The Need for Equivalence Testing in Economics

Jack Fitzgerald, Vrije Universiteit Amsterdam*

October 3, 2024

Abstract

I introduce equivalence testing procedures that can provide statistically significant evidence that economic relationships are practically equal to zero. I then demonstrate their necessity by systematically reproducing the 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 wants to know the relationship between two variables, so they run a regression. As it turns out, the regression estimate is not statistically significantly different from zero. Assuming that this finding 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 a finding by claiming that there is 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-known to be 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*. Under this framework, the researcher first sets a *region of practical equivalence (ROPE)* around zero, denoting the range of values for the relationship of interest that are 'practically equal to zero', or in economic parlance, 'economically insignificant'. Once the ROPE is set, equivalence testing assumes in the null hypothesis that the estimate is *not* bounded within the ROPE. If the estimate is significantly bounded within the ROPE, then one has credible evidence that the relationship of interest is practically equal to zero. Equivalence testing is routinely applied in medicine, and is being rapidly adopted by 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 show that the standard testing procedures that economists use to make and defend null claims likely tolerate unacceptably high Type II error rates. In particular, I systematically reproduce and standardize the estimates that defend 135 null claims

made by 81 articles published in Top 5 economics journals from 2020-2023, and subject these estimates to equivalence testing. 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 the performance of these estimates under equivalence testing, I set symmetric ROPEs with boundaries defined by Cohen's (1988) widely-used small effect size benchmarks. These are very lenient ROPEs, 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.

ETFRs 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 deems 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 deems 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 acceptable levels by around 23 percentage points. Though researchers distrust many null results in the current economic literature, this mistrust appears to be relatively well-placed. These results together imply a strong need for equivalence testing in future economic research.

Given this clear need, I provide guidelines for credible equivalence testing in economic research. To reduce researcher degrees of freedom and 'ROPE-hacking', I rec-

ommend that researchers aggregate ROPEs by surveying independent parties, such as experts or relevant stakeholders, regarding the smallest relationships that they would consider to be practically meaningful. Such surveys are practical to conduct using centralized 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 practically significant *or* practically equal to zero. In such cases, the testing frameworks I advocate for in this paper require that researchers concede that their results are *inconclusive*. This ensures that imprecise estimates are not considered definitive evidence of null relationships. It also ensures that relationships are only considered *statistically* significant when there is highly certain evidence that they are *practically* significant.

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 data underlying my empirical analysis. In Section 3, I leverage this data to document problems with current economic practice for evidencing null claims; Section 4 provides equivalence testing frameworks and procedures that correct for these issues. Section 5 provides methodological details for my empirical analysis, and Section 6 details my empirical results. Section 7 offers 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 my 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 term this full sample of articles, claims, and estimates the *intermediate sample*.

The *final sample* contains all estimates in the intermediate sample that are conformable and computationally reproducible using publicly-available data.[3] The final sample is comprised of 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 sample of articles represented in the final sample, alongside additional data repositories attached to these articles (when applicable). In Online Appendix C, I provide the sample of articles in the intermediate sample that are excluded from the final sample.

Table 1 displays summary statistics. The majority of articles make only one null claim, and more than 90% make between one and three null claims. The median null

[2]This includes articles not yet published in print, but digitally published as corrected proofs at the time of 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 of 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

claim is defended by four estimates. Effect sizes are quite small throughout the final sample, 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 sample 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 virtually always arise alongside other statistically insignificant estimates that together defend their null claim.[6] Initially significant estimates are more common for null claims made about directional hypotheses.

There are also a few important binary variables whose summary statistics are not reported in Table 1. 8.3% of estimates in the final sample are not fully replicable, in the sense that my best attempts to reproduce the article's findings using its replica-

[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 statistically significant result (Fuster, Kaplan, & Zafar 2021).

tion 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, implying that the model used to obtain the estimate in the final sample differs from the model used to obtain the estimate in the published article.[7] Both the outcome and exposure variable are continuous for 22.9% of estimates in the final sample, while 25.5% of estimates in the final sample correspond to binary outcome and exposure variables. The most frequent type of estimate corresponds to a continuous outcome variable and a binary exposure variable, representing 35.7% of estimates in the final sample.

2.1 Prediction Platform Data

In addition to my main replication data, I administered a Qualtrics-based survey on the Social Science Prediction Platform (SSPP) from 30 March to 30 April 2024 (see DellaVigna, Pope, & Vivalt 2019). The survey and the original Qualtrics file can be found at https://socialscienceprediction.org/s/602202. The SSPP survey asks social science researchers to provide their predictions and judgments concerning equivalence testing results in the final sample.[8] I also ask researchers to provide judgments on acceptable Type I and Type II error rates in Top 5 economics journals. After screening out respondents who reported familiarity with the results of my analysis or gave incomplete responses, I possess a sample of judgments and predictions from 62 researchers. Online Appendix D details this sample of researchers.

[7]For example, marginal effects must be estimated in the case of probit or logit models for estimands to be appropriately interpreted in standardized units of the outcome variable.

[8]I specifically ask respondents to provide their predictions and judgments of TOST/ECI ETFRs in the final sample for a ROPE of $[-0.2\sigma, 0.2\sigma]$ at a 5% significance level (see Sections 5.1 and 5.2 for more details). To minimize confusion, I then ask each respondent whether they anticipate that these ETFRs will be different within a ROPE of $[-0.1r, 0.1r]$ than they will be within a ROPE of $[-0.2\sigma, 0.2\sigma]$. If they answer yes, then the respondent is asked to provide these same predictions and judgments of ETFRs within a ROPE of $[-0.1r, 0.1r]$. If they answer no, then the respondent is not shown these new questions, and the respondent's predictions and judgments of ETFRs within a ROPE of $[-0.1r, 0.1r]$ are imputed, using their predictions and judgments of ETFRs within a ROPE of $[-0.2\sigma, 0.2\sigma]$.

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 do so using a two-tailed test under the standard NHST framework (Imbens 2021).[9]

Definition 3.1 (The Standard Null Hypothesis Significance Testing Framework)**.** *The researcher wishes to assess whether $\delta \neq 0$ using a test with Type I error rate $\alpha \in (0, 1]$. They thus formulate null and alternative hypotheses as*

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

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, and that there is no relationship between Y and D if H_0 is not rejected (Romer 2020; Imbens 2021). Table 2 details the

[9] Though economists are sometimes interested in testing whether δ significantly differs from some non-zero point null, $\delta = 0$ is by far the most frequent null hypothesis. For ease of exposition, my definition of the standard NHST framework here is thus limited to this typical use case.

