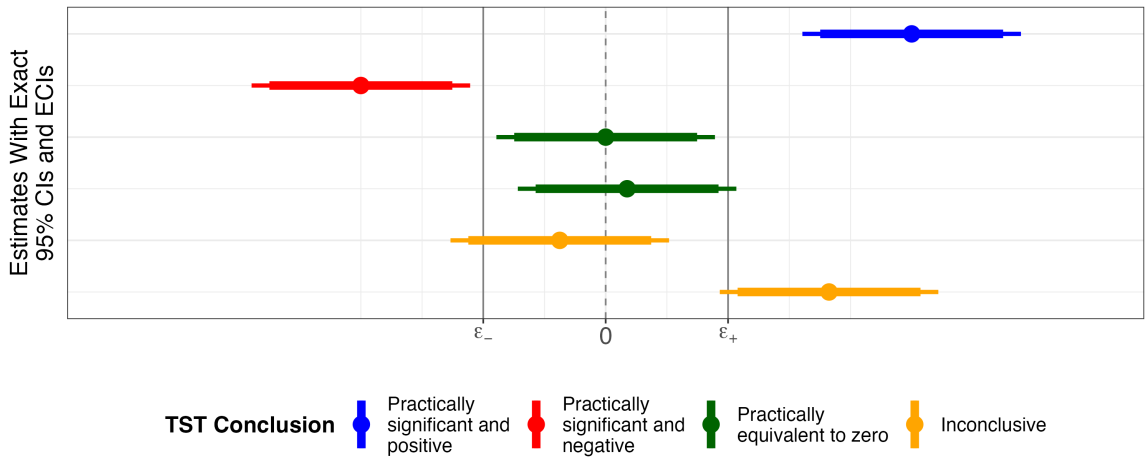
The researcher concludes that δ is significantly bounded above the ROPE if and only if $t_P > t^*_{\alpha/2,df}$. The researcher concludes that δ is significantly bounded below the ROPE if and only if $t_N < -t^*_{\alpha/2,df}$. As in Definition 4.2, if $t_{TOST} = t_-$, then the researcher concludes that δ is significantly bounded within the ROPE if and only if $t_{TOST} \geq t^*_{\alpha,\ df}$, but if $t_{TOST} = t_+$, then the researcher concludes that δ is significantly bounded within the ROPE if and only if $t_{TOST} \leq -t^*_{\alpha,\ df}$. If the researcher does not find that δ is

significantly bounded above the ROPE, below the ROPE, or within the ROPE, then the researcher concludes that the practical significance of δ is inconclusive.

The TST framework combines tests for practical equivalence with tests for practical significance, addresses all aforementioned concerns with the CET procedure, and still retains the CET procedure's positive properties. Principally, under the TST framework, δ is never declared to be practically significant unless there is significant evidence that δ is bounded outside its ROPE. Further, even though the TST framework conducts three simultaneous hypothesis tests, the family-wise error rate of these three tests for a single application of the TST framework is controlled at α without any multiple hypothesis testing adjustments, even if combined with standard NHST (Goeman, Solari, & Stijnen 2010; Isager & Fitzgerald 2025). However, like CET, the TST framework also still retains the possibility for inconclusive results. Such results arise if $\hat{\delta}$ is too close to one of the ROPE boundaries to say that δ is significantly bounded inside or outside of the ROPE given the precision of $\hat{\delta}$.

The empirical findings of this paper provide an example of how conclusions can be made using the TST framework. The question of whether ETFRs are significantly greater than zero is uninteresting; ETFRs are greater than zero almost by construction. However, as discussed in Section 7.1, thresholds for maximum acceptable ETFRs are a relevant measure of 'practically (in)significant' effect sizes for the purposes of this paper. After eliciting judgments on these thresholds in the SSPP survey (see Sections 5.3 and 6.1), for each effect size measure, I set a ROPE of $[0, \epsilon]$, where ϵ is the median of these threshold judgments for that given effect size measure. In Section 6.2, I then show that the 95% confidence intervals of my main ETFR estimates are bounded above these ϵ thresholds. Under the TST framework, this is statistically significant evidence that the ETFRs in my final sample are practically significant.

