The Need for Equivalence Testing in Economics

Jack Fitzgerald, Vrije Universiteit Amsterdam*

May 30, 2024

Abstract

Equivalence testing methods can provide statistically significant evidence that relationships are practically equal to zero. I demonstrate their necessity in a systematic reproduction of estimates defending 135 null claims made in 81 articles from top economics journals. 37-63% of these estimates cannot be significantly bounded beneath benchmark effect sizes. Though prediction platform data reveals that researchers find these equivalence testing 'failure rates' to be unacceptable, researchers actually expect unacceptably high failure rates, accurately predicting that failure rates exceed acceptable thresholds by around 23 percentage points. To obtain failure rates that researchers deem acceptable, one must contend that over 75% of published effect sizes in economics are practically equivalent to zero, implying that Type II error rates are likely quite high throughout economics. This paper provides economists with empirical justification, guidelines, and commands in Stata and R for conducting credible equivalence 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, Andre Lucas, Derek Mikola, and Jonathan Roth for valuable input on this paper. I also thank the multiple authors who answered my questions about their research and replication data. I currently hold a 12-month term as a member of the Superforecaster Panel for the Social Science Prediction Platform (SSPP; DellaVigna, Pope, & Vivalt 2019). The views expressed in this paper do not necessarily represent the views of the SSPP, or 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 significant. Though many such findings go unpublished (Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019), suppose this one makes it to print. How would most economists report this finding? I show in this paper 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 (McShane & Gal 2016) and even statisticians (McShane & Gal 2017).

This is widely-known to be bad scientific practice (Altman & Bland 1995; Imai, King, & Stuart 2008; Wasserstein & Lazar 2016). Statistically insignificant estimates do not necessarily imply that a relationship of interest is negligible; such results may simply reflect imprecision (i.e., low statistical power). This conflation between imprecision and null results arises because the standard null hypothesis significance testing (NHST) framework assumes in the null hypothesis that there is no relationship between the variables of interest. Thus when economists use the standard NHST framework to show the absence of a relationship, the burden of proof is reversed: the researcher begins by assuming that what they want to show is true. Therefore, for an economist trying to show the absence of a relationship, *imprecision is 'good'*: the probability of finding no significant relationship declines as the researcher accrues more power, even if the regression estimate remains negligibly small. This conflation between null results and imprecision under the standard NHST framework contributes to a myriad of problems for the credibility of economic research, including 'reverse p-hacking' (Dreber, Johanneson, & Yang 2024), low faith in the quality and publishability of null results (McShane & Gal 2016; McShane & Gal 2017; Chopra et al. 2024), and publication bias from non-publication of null findings (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019). Additionally, because power throughout the economics literature is quite low (Ioannidis, Stanley, & Doucouliagos

2017), Type II error rates are also likely quite high amongst published null findings.

Fortunately, there is a better frequentist testing framework – known as the *equivalence testing* framework – that corrects for all of these issues, and can provide credible evidence of null relationships. The equivalence testing framework begins by specifying a range of values around zero called the *region of practical equivalence (ROPE).* The ROPE denotes the range of values for the relationship of interest that one would consider to be 'practically equivalent to zero'. This is a subjective effect size judgment, and is effectively a way of specifying what values of a relationship are 'economically insignificant'. Naturally, these boundaries should vary for different relationships of interest. Fortunately, relatively recent online platforms make it practical for researchers to aggregate credible ROPEs from independent parties, such as experts or relevant stakeholders. Once the ROPE is set, equivalence testing restores a proper burden of proof by assuming in the null hypothesis that the estimate is *not* bounded within the ROPE. Statistically significant evidence under the equivalence testing framework provides credible evidence that the relationship of interest is bounded within the ROPE, and thus that this relationship is practically equivalent to zero. Equivalence testing is routinely applied in medicine (Piaggo et al. 2012), and is being rapidly adopted by psychology (see e.g., Lakens, Scheel, & Isager 2018) and political science (see e.g., Hartman & Hidalgo 2018). To facilitate equivalence testing's use in economics, this paper thoroughly details *what* frequentist equivalence testing procedures are available, *why* they are necessary, and *how* to credibly use them in future research.

1.1 What Equivalence Testing Methods Are Available?

First, I detail the *two one-sided tests (TOST)* procedure. The TOST procedure employs two one-sided tests to assess whether an estimate is significantly bounded *above* the ROPE's *lower* bound and significantly bounded *beneath* the ROPE's *upper* bound. If both are true, then there is significant evidence that the estimate is practically equivalent to zero. Second, I discuss the *equivalence confidence interval (ECI)* approach, which produces identical conclusions to the TOST procedure. The $(1-\alpha)$ ECI is simply the $(1-2\alpha)$ ordinary CI; one can conclude that an estimate is significantly bounded within a given ROPE at significance level α if the $(1-\alpha)$ ECI is entirely bounded within that ROPE. An extension of this latter approach is of great interest to many applied economists, as the furthest bound of the ECI from zero is the smallest effect size that one can 'rule out' with statistically significant evidence.

1.2 Why Is Equivalence Testing Necessary?

Using these methods, I show that the standard testing procedures which economists use to make and defend null claims likely tolerate unacceptably high Type II error rates. I systematically reproduce and standardize results from the models defending 135 null claims made by 81 articles published in Top 5 economics journals from 2020-2023. I then estimate *failure rates* by testing how frequently results in this sample fail to be statistically significantly bounded within 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). It should thus be expected that models defending null claims made in top economics journals are significantly bounded within these ROPEs. However, failure rate estimates are unacceptably high.

At the 5% significance level, failure rates within these ROPEs range from 37-63%. Examining the distributions of failure rates across varying ROPE sizes shows that in order to obtain failure rates that surveyed researchers deem acceptable, one must be willing to claim that more than 75% of all published effect sizes in economics are practically equivalent to zero. Since such a claim is ludicrous, these results imply that null claims in top economics journals exhibit unacceptably high error rates.

Critically, prediction platform survey data shows that though failure rates for null claims are unacceptably high, *researchers expect this.* The median researcher deems failure rates of 10.65-12.95% to be acceptable, but predicts failure rates from

35.1-38.35%, roughly in line with the lower bound of actual failure rate estimates. On average, researchers expect failure rates 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.

1.3 How Should Equivalence Testing Be Done?

Given this clear need, I provide guidelines for credible equivalence testing in economic research. To reduce researcher degrees of freedom and 'ROPE-hacking', I recommend that researchers aggregate ROPEs by surveying independent parties, such as experts or stakeholders relevant to the research question, 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).

Once a credible ROPE is set, not only is it no longer construct-valid to employ the standard NHST framework, but simultaneously employing standard NHST alongside equivalence testing sacrifices either power or Type I error coverage. I thus advocate for researchers to approach hypothesis testing using the *three-sided testing (TST)* framework (Goeman, Solari, & Stijnen 2010). The TST framework consists of three simultaneous hypothesis tests on an estimate: a one-sided test assessing whether the estimate is significantly bounded *below* the ROPE, a one-sided test assessing whether the estimate is significantly bounded *above* the ROPE, and a TOST procedure. Significance conclusions can effectively be derived from the smallest p-value among these three tests. Family-wise error rates remain controlled at nominal significance levels despite the use of multiple hypothesis tests because only one of the three alternative hypotheses under the TST framework can be true at a time. Critically, under the TST framework, a result can be too imprecise to be reliably classified as either practically significant *or* practically equivalent to zero. In such cases, the TST framework requires that researchers concede that their results are inconclusive, ensuring that imprecise estimates are not considered definitive evidence of null relationships.