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

ways in which 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.[10] 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, the magnitude of estimates, 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

[10]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.

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 arising from low power. As s grows arbitrarily large, any arbitrarily large $\hat{\delta}$ can be deemed to be 'insignificant' under the standard NHST framework. Therefore, generally inferring a null result from a statistically insignificant estimate can often result in erroneously deeming 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 rejects the null hypothesis that $\delta = 0$ when one should not, 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 traditionally set at 0.05.[11] Type II error rate $\beta_{\text{NHST}} \in (0, 1]$ relates to the power $(1 - \beta_{\text{NHST}})$ with which one can detect a relationship with a magnitude at or above $\epsilon \geq 0$ under standard NHST. As the complement of the standard NHST Type II error rate for effect size ϵ, $(1 - \beta_{\text{NHST}})$ 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 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[12]

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

Power levels above (below) 0.8 are generally considered to be (in)sufficient in

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

[12] 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).

economics and the social sciences more broadly (Ioannidis, Stanley, & Doucouliagos 2017). The classical thresholds of $\alpha = 0.05$ and $\beta_{\text{NHST}} = 0.2$ reflect a presumption 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 ϵ for which 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 economic estimates exhibit sufficiently high power to detect reasonably small ϵ values, then insignificant results in the economics literature usually reflect true nulls, and there is no need to change current testing practices in economics.

Unfortunately, power is usually remarkably low throughout the economics literature. Ioannidis, Stanley, & Doucouliagos (2017) estimate median power to observe 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 median power estimates of 5% in Top 5 economics journals. These low power levels are not necessarily due to poor research practices. Answering important economic questions often requires researchers to examine pre-existing datasets where the researcher has no control over the data-generating process, leaving economists 'at the mercy' of existing sample sizes. However, these low power levels are still a flaw of economic research that must be acknowledged.

This low power poses serious challenges for the credibility of null claims in economics. When researchers interested in claiming that $\delta = 0$ use the standard NHST framework in Definition 3.1, the hypotheses are organized such that the researcher begins by assuming that what they want to show is true – that $\delta = 0$ – only concluding otherwise if the estimate is statistically significant enough to force them to abandon their claim. This shifts the burden of proof off of the researcher, which implies that for researchers trying to show that $\delta = 0$, imprecision is 'good', as the probability of finding a null relationship is inversely related to statistical precision. This is the

key motive for 'reverse p-hacking', which economists often engage in when performing placebo tests (Dreber, Johanneson, & Yang 2024).

Because the burden of proof is shifted off of the researcher in such settings, 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 reviews of the economics literature combined with the high frequency of unqualified null claims documented in Table 2 thus imply that economists often tolerate large Type II error rates.

The standard NHST framework is ultimately an untenable framework through which to reach conclusions that relationships are null, because one often cannot reliably discern whether an estimate is statistically insignificant due to small size or due to imprecision. This conflation between imprecision and null findings contributes to widespread beliefs that null results are 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, leading to publication bias throughout the economics literature (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019). Worse yet, the high Type II error rates that are effectively tolerated by current economic practice imply that even amongst the null findings that are prominently published, a considerable proportion are false negative results 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 aforementioned 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 arise at least in part from rational preferences for minimizing error rates. Therefore, if a testing framework can provide better control over error rates for null claims, then this testing framework may also yield the added benefit of mitigating the null result penalty, and in turn publication bias against null results.

4 Equivalence Testing

A credible framework for testing whether relationships are practically null can be constructed by making two modifications to the standard NHST framework. First, the null and alternative hypotheses in Equation 1 can be flipped, restoring the burden of proof on 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 resulting hypotheses 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.[13]

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* $\delta \approx$

[13]Though equivalence testing can be used to test a relationship's practical equivalence to any value, for ease of exposition, I limit my definition of the equivalence testing framework here to the typical use case where a researcher wants to show that there is virtually zero relationship between Y and D.

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* $[\epsilon_-, \epsilon_+]$.

$[\epsilon_-, \epsilon_+]$ is the *region of practical equivalence (ROPE)*, which is the range of δ values that one would deem to be practically equal to zero. ROPE boundaries thus effectively designate the range of δ values that are 'economically insignificant'. ROPEs are often (though not always) symmetric around zero such that $\epsilon_- = -\epsilon_+$.[14] A symmetric ROPE around zero can be said to have a *length* of $\epsilon > 0$ and written as $[-\epsilon, \epsilon]$. I discuss several tests that can assess whether $\hat{\delta}$ is statistically significantly bounded within the ROPE throughout the remainder of this section.

Though equivalence testing has historically been challenged by difficulties with establishing credible ROPES (see Ofori et al. 2023), relatively new online resources make the aggregation of credible ROPEs quite feasible for researchers. These resources are discussed further in Section 7.1. Further, though hypothesis tests based upon practically relevant intervals rather than point nulls are a common feature in Bayesian inference (Linde et al. 2023), the tests I discuss further in this section do not require reorienting to Bayesian methods, as all tests in this paper are frequentist in nature.[15]

[14]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 aim of such interventions is to increase Y, then even small negative effects of D are practically meaningful after factoring in the opportunity cost of abandoning other interventions. In this setting, it may be reasonable to set the ROPE such that $|\epsilon_-| < |\epsilon_+|$.

[15]Simulation evidence shows that conclusions reached under frequentist and Bayesian equivalence testing are relatively similar (Campbell & Gustafson 2018), though Bayesian equivalence tests can be better-powered (Linde et al. 2023).

4.1 Two One-Sided Tests Procedure

The hypotheses in Equation 4 can be rewritten as

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

This joint alternative hypothesis can be assessed using two one-sided tests:

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

Statistically significant evidence for H_A in Definition 4.1 can be obtained by showing statistically significant evidence for both H_A 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*

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

and compute

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

The exact critical value for this test can be written as

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

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 TOST procedure's name and modern form arises from Schuirmann (1987), who demonstrates that the TOST procedure often provides better power and error rate control than the traditional 'power approach' discussed in Section 3. The TOST procedure's size is preserved at nominal level α despite the use of simultaneous testing because the relevant test statistic is the smaller of its two t-statistics, and 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

