The Need for Equivalence Testing in Economics

Jack Fitzgerald, Vrije Universiteit Amsterdam*

October 6, 2024

Abstract

I introduce equivalence testing procedures that can provide statistically significant evidence that economic relationships are practically negligible. I then demonstrate their necessity in a large-scale replication of estimates that defend 135 null claims made in 81 articles from top economics journals. 36-63% of these estimates fail lenient equivalence tests. Though prediction platform data reveals that researchers find these equivalence testing failure rates (ETFRs) to be unacceptably high, researchers actually anticipate unacceptably high ETFRs, accurately predicting that ETFRs exceed acceptable thresholds by around 23 percentage points. To obtain ETFRs that researchers deem acceptable, one must contend that nearly 75% of published effect sizes in economics are practically equal to zero. This implies that Type II error rates are unacceptably high throughout economics. This paper provides economists with empirical justification, 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 thank Abel Brodeur, Katharina Brütt, Eve Ernst, Jelle Goeman, Yi He, Florian Heine, Peder Isager, Nick Koning, Stan Koobs, Andre Lucas, Derek Mikola, Jonathan Roth, Martin Schumann, and Arjen van Witteloostuijn for valuable input on this paper, alongside conference and seminar participants from the European Commission CC-ME COMPIE Conference, KVS New Paper Sessions, MAER-Net Colloquium, PhD-EVS Seminar, Technische Universiteit Eindhoven, Tinbergen Institute, and Vrije Universiteit Amsterdam for comments and feedback. I also thank the multiple authors who answered my questions about their research and replication data. I am grateful to the Amsterdam Law and Behavior Institute for financial support. At time of writing, I currently hold 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 can be found at https://jack-fitzgerald.github.io/files/The_Need_for_Equivalence_Testing_in_Economics_Online_Appendix.pdf.

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 over 72% of article abstracts in top economics journals report such results by claiming that the two variables have no meaningful relationship at all. Readers also interpret such findings in this way, including researchers and even statisticians (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 the standard null hypothesis significance testing (NHST) framework, 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 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 why economics must adopt equivalence testing as well, and demonstrates how to credibly apply this testing framework.

I conduct a large-scale replication to show that the standard testing procedures economists use to make and defend null claims likely tolerate unacceptably high Type II error rates. I reproduce 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 the estimates in my replication sample perform under equivalence testing, I set symmetric ROPEs with boundaries based on Cohen's (1988) widely-used small effect size benchmarks. These ROPEs are quite lenient, with boundaries larger than a substantial proportion of published estimates in economics (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. I estimate 'equivalence testing failure rates (ETFRs)' by computing the proportion of estimates that 'fail' these lenient tests.

Equivalence testing failure rates are unacceptably high. At a 5% significance level, ETFRs within these lenient ROPEs range from 36-63%. To obtain ETFRs that my prediction platform sample considers acceptable, one must be willing to claim that nearly 75% of all published effect sizes in economics are practically equal to zero. Because such a claim is ludicrous, 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%, roughly in line with the lower bound of my actual ETFR estimates. On average, researchers expect ETFRs to exceed their personal acceptability thresholds by around 23 percentage points. Though researchers distrust many null results in the current economics literature, this mistrust appears to be relatively well-placed. These results imply that equivalence testing is a necessary addition to methodological practice in economics.

Given the clear need for equivalence testing in economics, I provide guidelines for conducting credible equivalence testing in economic research. To minimize researcher

degrees of freedom and 'ROPE-hacking', I recommend that researchers aggregate ROPEs by surveying independent parties, such as experts or relevant stakeholders, regarding the smallest relationships that they would deem practically meaningful. Such surveys are practically feasible to conduct using research-centric belief elicitation platforms such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019). I also introduce 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 I discuss in this paper require researchers to concede that their results are 'inconclusive'. This ensures that imprecise estimates are not misinterpreted as definitive evidence of null relationships.

Finally, I 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. Both the `tsti` command and the `eqtesting` package can be downloaded from Github.[1]

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

[1]For `tsti`, see https://github.com/jack-fitzgerald/tsti, and for `eqtesting`, see https://github.com/jack-fitzgerald/eqtesting.

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*).[2] The systematic selection procedure is detailed in Online Appendix A. 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 are computationally reproducible using publicly-available data.[3] The final sample comprises 876 estimates that defend 135 null claims made in the abstracts of 81 articles. For each estimate, the final sample stores the corresponding standardized regression coefficient σ, standard error s, sample size N, residual degrees of freedom df,[4] replicability status, conformability status, outcome and exposure variables (with dummies indicating if each is binary), and the initial standard NHST p-value (without conformability changes, if applicable). The standardization procedure for σ and s is detailed in Section 5.1. In Online Appendix B, I provide the articles represented in the final sample, alongside additional data repositories attached to these articles (when applicable). In Online Appendix C, I provide the articles in the intermediate sample that are excluded from the final sample.

Table 1 displays summary statistics for the final sample. The majority of articles make only one null claim, and over 90% make between one and three null claims.

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

[3]For the purposes of this paper, 'publicly-available' data includes data stored in repositories of the Inter-university Consortium for Political Science Research (ICPSR), whose data is freely available to anyone who creates an ICPSR account.

[4]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 median null claim 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 $df = 180$.[5] At a 5% significance level, the majority of these estimates have at least 80% power to detect an effect size of 0.2σ under the standard NHST framework. However, there is a concentrated group of underpowered estimates. 32% of estimates in the final sample lack even 50% power to detect a 0.2σ effect. 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.[6].