Finally, I provide the `tsti` command in Stata and the `equivtest` package in R, which enable computation of immediate testing results under the TST framework for a given estimate, standard error, and ROPE. Because TOST is nested in the TST framework, both `tsti` and `equivtest` can be used to obtain equivalence testing results using the TOST procedure. Both suites are available for download from Github.[1]

2 Data

I obtain a systematically-selected sample of 2346 models defending 279 null claims made in the abstracts of 158 articles published in Top 5 economics journals (specifically *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*) from 2020-2023.[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), and are defended by at least one statistically insignificant estimate. I term this full sample of articles, claims, and models the *intermediate sample*.

The *final sample* contains all models in the intermediate sample that are conformable and computationally reproducible using publicly-available data.[3] The final sample is comprised of 876 models across 135 null claims in 81 articles, and stores each model's standardized regression coefficient σ, standard error s, sample size N, degrees of freedom df,[4] replicability status, conformability status, outcome and ex-

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

[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 $df = N - b$ (see Section 3).

	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 Models, 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 Models, Final Sample	1	1	3	6	14	24	82	10.815	13.145	81
Panel B: Claim-Level										
# of Models, Intermediate Sample	1	1	2	4	8	16	288	8.409	22.372	279
# of Models, Final Sample	1	1	2	4	7.5	14.6	55	6.489	8.128	135
Panel C: Model-Level										
σ	-1.671	-0.12	-0.026	0.004	0.044	0.122	1.817	0.002	0.201	876
$\|\sigma\|$	0	0.004	0.013	0.036	0.102	0.244	1.817	0.096	0.177	876
s	0	0.012	0.027	0.069	0.13	0.208	5.783	0.108	0.259	876
Initial NHST p-value	0	0.05	0.229	0.483	0.739	0.899	1	0.481	0.303	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 $\|\sigma\| = 0.2$	0.031	0.157	0.33	0.825	1	1	1	0.684	0.341	876

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

Table 1: Summary Statistics

posure variables with dummies indicating if each is binary, and the initial p-value implied by the model (without conformability changes, if applicable). The standardization procedure for σ and s is detailed in Section 5.1. The sample of articles with claims and models included in the final sample, alongside additional data repositories attached to these articles (when applicable), is provided in Online Appendix B, whereas the sample of articles in the intermediate sample that are excluded from the final sample are provided in Online Appendix C.

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 claim is defended by four models. Effect sizes are quite small throughout the final sample, with the median standardized coefficient magnitude at 0.036σ. The median model in the final sample is estimated with $N = 3558$ and $df = 180$. 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: 32.1% of estimates in the final sample lack even 50% power to detect a 0.2σ effect.

This imputation is conservative for the purposes of this paper, if anything deflating failure rates for partial correlations (see Sections 5.1 and 5.2).

Over 90% of the 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 always arise alongside other statistically insignificant estimates that together defend their null claim. 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 models in the final sample are not fully replicable, in the sense that my best attempts to reproduce the article's findings using its replication repository do not yield the exact same results as those published in the article. Further, 7.9% of the models in the final sample 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.[5] Both the outcome and exposure variable are continuous for 22.9% of models in the final sample, while 25.5% of the models in the final sample have binary outcome and exposure variables. The most frequent type of model is that with a continuous outcome and a binary exposure, representing 35.7% of models 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; see DellaVigna, Pope, & Vivalt 2019) from 30 March to 30 April 2024. 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.[6] I also ask researchers to provide judgments on acceptable

[5] 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.

[6] I specifically ask respondents to provide their predictions and judgments of TOST/ECI failure rates in the final sample for a ROPE of $[-0.2\sigma, 0.2\sigma]$ at a 5% significance level (see Section 5.2 for more details). To minimize confusion, I then ask each respondent whether they anticipate that these failure rates will be different within a ROPE of $[-0.1r, 0.1r]$ than they will be within a ROPE of $[-0.2\sigma, 0.2\sigma]$ (see Section 5.1 for more details). If they answer yes, then the respondent is asked to

Type I and Type II error rates in Top 5 economics journals. After screening incomplete responses and responses from respondents who reported familiarity with the results of my analysis, I possess a sample of judgments and predictions from 62 researchers. Online Appendix D details this sample of researchers.

3 Null Claims in Economics: Theory and Practice

In practice, economists estimate relationships of interest using linear models that presume that such relationships arise from data-generating processes 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.[7] The parameter of interest is δ, the linear association between Y and D. Point estimate $\hat{\delta}$ and standard error (SE) $s > 0$ 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).[8]

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

provide these same predictions and judgments of failure rates 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 failure rates within a ROPE of $[-0.1r, 0.1r]$ are imputed as their predictions and judgments of failure rates within a ROPE of $[-0.2\sigma, 0.2\sigma]$.

[7] In general, X includes a constant term. D itself need not be linear in its underlying variable.

[8] 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.

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

In practice, economists using the standard NHST framework conclude that there is (not) a relationship between Y and D if H_0 is (not) rejected (Romer 2020; Imbens 2021). Table 2 details the ways in which economists make null claims when H_0 is not rejected. Specifically, I use a slightly modified version[9] of the categorization from Gates & Ealing's (2019) survey of null claims in medical journals to classify all null claims in my intermediate sample. Table 2 shows that economists frequently make null claims based on statistically insignificant models. Though this practice is not unique to economics (Gates & Ealing 2019), a striking feature of the way 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 claims are unambiguous assertions that the relationship of interest is negligible or nonexistent.

Of course, if $\hat{\delta}$ is statistically insignificant, this does not necessarily imply that δ is negligibly small. A statistically insignificant result could simply reflect imprecision arising from low power. Under the standard NHST framework, as s grows arbitrarily large, $\hat{\delta}$ will always be statistically insignificant even if $\hat{\delta}$ is arbitrarily large. Therefore, generally inferring a null result from a statistically insignificant estimate under the standard NHST framework can often result in erroneously deeming that a genuinely meaningful relationship does not exist, among other negative consequences.

[9] 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.

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

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. The rate of Type I errors is largely controlled by the significance level α, which is traditionally set at 0.05.[10] The rate of Type II errors $\beta_{\text{NHST}} \in (0, 1]$ relates to the power $(1 - \beta_{\text{NHST}})$ with which one can detect an effect size at or above some effect size $\epsilon \geq 0$ under the standard NHST framework. As the complement of the standard NHST Type II error rate for effect size ϵ, $(1 - \beta_{\text{NHST}})$ represents the probability that $\hat{\delta}$ is statistically significant under the standard NHST

[10] 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.

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 $t^{*}_{\alpha,df}$ as defined in Equation 2. Then given α, power to detect an effect size of $|\delta| \geq \epsilon$ can be written as[11]

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

Power levels above (below) 0.8 are generally considered to be (in)sufficient in 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 show that $\delta < \epsilon$ with high probability. In principle, if power is sufficiently high in the economics literature, then insignificant results in that 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. This low power poses serious challenges for the credibility of null claims in economics. When the researcher uses the standard NHST framework in Definition 3.1 when interested in claiming that $\delta = 0$, 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

[11] 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).

show that $\delta = 0$, power is 'bad' because imprecision is 'good', as the probability of finding a null relationship is inversely related to statistical precision.

