NBER WORKING PAPER SERIES

HOW CREDIBLE IS THE CREDIBILITY REVOLUTION?

Kevin Lang

Working Paper 31666
http://www.nber.org/papers/w31666

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2023

This research was supported in part by NSF grant SES-1851636. I am indebted to Qingyuan Chai and Xinze Liu for superb research assistance. This paper is based on my Presidential address to the Society of Labor Economists. I thank the participants there and at the Boston University empirical microeconomics workshop, the Hong Kong Baptist University Political Economy seminar, the Canadian Labour Economics Forum annual conference, and the (Australian) Labour Econometrics Workshop, and Isaiah Andrews, Henry Braun, and James MacKinnon for their helpful comments and questions. The usual caveat applies. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

How Credible is the Credibility Revolution?
Kevin Lang
NBER Working Paper No. 31666
September 2023
JEL No. A10,C12

ABSTRACT

When economists analyze a well-conducted RCT or natural experiment and find a statistically significant effect, they conclude the null of no effect is unlikely to be true. But how frequently is this conclusion warranted? The answer depends on the proportion of tested nulls that are true and the power of the tests. I model the distribution of t-statistics in leading economics journals. Using my preferred model, 65% of narrowly rejected null hypotheses and 41% of all rejected null hypotheses with $|t|<10$ are likely to be false rejections. For the null to have only a .05 probability of being true requires a t of 5.48.

Kevin Lang
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
and NBER
lang@bu.edu

How Credible is the Credibility Revolution?*

Kevin Lang†
Boston University, NBER, and IZA

July 19, 2023

Abstract

When economists analyze a well-conducted RCT or natural experiment and find a statistically significant effect, they conclude the null of no effect is unlikely to be true. But how frequently is this conclusion warranted? The answer depends on the proportion of tested nulls that are true and the power of the tests. I model the distribution of t-statistics in leading economics journals. Using my preferred model 65% of narrowly rejected null hypotheses and 41% of all rejected null hypotheses with $|t| < 10$ are likely to be false rejections. For the null to have only a .05 probability of being true requires a t *of* 5.48.

1 Introduction

Suppose you test a null hypothesis, and the t turns out to be 1.96. Assume the model is correctly specified and the $t - statistic$ is really distributed as t. What is the probability that the null hypothesis is actually true?

If you're a nihilist, you may respond that all nulls are false. That seems to be an argument against hypothesis testing. It is one you should eschew if you adhere to the frequentist tradition. The researcher who "knows their model is wrong" nevertheless expects us to believe that their model is approximately correct and, therefore, suitable for policy analysis. So either we should ignore the nihilist's policy analyses, or we should be willing to ask

*This research was supported in part by NSF grant SES-1851636. I am indebted to Qingyuan Chai and Xinze Liu for superb research assistance. This paper is based on my Presidential address to the Society of Labor Economists. I thank the participants there and at the Boston University empirical microeconomics workshop, the Hong Kong Baptist University Political Economy seminar and the Canadian Labour Economics Forum annual conference, and Isaiah Andrews, Henry Braun, and James MacKinnon for their helpful comments and questions. The usual caveat applies.

†Boston University, lang@bu.edu

whether its predictions are contradicted. I accept that sometimes it may be appropriate to view the null as "the coefficient is *approximately* 0." However, the increase in effective size from changing the null in this way will generally not be large, as I confirm in the empirical part of this paper.

On the other hand, if we don't stop to think, most of us trained in the frequentist tradition will respond "5 percent." As Colquhoun (2014) points out, this is obviously incorrect. The probability that the null is false depends on the likelihood of getting a t of 1.96 if the null is false and, thus, indirectly, on the power of the test. The probability also depends on the *ex-ante* probability that the null was true, your prior if you are a Bayesian. If we are almost sure the null hypothesis is false, we should continue believing that the null is false even when we fail to reject. This is the message of DeLong and Lang (1992), who find that at least two-thirds of published unrejected nulls are false and cannot reject that 100% of the unrejected nulls are false when the unrejected hypothesis is central to the paper's message. They conclude that journals publish unrejected nulls only when failing to reject them is very surprising.

This paper makes two contributions:

1. It presents a set of examples that flesh out the argument in the previous paragraph, and

2. It estimates the proportion of rejected nulls published in leading journals that are true and shows how this proportion relates to the reported $t - statistic$.

My message is related to, but distinct from, concerns about publication bias (Andrews and Kasy 2019, McCrary, Christensen, and Fanelli 2016, Dellavigna and Linos 2022). Sometimes there are or could be many studies using different data sets to test the same hypothesis. If so, publishing all properly executed studies would facilitate finding the correct answer. As is well-known, if we only publish significant findings, we may draw the wrong conclusion if we fail to correct for publication bias.

But, many settings are unique. Most social programs are rolled out only once. We can use the staggered rollout of a program to assess its effectiveness when it began. Replicating a study utilizing the rollout can check for mistakes and sensitivity to statistical technique or specification. However, we cannot examine twenty rollouts of the same program and note that we only found a significant effect for one of the twenty. A finding of a positive or null effect for a program providing food for pregnant women is not informative about whether a program providing high-quality daycare is beneficial. Publishing all results using a staggered rollout to identify policy effects will not tell us whether a particular program was effective.

Indeed, one of the credibility revolution's contributions has been to show how to determine causality in settings where we are unlikely to observe or perform multiple experiments.

My message is also distinct from concerns about $p-$hacking (Brodeur et al. 2016; Brodeur, Cook, and Heyes 2020; Elliott, Kudrin, and Wuthrich 2022). Subject to some caveats in the last of these papers, the practice of $p-$hacking creates excess density at values of t just above 1.96. Taking account of $p-$hacking lowers our confidence that narrowly rejected nulls are false. However, even if there were no $p-$hacking, the issues raised in this paper would remain. In fact, I find little evidence of $p-$hacking. Instead, I find weak evidence of a shortage of $t-statistics$ just above 1.96 or, in other words, that there is some evidence of publication bias against marginally significant results.

I begin by using some examples to discuss the probability that a rejected null is false and how that relates to power and the nature of the null and alternative hypotheses. I explore when our conclusions about the effectiveness of a policy are strong and when we should be extremely cautious about acting on a narrow rejection of ineffectiveness. I show that our confidence that a rejected null is false should often be relatively low. Moreover, the common practice of stating power against a point alternative often overstates the study's power. Recognizing that the alternative hypothesis typically places weight on a range of values generally increases the probability that a rejected null is a false discovery.

I then structurally model the distribution of $t-statistics$, conditional on $|t| > 1.96$. In the base model, I assume that some fraction of null hypotheses are true and that the $t-statistics$ for these hypotheses have a standard $t-distribution$. Given the power of the test and the true parameter value, the expected value of each of the remaining $t-statistics$ is drawn from an exponential distribution. Given its expected value, each realized t follows a noncentral $t-distribution$. My approach addresses the counterpart to the question in DeLong and Lang (1992): what proportion of rejected nulls are true?

I limit the sample to articles that measure causal effects using techniques associated with the credibility revolution (instrumental variables, randomized controlled trials, difference-in-differences, matching). This is not intended to disparage the contribution of the credibility revolution. Although I have been critical of some of the abuses of the techniques it promotes (Kahn-Lang and Lang 2020), these techniques have greatly influenced the profession, including me, in generally positive ways. However, studies drawing on credibility revolution techniques often claim "convincing evidence" of a causal effect such that we may draw a strong policy conclusion from a single study. My goal is to help us think more clearly about hypothesis testing in policy research. I focus on credibility revolution techniques because, as I have noted elsewhere (Lang and Palacios 2018), structural labor economists rarely put standard errors on their policy estimates. Moreover, most structural papers do not test a

clearly stated hypothesis.

Using the model, I ask what proportion of rejected nulls are, in fact, true. Under my preferred specification, I estimate that 41% of published rejected nulls are false rejections. Almost two-thirds of narrow rejections, those with t just above 1.96, are false rejections. To get to the conventional .05 level requires a $|t|$ greater than 5.48. Only 18% of rejected nulls, including those with $|t| > 10$, satisfy this requirement. In a policy context, unless the level of statistical significance dramatically exceeds current conventional levels, this will generally require us to be cautious about applying the findings of a single study, even one conducted honestly and carefully. Of course, in a decision-theoretic context, how certain we need to be depends on the costs of type 1 and type 2 errors.

The fraction of true but rejected nulls that I find is consistent with very high levels of publication bias, although possibly at lower rates than in Andrews and Kasy (2019). If half of studies investigate true nulls, tests have power .6, and only significant results are published, about 8% of published papers will have rejected true nulls. My estimate is substantially higher, suggesting that researchers generally study, or that good journals more commonly publish, rejected null hypotheses that are unlikely to be false.

Finally, by comparing the results in ranges where p-hacking is and is not likely to be problematic, I conclude that few marginal rejections can be attributed to this practice. The results change little if we exclude findings with $p - values$ close to .05. This suggests that p-hacking may be less of a problem than suggested by Brodeur et al. (2016) and Brodeur, Cook, and Heyes (2020), a conclusion in line with Elliott, Kudrin, and Wuthrich (2022). In fact, I find evidence of continued publication bias against $t - statistics$ marginally above 1.96.

2 Close Rejections Generally Imply a High False Discovery Rate

Proposals for funding for randomized controlled trials (RCTs) typically include power calculations. The researcher makes a statement of the form, "My sample size permits me to detect an effect size of c with power $1 - \beta$."

I am open to the argument that this power calculation answers an uninteresting question. If the policy is only worth undertaking if the effect size is .1, knowing that you have power .8 to reject 0 if the effect size is .1 is not directly helpful. More generally, Romer (2020) is undoubtedly correct that in many settings, we are interested in the coefficient estimate and confidence interval rather than whether we can reject 0. This will partially justify ignoring

very large $t-statistics$ in the empirical section. Many very large $t-statistics$ will be in settings where the researcher believes a parameter value of 0 is implausible.

However, there are also settings in which rejecting 0 or at least values very close to 0 is important. Sometimes, a theory predicts that some coefficient will be positive (or negative) but provides little guidance regarding the magnitude of the effect. In such settings, rejecting 0 in the right direction supports the theory, but we want to know how confident we should be that the null of 0 is wrong.

2.1 Back to Ba(ye)sics

For the moment, however, let's take the discussion of power at face value and consider an example of the false discovery rate in the spirit of Colquhoun (2014). For simplicity, assume the researcher knows that the standard error of the estimate is 1. Therefore, using a one-sided test, she rejects the null at the .05 level if the coefficient estimate is 1.645. For the test's power, $1-\beta$, to be .8, requires the alternative hypothesis to be $c = 2.487$.