Figure 9 illustrates how TST can be applied using a confidence interval approach at a 5% significance threshold. The top four estimates shown in Figure 9 depict estimates for which the researcher can make highly certain practical significance conclu-



Note: Estimates are displayed along with 95% ECIs (thicker bands) and confidence intervals (thinner bands). ϵ_- and ϵ_+ respectively denote the lower and upper bounds of the ROPE for these estimates.

Figure 9: Research Conclusions in the TST Framework

sions. The first and second estimates' 95% confidence intervals are bounded outside of the ROPE, so these estimates are practically significantly positive and negative (respectively). The third and fourth estimates' entire 95% ECIs are bounded inside the ROPE, so there is significant evidence that these estimates are practically equal to zero. In contrast, the practical significance of the last two estimates in Figure 9 is inconclusive. The first of these two estimates has a point estimate bounded within the ROPE, but its 95% ECI intersects the ROPE. The last estimate has a point estimate bounded outside of the ROPE, but its 95% confidence interval intersects the ROPE.

The bottom estimate in Figure 9 is particularly important for understanding how TST augments standard NHST. This estimate is statistically significantly different from zero. Because its point estimate exceeds the ROPE, most economists would likely conclude that the relationship is 'economically significant'. However, under TST, this estimate would still be deemed too noisy to yield highly certain practical significance conclusions. This bottom estimate lacks sufficient precision to rule out the possibility that its point estimate falls outside of the ROPE simply due to sampling variation. The TST framework only deems estimates whose $(1 - \alpha)$ confidence intervals are fully bounded outside of the ROPE, such as the top two estimates, to be practically

significant. These sorts of estimates are both large enough and precise enough to instill strong confidence that these relationships are practically meaningful. For a more detailed tutorial on TST, see Isager & Fitzgerald (2025).

8 Conclusion

I introduce the economics literature to a suite of simple equivalence testing methods. I then demonstrate their necessity, showing that many estimates defending published null claims in top economics journals fail lenient equivalence tests. At a 5% significance level, equivalence testing failure rates for these estimates range from 36-63% within lenient ROPEs. To obtain acceptable equivalence testing failure rates, one must claim that nearly 75% of all published effect sizes in economics are practically equal to zero. Economists' current testing practices for making and defending null claims thus likely tolerate unacceptably high error rates.

These results demonstrate that testing practices in economics need to change, and I provide a practical blueprint for how researchers can make this change. Specifically, researchers should specify the smallest practically important effect size for each relationship that they are interested in estimating. Such effect sizes can be determined by surveying experts or relevant stakeholders for their judgments of minimal meaningful effect sizes, or through credible cost-benefit analysis. These effect sizes can be used to set ROPEs, which can then be used to test estimates using the TST framework.

The TST framework has several advantageous properties. First, TST permits researchers to simultaneously test for an estimate's practical significance and practical equivalence to zero. Error rates from these simultaneous tests remain controlled at nominal significance levels. Second, the TST framework disentangles estimates' precision from their practical significance, ensuring that relationships are not deemed practically significant unless there is credible evidence that such relationships are larger than their smallest effect size of interest. Third and finally, the TST framework

makes it possible for inconclusive results to arise. When the researcher lacks enough power to make definitive claims about the practical significance of the relationship, they should assert that their results are inconclusive. The TST framework requires such conclusions in these settings.

Adoption of these techniques would have many positive effects on economic research. First, credible equivalence testing can help assuage existent concerns about the quality and publishability of null results, helping reduce publication bias against null findings in the economics literature. Further, equivalence testing makes economic theories credibly falsifiable by making it possible to obtain significant evidence that a theorized economic relationship is practically equal to zero. Additionally, there is immense potential for further applications of equivalence testing in placebo tests, which are critical for evidencing identification assumptions but overwhelmingly applied fallaciously (e.g., see Fitzgerald 2025). Equivalence testing places the burden of proof back on the researcher to demonstrate that placebo test results are practically equal to zero before making broader inferences from their statistical findings. There is a wealth of potential for future methodological research on this topic. Finally, ROPE-setting and the TST framework can help ensure that both null results and significant results published in economics are credible and practically relevant. These procedures can be implemented using the `tsti` Stata command (available on SSC) and the `tst` command in the `eqtesting` R package (available on CRAN).