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-powered results. Generally inferring null relationships from statistically insignificant results without any regard to the Type II error control implied by the power of the results 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 in many cases relevant to empirical economics, one can not 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 belief that null results are low-quality and unpublishable (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 than positive, statistically significant results, leading to widespread publication bias throughout the economics literature (Fanelli 2012; Franco, Malhotra, & Simonovits 2014; Andrews & Kasy 2019). Worse yet, the high Type II error rate tolerance suggested by current economic practice implies 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 Type II error control 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

For economic hypotheses to be practically falsifiable, economists must be able to credibly demonstrate that relationships and phenomena are negligible or nonexistent. However, Section 3 demonstrates that the standard NHST framework is not credible when the researcher is trying to show that $\delta = 0$, as this desired conclusion is assumed to be true in the null hypothesis. This issue can be corrected in two steps. First, the constraints in Equation 1 can be relaxed; rather than assessing whether $\delta = 0$ strictly, one can test instead whether $\delta \approx 0$. Thereafter, the null and alternative hypotheses in Equation 1 can be flipped, restoring the burden of proof on researchers trying to show that $\delta = 0$. The resulting hypotheses take the form

$$\begin{aligned} H_0 &: \delta \not\approx 0 \\ H_A &: \delta \approx 0. \end{aligned}$$

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.[12]

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

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

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

The different tests assessing whether $\hat{\delta}$ is statistically significantly bounded within $[\epsilon_-, \epsilon_+]$ are discussed throughout this section. In equivalence testing, $[\epsilon_-, \epsilon_+]$ is referred to as the *region of practical equivalence (ROPE)*, which is the range of values within which one would deem δ to be practically equivalent to zero. ROPE boundaries thus effectively designate which values of δ are economically (in)significant. ROPEs are often (though not always) symmetric around zero such that $\epsilon_- = -\epsilon_+$.[13] A symmetric ROPE around zero can be said to have a *length* of $\epsilon > 0$ and written as $[-\epsilon, \epsilon]$.

Though equivalence testing has historically been challenged by difficulties with establishing credible ROPES (see Ofori et al. 2023), relatively new virtual 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

[12] Though equivalence testing can be used to test a relationship's practical equivalence to any value, here I follow similar practice to Definition 3.1, in the sense that my definition of the equivalence testing framework here is limited to the typical use case where a researcher wants to show that there is virtually zero relationship between Y and D for ease of exposition.

[13] E.g., 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, 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 $|\epsilon_-| < |\epsilon_+|$.

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.[14]

4.1 Two One-Sided Tests Procedure

The hypotheses in Definition 4.1 can be rewritten as

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

Further, this joint hypothesis can be separated into two one-sided hypotheses:

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

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 TOST procedure.

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

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

and compute

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

[14] 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).

The exact critical value for this test can be written as

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

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

Put simply, at a 5% significance level, the TOST procedure deems $\hat{\delta}$ to be significantly bounded within a ROPE if it is bounded 1) ≈ 1.645 SEs *above* the ROPE's *lower* bound, and 2) ≈ 1.645 SEs *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 Type I error rate control than inferring null results from adequately-powered statistically insignificant results under the standard NHST framework. The TOST procedure's Type I error rate 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

The TOST procedure at a significance level of α yields equivalent results to a CI-based approach that makes use of the symmetric $(1-2\alpha)$ CI (Berger & Hsu 1996). Following Hartman (2021), I term this interval the ECI.

Definition 4.3 (The Equivalence Confidence Interval)**.** *The researcher wants to test the hypotheses in Definition 4.1 using a test with Type I error rate α. They thus formulate a real interval*

$$ECI_{1-\alpha} = \left[\Delta_-, \Delta_+\right], \tag{9}$$

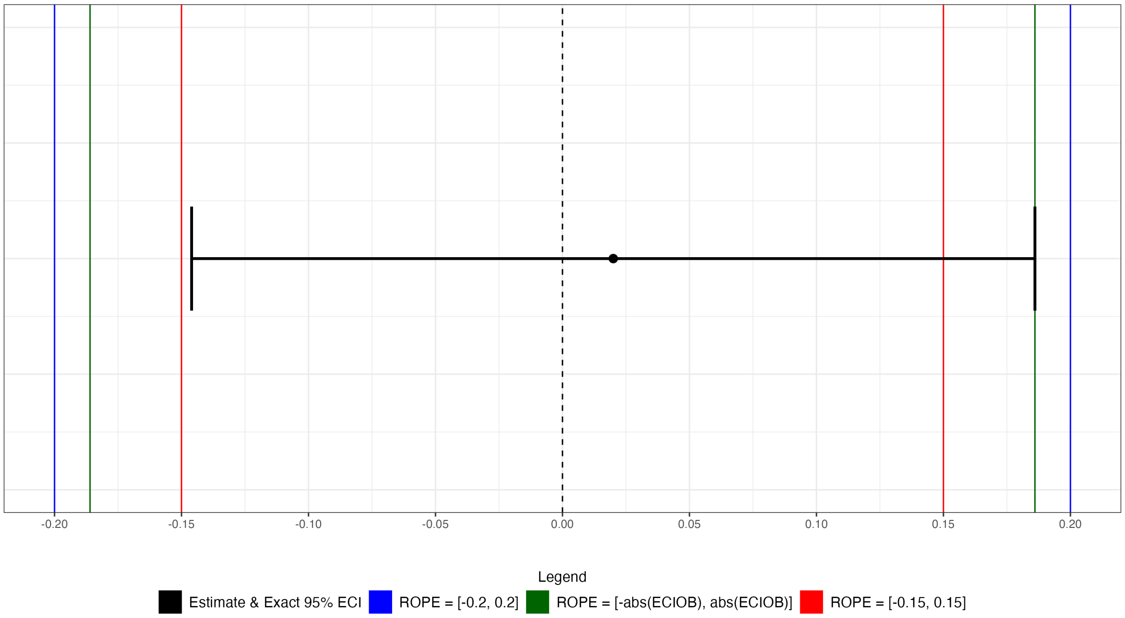
where Δ_- *and* Δ_+ *are calculated as*

$$\begin{aligned} \Delta_-(1-\alpha, df) &= \hat{\delta} - t^*_{\alpha,df}s \\ \Delta_+(1-\alpha, df) &= \hat{\delta} + t^*_{\alpha,df}s \end{aligned} \tag{10}$$

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

Because the $(1-\alpha)$ ECI is simply the $(1-2\alpha)$ CI, its computation is trivially simple. E.g., the 95% ECI is the 90% CI, the 90% ECI is the 80% CI, and so forth. The key differences between ECIs and CIs are the ways in which they can be used to judge statistical significance. Standard NHST significance judgments are derived from CIs based on the CI's relationship with zero. In contrast, equivalence testing significance judgements 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.