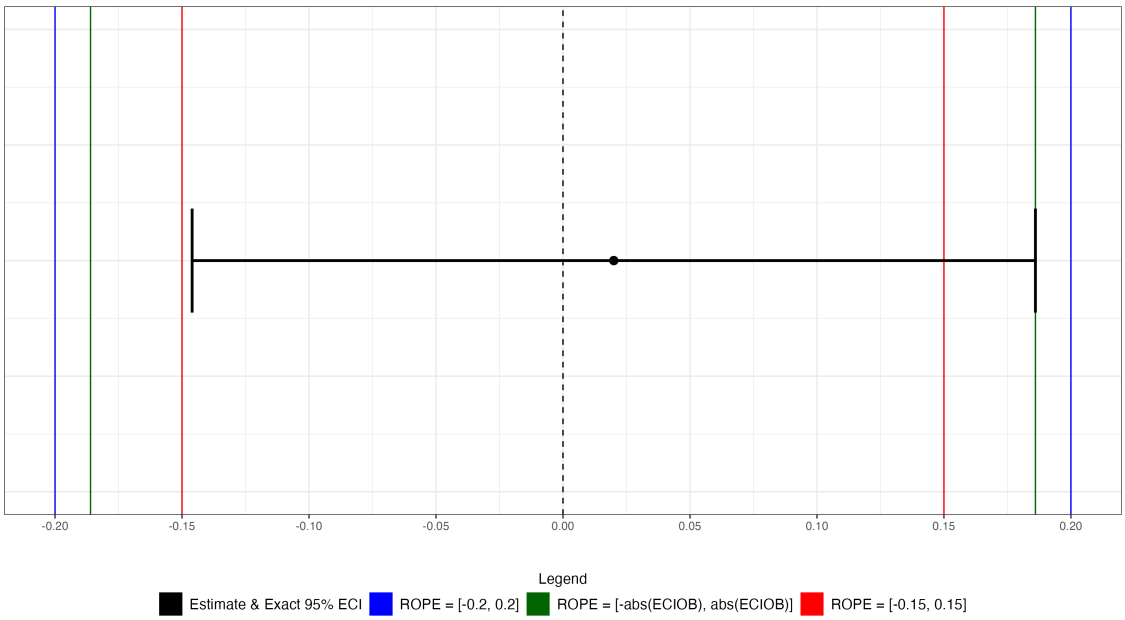
At a significance level of α, the TOST procedure can be inverted into an identical confidence interval-based approach that makes use of the symmetric $(1 - 2\alpha)$ confidence interval (Berger & Hsu 1996). Following Hartman (2021), I term this interval the *equivalence confidence interval (ECI)*.

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 calculated as*

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

and $t^*_{\alpha,df}$ *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, it is trivially simple to compute ECIs. For instance, the 95% ECI is just the 90% confidence interval. The



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

Figure 1: An ECI Example

key differences between ECIs and confidence intervals are the ways in which they can be used to judge statistical significance. Standard NHST significance judgments are derived from confidence intervals based on the confidence interval's relationship with zero. In contrast, significance judgments in equivalence testing are derived from ECIs 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, as the ECI approach is simply an inversion of the TOST procedure.

Figure 1 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 this estimate is 0.186, because the upper bound of this estimate's ECI is further away from zero than the lower bound of that ECI. The magnitude of the ECIOB is the length of the smallest symmetric ROPE around zero wherein one could find statistically significant evidence that $\delta \approx 0$. Therefore, the ECIOB's magnitude serves as a measure of how closely one can significantly bound $\hat{\delta}$ to zero. ECIOB magnitudes are thus of great interest to many applied economists, as the ECIOB magnitude is the smallest effect size that one can 'rule 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 effect size used 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 used is the *partial correlation coefficient* r, a widely-used effect size measure in meta-analyses. Per van Aert & Goos (2023), regression co-

efficients can be sequentially converted first into partial correlations and then into corresponding standard errors as

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

Here t_{NHST} is the standard NHST t-statistic as described in Definition 3.1, where $\hat{\delta} = \sigma$ and s is the standard error of σ.[16]

As Section 5.2 details further, equivalence testing failure rates measure how often estimate magnitudes 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 E shows that both $|r| = 0.1$ and $|\sigma| = 0.2$ are large effect sizes even amongst a benchmark sample of plausibly large economic effects. Thus 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.

5.2 Measuring Equivalence Testing Failure

The *equivalence testing failure rate (ETFR)* is defined here 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. For example, 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. The average

[16]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.

claim-level ETFR in this toy dataset for a ROPE of $[-0.2\sigma, 0.2\sigma]$ would be $(20\% + 100\% + 0\%)/3 = 40\%$.

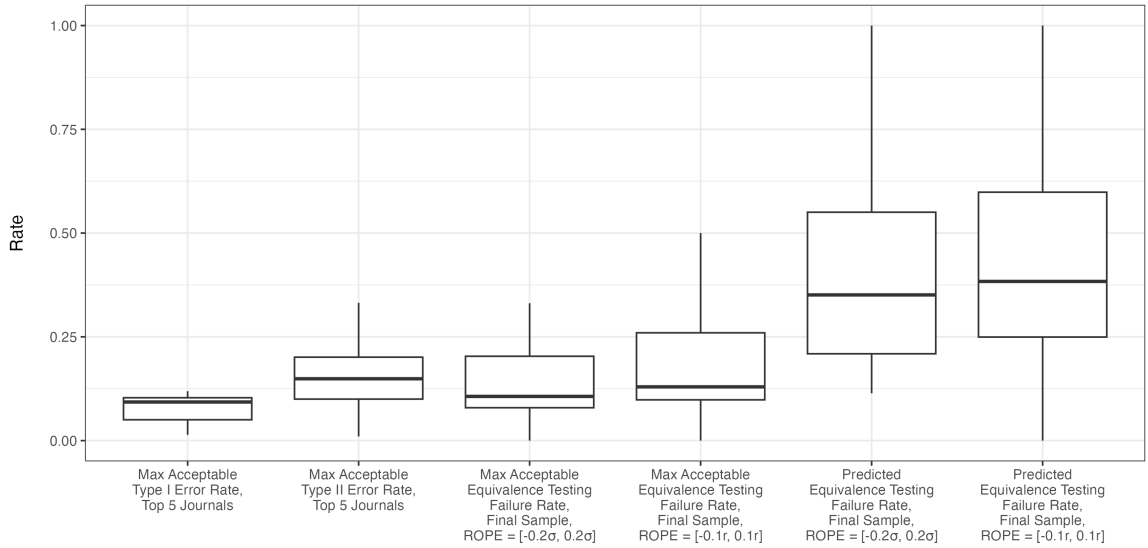
I calculate average claim-level and article-level ETFRs. I also calculate an average inverse-weighted claim-level ETFR that ensures all articles receive the same weight in the sample. Because these average ETFRs are calculated by taking a mean of partition-level ETFRs over all partitions, my precision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

6 Results

6.1 Predictions and Judgments

Figure 2 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 concerning 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 greatly 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 predictions are fairly accurate, though the median prediction of ETFRs for a ROPE of $[-0.1r, 0.1r]$ is an underestimate.