There are also several important binary variables whose summary statistics are not reported in Table 1. 8.3% of estimates in the final sample are not fully replicable; i.e., my best attempts to reproduce the article's findings using its replication repository

[5]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.

[6]One claim – the only null claim in its article – is defended with a single statistically significant result (Fuster, Kaplan, & Zafar 2021). Initially significant estimates are more common for 'directional' null claims. E.g., see categories 3 and 6 in Table 2.

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.[7] 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 $Y = \delta D + X\phi$, 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 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* $\alpha \in (0, 1]$*. They formulate null and alternative hypotheses as*

$$\begin{aligned} H_0 &: \delta = 0 \\ H_A &: \delta \neq 0 \end{aligned} \tag{1}$$

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

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

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

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

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

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* $|t_{NHST}| \geq t^*_{\frac{\alpha}{2},\ df}$.

Economists using the standard NHST framework typically conclude that there is a relationship between Y and D if H_0 is rejected, or 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.[8] Table 2 shows that economists frequently make null claims based on statistically insignificant estimates. Though Gates & Ealing (2019) show that this practice is not unique to economics, a striking feature of the way that economists communicate null claims is how definitively the claims are made. Fewer than 28% of such claims are qualified with references to statistical significance, estimate magnitude, or a lack of evidence. More than 72% of all null claims in the intermediate sample are in this sense 'unqualified'. These unqualified null claims are unambiguous assertions that the relationship of interest is negligible or nonexistent.

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 a null result from a statistically insignificant estimate can often result in erroneously concluding that a genuinely meaningful relationship does 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.[9] Type II error rate $\beta \in (0, 1]$ relates to the power $(1-\beta)$ that a model has to detect a relationship with a magnitude of at least $\epsilon \geq 0$ under the standard NHST framework. 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 $\left|\hat{\delta}\right| \geq \epsilon$. Let $F_\alpha(t, df)$ represent the CDF of the noncentral t-distribution with df degrees of freedom and noncentrality

[8] 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.

[9] Of course, when multiple hypothesis tests are performed simultaneously, false positive rates can exceed α. The subsequent analysis remains valid in the special case where only one hypothesis test is performed.

parameter $t^*_{\alpha,df}$, where $t^*_{\alpha,df}$ is defined in Equation 2. Then given α, power to detect an effect size of $|\delta| \geq \epsilon$ can be written as[10]

$$\begin{aligned} 1 - \beta &= \Pr\left(|t_{\text{NHST}}| \geq t^*_{\frac{\alpha}{2},df} \mid |\delta| \geq \epsilon\right) \\ &= F_{\frac{\alpha}{2}}\left(\frac{\epsilon}{s}, df\right) + F_{\frac{\alpha}{2}}\left(-\frac{\epsilon}{s}, df\right). \end{aligned} \tag{3}$$

Power levels above 0.8 are generally considered to sufficient in the social sciences, whereas power levels below 0.8 are considered insufficient (Ioannidis, Stanley, & Doucouliagos 2017). The classical thresholds of $\alpha = 0.05$ and $\beta = 0.2$ imply that Type I errors are four times as costly as Type II errors (Cohen 1988, pg. 56). Because one can never achieve adequate power for $\epsilon = 0$, the researcher must choose a reasonable effect size benchmark $\epsilon > 0$ to calculate power. When $\hat{\delta}$ is statistically insignificant, ϵ is ordinarily set to a small effect size benchmark, as the goal of power analysis in this setting is typically to assess whether $\delta < \epsilon$ with high probability. 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, power levels are usually remarkably low throughout the economics literature. As discussed in Section 2, 32% of the estimates in my final sample lack even 50% power to detect a 0.2σ effect. 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 and 5% in Top 5 economics journals.

These low power levels are not necessarily due to poor research practices, and can naturally arise from the inherent constraints of economic research. Answering important economic questions often requires researchers to work with pre-existing

[10]This is simply a generalized extension of the power equation for a two-sided test employed by Stata's `power oneslope` command (StataCorp 2023, pg. 433).

datasets whose data is generated through a process that the researcher cannot control. As a result, economists are often 'at the mercy' of existing sample sizes, and usually cannot summon new data at will to meet power constraints.

This low power challenges the credibility of null claims in economics. When a researcher interested in claiming 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 is the key motive for 'reverse p-hacking', a common practice in placebo tests where null results are desirable (Dreber, Johanneson, & 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 the power of the model can result in researchers unwittingly tolerating unacceptably 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 due to small size or due to imprecision under the standard NHST framework, null results are frequently

conflated with imprecise results, which contributes to widespread publication biases. Researchers assign strong 'null result penalties', viewing null results as low-quality and unpublishable (see McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024). This in turn leads to null results being far less likely to be published in economics journals (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019). Even amongst the null findings that are published, the high Type II error rates effectively tolerated by current economic practice imply that many null findings in economics are 'false negatives' that wrongfully declare meaningful economic relationships to be nonexistent.

Fortunately, testing frameworks that provide better error control for null results can mitigate or eliminate all of these problems. If researchers inherently understand these dynamics in the current research landscape, then aesthetic preferences for pattern-finding may not entirely explain the null result penalty (see Chopra et al. 2024). Rather, the null result penalty may partly reflect rational preferences for minimizing error rates. Therefore, testing frameworks that provide better error control for null claims may yield the added benefit of mitigating the null result penalty, and could in turn reduce publication bias against null results.

4 Equivalence Testing

A credible framework for testing whether relationships are practically null can be constructed by making two adjustments to the standard NHST framework. 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

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

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.