Figure 1 provides an illustration 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 $\hat{\delta}$ can thus be roughly written as $[-0.146, 0.186]$, as $t^*_{0.05,100} \approx 1.66$. This implies that the *ECI outer bound (ECIOB)* is $\text{ECIOB}_\alpha \approx 0.186$, as the upper bound is further away from zero than the lower bound. If the ROPE is set as $[-0.2, 0.2]$ – the blue lines in the figure – then $\hat{\delta}$ is statistically significantly bounded within the ROPE at a significance level of $\alpha = 0.05$, because the entire 95% ECI is bounded within the specified ROPE. However, the same conclusion can not be reached if the ROPE is instead specified as $[-0.15, 0.15]$ (the red lines in Figure 1). $\hat{\delta}$'s $(1-\alpha)$ ECI is the smallest ROPE wherein one can find statistically significant evidence that δ is practically equivalent to zero. Thus the magnitude of ECIOB_α is the length of the smallest symmetric ROPE around zero wherein one could find statistically significant evidence that $\delta \approx 0$ at significance level α (this is



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

Figure 1: An ECI Example

illustrated by the green lines in Figure 1). Therefore, the magnitude of ECIOB$_\alpha$ 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

$$\sigma = \begin{cases} \frac{\delta}{\sigma_Y} \text{ if } D \text{ is binary} \\ \frac{\delta\sigma_D}{\sigma_Y} \text{ otherwise} \end{cases} \qquad s = \begin{cases} \frac{\mathrm{SE}(\delta)}{\sigma_Y} \text{ if } D \text{ is binary} \\ \frac{\mathrm{SE}(\delta)\sigma_D}{\sigma_Y} \text{ otherwise} \end{cases}, \tag{11}$$

where σ_D and σ_Y respectively represent the standard deviations of the exposure and outcome variables of interest within the estimation sample, and δ is the linear association between D and Y. For binary exposure variables, standardized coefficients closely relate to the widely-used Cohen's d effect size metric, and are in fact exactly equivalent to Cohen's d in the case where no covariates are added to the model besides D (Cohen 1988, pg. 20). This also implies that my standardization produces valid estimates for cases where both outcome and exposure variables are binary, as Cohen's d values have a close effect size correspondence with odds ratios (Chen, Cohen, & Chen 2010). The standardization in Equation 11 is also a natural choice for regressions with continuous exposure variables, effectively ensuring that all exposure variables in such regressions share the same variability and scale.

The second effect size used is the *partial correlation coefficient (PCC)*, a widely-used effect size measure in meta-analyses. Per van Aert & Goos (2023), regression coefficients can be sequentially converted first into PCCs and then into PCC SEs as

$$r = \frac{t_{\text{NHST}}}{\sqrt{t^2_{\text{NHST}} + df}} \qquad\qquad \text{SE}(r) = \frac{1 - r^2}{\sqrt{df}}. \tag{12}$$

Here t_{NHST} is the usual NHST t-statistic as described in Definition 3.1, where $\delta = \sigma$ and s is the SE of σ. Note that per Equation 11, 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.

As Section 5.2 details further, failure rates measure how often estimand magnitudes in the final sample can be significantly bounded beneath classical small effect size benchmarks. I specifically use Cohen's (1988) 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 model(s) defending that claim can significantly bound their estimates beneath $|\sigma| = 0.2$ or $|r| = 0.1$ should be easy. However, the results in Section 6.2 show that many such models in the recent economics literature fail even this lenient test.

5.2 Failure Measures

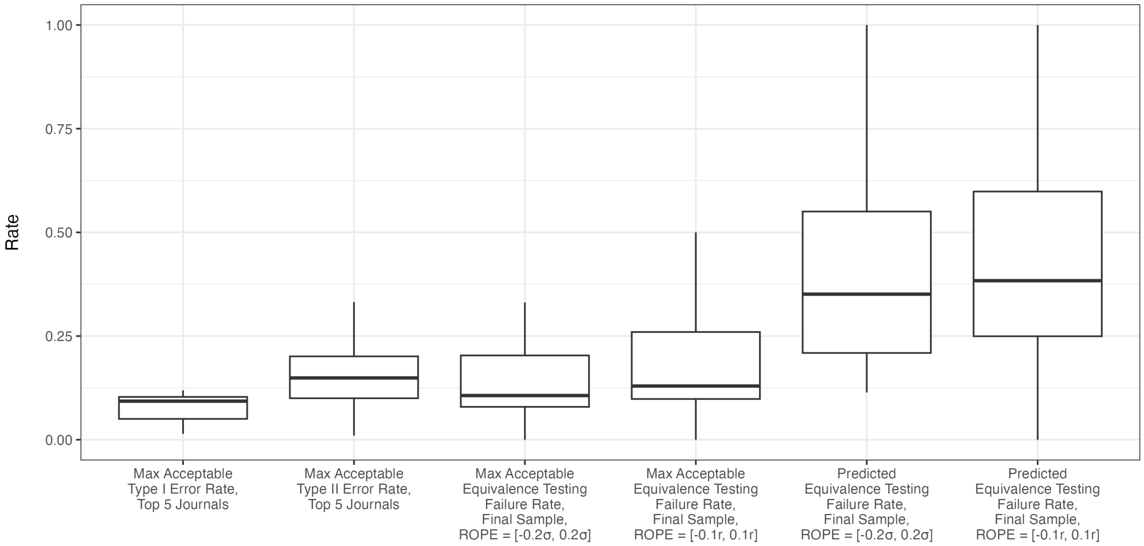
The equivalence testing *failure rate* is defined here as the average partition-level proportion of model estimates that fail to be statistically significantly bounded within a given ROPE at a 5% significance level for a given aggregation level. For example, consider a toy dataset of estimates defending three null claims. Suppose that 20% of the estimates defending the first claim can not 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 claim-level failure rate 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 failure rates. I also calculate an average inverse-weighted claim-level failure rate that ensures all articles receive the same weight in the sample. Because these average failure rates are calculated by taking a mean of partition-level failure rates over all partitions, my precision measure is the SE of that mean (SEM). Online Appendix G provides precise computational details for partition-level failure rates and their standard errors.

6 Results

6.1 Predictions and Judgments

Figure 2 presents box plots displaying descriptive statistics of the SSPP sample's predictions and judgments of failure rates in the final sample, along with their judgments

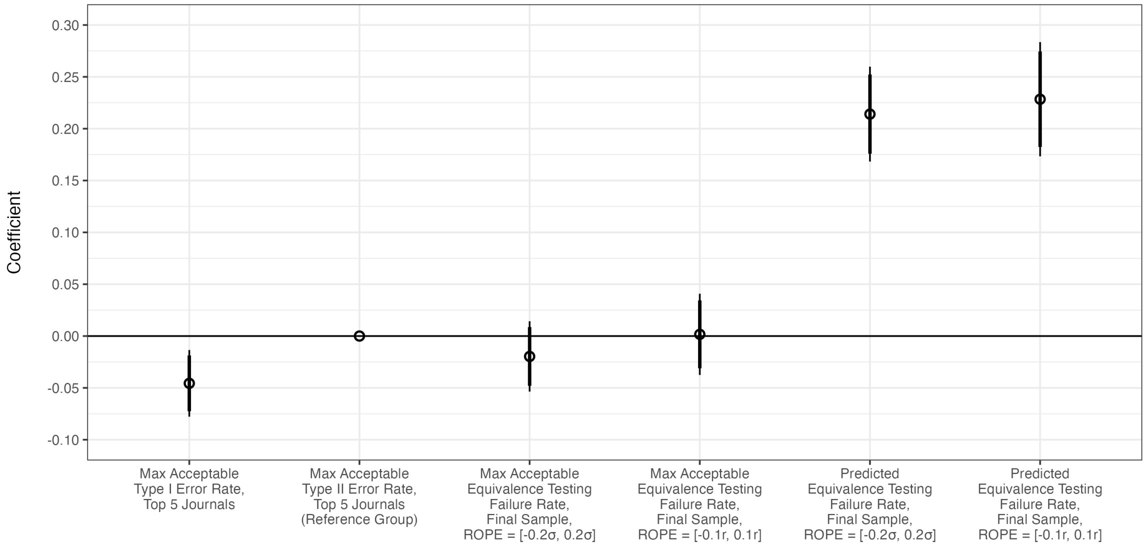


Note: Each box plot displays the 25th, 50th, and 75th percentile of its respective rate in the SSPP sample, along with whiskers that extend to the largest (smallest) point that lies within 1.5 interquartile ranges above (below) the box.