We know in this setting that if the researcher's assumptions are correct, the probability of a type II error is .2 if the null hypothesis is false. Similarly, the probability of a type I error is .05 if the hypothesis is true and the researcher sets $\alpha = .05$.

But I want the probability of falsely rejecting the null hypothesis and, therefore, of falsely accepting the alternative when the $t-statistic$ is precisely 1.645. By Bayes rule, this probability is

$$(p_{true}|t = 1.645) = \frac{p_{true} * (density|true \wedge t = 1.645)}{p_{true} * (density|true \wedge t = 1.645) + (1 - p_{true}) * (density|false \wedge t = 1.645)}. \quad (1)$$

Note that (1) holds even if, like me, you are, broadly speaking, a frequentist. Later in this paper, I will estimate p_{true} empirically, allowing me to avoid going "full-blown Bayesian."

The formula tells us what a moment's reflection about my opening question would have; the probability that the null hypothesis is true given a $t-statistic$ of 1.645 depends on the unconditional probability that the null is true and how likely a $t-statistic$ of 1.645 is under the alternative. Applied to the example, if the *ex-ante* probability that the null is true is .5, and the estimates are normally distributed, the probability of a false discovery when $t = 1.645$ is .27.

2.2 Moving from point to interval alternatives does not help

Some readers may object that assuming a single point for the alternative hypothesis roughly captures power over some interval. Perhaps treating the alternative as uniform over some

range centered on 2.487 would be fairer. For simplicity, let us assume that the alternative is $U(0, 2 * 2.487)$.

This interval interpretation of the alternative means that the researcher has notably exaggerated the test power. If the alternative is true, the power to detect a statistically significant effect is

$$1 - \beta = \frac{1}{4.974} \int_0^{4.974} (1 - \Phi(1.645 - x))\, dx = .67 \tag{2}$$

where Φ is the standard normal distribution function.

To understand why actual power is less than claimed power, consider an extreme case where the effect size is 4.974 or arbitrarily close to 0 with equal probability. When the true effect size is approximately 0, we reject 0 about 5% of the time. In contrast, we reject almost 100% of the time when the effect size is 4.974.[1] So our power would be somewhat more than .52. More generally, moving towards 0 from the point at which we calculated our power lowers the probability of rejection by more than an equal move away from 0.

Perhaps more disturbing, the probability of a false discovery when $t = 1.645$ is given by

$$(p_{true}|t = 1.645) = \frac{p_{true} * \phi(1.645)}{p_{true} * \phi(1.645) + \frac{(1-p_{true})}{4.974} * \int_0^{4.974} \phi(1.645 - x)\, dx}$$

and, therefore, rises to .35 when we use the interval. The reason is that the probability of a false discovery when $t = 1.645$ is high if the true coefficient, c, is very low or if c is very high. As c approaches 0, at any t, the probability of a false discovery approaches the *ex-ante* probability that the null is true (.5 in the example). Similarly, when c is very high (4.974 in this example), it is unlikely that we will observe $\widehat{c} = 1.645$. Therefore, by (1), the likelihood of false discovery is high. In the example, the false discovery probability is .99 when the alternative is that $c = 4.974$ and $t = 1.645$.

2.3 Lindley's Paradox: Rejection May Support the Null

The high false discovery rate when $c = 4.974$ is an example of what is called Lindley's (1957) paradox, although it was first discussed by Jeffreys (1936). The essential point is that a narrow rejection is highly unlikely if a test is very powerful. Therefore, a narrow rejection is evidence *in favor of* the rejected null. Note that Lindley's paradox does not state that increasing power always increases the false discovery rate associated with a narrow rejection.

[1] Here and elsewhere in this section, I am cheating slightly by assuming that the variance of the coefficient estimate is known and always equals 1. Actual rejection rates for a fixed sample size would vary with the coefficient estimates if we use the standard Wald test.

There are trivial counterexamples.

To understand Lindley's paradox, consider the following extreme example. Suppose we test the null hypothesis of a fair coin against the alternative that the coin has two heads. We toss the coin 18 times and get 13 heads and 5 tails. If the coin is fair, the probability of getting heads 13 or more times is just under .05. Therefore, our result is unlikely under the null. Using classical hypothesis testing, we reject it. However, even a frequentist like me is unlikely to conclude that the alternative hypothesis is correct. The outcome is impossible under the alternative. Therefore, I must conclude that the coin is fair.

As an aside, the example also shows how a Bayesian prior can drive results. If I did not begin with the prior that the coin was fair with some probability and two-headed with one minus that probability, I would not have to conclude that the coin is definitely fair. I may be comfortable rejecting the null if the alternative is less stark. Still, I would like to know more precisely how likely I am to reject the null incorrectly.

Despite the example, spreading the distribution under the alternative symmetrically around the mean does not always increase the false discovery rate. To see this intuitively, start with an alternative with power close to 1. Therefore, the false discovery rate is approximately 1 if the rejection is narrow (e.g, $t = 1.645$). Now change the alternative so that with probability .5 each, it is twice the original and arbitrarily close to 0. When the rejection is narrow, the false discovery rate remains close to one one when the original parameter is doubled and is approximately equal to the probability the null is true when the true parameter is close to 0. Therefore, a narrow rejection is associated with a false discovery rate of approximately $.5(1 + p_{true})$.

This counterexample notwithstanding, the probability of a false discovery typically increases as the alternative hypothesis spreads out. In the appendix, I consider replacing a point alternative hypothesis as in (1) with a uniform alternative centered on the point alternative and bounded away from 0 as in (4) below.

$$(p_{true}|t = 1.645) = \frac{p_{true} * \phi(1.645)}{p_{true} * \phi(1.645) + \frac{(1-p_{true})}{2a} * \int_{c^*-a}^{c^*+a} \phi\left(1.645 - x\right) dx} \tag{3}$$

$$= \frac{p_{true} * \phi(1.645)}{p_{true} * \phi(1.645) + \frac{(1-p_{true})}{2a} * \left(\Phi\left(1.645 - c^* + a\right) - \Phi\left(1.645 - c^* - a\right)\right)}. \tag{4}$$

I show that if the power lies between roughly .16 and .84, the latter always has a higher false discovery rate. Note that this condition is sufficient, not necessary. I note that Ioannidis, Stanley, and Doucouliagos (2017) estimate that the median power of economics studies is .18.

2.4 What if we think the point alternative is a lower bound?

You may object that we perform power calculations at parameter values we believe to be conservative. Therefore, your argument continues, if we calculated the false discovery rate using a parameter at or near our best guess of the effect size, the false discovery rate would be lower. However, this objection runs smack into Lindley's paradox. When the true parameter is above this lower bound, our test will be more powerful. Assuming that we began with a reasonable degree of power at our single point, the increased power will generally make a narrow rejection less likely, and, therefore, should increase the probability that a narrow rejection is a false discovery. Of course, the greater power will make narrow rejection less likely if the alternative is true. In that sense, power is always good.

2.5 But we don't believe the null is exactly true

The nihilist answered my question in the introduction by responding that all null hypotheses are false. Given the so-called butterfly effect, any x we can find probably has a non-zero relation with our dependent variable. We simply lack sufficient data to unearth effects near zero. Regardless of whether the butterfly effect is real, it is hard to believe that most education policies, for example, have no effect on student outcomes. The effect may be small and positive or negative, but exactly no effect is unlikely.

Unfortunately, the nihilist argument only hurts us. The problem is that the density is convex to the origin in the range in which we narrowly reject. Therefore, if we allow the true parameter to have a symmetric distribution centered on 0 and with bounds within $\pm 1.96\sigma$, the false discovery rate is greater than with a null fixed at 0.

2.6 Setting the False Discovery Rate at .05 Requires a Much Higher Critical Value

Let's return to the example in which the researcher claims power .8 to reject 0 and find the $t - statistic$ required for a false discovery rate of 5%. The answer clearly depends on the probability that the null hypothesis is true. If this probability is 0, there is no risk of falsely rejecting the null regardless of the t. If it is 1, all rejections must be false.

Table 1 shows the $t-statistics$ at which the false discovery probability is .05. Recall that the table is for the case where claimed power to reject 0 is .8 against a point alternative. For the reasons discussed above, the critical $t - statistics$ may be higher if the alternative or null is more diffuse.

In a sense the table supports the practice of demanding stronger evidence when a result

Table 1: Critical Values Required for False Discovery Rate < .05

Prob. true	Critical t
.1	1.54
.125	1.645
.2	1.87
.3	2.09
.4	2.26
.5	2.43
.6	2.59
.7	2.77
.8	2.98
.9	3.31

Notes: Each row shows the prior probability that the null is true and the critical t at which the probability of a false discovery is no more than .05.

contradicts our expectations. A researcher who accepts a 5% chance of falsely rejecting the null may be quite sanguine about rejecting based on a $p - value$ derived from the normal distribution that is higher than .05 if they are already fairly confident that the null is false. One who views the probability of the null being correct as very high may require p to be below .001. Researchers and others generally decry confirmatory bias in the review process (Mahoney 1977). While confirmatory bias is beyond the scope of this article, spiritual Bayesians, to use John Rust's (1988) terminology, should indulge their bias to some degree. I do not claim that they do so optimally.

2.7 Lessons for Reading the Literature

While confirmation bias suggests that it should be easier to publish a paper rejecting a null that most economists believe to be false, editors and reviewers are unlikely to find much value in rejections of obviously false nulls. In a parallel to the current paper, DeLong and Lang (1992) investigate published unrejected nulls in cases where the failure to reject the null is the article's main point. Our point estimate in that paper is that none of the unrejected nulls is true, and the confidence interval excludes more than one-third being true. Our interpretation of the results is that economists find the failure to reject a null interesting only when they expected the null to be rejected because it was false. In most cases, their expectation was correct.

In parallel fashion, rejections of null hypotheses with low values of p_{true} should be rare among published articles; they simply aren't interesting. On the other, only an intrepid researcher would hope to reject a null hypothesis that is very likely to be true. The paper

will be of little interest in the highly likely case that they fail. If the researcher does reject the null, they should anticipate being subjected to considerable scrutiny due to confirmation bias. Colquhoun argues that $p_{true} < .5$ is implausible. I accept his argument for "exciting discoveries," but much normal science should examine hypotheses with moderate probabilities of being correct. On the other hand, Abadie (2020) shows that statistical insignificance is more informative than significance when the probability of rejection exceeds .5. This suggests that journals should favor publishing rejections when p_{true} is relatively low. I make no claim that this argument explains publication bias, but there may well be an imperfect understanding of Abadie's point that helps determine which rejected nulls are published.