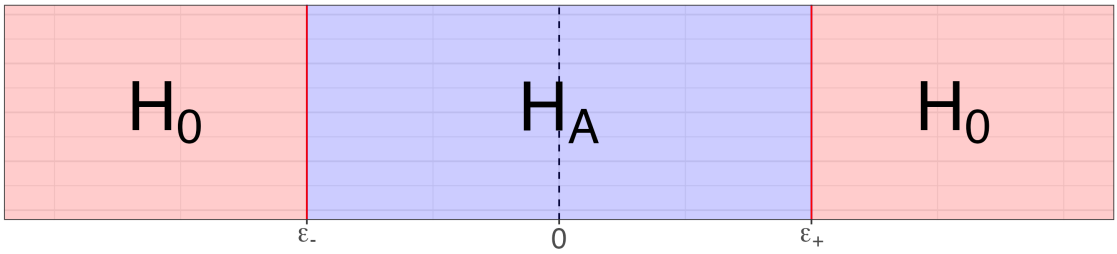
Definition 4.1 (The Equivalence Testing Framework)**.** *The researcher wants to test whether* $\delta \approx 0$*. Let* $[\epsilon_-, \epsilon_+]$ *be a range where* $\epsilon_- < \epsilon_+$*, where* $0 \in [\epsilon_-, \epsilon_+]$*, and where* $\delta \approx 0$ *when* $\delta \in [\epsilon_-, \epsilon_+]$*.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

$$\begin{aligned} H_0 &: \delta \notin [\epsilon_-, \epsilon_+] \\ H_A &: \delta \in [\epsilon_-, \epsilon_+]\,. \end{aligned} \tag{4}$$

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 discuss in Section 7.1, these boundaries can be credibly set by surveying other researchers for their independent judgments on the smallest effect size of interest for a given relationship. ROPEs are often symmetric around zero such that $\epsilon_- = -\epsilon_+$, but ROPEs can also be asymmetric.[11] 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

[11] For instance, asymmetric ROPEs can arise when estimates of interest are mechanically bounded above or below zero. Asymmetric ROPEs can also arise when D represents a costly intervention chosen from among many. If the goal of such interventions is to increase Y, then even small negative effects of D become practically meaningful after factoring in the opportunity cost of abandoning other interventions. In this context, it may be reasonable to set the ROPE such that $|\epsilon_-| < |\epsilon_+|$.



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

Figure 1: Visualization of the Equivalence Testing Hypothesis Framework

bounded within the ROPE throughout the rest of this section.

4.1 Two One-Sided Tests Procedure

The hypotheses in Equation 4 can be rewritten as

$$\begin{aligned} H_0 &: \delta < \epsilon_- \quad \text{or} \quad \delta > \epsilon_+ \\ H_A &: \delta \geq \epsilon_- \quad \text{and} \quad \delta \leq \epsilon_+. \end{aligned}$$

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

$$\begin{aligned} H_0 &: \delta < \epsilon_- & H_0 &: \delta > \epsilon_+ \\ H_A &: \delta \geq \epsilon_- & H_A &: \delta \leq \epsilon_+. \end{aligned} \tag{5}$$

Under Definition 4.1, statistically significant evidence that $\delta \approx 0$ can be obtained by showing statistically significant evidence against both H_0 statements in Equation 5. This is the principle underlying 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*

$$t_- = \frac{\hat{\delta} - \epsilon_-}{s} \qquad t_+ = \frac{\hat{\delta} - \epsilon_+}{s} \tag{6}$$

and compute

$$t_{TOST} = \underset{t \in \{t_-, t_+\}}{\arg\min} \{|t|\}. \tag{7}$$

The exact critical value for this test can be written as

$$t^*_{\alpha,df} = F^{-1}(1-\alpha, df). \tag{8}$$

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 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 procedure's name and modern form are established by Schuirmann (1987), who shows that the TOST procedure has more power to significantly bound estimates than the traditional 'power approach' discussed in Section 3. 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 (Schuirmann 1987; Berger & Hsu 1996; Lakens, Scheel, & Isager 2018).

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)**.** *The researcher wants to test the hypotheses in Definition 4.1 using a size-α test. They thus formulate a real interval $\left[\Delta_-, \Delta_+\right]$, where Δ_- and Δ_+ are computed as follows:*

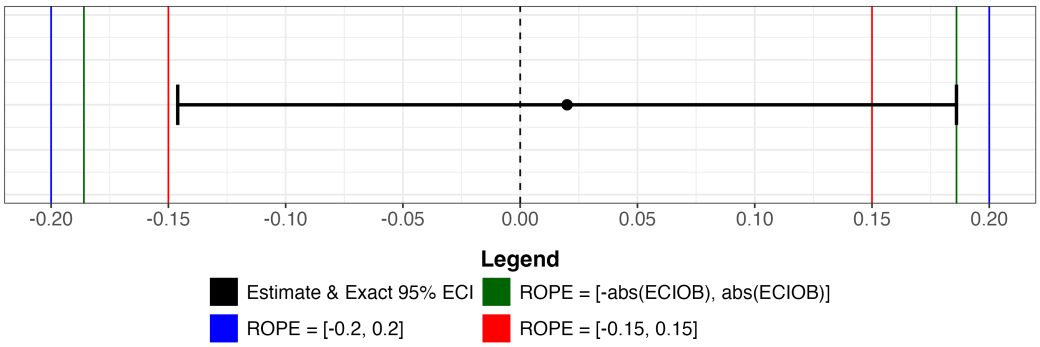
$$\begin{aligned} \Delta_-(1-\alpha, df) &= \hat{\delta} - \left(s \times t^*_{\alpha,df}\right) \\ \Delta_+(1-\alpha, df) &= \hat{\delta} + \left(s \times t^*_{\alpha,df}\right). \end{aligned} \tag{9}$$

$t^*_{\alpha,df}$ is defined as in Equation 8. 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 straightforward. For example, the 95% ECI is simply the 90% confidence interval. The key differences between ECIs and confidence intervals lie in how they can be used to judge statistical significance. In the standard NHST framework, 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 statistically 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 thus be roughly written as $[-0.146, 0.186]$, as $t^*_{0.05,100} \approx 1.66$. If the ROPE is set as $[-0.2, 0.2]$, then $\hat{\delta}$ is statistically significantly bounded within the ROPE at a 5% significance level, 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



Note: The coefficient of the estimate in this figure has arbitrary scale.