Figure 2: Distributions of SSPP Predictions and Judgments

of acceptable Type I and Type II error rates in Top 5 economics journals. 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 equivalence testing failure rates is somewhere between their median tolerances for Type I and Type II errors. The median respondent deems equivalence testing failure rates up to 10.65% to be acceptable for a ROPE of $[-0.2\sigma, 0.2\sigma]$, while this tolerance increases to 12.95% for a ROPE of $[-0.1r, 0.1r]$. However, respondents predict that failure will greatly exceed these thresholds. Median predictions for equivalence testing failure rates 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 failure rates for a ROPE of $[-0.1r, 0.1r]$ is an underestimate.

Figure 2 displays substantial dispersion across many variables, particularly predicted failure rates. Though this partially reflects considerable disagreement on pre-



Note: γ_r estimates from Equation 13 are provided along with 95% CIs and ECIs. SEs 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

dictions and judgments on failure rates, 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 rate predictions/judgments using a panel data regression model that controls for researcher fixed effects. I.e., letting i index the researcher and r index one of the six rates displayed in Figure 2, I estimate the model

$$\text{Rate}_{i,r} = \theta + \gamma_r + \lambda_i + \mu_{i,r}. \tag{13}$$

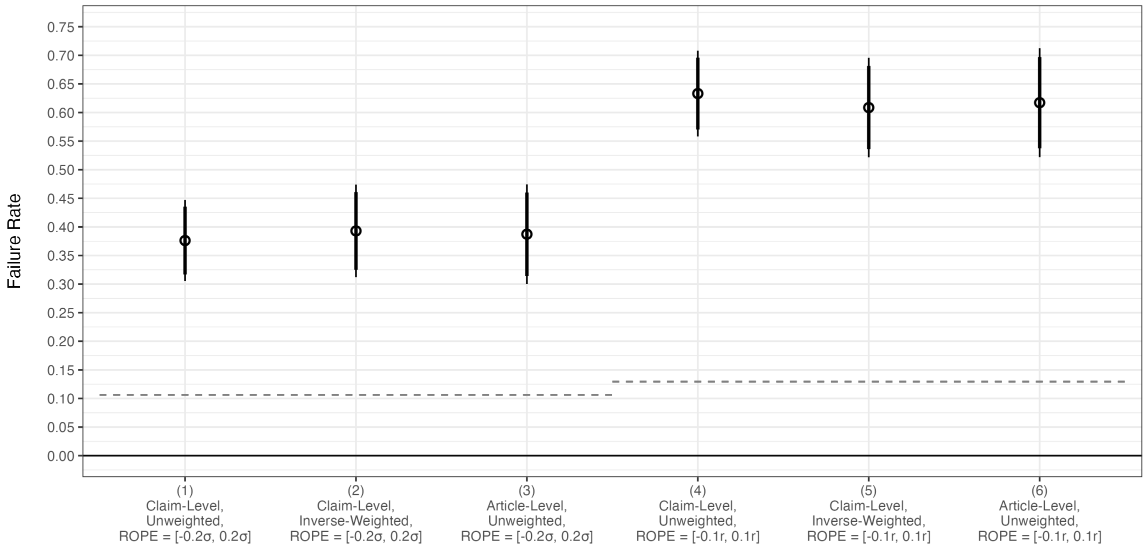
Figure 3 displays the estimates of γ_r from a model of Equation 13 that treats judgments on Type II error rates as the reference group.[15] On average, a given researcher reports that for results in Top 5 economics journals, their tolerance for Type I error rates is 4.561 percentage points lower than their tolerance for Type II error rates. This is direct evidence of a preference-based null result penalty (see Chopra et al.

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

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 equivalence testing failure rate tolerance is quantitatively close to Type II error rate tolerance. The average researcher's tolerance for Type II errors is 1.966 percentage points higher than their tolerance for equivalence testing failure rates within a ROPE of $[-0.2\sigma, 0.2\sigma]$, and 0.165 percentage points lower than their tolerance for failure rates 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 rates, it is not clear that this difference is practically equivalent to zero; it in fact exceeds the difference I detect between Type I and Type II error rate tolerances. There is thus insufficient power to say that equivalence testing failure rate tolerances are practically equivalent to Type II error rate tolerance.

However, researchers' predictions of equivalence testing failure rates in my final sample far exceed any of these acceptability thresholds. The average researcher predicts that equivalence testing failure rates will exceed their Type II error rate tolerance by 21.406 percentage points within a ROPE of $[-0.2\sigma, 0.2\sigma]$, and by 22.842 percentage points within a ROPE of $[-0.1r, 0.1r]$. Accounting for the aforementioned differences between Type II error rate tolerance and equivalence testing failure rate tolerance, these estimates imply that the average researcher predicts that equivalence testing failure rates will be around 23 percentage points higher than the maximum levels they would find acceptable. 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 failure rate estimates in the remainder of this section show that this prediction is quite accurate.



Note: Equivalence testing failure rates are provided along with 95% ECIs and CIs, based on the SEM for unweighted failure rates and the weighted SEM for weighted failure rates (see Online Appendix G). Dashed lines represent the median SSPP respondent's maximum acceptable claim-level failure rate for the given ROPE at a 5% significance level.

Figure 4: Main Failure Rate Estimates

6.2 Failure Rates

Figure 4 displays the main failure rate estimates.[16] The dotted lines represent the median SSPP respondent's threshold for acceptable failure rates (see Section 6.1). Failure rates lie significantly above both zero and these thresholds. For a ROPE of $[-0.2\sigma, 0.2\sigma]$, equivalence testing failure rates range from 37.6-39.3%. These failure rates are even higher for a ROPE of $[-0.1r, 0.1r]$, ranging from 60.9-63.3%. Therefore, equivalence testing failure rates within lenient benchmark ROPEs range from 37-63% for recent null claims in top economics journals.

The significance of these failure rates is robust to a wide range of checks. Principally, failure rates are not sensitive to the choice of aggregation procedure. Within each effect size metric, failure rates vary by less than 1.8 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

[16] A table version of these failure rate estimates is provided in Online Appendix Table A3.