Outside economics, 35% of phase 2 clinical trials lead to drug approval (Wong, Siah, and Lo 2019). While there are undoubtedly type I and type II errors, this suggests that even when the cost of hypothesis testing is high, the probability that the null is (approximately) true is high.

3 A Model of Published Significant t-statistics

In this section, I ignore unrejected nulls. Either these nulls are simply not published, or the motivation for publishing them is different from, although possibly analogous to, that for publishing rejections. Brad DeLong and I have argued that journals publish unrejected nulls only when failure to reject is surprising. I will reach a similar, although somewhat weaker, conclusion about rejected nulls. Most narrowly rejected nulls are probably true.

3.1 The Base Model

Denote the fraction of true null hypotheses, for which the expected t is 0, by $(1-q)$. Given typical sample sizes, the $t-distribution$ for these hypotheses is well approximated by the standard normal density, ϕ, and distribution, Φ, functions.[2] For each false hypothesis, let t^* be the expected $t-statistic$ given the power of the test and the true coefficient. For large samples, the distribution of each t with expected value t^* is approximately a noncentral t with location parameter t^*.[3] Assume that t^* is drawn from some distribution with density $f(t^*)$.

[2]One notable caveat is that when using the Liang-Zeger cluster correction, the $t-statistic$ has degrees of freedom equal to the number of clusters minus 1. Since many papers use 50 state clusters, the $t-dsitribution$ will often be somewhat more distant from the normal than implied by the large sample size.

[3]For a fixed sample size, as t^* gets large, this approximation worsens. Fortunately, the density of t when t^* is very large will be low in the range used in the empirical estimation.

The probability (density) that we observe $|t|$ is

$$L(|t|) = (1-q)\,\phi(t) + q\int \phi(t-t^*)\,f(t^*)\,dt^* + (1-q)\,\phi(-t) + q\int \phi(-t-t^*)\,f(t^*)\,dt^* \quad (5)$$

where $f(t^*)$ is the density of the distribution of t^* in the population of studies of false null hypotheses.

The $f(t^*)$ I use is symmetric about the y-axis, and the density of t is symmetric with respect to t^*. I focus on $|t|$ for the analysis. To simplify notation, from now on, I refer to this as t rather than $|t|$ whenever doing so is unlikely to be confusing.

Therefore, we have

$$L(t) = 2 * [(1-q)\,\phi(t) + q\int_{-\infty}^{\infty} \phi(t-t^*)\,f(t^*)\,dt^*]. \quad (6)$$

This density has to be adjusted for publication bias. I ignore observed $t-statistics$ for which $t < 1.96$, including those that may not be published. I also drop the observations for which $t > 10$**.** As discussed in the introduction to section 2, many, although admittedly not all, cases where articles report very high values of t are cases where the researchers put little weight on the possibility that the true coefficient was 0. Given these restrictions, the likelihood of an observation is

$$L(t) = \frac{(1-q)\,\phi(t) + q\int_{-\infty}^{\infty} \phi(t-t^*)\,f(t^*)\,dt^*}{(1-q)\,[\Phi(10)-\Phi(1.96)] + q\int_{-\infty}^{\infty}[\Phi(10-t^*)-\Phi(1.96-t^*)]f(t^*)\,dt^*} \quad (7)$$

for $10 > t > 1.96$. For a given choice of $f(t^*)$, (7) is estimable by maximum likelihood.

Assume that t^* has a Laplace distribution with mean zero, which means t^* is exponentially distributed and t^* is symmetric about the y-axis. The density is thus $.5\lambda \exp(-\lambda\,|t^*|)$. This assumes that alternatives to tested nulls are as likely to be positive as negative. In principle, we could allow for asymmetry, which seems to be present in the data, by dropping the assumption that the mean is 0, but I have not done this.

Given a set of estimates, the probability a rejected null hypothesis is true given t is

$$P(true|t) = \frac{(1-q)\phi(t)}{(1-q)\phi(t) + q\int_{-\infty}^{\infty} \phi(t-t^*)\,f(t^*)\,dt^*}. \quad (8)$$

Note that (8) treats any departure from an expected t of 0 as a false null. Thus, it implicitly treats all tests as against a two-sided alternative. If the alternative is one-sided, cases where t^* has the wrong sign support the null. However, since it is implausible that the alternative is symmetric around 0, when the alternative is one-sided, and because it is often

difficult to determine whether the researcher viewed the test as one-sided, I retain $|t| \geq 1.96$ as the cutoff for including the test in the analysis and have not attempted to adjust the analysis to allow for one-sided tests.

4 Identification

4.1 Identifying q

For any finite value of λ, the density of the $t - statistics$ of the false nulls declines less sharply above 1.96 than it does when the null is true. Therefore, for a fixed value of λ, sharper declines in the density near 1.96 indicate a higher proportion of true nulls that have been rejected. Figure 1 shows the density for $\lambda = .5$ when q is 1/3 (one-third of null hypotheses are false, the solid line) and q is 2/3 (the dashed line). As expected, as t gets large, the proportion of false null hypotheses has little effect on the shape of the density. However, at values near 1.96, higher values of q (false nulls) are associated with the density declining less steeply.

4.2 Identifying λ

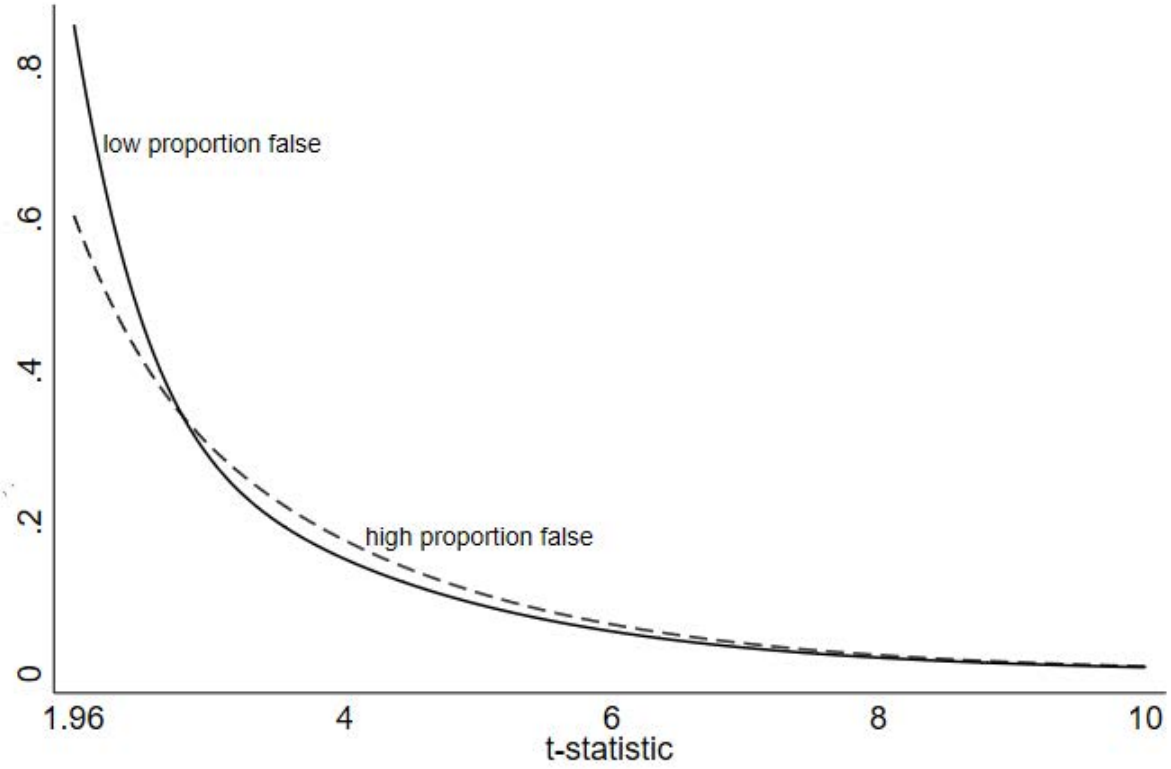
When the null is false, the density of the $t - statistic$ near 1.96 declines more sharply at higher values of λ. Intuitively, as $\lambda \to \infty$, the distribution of expected $t - statistics$ of false nulls degenerates to 0, and the distributions of the $t - statistics$ associated with false and true null hypotheses are indistinguishable. As $\lambda \to 0$, the density of the expected t when the null is false becomes flat.

Therefore, for a fixed q, the $t - distribution$ is flatter when λ is lower. Figure 2 shows the distribution of $t - statistics$ for false null hypotheses for $\lambda = .25$ (dotted line), .5 (dashed line), and 1 (solid line). Higher values of λ are associated with a much sharper decline in the density of the $t - statistic$ even between t equal to 4 and 5, values that are highly unusual if the null hypothesis is true. Thus, λ is most clearly identified by the part of the distribution in which false rejections are relatively unlikely. Of course, λ will not be well identified if false nulls are extremely rare.

5 Data

From a separate project, I have data from all the empirical papers published in the top five journals (*American Economic Review, Econometrica, Journal of Political Economy,*

Figure 1: Density of the t-statistic is more convex when the proportion of false hypotheses is low