Figure 2: An ECI Example

is 0.186, because the upper bound of this estimate's ECI is further away 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. This makes ECIOB magnitudes interesting for many applied economists, as the ECIOB magnitude is the smallest effect size that can be 'ruled out' with statistically significant evidence.

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).

The second effect size 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.[12]

As Section 5.2 details further, equivalence testing failure rates measure how often the magnitudes of estimates in the final sample can be significantly bounded beneath classical benchmarks. I specifically use Cohen's (1988) small effect size benchmarks, separately testing whether $\sigma \in [-0.2, 0.2]$ and $r \in [-0.1, 0.1]$. These ROPEs are quite lenient. $|r| = 0.1$ is larger than more than 25% of all published estimates in economics (Doucouliagos 2011), and Online Appendix D shows that both $|r| = 0.1$ and $|\sigma| = 0.2$ are large effect sizes even amongst a benchmark sample of plausibly large economic effects. Therefore, when an article in a top economics journal claims that a relationship is null or negligible, showing that the estimates defending that claim are significantly bounded beneath $|\sigma| = 0.2$ or $|r| = 0.1$ should be easy, as these are lenient thresholds.

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

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 statistically significantly bounded within a given ROPE at a 5% significance level for a given aggregation level. To make this concrete, consider a toy dataset of estimates defending three null claims. Suppose that 20% of estimates defending the first claim cannot be significantly bounded within a ROPE of $[-0.2\sigma, 0.2\sigma]$ at a 5% significance level, and that the same is true of all estimates defending the second claim and no estimates defending the third claim. In this toy dataset, the average claim-level ETFR for a ROPE of $[-0.2\sigma, 0.2\sigma]$ would be $(20\% + 100\% + 0\%)/3 = 40\%$.

I calculate average claim-level and article-level ETFRs. I also compute an average inverse-weighted claim-level ETFR that ensures all articles receive the same weight in the sample. Because these average ETFRs are calculated by taking a mean of partition-level ETFRs over all partitions, my precision measure is the standard error of that mean. Online Appendix G provides precise computational details for partition-level ETFRs and their standard errors.

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).[13] The SSPP survey asks social science researchers to provide their predictions and judgments concerning equivalence testing results in the final sample. Specifically, I ask respondents to predict ETFRs in the final sample for a ROPE of $[-0.2\sigma, 0.2\sigma]$ at a 5% significance level. Thereafter, I ask respondents to provide the smallest ETFRs that they would consider to be acceptable. To minimize confusion, I then ask each respondent whether they anticipate

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

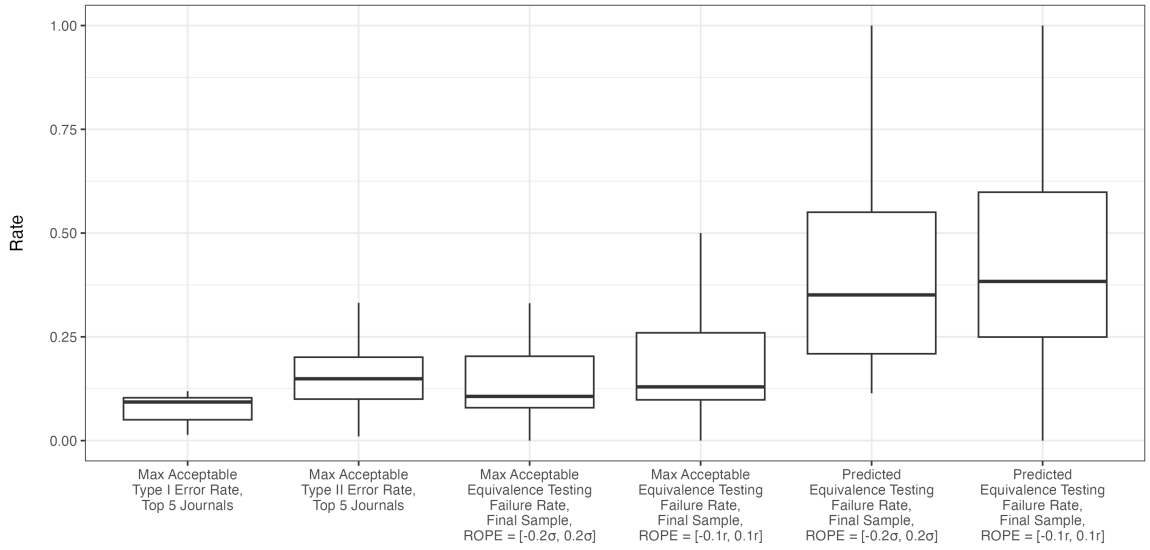
that these ETFRs will differ within a ROPE of $[-0.1r, 0.1r]$. If they answer 'yes', then they are asked to provide these same predictions and judgments of ETFRs within a ROPE of $[-0.1r, 0.1r]$. If they answer 'no', then they are not shown these new questions, and their predictions and judgments of ETFRs within a ROPE of $[-0.1r, 0.1r]$ are imputed based on their responses for the ROPE of $[-0.2\sigma, 0.2\sigma]$. 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 the results of my analysis 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 F.