References

Abadie, Alberto (2020). "Statistical nonsignificance in empirical economics". *American Economic Review: Insights* 2.2, pp. 193–208. DOI: 10.1257/aeri.20190252.

Altman, D. G. and J. M. Bland (1995). "Statistics notes: Absence of evidence is not evidence of absence". *BMJ* 311.7003, pp. 485–485. DOI: 10.1136/bmj.311.7003.485.

Andrews, Isaiah and Maximilian Kasy (2019). "Identification of and correction for publication bias". *American Economic Review* 109.8, pp. 2766–2794. DOI: 10.1257/aer.20180310.

Arel-Bundock, Vincent, Noah Greifer, and Andrew Heiss (2024). "How to interpret statistical models using marginaleffects for R and Python". *Journal of Statistical Software* 111.9, pp. 1–32. DOI: 10.18637/jss.v111.i09.

Askarov, Zohid et al. (2023). "Selective and (mis)leading economics journals: Meta-research evidence". *Journal of Economic Surveys,* Forthcoming. DOI: 10.1111/joes.12598.

Berger, Roger L. and Jason C. Hsu (1996). "Bioequivalence trials, intersection-union tests and equivalence confidence sets". *Statistical Science* 11.4. DOI: 10.1214/ss/1032280304.

Bloom, Howard S. (1995). "Minimum detectable effects: A simple way to report the statistical power of experimental designs". *Evaluation Review* 19.5, pp. 547–556. DOI: 10.1177/0193841x9501900504.

Boardman, Anthony E. et al. (2018). *Cost-benefit analysis: Concepts and practice.* 5th ed. Cambridge University Press.

Camerer, Colin F., Anna Dreber, Eskil Forsell, et al. (2016). "Evaluating replicability of laboratory experiments in economics". *Science* 351.6280, pp. 1433–1436. DOI: 10.1126/science.aaf0918.

Camerer, Colin F., Anna Dreber, Felix Holzmeister, et al. (2018). "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015". *Nature Human Behaviour* 2.9, pp. 637–644. DOI: 10.1038/s41562-018-0399-z.

Campbell, Harlan and Paul Gustafson (2018). "Conditional equivalence testing: An alternative remedy for publication bias". *PLOS ONE* 13.4. DOI: 10.1371/journal.pone.0195145.

Campbell, Harlan and Paul Gustafson (2021). "What to make of equivalence testing with a post-specified margin?" *Meta-Psychology* 5. DOI: 10.15626/mp.2020.2506.

Chopra, Felix et al. (2024). "The null result penalty". *The Economic Journal* 134.657, pp. 193–219. DOI: 10.1093/ej/uead060.

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

DellaVigna, Stefano, Devin Pope, and Eva Vivalt (2019). "Predict science to improve science". *Science* 366.6464, pp. 428–429. DOI: 10.1126/science.aaz1704.

Doucouliagos, Hristos (2011). *How large is large? Preliminary and relative guidelines for interpreting partial correlations in economics*. Working Paper SWP 2011/5. Geelong, Australia: Deakin University. URL: https://core.ac.uk/download/pdf/6290432.pdf (visited on 05/13/2024).

Dreber, Anna, Magnus Johannesson, and Yifan Yang (2024). "Selective reporting of placebo tests in top economics journals". *Economic Inquiry,* Forthcoming. DOI: 10.1111/ecin.13217.

Fanelli, Daniele (2012). "Negative results are disappearing from most disciplines and countries". *Scientometrics* 90.3, pp. 891–904. DOI: 10.1007/s11192-011-0494-7.

Fitzgerald, Jack (Feb. 2025). *Manipulation tests in regression discontinuity design: The need for equivalence testing*. DOI: 10.31222/osf.io/2dgrp_v1. URL: https://osf.io/preprints/metaarxiv/2dgrp_v1.

Franco, Annie, Neil Malhotra, and Gabor Simonovits (2014). "Publication bias in the social sciences: Unlocking the file drawer". *Science* 345.6203, pp. 1502–1505. DOI: 10.1126/science.1255484.