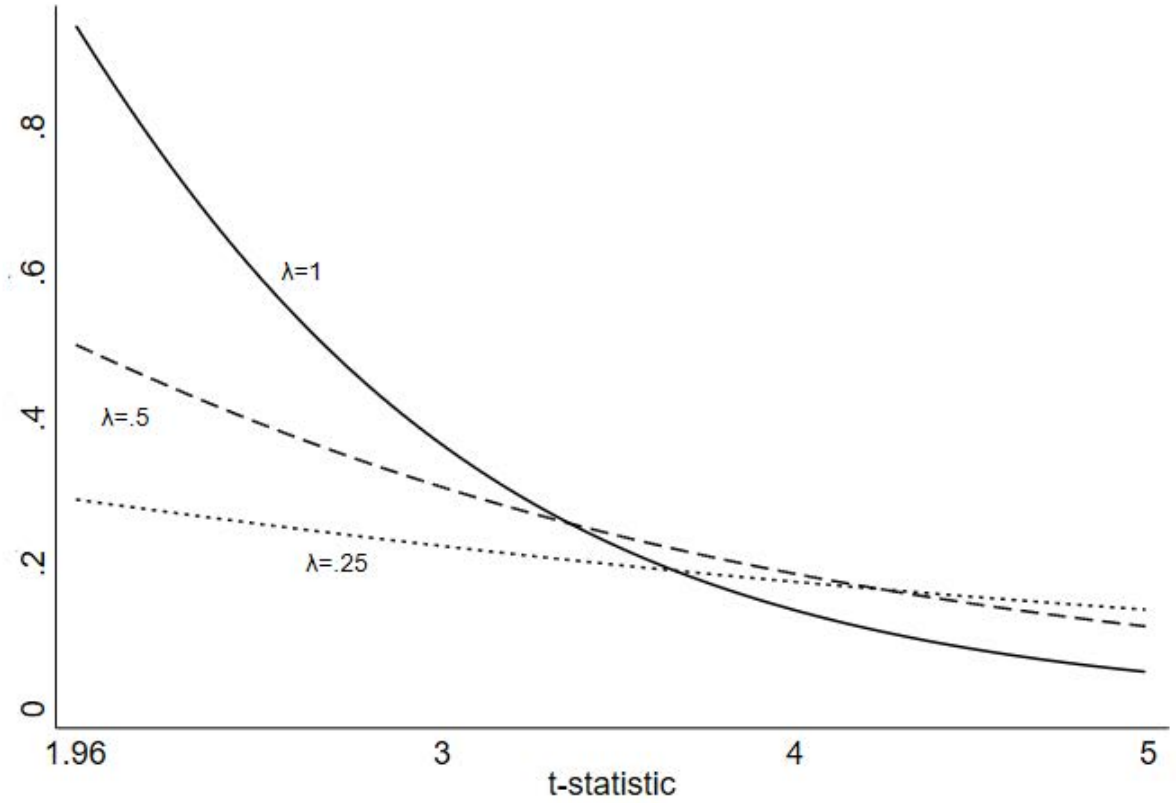
Notes: This figure shows the density of the $t-statistic$ for q, the fraction of false nulls, when q is 1/3 (one-third of null hypotheses are false, the solid line) and q is 2/3 (the dashed line). The spread parameter of the Laplace distribution, $\lambda = .5$.

Quarterly Journal of Economics, Review of Economic Studies) in 2021. I add the sample from Brodeur et al. (2020), which consists of articles in 25 top journals published in 2015 and 2018.[4]

The sample consists of studies exploiting natural, laboratory, or field experiments. Articles that focus on theory, structural estimation, or econometric approach are excluded. The included articles primarily use 'credibility revolution' or 'natural' or actual experiment techniques (difference-in-differences, instrumental variables, regression discontinuity, randomized controlled trials), although the team judged some (about 7% of the sample) ordinary least

[4] *AEJ: Applied Economics, AEJ: Economic Policy, AEJ: Macroeconomics, American Economic Review, Econometrica, Economic Journal, Economic Policy, Experimental Economics, Journal of Applied Econometrics, Journal of Development Economics, Journal of Economic Growth, Journal of Finance, Journal of Financial Economics, Journal of Financial Intermediation, Journal of Human Resources, Journal of International Economics, Journal of Labor Economics, Journal of Political Economy, Journal of Public Economics, Journal of Urban Economics, Journal of the European Economic Association, Quarterly Journal of Economics, Review of Economic Studies, Review of Economics and Statistics, Review of Financial Studies.*

Figure 2: Density of t-statistics for false nulls falls faster at higher values of λ



Notes: This figure shows the density of $t-statistics$ for false null hypotheses for different values of λ.

squares estimates to fit this definition. Roughly one-quarter of the hypothesis tests used RCTs; a similar proportion involved instrumental variables.[5] I further restrict the sample to rejected nulls.

Although I drew on the Brodeur et al. sample, I did not rely on their data. The research assistants checked abstracts and read the text describing the results. We assessed which estimates the author(s) emphasized by including them in the abstract or claiming in the text that these were the principal results. If multiple specifications tested the same hypothesis, we relied on the one the authors 'preferred' or on whose magnitude they relied in subsequent analysis. We excluded robustness checks or heterogeneity analysis unless heterogeneity was a focus of the article. Thus, if the main specification was difference-in-differences, but the authors included triple differences as a robustness check, we retained only the former. Similarly, if the authors provided separate estimates for men and women, we kept these estimates only if the separate analyses were core elements of the article. We

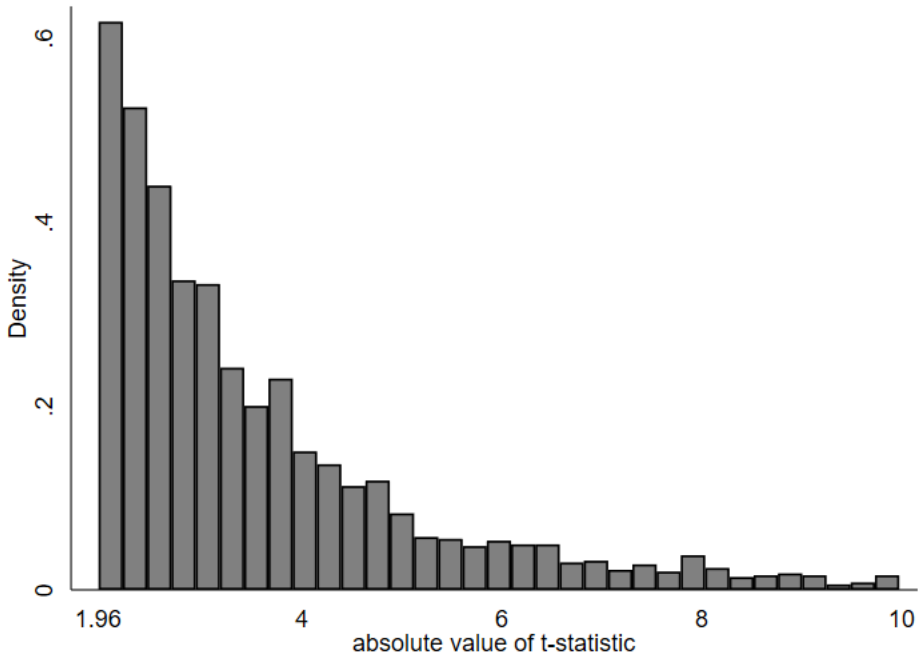
[5]There is some double counting since an RCT may use IV in the event of noncompliance or OLS to increase the efficiency of the estimates.

have 2,082 hypothesis tests from 663 articles.[6]

Most, but not all, articles report $t-statistics$ or standard errors. If an article reports p-values, we draw the p from the appropriate rounding range (e.g. if the reported value is .03 , it is drawn from $U(.025, .035)$) and then assign the $t-statistic$ associated with the p-value we draw. However, fully 97% of the test statistics are either reported as t or report the standard error, allowing us to calculate the t. We perform a similar derounding exercise on the calculated or reported $t-statistic$. One article reports a chi-squared test of multiple coefficients. I drop this test.

Finally, I remind the reader that this paper focuses on settings in which the researcher viewed the null hypothesis of 0 as plausible. As the $t-statistic$ increases, a larger portion of the sample will consist of settings in which the researcher was primarily interested in the coefficient's magnitude, not its statistical significance. Therefore, somewhat arbitrarily, I limit the sample to $t-statistics$ less than 10 in absolute value. Moreover, since the paper's topic is rejected null hypotheses, I further limit the sample to $t-statistics$ greater than or equal to 1.96 (in absolute value).

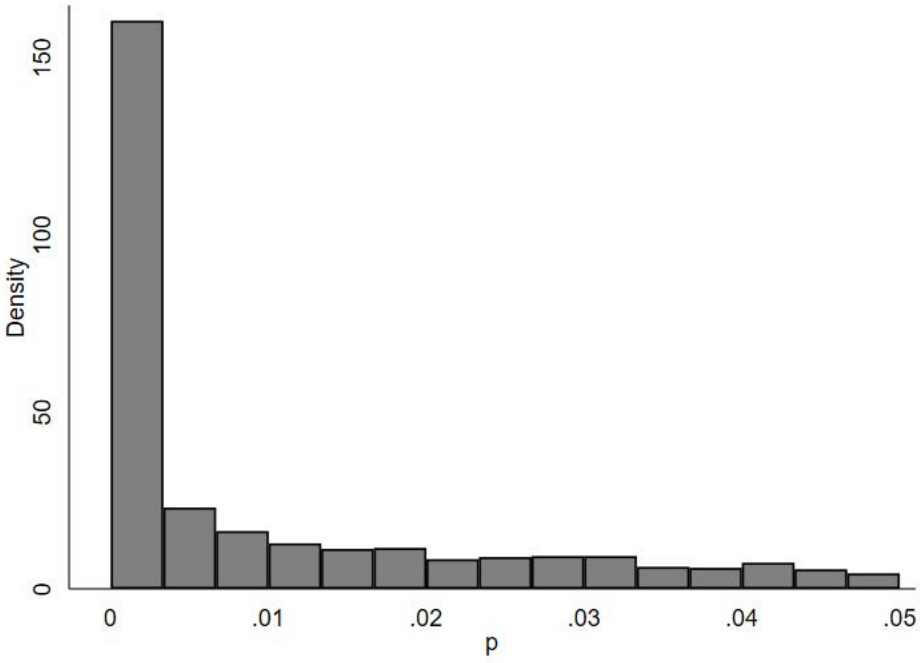
Figure 3: Distribution of absolute values of t-statistics: 1.96-10.0



Notes: This figure shows the distribution of the absolute values of t-statistics. The sample includes all rejected null hypotheses in the range where $|t| > 1.96$ and $|t| < 10$.

[6]In contrast, Brodeur et al. (2020) includes 21,740 tests from 684 articles.

Figure 4: Distribution of p-values below .05



Notes: This figure shows the distribution of the p-values in the sample with all rejected null hypotheses for which $p < .05$.

Figure 3 shows the distribution of the absolute value of t-statistics in the range where $p < .05$. Most $t - statistics$ are close to 1.96, but there is a long tail of values substantially farther from 0. This is consistent with at least some rejected null hypotheses being false. By presenting the same data in terms of p-values (see Figure 4), we get a first look at the importance of p-hacking. The distribution of p is generally monotonically decreasing as we move from left to right. Violations of monotonicity in expected locations, such as from close to .05 to somewhat lower, are strong evidence of p-hacking. However, the figure provides no such evidence. The biggest failure of monotonicity comes just above and just below .04, which is not a location where we expect to find p-hacking. Moreover, the density just above .04 is not statistically significantly greater than the density just below, even without correcting for multiple hypothesis testing.

Interestingly, the data pass all the tests for p-hacking/publication bias in Elliott, Kudrin, and Wüthrich (2022) except for their CS2B test. This will turn out to be consistent with some continued publication bias at values of p below .05 for which I provide evidence later.