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). 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 ROPE of $[-0.2\sigma, 0.2\sigma]$. This ETFR tolerance increases to 12.95% for a ROPE of $[-0.1r, 0.1r]$. However, respondents predict that ETFRs will substantially exceed these thresholds. Median predictions for ETFRs are 35.1% for a ROPE of $[-0.2\sigma, 0.2\sigma]$ and 38.35% for a ROPE of $[-0.1r, 0.1r]$. Section 6.3 shows that these



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.

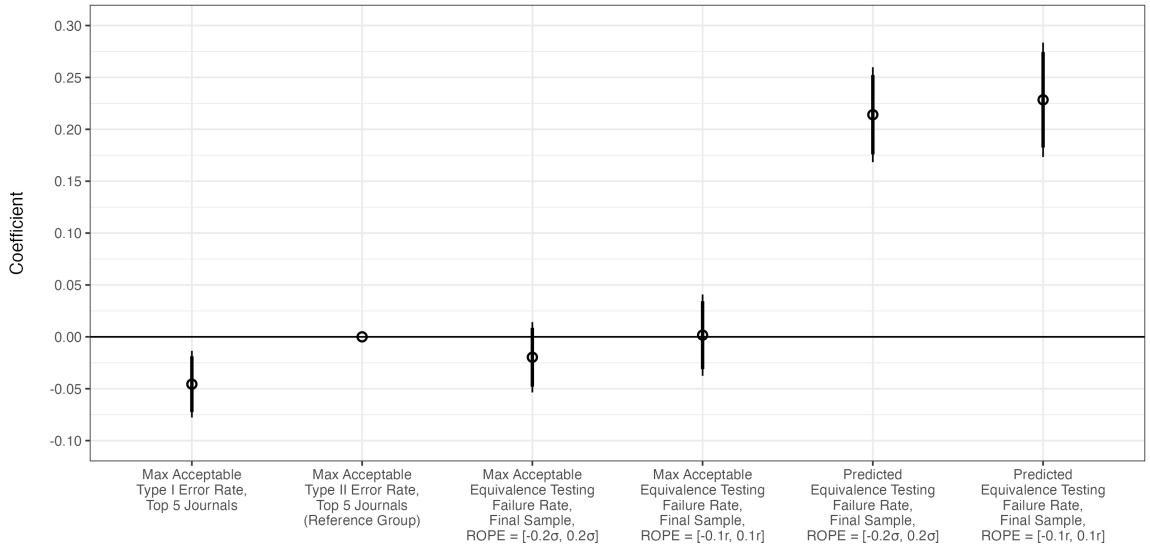
Figure 3: Distributions of SSPP Predictions and Judgments

predictions are fairly accurate, although the median ETFR prediction for a ROPE of $[-0.1r, 0.1r]$ is an underestimate.

Figure 3 displays large dispersion in predicted ETFRs, which reflects both between-respondent disagreement and relatively low power in my SSPP sample ($N = 62$). Fortunately, the within-subject design of my survey allows much greater power to be achieved by constructing a respondent-rate panel dataset. This panel dataset also makes it possible 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 better-powered within-respondent estimates of differences between rates. Specifically, Figure 4 shows γ_r estimates from a model of Equation 12



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 (see Cameron & Miller 2015).

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

that treats judgments on Type II error rates as the reference group.[14] 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 direct evidence of a preference-based null result penalty (see Chopra et al. 2024). Researchers in my SSPP sample care more about Type I errors than Type II errors, implying that they care more about articles in top economics journals claiming that relationships exist than about such articles claiming that relationships do not exist.

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 2 percentage points higher than their tolerance for ETFRs within a ROPE of $[-0.2\sigma, 0.2\sigma]$, and is 0.2 percentage points lower than their tolerance for ETFRs within a ROPE of $[-0.1r, 0.1r]$. Though one could use equivalence testing to significantly bound these two estimates within a five percentage point difference of Type II error

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

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.

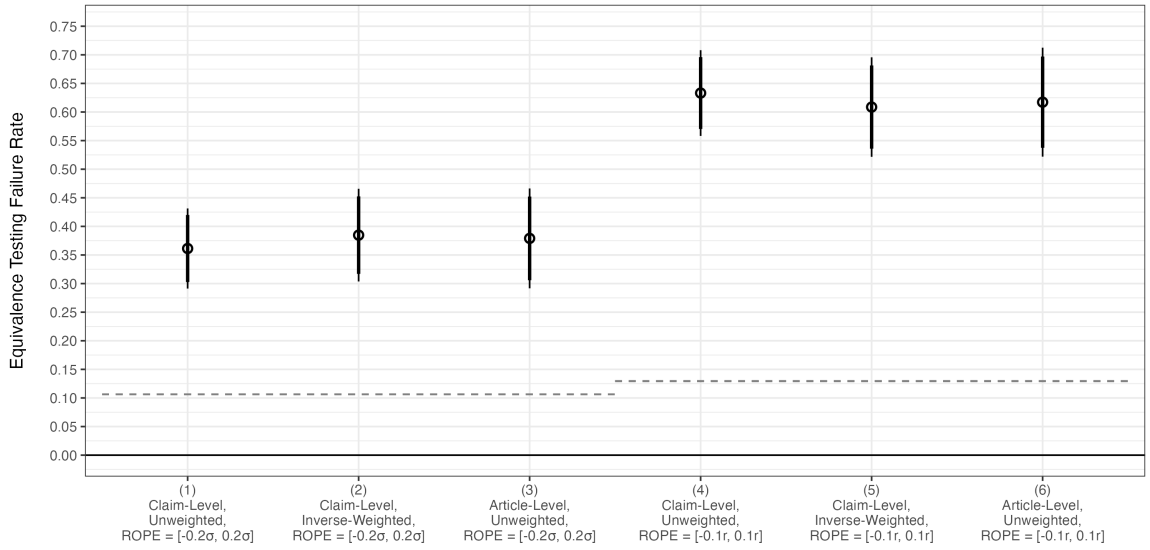
However, researchers predict that ETFRs in my final sample will far exceed any of these acceptability thresholds. The average respondent predicts that ETFRs will exceed their personal Type II error rate tolerance by 21.4 percentage points within a ROPE of $[-0.2\sigma, 0.2\sigma]$, and by 22.8 percentage points within a ROPE of $[-0.1r, 0.1r]$. 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 show that this prediction is quite accurate.