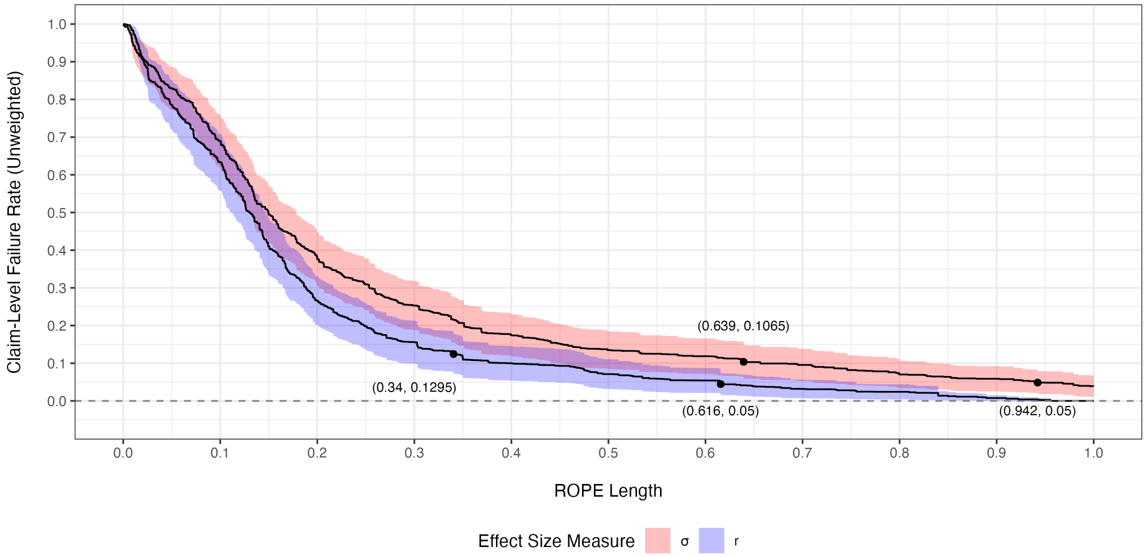
failure rates or by applying inverse weighting, increases failure rates for standardized coefficients but decreases failure rates for PCCs. It thus poses no threat to robustness to prefer one aggregation level when interpreting results. I therefore primarily reference unweighted claim-level failure rates in my discussion of results, largely due to relative ease of interpretability. For instance, Model 1 in Figure 4 implies that the average proportion of estimates defending a null claim in the final sample that are not significantly bounded beneath an effect size of 0.2σ is 37.6%.

Online Appendix Table A4 shows that failure rates remain significantly bounded above acceptability thresholds regardless of whether models that are initially statistically significant under NHST are removed from the sample. Additionally, Online Appendix Table A5 shows that failure rates 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 failure rates are robust to coding choices; using the same leave-one-out approach, I show that failure rates remain significantly above acceptability thresholds after removing models that are not fully replicable, and after removing models that require conformability modifications.

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 failure rates in the final sample vary with the choice of ROPE length ϵ. The shapes of the failure curves reflect the intuition that failure rates decline when one is willing to tolerate larger ROPEs. Figure 5 shows that failure rates remain significantly above nominal and acceptable levels even as ROPE lengths grow quite large. These general findings hold regardless of which effect size measure is used (i.e., σ or 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



Note: Failure curves are annotated by points indicating the ROPES that must be tolerated to bound failure rates beneath 1) 5% and 2) the median SSPP respondent's maximum acceptable level for claim-level failure rates within the benchmark ROPEs tested when producing Figure 4's estimates. Uncertainty bands represent 95% CIs based on the claim-level failure rate's SEM (see Online Appendix G).

Figure 5: Failure Curves

practices for null claims in economics are sufficient, and that failure rates are in fact 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 failure rates beneath 12.95% – the median SSPP respondent's maximum equivalence testing failure rate tolerance for a ROPE of $[-0.1r, 0.1r]$ – one must argue that a PCC magnitude of 0.34 is practically equivalent to zero. This is larger than over 75% of published results in economics (Doucouliagos 2011). To obtain claim-level failure rates beneath 5%, one must be willing to claim that a PCC magnitude of 0.616 is practically equivalent to zero, which is obscenely large.

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 failure rate is unreasonably large. The same is true of the 0.639σ effect size that is necessary to bound claim-level failure rates beneath 10.65%, the equivalence testing failure rate which the median SSPP respondent would tolerate for a ROPE of $[-0.2\sigma, 0.2\sigma]$.

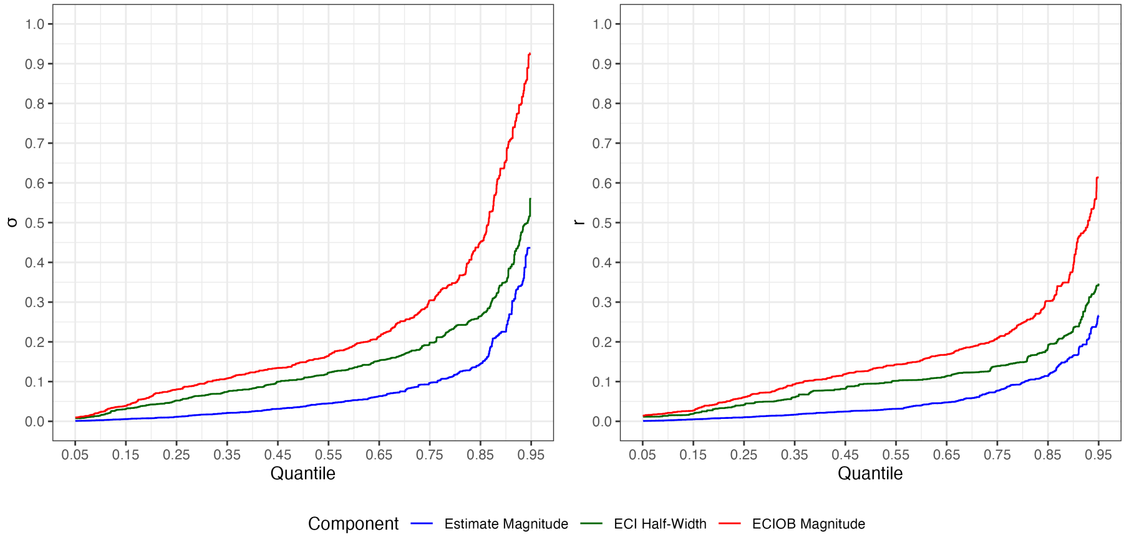
Given the absurdity of contentions that effect sizes this large are practically equivalent to zero, one is compelled to accept the more sensible alternative conclusion that the current testing paradigm used by economists to make and defend null claims tolerates unacceptably high error rates. Many meaningful economic relationships are thus likely erroneously dismissed as negligible or nonexistent under standard NHST.

6.4 Mechanisms

Are these high failure rates caused more by large effect sizes or by imprecision? Recall from Section 4.2 that the magnitude of the ECI outer bound (ECIOB) is the length of the smallest symmetric ROPE 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, and thereby directly determine failure rates for a given ROPE. This question of failure rate mechanisms can therefore be answered by decomposing the exact 95% ECIOB magnitude into its two constituent parts (see Definition 4.3): the estimate's magnitude $\left|\hat{\delta}\right|$, which measures effect size, and the estimate's 95% ECI half-width $st^*_{0.05,df}$, which measures imprecision.

Figure 6 plots the distributions of ECIOB magnitudes and their components. If estimate magnitudes were the only driver of failure rates, 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. Therefore, though both large effect sizes and low precision substantively contribute to high failure rates, 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 ECI half-widths. Both effect sizes and ECI half-widths are highly statistically significantly



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

associated with ECIOB magnitudes in theoretically expected directions. However, ECI half-widths display noticeably stronger relationships with ECIOB magnitudes than effect sizes, particularly for PCCs. For standardized coefficients, the elasticity of ECIOB magnitudes with ECI half-widths is around 13% larger than that elasticity for effect size $|\sigma|$. However, for PCCs, the elasticity of ECIOB magnitudes with ECI half-widths is around 121% larger than that elasticity for effect size $|r|$. This provides additional evidence that though large effect sizes are an important factor high failure rates, imprecision is the dominant determinant.