6 Results

6.1 Base Model

Table 2 shows the results of estimating q, the proportion of false null hypotheses, and λ, the parameter of the distribution of the expected $t-statistics$ among the false nulls, assuming that these expected $t-statistics$ have a Laplace distribution with mean zero. The left panel includes all rejected null hypotheses, while the right panel includes one randomly selected rejected null per article.

Table 2: Model parameters

	All			One per article		
$p<$	q	λ	N	q	λ	N
.05	0.55	0.54	2082	0.39	0.47	501
	(0.07)	(0.02)		(0.09)	(0.04)	
.04	0.48	0.54	1963	0.28	0.45	481
	(0.07)	(0.02)		(0.06)	(0.04)	
.03	0.44	0.53	1815	0.27	0.44	438
	(0.07)	(0.02)		(0.07)	(0.04)	
.02	0.41	0.53	1631	0.23	0.44	394
	(0.08)	(0.02)		(0.07)	(0.04)	
.01	0.33	0.52	1384	0.18	0.43	331
	(0.10)	(0.02)		(0.07)	(0.04)	

Notes: Standard errors in parentheses. They have not been corrected for any within-article correlation. $p<$ identifies the maximum p-value of the hypothesis tests included in the sample. q is the proportion of false nulls. I assume $t^* \sim \text{Exponential}(\lambda)$ where $f(t^*) = .5\lambda \exp(-\lambda|t^*|)$. $\lambda > 0$ is the spread parameter of the exponential distribution.

Except when the sample contains all hypothesis tests and rejections at the .05 level, the estimates imply that the majority of null hypotheses are true. This estimate is noticeably higher when I restrict the sample to one hypothesis per article. Recall that we estimate that q is lower when the density drops more rapidly above 1.96. Thus the lower estimated q when I use only one hypothesis per article suggests that when researchers present additional results as key findings, they are likely to have more results with relatively high p-values. This seems especially likely if the "key results" are based on subsamples. Note also that the standard deviation of the distribution of the expected $t-statistics$ for the false nulls is somewhat larger (λ is somewhat smaller) when I restrict the sample to a single test per article, but the difference is not large.

Recall that q is *not* the fraction of rejected nulls that are false or even the fraction of published nulls that are false. Instead, it is the proportion of false null hypotheses among

hypotheses that economists might have published in leading journals had they found significant results. Some of these were published. Some were rejected because reviewers and editors prefer significant results. Others were never submitted because the authors did not find the failure to reject worth pursuing.

The critical question is not whether economists were sufficiently clever to investigate a high proportion of false null hypotheses but the proportion of published rejections that are falsely rejected. Applying (8) to the results with the sample with all $t-statistics > 1.96$ reveals that 9.56% (standard error= 2.31) of published rejected nulls are false rejections. This rises to 15.5% ($s.e. = 4.4$) when limiting the sample to one null per article. I leave it to the reader to decide whether these figures are disturbing or reassuring.

As we restrict the sample to increasingly strong rejections, the proportion of false nulls declines. The standard deviation of the $t-distribution$ widens slightly**.** In other words, in order to fit the data when using observations near $p = .05$, the model has to place less weight on the standard normal distribution. If anything, relative to the implication of widespread p-hacking, the drop in the density resembles the normal distribution less closely very near to .05 than when it is somewhat farther away. This pattern holds even when we impose restrictions far from where we expect p-hacking to be prevalent.

6.1.1 The model fits reasonably well

The model's credibility partially depends on how well it can reproduce the actual distribution of $t-statistics$. In practice, economists show a slight preference for alternative hypotheses involving positive coefficients. The reported $t-statistics$ are, therefore, somewhat asymmetric. The model does not try to reproduce this. Therefore, Figure 5 plots the predicted and actual distributions of $|t|$ when the sample includes all $t-statistics$ between 1.96 and 10.0 in absolute value. The density is adjusted to account for the number of rejected nulls in the data in this range and for the use of 20 bins.

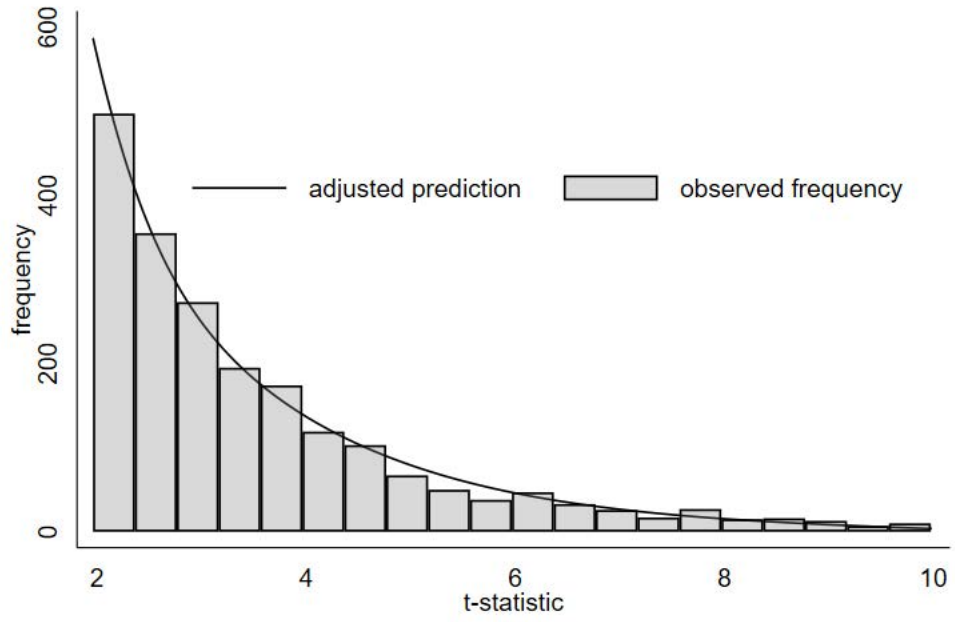
The model fits the data fairly closely. It somewhat over-predicts $t-statistics$ very close to 1.96 ($1.96 \leq t < 2.1109$) and somewhat underpredicts slightly higher values ($2.1109 \leq t < 2.3858$). If p-hacking were important, we would expect the model to under-predict the density just below .05 and over-predict it at slightly lower levels. There is little reason to expect p-hacking to produce an unusually large proportion of results significant at between the .017 and .035 levels. Therefore, I do not adjust the model to allow for additional density just below $p = .05$.

I use the formula in Moore (1977), which adjusts for the fact that the estimated parameters maximize the model's fit, to perform a χ^2 goodness of fit test. The bin widths are designed to equalize the predicted number of observations across bins. As shown in Table 3,

the test statistic with twenty bins is 26.9, which falls just short of significance at the .1 level. With forty bins, the test statistic is 43.6, which is insignificant at any conventional level.

The failure to reject the model is not entirely due to lack of power. Table 3 also shows the goodness-of-fit test statistics assuming that the distribution of the expected $t-statistics$ for false hypotheses is normally distributed, my initial assumption. We can easily reject that the model fits well. Note that my failure to pursue the normal distribution version of the model is a form of publication bias. Tables A.1 and A.2 in the appendix provide more detail about the goodness of fit calculations.

Figure 5: Model predictions v. actual frequencies



Notes: This figure shows the predicted and actual distributions of the absolute values for $t-statistics$ between 1.96 and 10. The predicted distribution is generated using the estimates in the left panel of Table 2 for the sample with all rejected null hypotheses. The prediction is adjusted to account for the number of rejected nulls in the data in this range and for the use of 20 bins

Table 3: χ^2 Goodness of Fit Test Results

	20 bins	40 bins
Exponential	26.88	43.63
Normal	57.74	98.22

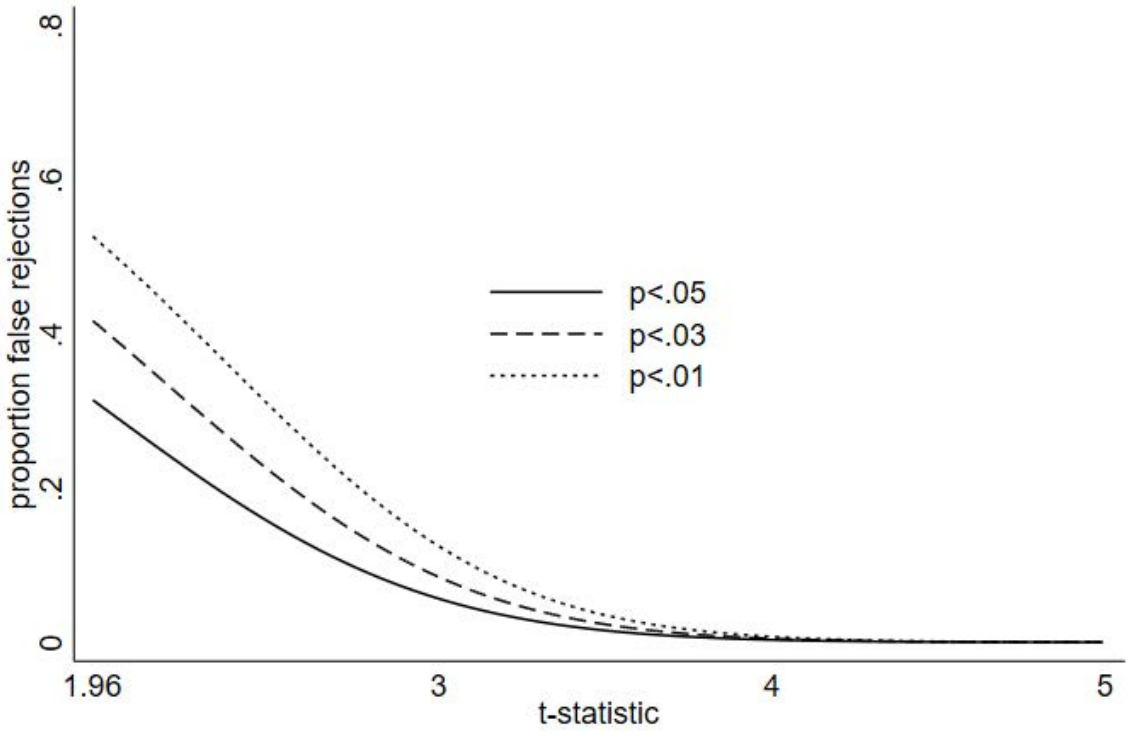
Notes: This table presents the Moore (1977) goodness of fit test statistics for models estimated under the assumption of an Exponential or Normal distribution of t^*.