6.2 Equivalence Testing Failure Rates

Figure 5 displays the main ETFR estimates.[15] The dashed lines represent the median SSPP respondent's thresholds for acceptable ETFRs (see Section 6.1). ETFRs are significantly above both zero and these thresholds. For a ROPE of $[-0.2\sigma, 0.2\sigma]$, ETFRs range from 36.1-38.5%. These ETFRs are even higher for a ROPE of $[-0.1r, 0.1r]$, ranging from 60.9-63.3%. Therefore, equivalence testing failure rates within lenient ROPEs range from 36-63% for recent null claims in top economics journals.

The significance of these ETFRs is robust to a wide range of checks. Principally, ETFRs are not sensitive to the choice of aggregation procedure. Within each effect size measure, ETFRs vary by less than 2.5 percentage points across aggregation levels. Further, no single aggregation strategy is uniformly stricter or more lenient than another. Giving all articles the same weight, either by using article-level ETFRs or by applying inverse weighting, increases ETFRs for standardized coefficients

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

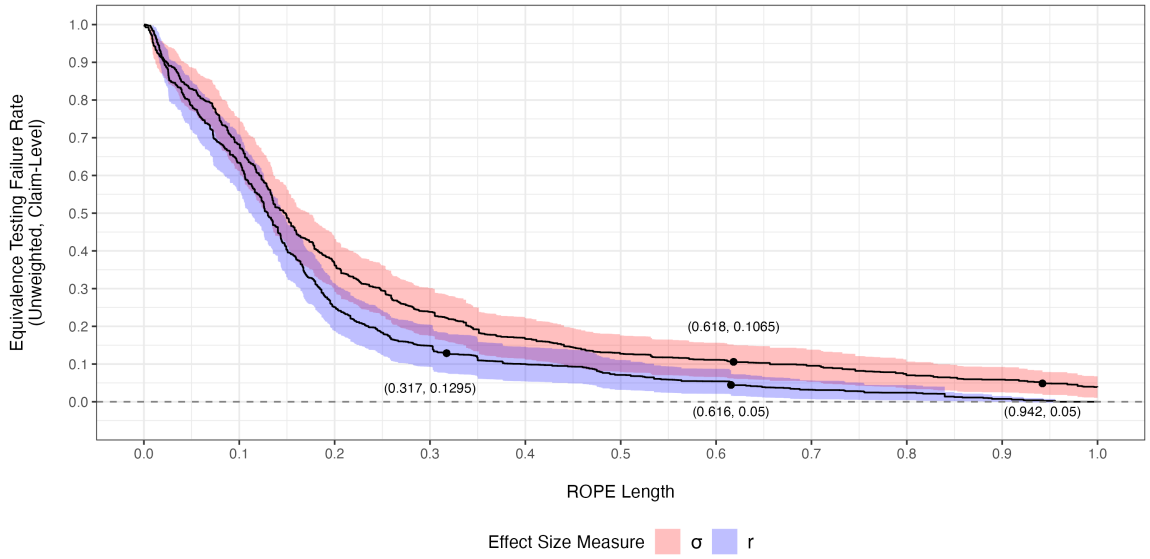


Note: ETFRs are provided along with 95% ECIs (thicker bands) and confidence intervals (thinner bands). These intervals are based on the standard error of the mean for unweighted ETFRs and the weighted standard error of the mean for weighted ETFRs (see Online Appendix G). Dashed lines represent the median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

Figure 5: Main Equivalence Testing Failure Rate Estimates

but decreases ETFRs for partial correlation coefficients. I thus primarily reference unweighted claim-level ETFRs when I discuss results, largely because they are relatively easy to interpret. For instance, Model 1 in Figure 5 implies that 36.1% of estimates defending the average null claim in the final sample cannot be significantly bounded beneath a 0.2σ effect.

Online Appendix Table A4 shows that ETFRs remain significantly bounded above acceptability thresholds regardless of whether I remove estimates from the sample that are initially statistically significant under the standard NHST framework. Additionally, Online Appendix Table A5 shows that ETFRs remain significantly above acceptability thresholds after employing a leave-one-out approach where subsamples of regressor type combinations are removed from the sample. Finally, Online Appendix Table A6 shows that ETFRs are robust to coding choices. Using the same leave-one-out approach, I show that ETFRs remain significantly above acceptability thresholds after removing estimates that are not fully replicable, and after removing



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 5's estimates. Uncertainty bands represent 95% confidence intervals based on the claim-level ETFR's standard error of the mean (see Online Appendix G).

Figure 6: Failure Curves

estimates from models that require conformability modifications.

