Essays in Econometrics

Patrick Vu

Thesis

Submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the Department of Economics at
Brown University

Providence, Rhode Island
May 2024

This dissertation by Patrick Vu is accepted in its present form by the Department of Economics as satisfying the dissertation requirement for the degree of Doctor of Philosophy.

Date of Signature

Jonathan Roth, Advisor

Recommended to the Graduate Council

Peter Hull, Reader

Toru Kitagawa, Reader

Approved by the Graduate Council

Thomas A. Lewis, Dean of the Graduate School

Curriculum Vitae

Patrick Vu received his Ph.D. in Economics at Brown University, where he was awarded the George Borts Prize for best doctoral dissertation in economics. He received an MPhil in economics from the University of Oxford and an undergraduate degree from the University of Western Australia, with majors in economics and classical music. In July 2024, Patrick will join the University of New South Wales Business School as an Assistant Professor.

Abstract

This dissertation contains three essays in econometrics. A common theme is the impact of publication bias on the statistical credibility of published research, reproducibility, and evidence-based policy.

The first chapter examines how adopting improved but enlarged standard errors for individual studies can inadvertently lead to higher bias in the studies selected for publication. Intuitively, this is because larger standard errors raise the bar on statistical significance, which exacerbates publication bias. Despite the possibility of higher bias, I show that the coverage of published confidence intervals unambiguously increases. I illustrate these phenomena using a newly constructed dataset on the adoption of clustered standard errors in the difference-in-differences literature between 2000 and 2009. Clustering is associated with a near doubling in the magnitude of published effect sizes. I estimate a model of the publication process and find that clustering led to large improvements in coverage but also sizable increases in bias.

The second chapter examines why replication rates for experimental studies are low in the social sciences. I emphasize that issues with common power calculations in replication studies may play an important role. In a simple model of the publication process, I show that issues with the way that replication power is commonly calculated imply we should always expect replication rates to fall below their intended power targets, even when original studies are unbiased and there is no p-hacking or treatment effect heterogeneity. Empirically, I find that a parsimonious model accounting only for issues with power calculations can fully explain observed replication rates in experimental economics and social science, and two-thirds of the replication gap in psychology.

The third chapter, which is joint work with Toru Kitagawa, examines how publication bias can impact evidence-based policy. For minimax regret policymakers, we characterize the optimal treatment rule with selective publication against statistically insignificant results. We then show that the optimal publication rule which minimizes maximum regret is non-selective. This means that the optimal publication regime for policy choice in the minimax regret framework is also consistent with valid statistical inference in scientific research.

Acknowledgements

I'm deeply grateful for my dissertation committee for their guidance and support. Peter Hull offered wise advice on both research and academic life more broadly. Toru Kitagawa, a co-author in Chapter 3, taught me the necessity of patience with proofs, especially as we wrestled with our own problem over several sessions, mostly coming up empty-handed, but eventually reaching the solution. Finally, I am incalculably grateful for my committee chair, Jon Roth, whose tremendous support throughout graduate school went over and above what I could ever have expected from an intellectual mentor.

I also benefited from comments and discussions with other faculty at Brown and elsewhere, including Johannes Abeler, Daniel Björkegren, Kenneth Chay, Pedro Dal Bó, Anna Dreber, Tomáš Havránek, Soonwoo Kwon, Susanne Schennach, Jesse Shapiro and Aleksey Tetenov.

Finally, I'm grateful for my family and friends. My parents and two older brothers, who have always supported me in everything I do. Jasmina, my sister-in-law, whose ambition for my intellectual endeavours has at times even exceeded my own. Peggy and Pierre, whose questions kept them with me every step along the way. Andy Lewis, my close friend and intellectual (and musical) collaborator. And lastly to my wife, Yelena, for whom it isn't possible to list all the things for which I'm grateful.

Contents

List of Figures

List of Tables

Chapter 1

Do Standard Error Corrections Exacerbate Publication Bias?