6.1.2 Close rejections are frequently false discoveries

Using the full sample of $t-statistics$ between 1.96 and 10.0, the model implies that when t equals 1.96, 31.4% of the 'rejected' nulls are true (standard error= 6.4). Using only a single null hypothesis per article, this proportion rises to 47.7% (standard error= 9.9). As we move down the left column of Table 2, the estimated proportion of null hypotheses that are false declines. Not surprisingly, this implies that the proportion of narrowly rejected hypotheses that are falsely rejected rises to 52.6% (standard error= 11.0). Similarly, as we go down the right column, which restricts the sample to a single null hypothesis per article, the proportion of false rejections among narrow rejections rises from to 72.6% (standard error= 10.3).

Of course, as the $t-statistic$ rises, the proportion of falsely rejected nulls declines. This is captured in Figure 6, which shows the proportion of falsely rejected nulls at each value of t for the sample with all rejected nulls at significance levels .05, .03, and .01. Still, to reach the standard 5% probability of a false rejection requires a t of 3.05 if we use the sample with rejections at the .05 level or better and 3.37 with the sample of rejections at the .01 level. For the sample with only one null per article, the corresponding values are 3.28 and 3.61.

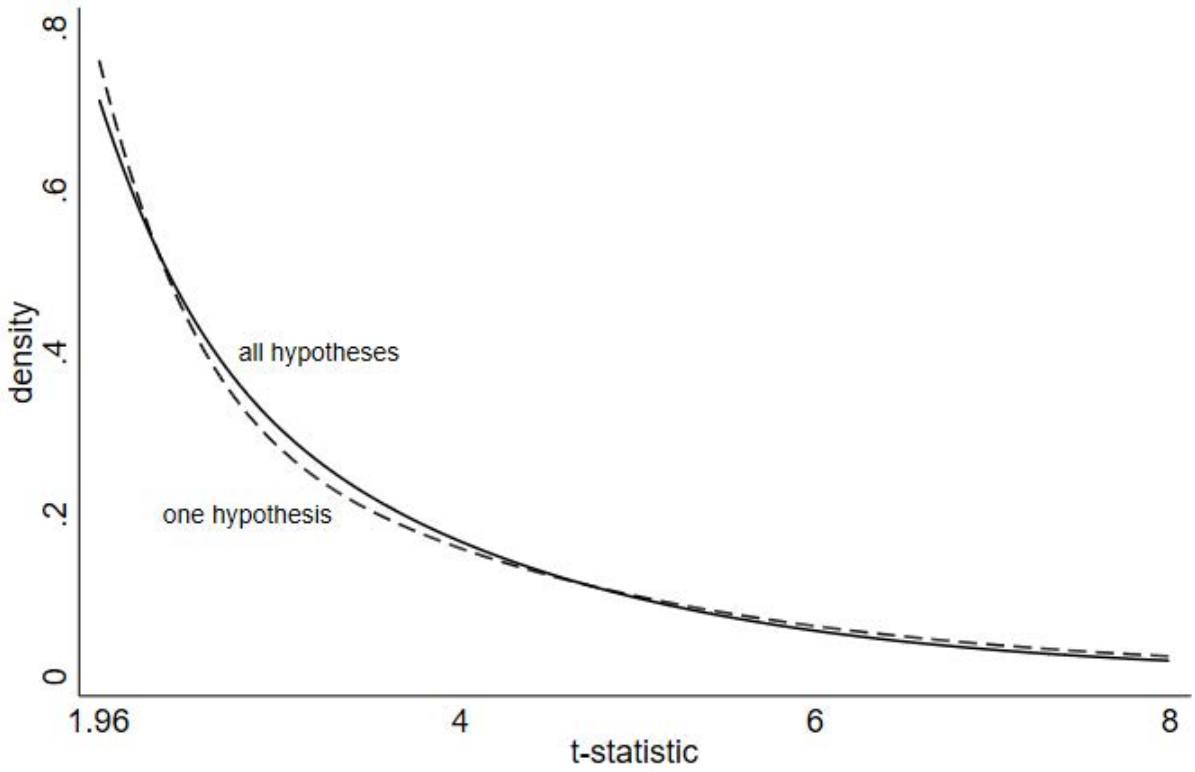
Figure 6: Proportion of falsely rejected nulls



Notes: This figure shows the posterior probability that the null hypothesis was falsely rejected at each value of t. The probabilities are calculated based on the estimates in Table 2 for the sample with all rejected null hypotheses at significance levels .05, .03 and .01.

In the relevant range, the predicted densities from using all hypotheses or one per article are similar, as shown in Figure 7 for all $t-statistics$ above 1.96. This is consistent with the modest differences in the proportion of narrowly rejected nulls that are false and the values of t at which only 5% of nulls are false rejected. Instead, the estimates from the two samples differ primarily in terms of their projected densities of the $t-statistics$ for unrejected nulls had they all been published.

Figure 7: The predicted densities of t-statistics are similar using all or one hypothesis per article



Notes: This figure shows the predicted densities generated using the estimates in Table 2 for all hypotheses and one per article for $|t| > 1.96$.

6.2 Extensions

Although the model fits well in sample, the differences across estimates with increasingly restricted samples should concern us. If the model is correctly specified, the estimates using the full sample of $t-statistics$ above 1.96 are efficient. The estimates using only $t-statistics$ with lower p-values are inefficient but remain consistent. Yet, we can tell by inspection, and a

formal Hausman test confirms, that the estimates using the full and most restrictive samples are significantly different.

Therefore, I consider three extensions of the model.

6.2.1 Almost true null hypotheses make little difference

Equation (8) assumes that any departure from 0 should be treated as a setting in which the null hypothesis is false. Sometimes, a very small effect justifies rejecting the null. However, we may often view a small effect as functionally equivalent to 0. Addressing this concern is somewhat challenging in my framework. We want to treat true coefficients near 0 as confirming the null hypothesis. In contrast, my approach focuses on the expected $t-statistic$, which can be close to 0 either because the coefficient is approximately 0 or because the expected standard error is large. In practice, I expect that researchers will not undertake projects where they anticipate low $t's$ due to highly imprecise estimates. Therefore, I approximate small coefficients by low expected $t-statistics$.

To address these "almost true null hypotheses," rewrite the distribution of false nulls as empty between $-\eta$ and $+\eta$ and allow the true null to have a triangular distribution centered on 0 and covering the same range. Thus,

$$L(t) = \frac{(1-q)\int_{-\eta}^{\eta}\phi(t-t')\frac{\eta-|t'|}{\eta^2}dt' + q\int_{|t^*|>\eta}\phi(t-t^*)f(t^*)dt^*}{(1-q)\int_{-\eta}^{\eta}[\Phi(10-t')-\Phi(1.96-t')]\frac{\eta-|t'|}{\eta^2}dt' + q\int_{|t^*|>\eta}[\Phi(10-t^*)-\Phi(1.96-t^*)]f(t^*)dt^*} \tag{9}$$

and

$$P(true|t) = \frac{(1-q)\int_{-\eta}^{\eta}\phi(t-t')\frac{\eta-|t'|}{\eta^2}dt'}{(1-q)\int_{-\eta}^{\eta}\phi(t-t')\frac{\eta-|t'|}{\eta^2}dt' + q\int_{|t^*|>\eta}\phi(t-t^*)f(t^*)dt^*}. \tag{10}$$

Setting η to .1 or .2 has little effect on the estimates. Using the full sample, the estimated proportion of false nulls declines from .55 to .53 or .52, as we would expect since some false nulls are redefined as true. λ declines trivially from .547 to .544 (table not shown). Similarly, with only one null per article, q declines from .39 to .37 and .38 while λ is unchanged to two decimal places. Most significantly, this extension fails to make the estimates independent of the p-value cutoff for including $t-statistics$ in the sample.

6.2.2 Allowing for publication bias at $t > 1.96$ helps

Although the base model fits well in sample, it somewhat overestimates the density of narrow rejections. This suggests that authors may be reluctant to publish or editors and reviewers reluctant to accept narrow rejections. Economists are increasingly sensitive to issues of

multiple hypothesis testing. Some use formal adjustments for multiple hypothesis testing (Romano and Wolf 2005, Benjamini and Hochberg 1995), which means that $t-statistics$ just above 1.96 may not be treated as rejections and may, therefore, be less likely to be published. Others recognize informally that researchers have considerable freedom to choose specifications that favor rejection even without consciously engaging in p-hacking, making them skeptical of narrow rejections.

To address this possibility, I model the probability that a 'rejected' null is published as[7]

$$P(t) = c_0 + c_1(|t| - 1.96) \; if \; |t| < \frac{1-c_0}{c_1} + 1.96 := M \tag{11}$$
$$P\,(t) = 1 \qquad\qquad otherwise.$$

The resulting likelihood is

$$L = \frac{[(1-q)\phi(t) + q\int_{-\infty}^{\infty}\phi(t-t^*)f(t^*)dt^*] * P(t)}{\int_{t(p)}^{10}[(1-q)\phi(t) + q\int_{-\infty}^{\infty}\phi(t-t^*)f(t^*)dt^*] * P(t)dt}.$$

In a sense, this approach is successful (see Table 4, top panel). The estimated proportion of false nulls is roughly constant across cutoffs for p. It ranges from .14 when I include all $t-statistics$ above 1.96 to .18 when I use a cutoff of $p < .03$. Still, I cannot reject that any pair of estimates of q are equal. Estimates of λ range from .502 to .509.

Furthermore, the estimates suggest that a very high proportion of narrow rejections are, in fact, false rejections. Using the full sample, 77% of rejections at $t = 1.96$ are false. The probability of a false rejection falls to 5% only when t reaches 3.73.

However, the range over which publication bias is active is improbably large. Depending on the sample, the model implies that publication probability reaches 1 only when the $t-statistic$ exceeds $3.09 - 3.17$, depending on the sample. Notably, when I restrict the sample to $t-statistics$ above 3.17, I no longer find evidence of publication bias using the Elliott, Kudrin, and Wüthrich (2022) CS2B test.[8]

[7]Note that this approach assumes that the probability of publication depends only on the realized t and not whether the hypothesis is false. I tried allowing the probability to depend on t^* and t (setting $t^* = 0$ for true nulls), but the estimates were too imprecise to be useful.

[8]This conclusion is somewhat sensitive to the choice of the number of bins. I follow the authors in using 15 bins as they do for their smaller samples. Given the much smaller sample I use for this test, I could make a case for using 10 or fewer bins. The results are consistent for 5-15 bins except that I reject at the .1 level when using 13 bins. Given the multiple tests involved, I feel confident in claiming that the CS2B test fails to reject.