6.3 Failure Curves

Perhaps the most important sensitivity check concerns the choice of ROPE. Figure 6 plots 'failure curves', which show how claim-level ETFRs vary with the choice of ROPE length ϵ. The shapes of the failure curves reflect the intuition that ETFRs decline when one is willing to tolerate larger ROPEs. Figure 6 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 sufficient, and that ETFRs are bounded below 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?

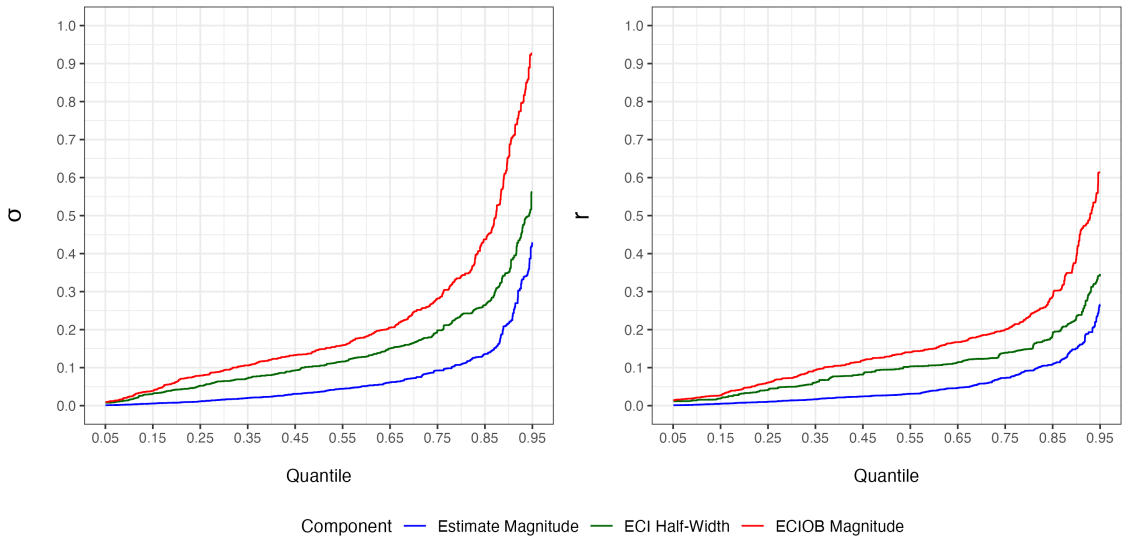
Figure 6's annotated points show that one must tolerate exceedingly large ROPEs to obtain acceptable ETFRs. As discussed in Section 6.1, the median SSPP respondent's maximum ETFR tolerance for a ROPE of $[-0.1r, 0.1r]$ is 12.95%. To obtain claim-level ETFRs beneath 12.95%, one must set a ROPE of $[-0.317r, 0.317r]$, which implies that one must argue that $|r| = 0.317$ is practically equal to zero. However, $|r| = 0.317$ is larger than nearly 75% of published results in economics (Doucouliagos 2011). To obtain claim-level ETFRs beneath 5%, one must be willing to claim that $|r| = 0.616$ is practically equal to zero, which is an extremely large effect size.

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 ROPE of $[-0.2\sigma, 0.2\sigma]$ 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 D shows that these magnitudes are large even in a sample of plausibly large economic effects.

It is absurd to argue that effect sizes this large are practically equal to zero. Given this, one is compelled to accept the more sensible alternative conclusion that the current testing paradigm that 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 statistically significantly bound $\hat{\delta}$. ECIOB magnitudes thus directly determine the ROPEs within which $\hat{\delta}$ fails



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

Figure 7: Inverse CDFs of ECIOB Magnitudes and Their Components

to be statistically significantly bounded. Therefore, ECIOB magnitudes directly determine ETFRs for a given ROPE. The mechanisms of ETFRs can thus 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 7 plots the distributions of 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.

Table 3 provides further evidence that imprecision is the main driver of ETFRs. Specifically, Table 3 displays constant elasticity estimates of the relationships between 95% ECIOB magnitudes, effect sizes, and 95% ECI half-widths. Both effect sizes and

	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 via a weighted univariate linear regression where the dependent variable is the ECIOB 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. 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

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. For standardized coefficients, the elasticity of ECIOB magnitudes with ECI half-widths is around 16% larger than that elasticity for effect size $|\sigma|$. For partial correlation coefficients, the elasticity of ECIOB magnitudes with ECI half-widths is around 127% larger than that elasticity for effect size $|r|$. 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 the empirical properties of equivalence testing. Section 3 notes that a key credibility issue with the standard NHST framework when the researcher wants to show that $\delta = 0$ is that imprecision is 'good'. This is because there is an inverse relationship between precision and the probability of obtaining a null result under the standard NHST framework. However, the second and fourth columns in Table 3 show that when using equivalence testing, one can bound an estimate significantly closer to zero when one has more precise estimates. This shows that when the researcher is trying to show a lack of association, equivalence testing restores the proportional relationship between precision and the probability of reaching this conclusion. This in turn demonstrates that in such

research settings, equivalence testing addresses many of the problems discussed in Section 3 by eliminating the conflation between imprecision and null findings.

7 The Future of Equivalence Testing in Economics

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 estimates of interest. Given the clear need for equivalence testing in economics, the rest of this section is dedicated to showing researchers how they can employ credible 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, so 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 estimate of in-

terest, 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, Johanneson, & Yang 2024). This is a problem that even pre-registration cannot fix, as researchers interested in obtaining evidence of null findings can simply pre-register an excessively wide ROPE. Unsurprisingly, this practice can inflate error rates in equivalence testing (Campbell & Gustafson 2021).