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.

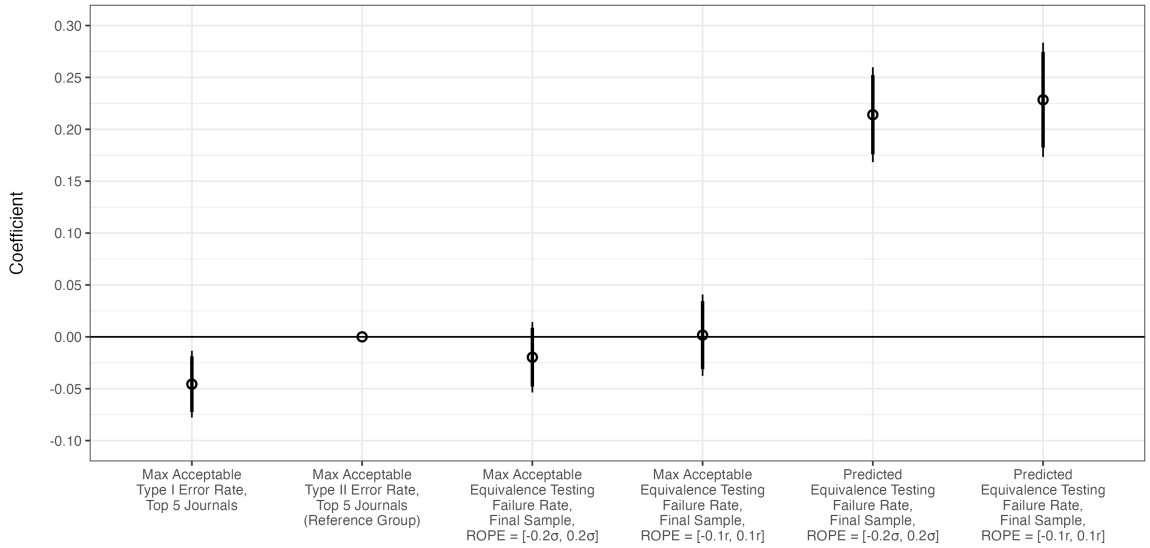
Figure 2: Distributions of SSPP Predictions and Judgments

Figure 2 displays substantial dispersion for predicted ETFRs. Though this partially reflects disagreement, it also reflects relatively low power in the SSPP sample ($N = 62$). Fortunately, the within-subject design of my SSPP survey allows much greater power to be achieved by constructing a researcher-rate panel dataset. This panel dataset also makes it possible to obtain within-researcher estimates of differences between rates using a panel data regression model that controls for researcher fixed effects. Specifically, letting i index the researcher and r index one of the six rates displayed in Figure 2, I estimate the model

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

Figure 3 displays estimates of γ_r from a model of Equation 12 that treats judgments on Type II error rates as the reference group.[17] On average, a given researcher reports that for results in Top 5 economics journals, their tolerance for Type I error

[17] A table version of these within-researcher 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 researcher level using a CR3 cluster-robust variance estimator (see Cameron & Miller 2015).

Figure 3: Within-Researcher Estimates of Differences in Predictions/Judgments

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 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 3 again show that ETFR tolerance is quantitatively close to Type II error rate tolerance. The average researcher'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 can 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' predictions of ETFRs in my final sample far exceed any

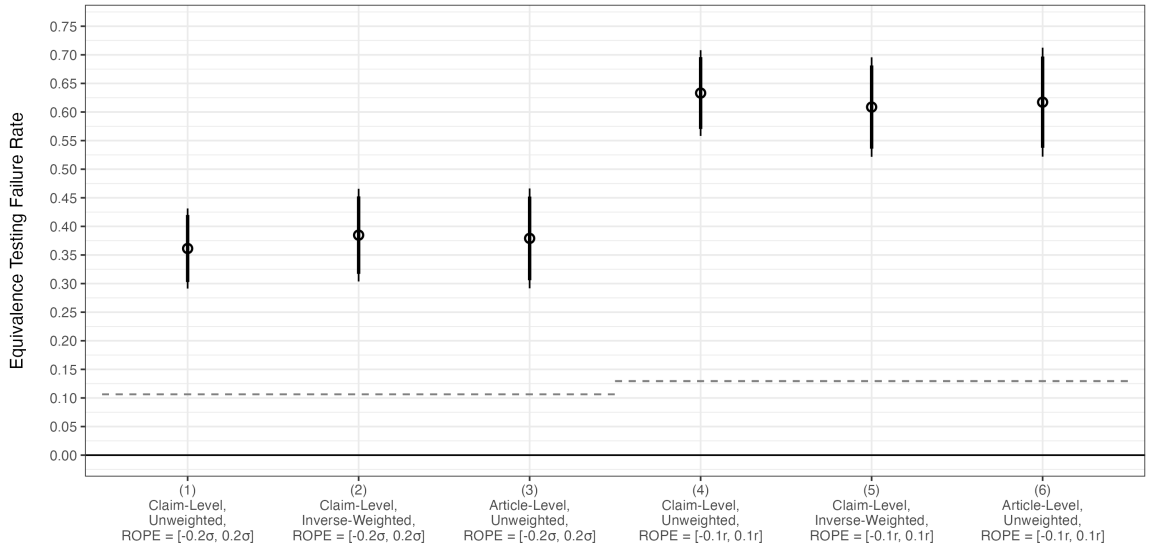
of these acceptability thresholds. The average researcher predicts that ETFRs will exceed their Type II error rate tolerance by 21.4 percentage points within a ROPE of $[-0.2\sigma, 0.2\sigma]$, and by 22.8 percentage points within a ROPE of $[-0.1r, 0.1r]$. Accounting for the aforementioned differences between Type II error rate tolerance and ETFR tolerances, these estimates imply that the average researcher predicts that ETFRs will exceed the maximum levels that they would find acceptable 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 Type II error rates. My ETFR estimates in the remainder of this section show that this prediction is quite accurate.