Fuster, Andreas, Greg Kaplan, and Basit Zafar (2021). "What would you do with $500? Spending responses to gains, losses, news, and loans". *The Review of Economic Studies* 88.4, pp. 1760–1795. DOI: 10.1093/restud/rdaa076.

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

Goeman, Jelle J., Aldo Solari, and Theo Stijnen (2010). "Three-sided hypothesis testing: Simultaneous testing of superiority, equivalence and inferiority". *Statistics in Medicine* 29.20, pp. 2117–2125. DOI: 10.1002/sim.4002.

Hartman, Erin and F. Daniel Hidalgo (2018). "An equivalence approach to balance and placebo tests". *American Journal of Political Science* 62.4, pp. 1000–1013. DOI: 10.1111/ajps.12387.

Imai, Kosuke, Gary King, and Elizabeth A. Stuart (2008). "Misunderstandings between experimentalists and observationalists about causal inference". *Journal of the Royal Statistical Society Series A: Statistics in Society* 171.2, pp. 481–502. DOI: 10.1111/j.1467-985x.2007.00527.x.

Imbens, Guido W. (2021). "Statistical significance, *p*-values, and the reporting of uncertainty". *Journal of Economic Perspectives* 35.3, pp. 157–174. DOI: 10.1257/jep.35.3.157.

Ioannidis, John P., T. D. Stanley, and Hristos Doucouliagos (2017). "The power of bias in economics research". *The Economic Journal* 127.605. DOI: 10.1111/ecoj.12461.

Isager, Peder M and Jack Fitzgerald (Dec. 2025). *Three-sided testing to establish practical significance: A tutorial*. DOI: 10.31234/osf.io/8y925_32. URL: osf.io/preprints/psyarxiv/8y925_v3.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager (2018). "Equivalence testing for psychological research: A tutorial". *Advances in Methods and Practices in Psychological Science* 1.2, pp. 259–269. DOI: 10.1177/2515245918770963.

McShane, Blakeley B. and David Gal (2016). "Blinding us to the obvious? The effect of statistical training on the evaluation of evidence". *Management Science* 62.6, pp. 1707–1718. DOI: 10.1287/mnsc.2015.2212.

McShane, Blakeley B. and David Gal (2017). "Statistical significance and the dichotomization of evidence". *Journal of the American Statistical Association* 112.519, pp. 885–895. DOI: 10.1080/01621459.2017.1289846.

Nikiforakis, Nikos and Robert Slonim (2015). "Editors' preface: Statistics, replications and null results". *Journal of the Economic Science Association* 1.2, pp. 127–131. DOI: 10.1007/s40881-015-0018-y.

Ofori, Sandra et al. (2023). "Noninferiority margins exceed superiority effect estimates for mortality in cardiovascular trials in high-impact journals". *Journal of Clinical Epidemiology* 161, pp. 20–27. DOI: 10.1016/j.jclinepi.2023.06.022.

Olken, Benjamin A. (2015). "Promises and perils of pre-analysis plans". *Journal of Economic Perspectives* 29.3, pp. 61–80. DOI: 10.1257/jep.29.3.61.

Open Science Collaboration, The (2015). "Estimating the reproducibility of psychological science". *Science* 349.6251. DOI: 10.1126/science.aac4716.

Piaggio, Gilda et al. (2012). "Reporting of noninferiority and equivalence randomized trials". *JAMA* 308.24, pp. 2594–2604. DOI: 10.1001/jama.2012.87802.

Romer, David (2020). "In praise of confidence intervals". *AEA Papers and Proceedings* 110, pp. 55–60. DOI: 10.1257/pandp.20201059.

Schuirmann, Donald 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, pp. 657–680. DOI: 10.1007/bf01068419.

Stanley, T. D. and Hristos Doucouliagos (2012). "Identifying and coding meta-analysis data". *Meta-regression analysis in economics and business*. Ed. by T. D. Stanley and Hristos Doucouliagos. Routledge, pp. 12–37.

Wasserstein, Ronald L. and Nicole A. Lazar (2016). "The ASA statement on p-values: Context, process, and purpose". *The American Statistician* 70.2, pp. 129–133. DOI: 10.1080/00031305.2016.1154108.