To control researcher degrees of freedom and ensure credible, independently-set significance thresholds, I recommend that researchers set ROPEs by eliciting judgments on minimal meaningful effect sizes from independent parties, such as experts or relevant stakeholders. Such judgments are practical to elicit using recently-developed 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 survey 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?" 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, which 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 additional prediction data will still be useful to help inform posterior beliefs, and to evidence the novelty of research findings (DellaVigna, Pope, & Vivalt 2019). Of course, other survey platforms (for example, Qualtrics) are also appropriate for such belief elicitation. The only strict requirement is that the researcher has a credible sample of experts or stakeholders who can provide effect size judgments.

A key advantage of this ROPE-setting strategy is that even *post hoc*, non-pre-registered surveys can provide credible, independent 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). As for other analyses, the advantage of pre-registering ROPEs 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.

7.2 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 equiv-

alence 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 described by Campbell & Gustafson (2018).

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.*

The CET procedure is not ideal. In highly-powered research settings, $\hat{\delta}$ can simultaneously be 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 research conclusions in this setting. The CET framework would deem $\hat{\delta}$ 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 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 statistically significant evidence that the estimate is different from zero nor to find statistically 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. 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 4.1 along with test statistics

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

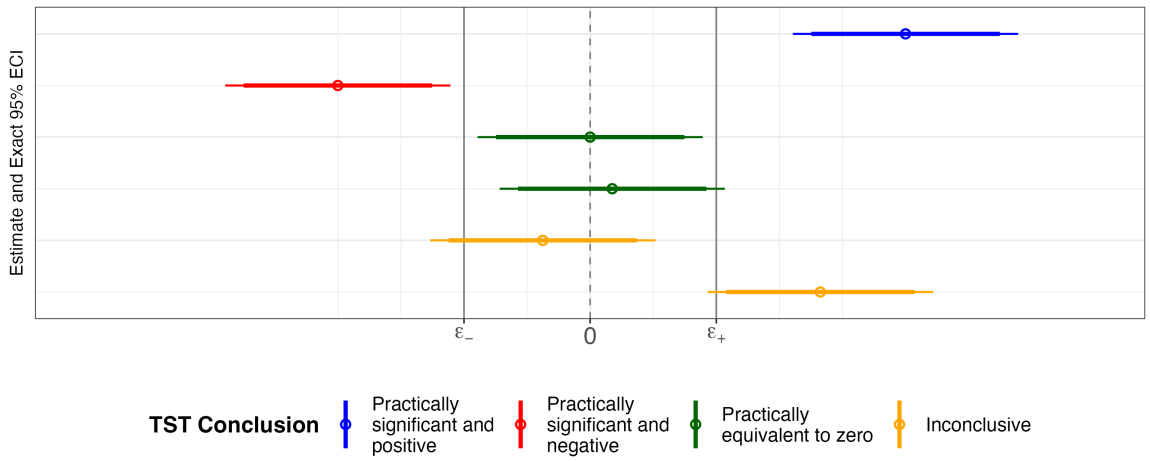
The researcher concludes that δ is significantly bounded above the ROPE if and only if $t_P > t^*_{\alpha/2,df}$. The researcher concludes that δ is significantly bounded below the ROPE if and only if $t_N < -t^*_{\alpha/2,df}$. As in Definition 4.1, if $t_{TOST} = t_-$, then the researcher concludes that δ is significantly bounded within the ROPE 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 statistically significantly different from zero unless there is statistically significant evidence that $\hat{\delta}$ is practically different from zero. 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 (Goeman, Solari, & Stijnen 2010). 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 8 illustrates how TST conclusions can be derived using a confidence interval approach. The top four estimates shown in Figure 8 depict estimates for which the researcher can make highly certain practical significance conclusions. 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 8 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 8 is particularly important for understanding how TST augments the standard NHST framework. 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 pre-



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

Figure 8: Research Conclusions in the TST Framework

cision 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 confidence intervals are fully bounded outside of the ROPE, such as the top two estimates, to be practically significant. These sorts of estimates are large and precise enough to instill strong confidence that these relationships are practically meaningful.

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. Because this claim is ludicrous, it is clear that economists' current testing practices for making and defending null claims 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 elicit independent judgments of the smallest practically important effect size for each relationship that they are interested in estimating. These judgments can either be elicited from other experts or from relevant stakeholders, and are practical to aggregate using centralized research-centric survey platforms such as the Social Science Prediction Platform (DellaVigna, Pope, & Vivalt 2019).

The ROPEs constructed from these judgments can then be used to test estimates using the three-sided testing framework, which 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 ensures that relationships are not deemed *statistically* significant unless there is credible evidence that such relationships are *practically* significant. 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 a myriad of positive effects on research findings in the economics literature. Credible equivalence testing can help assuage existent concerns about the quality and publishability of null results, helping reduce publication bias against null results 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. 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 testing procedures can be implemented using the `tsti` Stata command and the `tst` command in the `eqtesting` R package.[16]

[16]To access the repositories for both software suites, see https://github.com/jack-fitzgerald.

References

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.

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.

Cameron, Colin A. and Douglas L. Miller (2015). "A practitioner's guide to cluster-robust inference". *Journal of Human Resources* 50.2, pp. 317–372. DOI: 10.3368/jhr.50.2.317.

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.

— (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://www.deakin.edu.au/__data/assets/pdf_file/0003/408576/2011_5.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.

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.

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.

— (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.

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.

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.

StataCorp (2023). *Stata power, precision, and sample-size reference manual*. Vol. 18. Stata Press. URL: https://www.stata.com/manuals/pss.pdf.

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.