6.2 Equivalence Testing Failure Rates

Figure 4 displays the main ETFR estimates.[18] The dashed lines represent the median SSPP respondent's thresholds for acceptable ETFRs (see Section 6.1). ETFRs lie 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 metric, ETFRs vary by less than 2.5 percentage points across aggregation levels. Further, no one 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 but decreases ETFRs for partial correlation coefficients. It thus poses no threat to robustness to prefer one aggregation level when interpreting results. I therefore primarily reference unweighted claim-level ETFRs in my discussion of results, largely due to relative ease

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

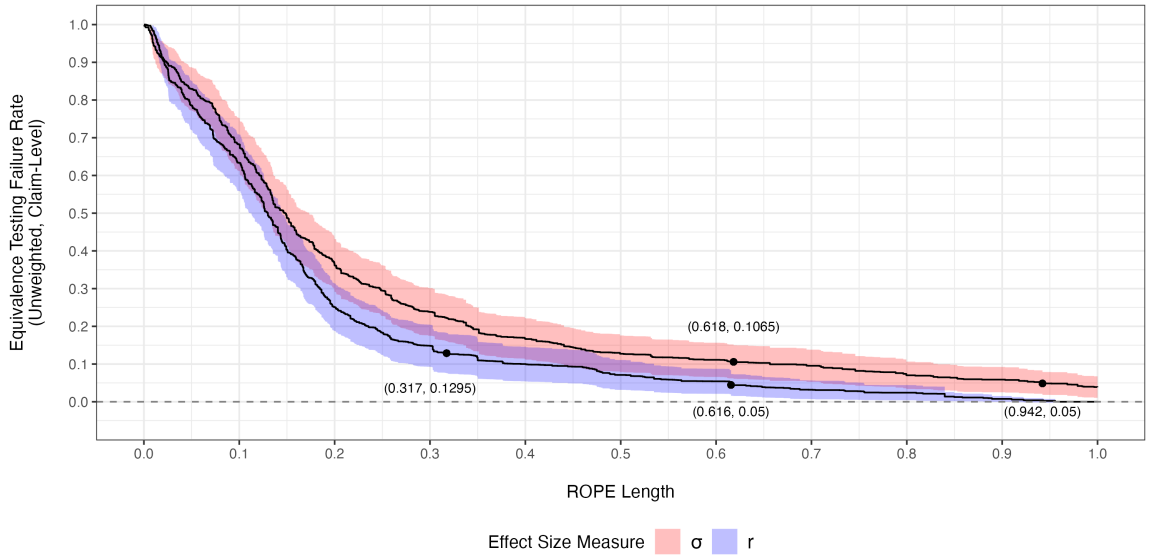


Note: ETFRs are provided along with 95% ECIs (thicker bands) and confidence intervals (thinner bands), based on the standard error of the mean for unweighted ETFRs and the weighted standard error of the mean for weighted ETFRs (see Online Appendix G). Dashed lines represent the median SSPP respondent's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

Figure 4: Main Equivalence Testing Failure Rate Estimates

of interpretability. For instance, Model 1 in Figure 4 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 estimates that are initially statistically significant under standard NHST are removed from the sample. 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 estimates from models that require conformability modifications.



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 4'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 5: Failure Curves

6.3 Failure Curves

Perhaps the most important sensitivity check concerns the choice of ROPE. Figure 5 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 5 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 that one wanted to assert that existing testing practices for null claims in economics are sufficient, and that ETFRs are bounded below some nominal level for reasonably-sized ROPEs. How large is the smallest ROPE that one would need to tolerate in order to make such a claim?

Figure 5's annotated points provide a sense of scale. To obtain claim-level ETFRs

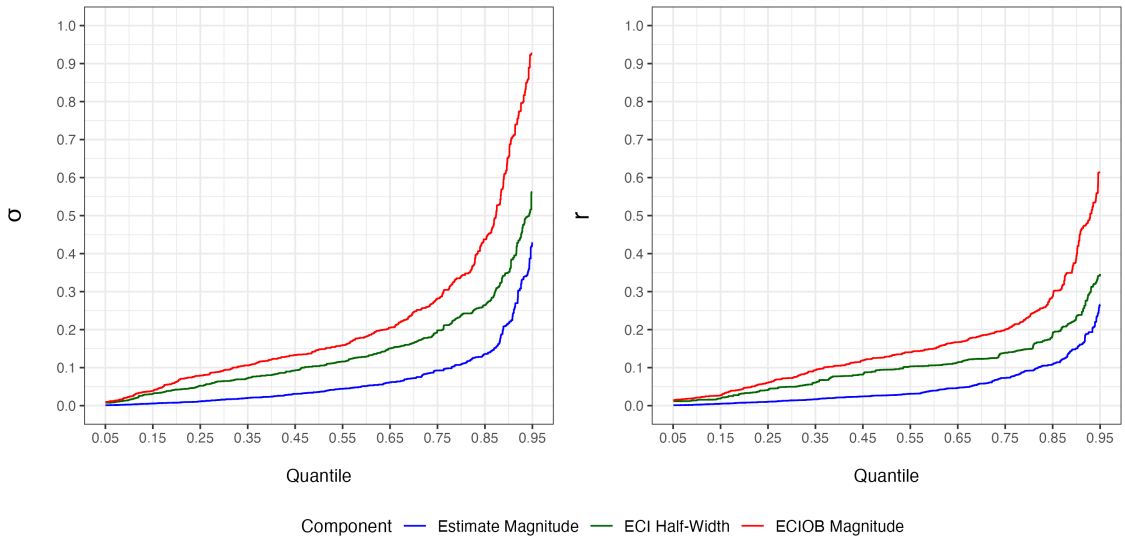
beneath 12.95% – the median SSPP respondent's maximum ETFR tolerance for a ROPE of $[-0.1r, 0.1r]$ – one must set a ROPE of $[-0.317r, 0.317r]$, which implies that one must argue that $|r| = 0.317$ is practically equal to zero. However, $|r| = 0.317$ is larger than nearly 75% of published results in 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.

Although the distribution of standardized coefficient magnitudes throughout the economics literature is not yet known, Online Appendix E shows that the 0.942σ ROPE length that one would need to tolerate to obtain a 5% claim-level ETFR is unreasonably large. The same is true of the 0.618σ effect size that is necessary to bound claim-level ETFRs beneath 10.65%, the ETFR which the median SSPP respondent would tolerate for a ROPE of $[-0.2\sigma, 0.2\sigma]$.

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 about zero wherein one can statistically significantly bound $\hat{\delta}$. ECIOB magnitudes thus directly determine the ROPEs within which $\hat{\delta}$ fails 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).



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

Figure 6: Inverse CDFs of ECIOB Magnitudes and Their Components

Figure 6 plots the distributions of ECIOB magnitudes and their components. If estimate magnitudes 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 substantively contribute to high ETFRs, low precision is the dominant driver.