6.2.3 Is the $t-statistic$ distributed as t?

There are multiple reasons for suspecting that researchers frequently overstate the precision of their estimates. The heteroskedasticity-robust variance estimator and the Liang-Zeger cluster correction can be severely downward biased in small samples (Chesher and Jewitt 1987). My personal experience suggests that 'small' can be pretty large. In addition, the Liang-Zeger $t-statistic$ has degrees of freedom equal to the number of groups minus 1, making the distribution somewhat more diffuse than the normal. Moreover, traditional (including heteroskedasticity-robust) standard error estimates for instrumental variables estimators can also be severely downward biased (Young 2022, Jiang 2017, Keane and Neal 2023). The standard errors reported in many RCTs appear to underestimate the true degree of uncertainty (Young 2019). The common practice of testing for pre-trends when using difference-in-differences invalidates the standard errors and makes finding a significant effect more likely (Roth 2022).

Therefore, I extend the base model by dropping the assumption that the $t-distribution$ has a standard deviation of 1. I write the likelihood function as

$$L = \frac{(1-q)\phi(\frac{t}{\delta}) + q\int_{-\infty}^{\infty}\phi(\frac{t-t^*}{\delta})f(t^*)dt^*}{\int_{t(p)}^{10}[(1-q)\phi(\frac{t}{\delta}) + q\int_{-\infty}^{\infty}\phi(\frac{t-t^*}{\delta})f(t^*)dt^*]dt}$$

where δ is the standard deviation of the central or noncentral $t-statistic$. I assume δ is constant regardless of the expected value of t. Self-evidently, this specification fails to account for a rich set of factors that may lead the $t-distribution$ to be misspecified. It is beyond this paper's scope to model such departures fully. Note that the density is $\delta^{-1}\phi(\frac{t}{\delta})$ or the equivalent for the noncentral t, but the δ^{-1} terms cancel.

The middle panel of Table 4 shows that we reject the null hypothesis that the standard deviation equals 1. The point estimate with the full sample implies that $t-statistics$ should be divided by roughly 1.7 (roughly 1.5 using only one null per article, see appendix table A.3) although the estimate is quite imprecise. Estimates are broadly similar across all specifications. With the full sample, the model implies that 68% of nulls are true when $t = 1.96$, or 72% when the sample is restricted to nulls rejected at the .01 level. Fully, 44.4% (standard error= 8.5), or 39.7% (standard error= 10.4) using only one null per article, of published rejected nulls are falsely rejected. It is reassuring that the results, although less precise, are similar when the sample is restricted to one hypothesis per article.

6.3 The Full Model

The results from each of the last two extensions are promising. Therefore, I simultaneously allow for some publication bias above $t = 1.96$ and relax the assumption that the $t-$ *distribution* has the anticipated variance. This gives the following likelihood function

$$L = \frac{[(1-q)\phi(\frac{t}{\delta}) + q\int_{-\infty}^{\infty}\phi(\frac{t-t^*}{\delta})f(t^*)dt^*] * P(t)}{\int_{t(p)}^{10}[(1-q)\phi(\frac{t}{\delta}) + q\int_{-\infty}^{\infty}\phi(\frac{t-t^*}{\delta})f(t^*)dt^*] * P(t)dt}. \tag{12}$$

The bottom panel of Table 4 gives the estimates of the combined model for the samples with $t > 1.96$ and $t > 2.05$; restricting the model further makes the results too imprecise. Interestingly, the parameters imply that only $t-$ *statistics* slightly above 1.96 are subject to publication bias. Results for which the value t is at least 2.02 are not denied publication simply because the t is too small. This is a plausible estimate. It rises to 2.09 when I restrict the sample to rejections at the .04 level, but the difference between the two estimates is insignificant at conventional levels. Note, that the $t-$*statistic* for a test of whether $M = 1.96$ is only 1.54. Since this test is on the boundary, it does not have a conventional $t-$*distribution*. However, it raises the question of whether there is any publication bias when $t > 1.96$ and whether we should prefer the full model or the model without additional publication bias.

Fortunately, consistent with the (at most) small rate of additional publication bias, the remaining estimates are similar to those obtained when ignoring publication bias for $t > 1.96$. The true standard deviation of the $t-$*statistic* is about 1.6. About 39% of all null hypotheses are false. Consequently, 65% (standard error= 8.82) of narrowly rejected null hypotheses and 41.1% (standard error= 9.4) of all published rejected nulls are falsely rejected. To ensure that the null you think you are rejecting has no more than a 5% probability of being true requires $t > 5.48$. Even the more classical norm of having only a 5% chance of falsely rejecting the null if it is true requires $t > 3.21$ $(1.96 * 1.64)$.

If I restrict the sample to one hypothesis per article, the model converges only with the largest sample. Still, the results are similar to those in the top line of the bottom panel of Table 4. The only substantive difference is that the estimated standard deviation of the $t-$*distribution* is 1.35 instead of 1.64 and not significantly different from 1. However, 67.5% (standard error= 9.5) of hypotheses are falsely rejected when $t = 1.96$ and 34.8% (standard error= 10.9) of all published rejected nulls are falsely rejected.

Camerer et al.'s (2016) replication of 18 RCTs in economics found 11 with significant results at the .05 level. Interestingly, if 41% of nulls are falsely rejected (59% are correctly rejected), given power of over 90%, the authors should have been able to replicate about 10 of the studies.

Table 4: Extended Models

$p <$	$\hat{q}$	$\hat{\lambda}$	$\hat{\delta}$	M	F	Obs
Publication Bias						
.05	0.14	0.50		3.17	0.77	2082
	(0.05)	(0.02)		(0.06)	(0.08)	
.04	0.16	0.51		3.13	0.74	1963
	(0.08)	(0.02)		(0.10)	(0.11)	
.03	0.18	0.51		3.11	0.72	1815
	(0.10)	(0.02)		(0.16)	(0.14)	
.02	0.17	0.51		3.11	0.74	1631
	(0.09)	(0.02)		(0.19)	(0.13)	
.01	0.15	0.50		3.09	0.76	1384
	(0.07)	(0.03)		(0.30)	(0.11)	
Estimated Variance of t						
.05	0.37	0.40	1.70		0.68	2082
	(0.09)	(0.10)	(0.12)		(0.11)	
.04	0.38	0.41	1.69		0.67	1963
	(0.09)	(0.10)	(0.13)		(0.11)	
.03	0.37	0.40	1.72		0.68	1815
	(0.09)	(0.11)	(0.14)		(0.11)	
.02	0.35	0.38	1.81		0.72	1631
	(0.09)	(0.12)	(0.15)		(0.12)	
.01	0.34	0.36	1.90		0.72	14
	(0.10)	(0.14)	(0.17)		(0.12)	
Publication Bias and Estimated Variance of t						
.05	0.39	0.42	1.64	2.02	0.65	2082
	(0.07)	(0.05)	(0.13)	(0.04)	(0.09)	
.04	0.38	0.41	1.67	2.09	0.66	1963
	(0.08)	(0.05)	(0.14)	(0.07)	(0.09)	

Notes: Standard errors in parentheses. They have not been corrected for any within-article correlation. $p <$ identifies the maximum p-value of the hypothesis tests included in the sample. q is the proportion of false nulls. $\lambda > 0$ is the spread parameter of the exponential distribution. δ is the standard error of the $t - statistic$ when the null hypothesis is true. M is the value of the $t - statistic$ above which all rejected nulls are reported. F is the proportion of false rejected null hypotheses when the $t - statistic = 1.96$.

7 Conclusion: A Strong Call for Caution

The credibility revolution in economics has unquestionably improved the quality of published research. It has also contributed to the explosion of empirical research focused on the effects of particular policies. Economics journals increasingly publish exciting new empirical findings. But this explosion and the increasing selectivity of what gets published contribute to an environment in which much of what gets published comes from authors who were lucky, in the sense that the results were sufficiently strong and unexpected to get published, and unlucky, in the sense that they are publishing incorrect information.

In some fields, this problem is at least partially self-correcting. Important experiments are repeated. Frequently, replications will either fail to reproduce the result or reduce its estimated effect size. Ideally, funders make replicating important studies sufficiently attractive that scientists undertake them.

The reward system in economics does not favor replication. Often there is no other data set on which to perform a replication. There is only one Hurricane Katrina or Mariel boatlift. If we find different effects from some other surge of immigrants or natural disaster, we are unsure if the 'experiments' differed or the original result was just an (un)fortunate draw. It is unclear that the distinction is even useful in the context of natural experiments.

This is not a call to abandon the credibility revolution, although I do believe that empirical work would often gain credibility from a closer tie to theory. Instead, it is a call to avoid the hubris of believing too strongly in limited results.

References

Abadie, Alberto. "Statistical nonsignificance in empirical economics." *American Economic Review: Insights* 2, no. 2 (2020): 193-208.

Andrews, Isaiah, and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review,* 109(8) (August 2019): 2766–2794.

Benjamini, Yoav and Yosef Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1) (1995):289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Brodeur, Abel, Nikolai Cook, and Anthony Heyes, "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 110(11) (November 2020): 3634–3660.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics,* 8(1) (January 2016): 1–32.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus

Johannesson, Michael Kirchler et al. "Evaluating replicability of laboratory experiments in economics." *Science* 351, no. 6280 (2016): 1433-1436.

Chesher, Andrew, and Ian Jewitt. "The bias of a heteroskedasticity consistent covariance matrix estimator." *Econometrica* (1987): 1217-1222.

David Colquhoun, "An investigation of the false discovery rate and the misinterpretation of P values," *Royal Society Open Science*, 1(3) (November 2014): 1-16.

Dellavigna, Stefano, and Elizabeth Linos. "RCTs to scale: Comprehensive evidence from two nudge units." Econometrica 90, no. 1 (2022): 81-116.

DeLong, J. Bradford and Kevin Lang, "Are All Economic Hypotheses False?" *Journal of Political Economy*, 100(6) (December 1992): 1257-72.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich. "Detecting p-hacking." *Econometrica* 90, no. 2 (2022): 887-906.

Ioannidis, John P. A., T. D. Stanley, Hristos Doucouliagos, "The Power of Bias in Economics Research," *Economic Journal*, 127(605) (October 2017): F236–F265.

Jeffreys, Harold , "Further Significance Tests," *Mathematical Proceedings of the Cambridge Philosophical Society* 32(3): 416–445.

Jiang, Wei. "Have instrumental variables brought us closer to the truth." *Review of Corporate Finance Studies* 6, no. 2 (2017): 127-140.

Kahn-Lang, Ariella and Kevin Lang, "The Promise and Pitfalls of Differences-in-Differences: Reflections on '16 and Pregnant' and Other Applications," *Journal of Business and Economic Statistics*, 38(3) (July 2020): 613-20.