Table 3 also yields encouraging evidence concerning the finite-sample properties of equivalence testing. Recall from Section 3 that a key problem 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 one can bound an estimate significantly closer to zero when one has more precise estimates. This provides clear evidence that when the researcher

	$\lvert\sigma\rvert$	ECI Half-Width, σ	$\lvert r\rvert$	ECI Half-Width, r
Elasticity **w/ |ECIOB|**	0.577 (0.131)	0.654 (0.05)	0.421 (0.03)	0.932 (0.054)
Adj. R^2	0.599	0.933	0.766	0.752

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

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 shows 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 equivalent 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. Given the high penalty against statistically insignificant results (McShane & Gal 2016, McShane & Gal 2017; Chopra et al. 2024), a key concern is *ROPE-hacking*, a practice whereby researchers 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 (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 can not fix, as researchers interested in obtaining evidence of null findings can simply pre-register an excessively wide ROPE. Unsurprisingly, this practice can inflate Type I error rates in equivalence testing (Campbell & Gustafson 2021).

To control researcher degrees of freedom and ensure credible, independently-set significance thresholds, I recommend ROPE-setting methods that elicit judgments on proper ROPEs from independent parties, such as experts or stakeholders related to the research question. Such judgments are practical to elicit using recently-developed research-centric survey platforms, such as the SSPP (see Section 2.1). 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 equivalent 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 failure rates they predict, I also asked the largest failure rates 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 data on predictions regarding the relationship(s) of interest will still be useful to such researchers to help inform posterior beliefs of such relationships’ distributions 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 ROPE 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 can not 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), implying 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. Principally, 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.

Additionally, because the CET procedure performs multiple hypothesis tests that are not mutually exclusive under closure, controlling Type I error rates of CET conclusions requires multiple hypothesis testing adjustments. Such adjustments can considerably reduce power. However, ignoring the need for such adjustments will inflate Type I error rates of conclusions derived from the CET procedure.

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 equivalent 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 $\delta \notin [\epsilon_-, \epsilon_+]$. 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 equivalent to zero. In such settings, researchers can not 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 equivalence testing failure rates 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.

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

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

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

$$\begin{array}{lll} H_0^{\{N\}}: \delta \geq \epsilon_- & H_0^{\{TOST\}}:\ \delta < \epsilon_- \textit{ or } \delta > \epsilon_+ & H_0^{\{P\}}: \delta \leq \epsilon_+ \\ H_A^{\{N\}}: \delta < \epsilon_- & H_A^{\{TOST\}}:\ \delta \geq \epsilon_- \textit{ and } \delta \leq \epsilon_+ & H_A^{\{P\}}: \delta > \epsilon_+. \end{array} \tag{14}$$

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

$$t_N = \frac{\hat{\delta} - \epsilon_-}{s} \qquad\qquad t_P = \frac{\hat{\delta} - \epsilon_+}{s}. \tag{15}$$

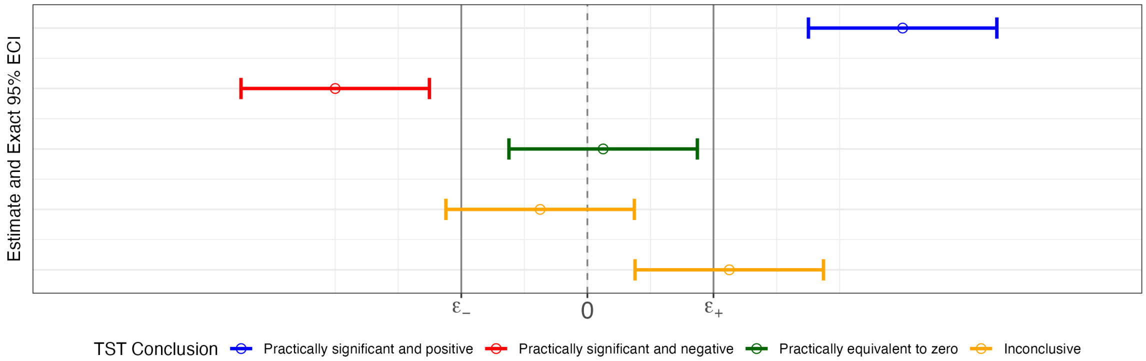
The critical value can be written as $t^*_{\alpha,df}$, as in Equation 8. The researcher concludes that δ is significantly bounded above the ROPE if and only if $t_P > t^*_{\alpha,df}$. The researcher concludes that δ is significantly bounded below the ROPE if and only if $t_N < -t^*_{\alpha,df}$. As 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). This is because the alternative hypotheses of each of the TST's three hypotheses in Equation 14 completely partition the parameter space of δ into disjoint regions. Because only one of the three alternative hypotheses in Equation 14 can thus be true at one time, the partitioning principle implies that if all three hypothesis tests in the TST framework are performed at significance level α, then the family-wise error rate of a single application of the TST framework is controlled at α, even without multiple hypothesis testing adjustments (Shaffer 1986; Finner & Strassburger 2002). 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 primary empirical findings of this paper provide an example of how conclusions can be made using the TST framework. The question of whether equivalence testing failure rates are significantly greater than zero is uninteresting; failure rates are greater than zero almost by construction. However, as aforementioned in Section 7.1, thresholds for maximum acceptable equivalence testing failure rates 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, I take the median of these judgments ϵ and set a ROPE of $[0, \epsilon]$. In Section 6.2, I then show that the 95% ECIs of my main failure rate estimates are bounded above these ϵ thresholds, which provides statistically significant evidence that the failure rates in my final sample are practically significant.[17]

The TST framework also accommodates ECIs for all of its hypotheses. ECIs can be applied to the TOST hypotheses of the TST framework in the usual way, as described in Section 4.2. However, these intervals can also be applied to the TST framework's other two tests that assess whether $\hat{\delta}$ statistically significantly exceeds the bounds of the ROPE. Figure 7 visualizes the ways in which research conclusions under the TST framework can be derived from ECIs. Recall from Definition 4.3 that estimate $\hat{\delta}$ is statistically significantly bounded above ϵ_- (below ϵ_+) at significance level α if and

[17] These conclusions are also supported under the standard NHST framework.



Note: The scale of these estimates and 95% ECIs is arbitrary. ϵ_- and ϵ_+ respectively denote the lower and upper boundaries of the ROPE for these estimates.

Figure 7: ECIs and Research Conclusions in the TST Framework

only if the $(1-\alpha)$ ECI of $\hat{\delta}$ is entirely above ϵ_- (below ϵ_+). In the same vein, $\hat{\delta}$ is statistically significantly bounded above ϵ_+ (below ϵ_-) at significance level α if and only if the $(1-\alpha)$ ECI of $\hat{\delta}$ is entirely above ϵ_+ (below ϵ_-).

Switching from the standard NHST framework to the TST framework generates a power tradeoff. On one hand, the one-sided tests under the TST framework require test statistics to exceed smaller critical values than those under the standard NHST framework before declaring that an estimate is statistically significant because tests under the TST framework are conducted using one-sided, rather than two-sided, critical values. On the other hand, the use of a nontrivial ROPE around zero implies that TST test statistics t_N and t_P have numerators that are almost always smaller than (and at most as large as) the numerator of standard NHST test statistic t_{NHST}. Thus depending on the ROPE and the precision of $\hat{\delta}$, the researcher may have more or less power to detect a practically significant relationship under the TST framework than they do to detect a statistically significant relationship under standard NHST. This implies that researchers may have greater ability to find statistically significant evidence under the TST framework, and also that practical significance testing under the TST framework is a useful robustness check for results that are statistically significant under standard NHST.