Table 3 provides further evidence of this dominance, displaying constant elasticity estimates of the relationships between 95% ECIOB magnitudes, effect sizes, and 95% ECI half-widths. Both effect sizes and ECI half-widths are highly statistically significantly associated with ECIOB magnitudes in theoretically expected directions. 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

	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

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 yields encouraging evidence on the finite-sample properties of equivalence testing. Section 3 notes that a key credibility issue with using the standard NHST framework when the researcher wants to show that $\delta = 0$ is that imprecision is 'good', in the sense that there is an inverse relationship between precision and the probability of obtaining a null result. 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 Type II error rates, and many null claims prominently made 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, and 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 remainder 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 effects 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 interest.

However, this practical need 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* to permit their estimates to be 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 (see Dreber, Johanneson, & Yang 2024), it is not difficult to imagine that ROPE-hacking could similarly emerge in economic applications of equivalence testing. 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, provided that the researcher has a credible sample of experts or stakeholders who can provide effect size judgments.

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 unfortunately often treated as an afterthought, utilized only when statistically significant evidence cannot be obtained for a given estimate 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 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, as 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, and thus readers (and potentially also the researcher) would not 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 the procedure can yield inconclusive results. The standard NHST framework currently results in a dichotomization of research findings – either a relationship is statistically 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*

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

Test statistic t_{TOST} is computed as in Definition 4.1 along with test statistics

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

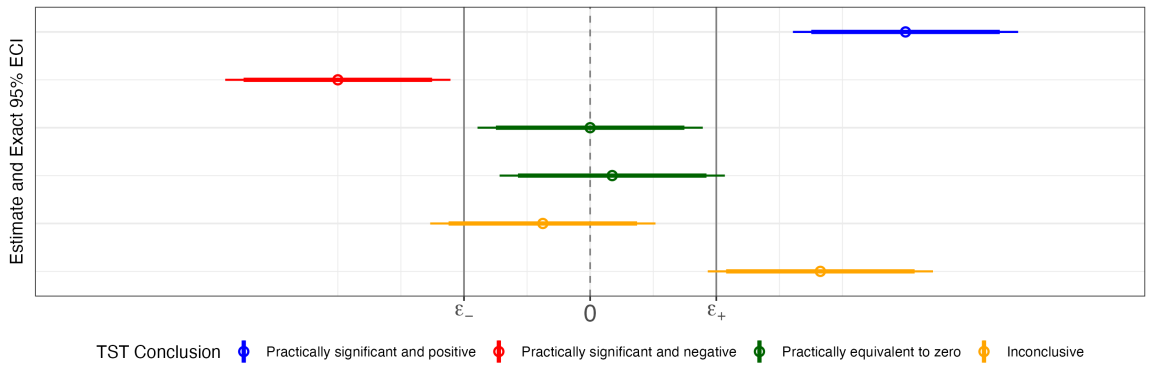
The researcher concludes that δ is significantly bounded above the ROPE if and only if $t_P > t^*_{\alpha/2,df}$. The researcher concludes that δ is significantly bounded below the ROPE if and only if $t_N < -t^*_{\alpha/2,df}$. As in Definition 4.1, if $t_{TOST} = t_-$, then the researcher concludes 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 consists of conducting 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 2.1 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 7 illustrates how TST conclusions can be derived using a confidence interval approach. The top three estimates shown in Figure 7 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 estimate's entire 95% ECIs are bounded inside the ROPE, so there is significant evidence that these estimates are practically equal to zero. In contrast, the



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 7: Research Conclusions in the TST Framework

practical significance of the last two estimates in Figure 7 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 7 is particularly important for understanding how TST augments the standard NHST framework. This estimate is statistically significantly different from zero, and because its point estimate exceeds the ROPE, this estimate would likely lead most economists to conclude that the relationship is 'economically significant'. However, under TST, this estimate would still be deemed to be too noisy to yield highly certain practical significance conclusions. This bottom estimate lacks sufficient precision to rule out the prospect 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 bounded outside of the ROPE.

8 Conclusion

I introduce the economics literature to a suite of simple equivalence testing methods. I then demonstrate their necessity, showing that a substantial proportion of 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 it is ludicrous to claim that the magnitudes of so many published economic estimates are practically equal to zero, it is instead 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, while controlling error rates from these simultaneous tests 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 frame-

work 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.[19]

[19]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 (2021). "Equivalence testing for regression discontinuity designs". *Political Analysis* 29.4, pp. 505–521. DOI: 10.1017/pan.2020.43.

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.

Linde, Maximilian et al. (2023). "Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor." *Psychological Methods* 28.3, pp. 740–755. DOI: 10.1037/met0000402.

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.

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.

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

van Aert, Robbie C. and Cas Goos (2023). "A critical reflection on computing the sampling variance of the partial correlation coefficient". *Research Synthesis Methods* 14.3, pp. 520–525. DOI: 10.1002/jrsm.1632.

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.

Online Appendix

A Systematic Review Process

My initial sample consists of all articles registered in Web of Science as published from 2015 onwards in a Top 5 economics journal (specifically *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*). I obtained bibliographic information on this initial set of 3732 articles, including digital object identifiers, titles, and abstracts from Web of Science on 28 July 2023. This bibliographic information is then loaded into ASReview, an interface that employs machine learning and text classification to assist with managing systematic literature reviews by sorting abstracts from most to least relevant (van de Schoot et al. 2021). I then manually reviewed the abstracts, classifying them as relevant if the abstract makes some claim that a phenomenon or relationship is either negligible or nonexistent. After reviewing 2987 abstracts, 50 consecutive abstracts were assessed to be irrelevant, and thus the remaining 745 articles are discarded as irrelevant based on ASReview's relevance probability ranking.[1] The abstract reviews yield 603 potentially relevant records, at which point all articles published prior to 2020 are discarded, ensuring that the sample reflects only the most recent practice in the economics literature and has the highest probability of reproducibility while still keeping the number of (attempted) reproductions down to a practically feasible level.[2] 287 potentially relevant articles published from 2020-2023 arise from this first phase of the systematic search.

I then examine the abstracts of each of these 287 potentially relevant articles, isolating every null claim made in each abstract and discarding an article if, upon

[1] This is an intended feature of ASReview – the probability ranking permits early cessation of the review process with a strong reassurance that the most relevant articles still remain in the sample (van de Schoot et al. 2021).