Keane, Michael and Timothy Neal, "Instrument strength in IV estimation and inference: A guide to theory and practice," Journal of Econometrics, 2023, https://doi.org/10.1016/j.jeconom.2022.12.

Lang, Kevin and Maria Dolores Palacios, "The Determinants of Teachers' Occupational Choice," *NBER Working Paper No. 24883*, August 2018.

Lindley, D. V. "A Statistical Paradox." *Biometrika,* 44, no. 1/2 (June 1957): 187–92. https://doi.org/10.2307/2333251.

Mahoney, Michael J., "Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System," *Cognitive Therapy and Research*, 1(2) (1977):161-73.

McCrary, Justin, Garret Christensen, and Daniele Fanelli. "Conservative tests under satisficing models of publication bias." *PloS one* 11, no. 2 (2016): e0149590.

Moore, David S. "Generalized inverses, Wald's method, and the construction of chi-squared tests of fit." *Journal of the American Statistical Association* 72, no. 357 (1977): 131-137.

Romano, Joseph P. and Michael Wolf, "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing," *Journal of the American Statistical Association*, 100(469)

(2005): 94-108, DOI: 10.1198/016214504000000539

Romer, David. "In praise of confidence intervals." In *AEA Papers and Proceedings*, vol. 110, pp. 55-60. 2014 Broadway, Suite 305, Nashville, TN 37203: American Economic Association, 2020.

Roth, Jonathan. "Pretest with caution: Event-study estimates after testing for parallel trends." *American Economic Review: Insights* 4, no. 3 (2022): 305-22.

Rust, John, "Comment on Poirier: The Subjective Perspective of a 'Spiritual Bayesian'," *Journal of Economic Perspectives*, 2(1) (Spring 1988): 145-151.

Wong, Chi Heem, Ken Wei Siah, and Andrew W. Lo, "Estimation of Clinical Trial Success Rates and Related Parameters," *Biostatistics*, 20(2) (April 2019):273-286. doi: 10.1093/biostatistics/kxx069. Erratum in: *Biostatistics*. 20(2) (April 2019):366.

Young, Alwyn, "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results," *Quarterly Journal of Economics*, Volume 134, Issue 2, May 2019, Pages 557–598, https://doi.org/10.1093/qje/qjy029

Young, Alwyn, "Consistency without Inference: Instrumental Variables in Practical Application," *European Economic Review*, Volume 147, 2022, 104112, https://doi.org/10.1016/j.euroecorev.20

A Appendix

The false discovery rate given in (4) will be decreasing in a if

$$\begin{aligned}
&\frac{d}{da}a^{-1}\left(\Phi\left(1.645-c^*-a\right)-\Phi\left(1.645-c^*+a\right)\right)\\
=&\frac{-1}{2a^2}\left(\text{NormalDist}\left(1.6449-x-a\right)-\text{NormalDist}\left(1.6449-x+a\right)\right)\\
&+\frac{1}{2a}\left(\text{NormalDen}\left(1.6449-x-a\right)+\text{NormalDen}\left(1.6449-x+a\right)\right)
\end{aligned} \tag{13}$$

is decreasing in a. Note that this expression is continuous in a for $a > 0$. Therefore, I solve numerically for combinations of c^* and a with $c^* > a > 0$ for which the derivative in (13) equals 0. There are no combinations with $.65 < c^* < 2.65$ and $a < c$ for which (13) is greater than 0.

A.1 Supplementary Tables

Table A.1: Fitted and actual values for 20 cells

Normal			Exponential		
cell interval	predicted # of obs	actual # of obs	cell interval	predicted # of obs	actual # of obs
[1.96,2.0256)	104.1	78	[1.96,2.0328)	104.1	86
[2.0256,2.0973)	104.1	92	[2.0328,2.1109)	104.1	99
[2.0973,2.1762)	104.1	102	[2.1109,2.1952)	104.1	110
[2.1762,2.2638)	104.1	117	[2.1952,2.2865)	104.1	118
[2.2638,2.3618)	104.1	108	[2.2865,2.3858)	104.1	108
[2.3618,2.4724)	104.1	110	[2.3858,2.4945)	104.1	103
[2.4724,2.598)	104.1	114	[2.4945,2.6138)	104.1	111
[2.598,2.7412)	104.1	115	[2.6138,2.7455)	104.1	105
[2.7412,2.9045)	104.1	115	[2.7455,2.8916)	104.1	107
[2.9045,3.09)	104.1	132	[2.8916,3.0548)	104.1	109
[3.09,3.2993)	104.1	126	[3.0548,3.2378)	104.1	121
[3.2993,3.5335)	104.1	103	[3.2378,3.4445)	104.1	96
[3.5335,3.7943)	104.1	103	[3.4445,3.68)	104.1	92
[3.7943,4.0866)	104.1	111	[3.68,3.9522)	104.1	127
[4.0866,4.4177)	104.1	102	[3.9522,4.273)	104.1	98
[4.4177,4.8027)	104.1	95	[4.273,4.663)	104.1	95
[4.8027,5.27)	104.1	72	[4.663,5.1588)	104.1	98
[5.27,5.88)	104.1	66	[5.1588,5.8391)	104.1	78
[5.88,6.8112)	104.1	92	[5.8391,6.9311)	104.1	101
[6.8112,10)	104.1	129	[6.9311,10)	104.1	120

Notes: This table shows the fitted and actual number of observations for each of the 20 cells. The cell widths are designed to equalize the predicted number of observations across cells. The table includes values for both normally and exponentially distributed t^*.

Table A.2: Fitted and actual values for 40 cells

Normal			Exponential		
cell interval	predicted # of obs	actual # of obs	cell interval	predicted # of obs	actual # of obs
[1.96,1.9921)	52.05	32	[1.96,1.9958)	52.05	39
[1.9921,2.0256)	52.05	46	[1.9958,2.0328)	52.05	47
[2.0256,2.0606)	52.05	48	[2.0328,2.0711)	52.05	54
[2.0606,2.0973)	52.05	44	[2.0711,2.1109)	52.05	45
[2.0973,2.1358)	52.05	42	[2.1109,2.1522)	52.05	56
[2.1358,2.1762)	52.05	60	[2.1522,2.1952)	52.05	54
[2.1762,2.2188)	52.05	61	[2.1952,2.2399)	52.05	64
[2.2188,2.2638)	52.05	56	[2.2399,2.2865)	52.05	54
[2.2638,2.3113)	52.05	46	[2.2865,2.335)	52.05	44
[2.3113,2.3618)	52.05	62	[2.335,2.3858)	52.05	64
[2.3618,2.4154)	52.05	53	[2.3858,2.4389)	52.05	46
[2.4154,2.4724)	52.05	57	[2.4389,2.4945)	52.05	57
[2.4724,2.5332)	52.05	53	[2.4945,2.5527)	52.05	52
[2.5332,2.598)	52.05	61	[2.5527,2.6138)	52.05	59
[2.598,2.6672)	52.05	54	[2.6138,2.678)	52.05	55
[2.6672,2.7412)	52.05	61	[2.678,2.7455)	52.05	50
[2.7412,2.8202)	52.05	55	[2.7455,2.8166)	52.05	49
[2.8202,2.9045)	52.05	60	[2.8166,2.8916)	52.05	58
[2.9045,2.9944)	52.05	64	[2.8916,2.971)	52.05	48
[2.9944,3.09)	52.05	68	[2.971,3.0548)	52.05	61
[3.09,3.1915)	52.05	63	[3.0548,3.1435)	52.05	63
[3.1915,3.2993)	52.05	63	[3.1435,3.2378)	52.05	58
[3.2993,3.413)	52.05	50	[3.2378,3.3378)	52.05	44
[3.413,3.5335)	52.05	53	[3.3378,3.4445)	52.05	52
[3.5335,3.6604)	52.05	48	[3.4445,3.558)	52.05	44
[3.6604,3.7943)	52.05	55	[3.558,3.68)	52.05	48
[3.7943,3.936)	52.05	70	[3.68,3.811)	52.05	60
[3.936,4.0866)	52.05	41	[3.811,3.9522)	52.05	67
[4.0866,4.2466)	52.05	58	[3.9522,4.1055)	52.05	49
[4.2466,4.4177)	52.05	44	[4.1055,4.273)	52.05	49
[4.4177,4.602)	52.05	45	[4.273,4.4574)	52.05	46
[4.602,4.8027)	52.05	50	[4.4574,4.663)	52.05	49
[4.8027,5.0235)	52.05	44	[4.663,4.8943)	52.05	58
[5.0235,5.27)	52.05	28	[4.8943,5.1588)	52.05	40
[5.27,5.551)	52.05	37	[5.1588,5.468)	52.05	42
[5.551,5.88)	52.05	29	[5.468,5.8391)	52.05	36
[5.88,6.283)	52.05	51	[5.8391,6.305)	52.05	51
[6.283,6.8112)	52.05	41	[6.305,6.9311)	52.05	50
[6.8112,7.6225)	52.05	44	[6.9311,7.889)	52.05	54
[7.6225,10)	52.05	85	[7.889,10)	52.05	66

Notes: This table shows the fitted and actual number of observations for each of the 40 cells. The cell widths are designed to equalize the predicted number of observations across cells. The table includes values for both normally and exponentially distributed t^*.

Table A.3: Add variance: 1 null per article

p <	t >	$\hat{q}$	$\hat{\lambda}$	$\hat{\delta}$	F	Obs
.05	1.96	0.3602	0.3539	1.5206	0.6861	501
		(0.0993)	(0.1280)	(0.2132)	(0.1400)	
.04	2.0537	0.3368	0.3807	1.3320	0.6746	481
		(0.0715)	(0.0896)	(0.1981)	(0.1030)	
.03	2.1701	0.3523	0.3690	1.4163	0.6766	438
		(0.0839)	(0.1061)	(0.2338)	(0.1181)	
.02	2.3263	0.3480	0.3677	1.4186	0.6815	394
		(0.0885)	(0.1077)	(0.2657)	(0.1159)	
.01	2.5758	0.3685	0.3621	1.4794	0.6715	331
		(0.1134)	(0.1198)	(0.3505)	(0.1312)	

Notes: Standard errors in parentheses. They have not been corrected for any within-article correlation. p< identifies the maximum p-value of the hypothesis tests included in the sample. q is the proportion of false nulls. $\lambda > 0$ is the parameter of the exponential distribution of $t - statistics$ when the null is false. $\hat{\delta}$ is the variance of the $t - distribution$. F is the implied proportion of falsely rejected null hypotheses when $t = 1.96$.