8 Conclusion

I introduce the economics literature to a suite of simple equivalence testing methods. I then use these methods to demonstrate that a substantial proportion of the models defending published null claims in top economics journals can not statistically significantly bound their estimates within lenient benchmark ROPEs. At a 5% significance level, failure rates for these models range from 26-39% within benchmark ROPEs. To obtain acceptable failure rates, one must claim that over half of all published effect sizes in economics are practically equivalent to zero. Because it is ludicrous to claim that the magnitudes of so many published economic estimates are practically equivalent 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 SSPP (DellaVigna, Pope, & Vivalt 2019).

The ROPEs constructed from these judgments can then be used to test estimates using the TST framework, which has several advantageous properties. First, the TST framework permits researchers to simultaneously test for an estimate's practical significance and practical equivalence to zero, while controlling Type I 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 framework requires such conclusions in these settings. I additionally provide statistical software suites for researchers to implement these recommendations in practice.

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 equivalent 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 equivalent to zero before making broader inferences from their statistical findings. 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.

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.

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.

Chen, Henian, Patricia Cohen, and Sophie Chen (2010). “How big is a big odds ratio? Interpreting the magnitudes of odds ratios in epidemiological studies”. *Communications in Statistics - Simulation and Computation* 39.4, pp. 860–864. DOI: 10.1080/03610911003650383.

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.* 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.

Finner, H. and K. Strassburger (2002). “The partitioning principle: A powerful tool in multiple decision theory”. *The Annals of Statistics* 30.4, pp. 1194–1213. DOI: 10.1214/aos/1031689023.

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.

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.

Shaffer, Juliet Popper (1986). "Modified sequentially rejective multiple test procedures". *Journal of the American Statistical Association* 81.395, pp. 826–831. DOI: 10.1080/01621459.1986.10478341.

StataCorp (2023). *Stata power, precision, and sample-size reference manual.* Vol. 18. Stata Press.

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 in a Top 5 economics journal (specifically *American Economic Review*, *Econometrica*, *Journal of Political Economy*, *Quarterly Journal of Economics*, and *Review of Economic Studies*) from 2015 onwards. I obtained bibliographic information on this initial set of 3732 articles, including digital object identifiers (DOIs), 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 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 model(s) used to support that claim within the article. I discard a claim if it is not defended by at least one statistically insignificant model, otherwise storing the main model(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 models across 279 claims in 158 articles. Thereafter, I attempt to reproduce every model in the intermediate sample. Models 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 models 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 is relatively unconfident in their predictions for this particular survey, rating their five-point Likert confidence in their predictions at a median of 2.5 (mean = 2.4, SD = 1). This is reflected by the fact that 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 model used to defend this claim. The benchmarking sample thus consists of ten articles, each with one claim and one model 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. Even amongst a benchmark sample of articles advertising plausibly large economic effects, six of ten estimates are either smaller than $\sigma = 0.2$ or $r = 0.1$.

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

Note: Effect sizes and initial p-values of each model are reported. The original estimate of each model 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 online versions of their submissions.

Table A1: Effect Size Benchmarking

F Benchmarking Sample

ree

G Failure Measures

Let j be an individual partition,[3] and let i index an individual model. Each model i belongs to exactly one partition j. Because all failure rates in this paper are calculated for symmetric ROPEs, it is sufficient to define failure rate $R(\epsilon, \tau, L)$ 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 standard TOST procedure described in Definition 4.2, I approach failure rate calculation by defining exact values for the 95% ECI outer bound $\text{ECIOB}_{i,j}(\tau)$ for each effect size measure τ of every model i belonging to every partition j. Let M_j represent the number of models i belonging to partition j, and let M be the total number of partitions j. One can then calculate the failure rate 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 failure rates that apply an inverse weighting approach ensuring that each article receives the same weight in the sample. Let U be a partition clustered in exactly one partition level H, and let $M^{\{U\}}$ be the total number of partitions U in the data. Then

$$W_{j,k} = \frac{1}{\sum_{j=1}^{M^{\{U\}}} \mathbb{1}\left[U_{j,k} \in H_k\right]}$$

is the inverse weight of partition $U_{j,k}$. In this setting, $U_{j,k}$ is claim j belonging to article k (H_k), so $W_{j,k}$ is simply one divided by the number of claims that belong to

[3] j represents an individual claim when calculating claim-level failure rates, whereas j represents an entire article when calculating article-level failure rates.

claim j's article. Then the inverse-weighted claim-level failure rate can be written as

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

where $M_{j,k}$ is now the number of models belonging to clustered partition $U_{j,k}$ – in this setting, this is simply the number of models belonging to claim j in article k.

I measure precision using standard errors of the mean for the unweighted failure rates in Equation A1 and standard errors of the weighted mean for the weighted failure rates in Equation A2. The standard error of the mean for a failure rate 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 the $R(\epsilon, \tau, L)$ vector. Let the failure rate 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

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

Here $\overline{W}_{j,k}$ is the mean inverse weight across all claims. The results in Section 6.2 show that this standard error derivation corresponds quite closely with simple standard errors for unweighted failure rates 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	$\cdot$	-0.02	0.002	0.214	0.228
	(0.016)	$(\cdot)$	(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)
Failure Rate	0.376	0.393	0.387	0.633	0.609	0.617
	(0.036)	(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 Failure Rate Estimates

I Robustness Checks

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

	Models	Claims	Articles	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Initially	788	132	80	0.345	0.36	0.353	0.612	0.587	0.594
Insignificant Models				(0.036)	(0.041)	(0.044)	(0.039)	(0.045)	(0.05)
Panel B: Initially	88	34	27	0.601	0.639	0.636	0.735	0.765	0.765
Significant Models				(0.084)	(0.085)	(0.089)	(0.077)	(0.075)	(0.081)
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: Models are deemed initially (in)significant if the standard NHST p-value of initial model estimate (before conformability changes, if applicable) is less than (greater than or equal to) 0.05. ROPEs are $[-0.2\sigma, 0.2\sigma]$ and $[-0.1r, 0.1r]$.

Table A4: Failure Rate Robustness: Initial Model Significance

	Models	Claims	Articles	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: CYCD	675	105	63	0.342	0.362	0.356	0.62	0.617	0.628
Removed				(0.04)	(0.046)	(0.049)	(0.044)	(0.049)	(0.054)
Panel B: CYBD	563	91	59	0.36	0.37	0.369	0.621	0.558	0.562
Removed				(0.045)	(0.049)	(0.054)	(0.047)	(0.053)	(0.058)
Panel C: BYCD	563	124	74	0.398	0.417	0.409	0.651	0.631	0.64
Removed				(0.038)	(0.043)	(0.047)	(0.04)	(0.046)	(0.051)
Panel D: BYBD	653	119	73	0.365	0.39	0.386	0.634	0.625	0.629
Removed				(0.038)	(0.043)	(0.046)	(0.04)	(0.046)	(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 models with continuous/binary outcome/exposure variables (respectively) are removed from the sample. For example, 'CYBD removed' implies that models with 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: Failure Rate Robustness: Regressor Type Combination

	Models	Claims	Articles	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Non-Replicable Models 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 Models Removed	807	130	77	0.374 (0.036)	0.379 (0.041)	0.373 (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: Models are non-replicable if my best attempts to replicate the exact published estimates using the article's replication repository do not succeed. Models are 'non-conformable' if they require conformability modifications before inclusion in the final sample. ROPEs are $[-0.2\sigma, 0.2\sigma]$ and $[-0.1r, 0.1r]$.

Table A6: Failure Rate Robustness: Replicability/Conformability