[2] The additional articles from 2015-2019 help ensure the quality of the relevance probability ranking, and thus the irrelevance of discarded articles.

further inspection, its abstract does not in fact make an identifiable null claim. This step produces 556 null claims across 285 articles. For each of these null claims, I attempt to locate the estimate(s) used to support that claim within the article. I discard a claim if it is not defended by at least one statistically insignificant estimate, otherwise storing the main estimate(s) being used to defend that claim. I discard articles if no null claims remain after this discarding process. This step yields my intermediate sample of 2346 estimates across 279 claims in 158 articles. Thereafter, I attempt to reproduce every estimate in the intermediate sample. Estimates are discarded when data is not available for reproduction or the reproduction is not conformable to my final analysis. After such discarding, my final sample consists of 876 estimates across 135 null claims in 81 articles.

B Final Sample

ree

C Intermediate Sample

ree

D SSPP Data

The SSPP survey was posted publicly to the SSPP website, and any interested respondent was free to take the survey. The survey was also publicly disseminated on Twitter/X by the SSPP. 58 of the 62 survey respondents (93.5%) are members of the SSPP's Superforecaster Panel, which is a sample of researchers that are pre-selected by SSPP and are paid a semi-annual flat rate for completing a sufficient proportion of the surveys that are posted to the SSPP website each month. The remaining four respondents are not part of the Superforecaster Panel, and are not incentivized to take the survey.

My SSPP sample is relatively young, with the median respondent being 32.5 years of age (mean = 34.6, SD = 10.8). Though much of the sample has ample experience with making predictions for social science research questions by virtue of being part of the Superforecaster Panel, my sample rates their five-point Likert confidence in their predictions at a median of 2.5 (mean = 2.4, SD = 1). This is sensible, as only nine respondents (14.5%) report conducting prior research on the topics discussed in my survey. The sample is male-dominated, with 53 respondents (85.5%) reporting a masculine gender identity. The SSPP sample also predominantly originates from WEIRD countries (Henrich, Heine, & Norenzayan 2010) – 42 respondents (67.7%) spent the majority of their time prior to starting university education in OECD member states, and 48 respondents (77.4%) have spent the majority of their time since starting university education in OECD member states.

E Effect Size Benchmarking

Table A1 shows the values of σ and r for a selected sample of ten highly-cited and recent results from the economics literature that represent plausibly large effects. I term this the *benchmarking sample*. All articles in this sample have publicly-available replication repositories and are published between 2015-2020. I isolate one main claim of each article and the primary estimate used to defend this claim. The benchmarking sample thus consists of ten articles, each with one claim and one estimate defending that claim. Appendix F provides citations for all articles in the benchmarking sample, along with associated replication repositories (when applicable).

Two features of Table A1 are worth noting. First, though σ and r are quite positively correlated and always share the same sign, they do not necessarily monotonically correspond, as σ is a measure of magnitude whereas r is a measure of fit. Second, though the estimates in this benchmarking sample are all statistically significant under the standard NHST framework, their effect sizes are also quite small in general.

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

Note: Effect sizes and initial p-values of each estimate are reported. Each original estimate can be found in its respective article at the specified location. Some articles are reproduced using data from repositories (Hanushek 2016; Benhassine et al. 2019; Berman et al. 2019; Deschênes, Greenstone, & Shapiro 2019; Duflo, Dupas, & Kremer 2019), whereas others are reproduced using files linked to the publisher's online webpage for the article.

Table A1: Effect Size Benchmarking

F Benchmarking Sample

ree

G Equivalence Testing Failure Rate Computation

Let j be an individual partition, and let i index an individual estimate. j represents an individual claim when calculating claim-level ETFRs, whereas j represents an entire article when calculating article-level ETFRs. Each estimate i belongs to exactly one partition j. Because all ETFRs in this paper are calculated for symmetric ROPEs, it is sufficient to define ETFR $R(\epsilon, \tau, L)$ as a function of ROPE length $\epsilon > 0$, effect size measure $\tau \in \{\sigma, r\}$, and aggregation level L. Further, because the ECI approach described in Definition 4.3 yields identical results to the TOST procedure described in Definition 4.2, I approach ETFR calculation by defining the exact 95% ECI's outer bound $\text{ECIOB}_{i,j}(\tau)$ for each effect size measure τ of every estimate i. Let M_j represent the number of estimates i belonging to partition j, and let M be the total number of partitions j. One can then calculate the ETFR as

$$R(\epsilon, \tau, L) = \sum_{j=1}^{M} \sum_{i=1}^{M_j} \frac{\mathbb{1}\left[|\text{ECIOB}_{i,j}(\tau)| > \epsilon\right]}{M_j M}. \tag{A1}$$

I also calculate claim-level ETFRs that apply an inverse weighting approach, ensuring that each article receives the same weight in the sample. Let $W_{j,k}$ be equal to 1 divided by the number of claims that belong to claim j's article, and let k be an individual article. Then the inverse-weighted claim-level ETFR can be written as

$$R_{\text{Wgt.}}(\epsilon, \tau) = \frac{1}{\sum_{j=1}^{M} W_{j,k}} \sum_{j=1}^{M} W_{j,k} \sum_{i=1}^{M_{j,k}} \frac{\mathbb{1}\left[|\text{ECIOB}_{i,j,k}(\tau)| > \epsilon\right]}{M_{j,k}}, \tag{A2}$$

where $M_{j,k}$ is now the number of estimates belonging to claim j in article k, and M is now the total number of articles.

I measure precision using standard errors of the mean for the unweighted ETFRs in Equation A1 and standard errors of the weighted mean for the weighted ETFRs

in Equation A2. The standard error of the mean for an ETFR is

$$\text{SE}\left[R(\epsilon, \tau, L)\right] = \frac{\text{SD}\left[R(\epsilon, \tau, L)\right]}{\sqrt{M}}, \tag{A3}$$

where $\text{SD}\left[R(\epsilon, \tau, L)\right]$ is just the within-sample standard deviation of $R(\epsilon, \tau, L)$. Let the ETFR for claim j in article k be defined as

$$R_{j,k}(\epsilon, \tau, L) = \sum_{i=1}^{M_{j,k}} \frac{\mathbb{1}\left[|\text{ECIOB}_{i,j,k}(\tau)| > \epsilon\right]}{M_{j,k}}.$$

Though Gatz & Smith (1995) note that there is no universally-agreed definition for the standard error of the weighted mean, they find that one formulation produces closer estimates to the bootstrap than other competing formulas. In this setting, the square of that optimal formula can be written as

$$\begin{aligned}\left(\text{SE}\left[R_{\text{Wgt.}}(\cdot)\right]\right)^2 = \frac{M}{(1-M)\,M^2}\Bigg[& \sum_{j=1}^{M}\left\{\left[W_{j,k}R_{j,k}(\cdot) - \overline{W}_{j,k}R_{\text{Wgt.}}(\cdot)\right]^2\right\} - \\ & 2R_{\text{Wgt.}}(\cdot)\sum_{j=1}^{M}\left\{\left(W_{j,k} - \overline{W}_{j,k}\right)\left[W_{j,k}R_{j,k}(\cdot) - \overline{W}_{j,k}R_{\text{Wgt.}}(\cdot)\right]\right\} + \\ & \left[R_{\text{Wgt.}}(\cdot)\right]^2\sum_{j=1}^{M}\left\{\left[W_{j,k} - \overline{W}_{j,k}\right]^2\right\}\Bigg],\end{aligned}$$

where $\overline{W}_{j,k}$ is the mean inverse weight $W_{j,k}$ across all claims and M is the total number of articles. The results in Section 6.2 show that this standard error derivation corresponds quite closely with simple standard errors for unweighted ETFRs as derived in Equation A3.

H Appendix Tables

This appendix provides table versions of two main figures in Section 6.

	(1)	(2)	(3)	(4)	(5)	(6)
γ_r	-0.046	·	-0.02	0.002	0.214	0.228
	(0.016)	(·)	(0.017)	(0.02)	(0.023)	(0.028)
Type	Judgment	Judgment	Judgment	Judgment	Prediction	Prediction
Rate	Type I Error	Type II Error	TOST/ECI Failure	TOST/ECI Failure	TOST/ECI Failure	TOST/ECI Failure
Effect Size Measure			σ	r	σ	r

Note: This table provides the numerical estimates displayed in Figure 3.

Table A2: Within-Researcher Estimates of Differences in Predictions/Judgments

	(1)	(2)	(3)	(4)	(5)	(6)
Equivalence Testing Failure Rate	0.361	0.385	0.379	0.633	0.609	0.617
	(0.035)	(0.041)	(0.044)	(0.038)	(0.044)	(0.048)
Effect Size Measure	σ	σ	σ	r	r	r
SSPP Tolerance	0.1065	0.1065	0.1065	0.1295	0.1295	0.1295
Aggregation Level	Claim	Claim	Article	Claim	Claim	Article
Inverse Weighting		x			x	

Note: This table provides the numerical estimates displayed in Figure 4.

Table A3: Main Equivalence Testing Failure Rate Estimates

I Robustness Checks

This appendix reports extended robustness checks on the main results in Section 6.2.

	Estimates	Claims	Articles	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: CYCD Removed	675	105	63	0.332 (0.04)	0.346 (0.045)	0.34 (0.048)	0.62 (0.044)	0.617 (0.049)	0.628 (0.054)
Panel B: CYBD Removed	563	91	59	0.338 (0.044)	0.358 (0.049)	0.358 (0.054)	0.621 (0.047)	0.558 (0.053)	0.562 (0.058)
Panel C: BYCD Removed	563	124	74	0.39 (0.038)	0.41 (0.043)	0.402 (0.047)	0.651 (0.04)	0.631 (0.046)	0.64 (0.051)
Panel D: BYBD Removed	653	119	73	0.348 (0.037)	0.381 (0.043)	0.377 (0.047)	0.634 (0.04)	0.625 (0.046)	0.629 (0.052)
Effect Size Measure				σ	σ	σ	r	r	r
SSPP Tolerance				0.1065	0.1065	0.1065	0.1295	0.1295	0.1295
Aggregation Level				Claim	Claim	Article	Claim	Claim	Article
Inverse Weighting					x			x	

Note: Estimates are deemed initially (in)significant if the standard NHST p-value of initial estimate is less than (greater than or equal to) 0.05 (before conformability changes, if applicable). ROPEs are $[-0.2\sigma, 0.2\sigma]$ and $[-0.1r, 0.1r]$.

Table A4: ETFR Robustness – Initial Estimate Significance

	Estimates	Claims	Articles	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: CYCD Removed	675	105	63	0.342 (0.04)	0.362 (0.046)	0.356 (0.049)	0.62 (0.044)	0.617 (0.049)	0.628 (0.054)
Panel B: CYBD Removed	563	91	59	0.36 (0.045)	0.37 (0.049)	0.369 (0.054)	0.621 (0.047)	0.558 (0.053)	0.562 (0.058)
Panel C: BYCD Removed	563	124	74	0.398 (0.038)	0.417 (0.043)	0.409 (0.047)	0.651 (0.04)	0.631 (0.046)	0.64 (0.051)
Panel D: BYBD Removed	653	119	73	0.365 (0.038)	0.39 (0.043)	0.386 (0.046)	0.634 (0.04)	0.625 (0.046)	0.629 (0.052)
Effect Size Measure				σ	σ	σ	r	r	r
SSPP Tolerance				0.1065	0.1065	0.1065	0.1295	0.1295	0.1295
Aggregation Level				Claim	Claim	Article	Claim	Claim	Article
Inverse Weighting					x			x	

Note: Panels denote whether estimates corresponding to continuous/binary outcome/exposure variables (respectively) are removed from the sample. For example, 'CYBD removed' implies that estimates corresponding to a continuous outcome variable and a binary exposure variable are removed from the sample. ROPEs are $[-0.2\sigma, 0.2\sigma]$ and $[-0.1r, 0.1r]$.

Table A5: ETFR Robustness – Regressor Type Combination

	Estimates	Claims	Articles	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Non-Replicable Estimates Removed	803	123	74	0.388 (0.038)	0.406 (0.043)	0.399 (0.047)	0.618 (0.04)	0.607 (0.046)	0.615 (0.051)
Panel B: Non-Conformable Estimates Removed	807	130	77	0.358 (0.036)	0.37 (0.041)	0.365 (0.044)	0.65 (0.038)	0.626 (0.044)	0.636 (0.049)
Effect Size Measure				σ	σ	σ	r	r	r
SSPP Tolerance				0.1065	0.1065	0.1065	0.1295	0.1295	0.1295
Aggregation Level				Claim	Claim	Article	Claim	Claim	Article
Inverse Weighting					x			x	

Note: Estimates are non-replicable if my best attempts to replicate the exact published estimates using the article's replication repository do not succeed. Estimates are 'non-conformable' if the models that produce them require conformability modifications before inclusion in the final sample. ROPEs are $[-0.2\sigma, 0.2\sigma]$ and $[-0.1r, 0.1r]$.

Table A6: ETFR Robustness – Replicability/Conformability