Abstract. Over the past several decades, econometrics research has devoted substantial efforts to improving the credibility of standard errors. This paper studies how such improvements interact with the selective publication process to affect the ultimate credibility of published studies. I show that adopting improved but enlarged standard errors for individual studies can inadvertently lead to higher bias in the studies selected for publication. Intuitively, this is because increasing standard errors raises the bar on statistical significance, which exacerbates publication bias. Despite the possibility of higher bias, I show that the coverage of published confidence intervals unambiguously increases. I illustrate these phenomena using a newly constructed dataset on the adoption of clustered standard errors in the difference-in-differences literature between 2000 and 2009. Clustering is associated with a near doubling in the magnitude of published effect sizes. I estimate a model of the publication process and find that clustering led to large improvements in coverage but also sizable increases in bias. To examine the overall impact on evidence-based policy, I develop a model of a policymaker who uses information from published studies to inform policy decisions and overestimates the precision of estimates when standard errors are unclustered. I find that clustering lowers minimax regret when policymakers exhibit sufficiently high loss aversion for mistakenly implementing an ineffective or harmful policy.

1.1 Introduction

Over the past several decades, econometrics research has devoted substantial efforts to improving the accuracy of estimated standard errors in a wide variety of settings (White, 1980; Moulton, 1986; Newey and West, 1987; Staiger and Stock, 1997). In practice, these improvements often lead to larger standard errors that increase the coverage of reported confidence intervals for a given study. However, larger standard errors also make statistical significance more difficult to obtain, and insignificant results are frequently censored in the publication process (Franco et al., 2014; Brodeur et al., 2016; Andrews and Kasy, 2019). Thus, the studies that are ultimately selected for publication may depend critically on how standard errors are calculated. This in turn can affect the statistical credibility of published research in unanticipated ways.

Little attention has been paid to the close connection between standard error corrections and selective publication. This paper studies how their interaction can affect true and estimated treatment effects in published research, bias, and overall coverage. A key insight is that increasing reported standard errors effectively raises the bar for statistical significance, which can exacerbate publication bias. Higher bias pushes toward undercoverage, raising questions about whether more robust inference methods actually meet their primary aim of improving coverage conditional on publication. I develop a theoretical framework to answer these questions and then apply it to the difference-in-differences (DiD) literature in the 2000's when clustering was growing in popularity.

I begin by extending the selective publication model in Andrews and Kasy (2019) to incorporate the possibility that reported standard errors are mismeasured. In the model, researchers draw an estimated treatment effect $\hat{\beta}_j$ from an $N(\beta_j, \sigma_j^2)$ distribution, where the true treatment effect and standard error (β_j, σ_j) are drawn from a joint probability distribution $\mu_{\beta,\sigma}$. Publication may depend on the statistical significance of the reported t-ratio, either because journals prefer publishing significant results or because researchers do not write them up in anticipation of low chances of publication. In contrast to the standard model, reported standard errors may be downward biased (and t-ratios upward biased). This makes it easier to obtain statistical significance, which can increase the probability of publication. The model applies to clustered standard errors to account for serial correlation, which is the empirical setting I analyze, but also more generally to any corrections that tend to enlarge reported standard errors e.g. heteroscedasticity-robust standard errors, heteroscedasticity and autocorrelation consistent standard errors, or corrections for weak instruments.

Using this framework, I show that average bias in published studies can either increase or decrease following standard error corrections, but that increases are inevitable when corrections are sufficiently large. Moreover, I show that analogous results hold for changes in true and estimated treatment effects. The case of large corrections is empirically relevant because uncorrected standard errors have been shown in many instances to be severely downward biased.[1] Intuitively, in a regime where standard errors are severely downward biased, a relatively high share of estimates will be reported as statistically significant (often erroneously). This means that relatively few studies are censored by selective publication, leading to little bias in published studies. By contrast, in a regime where standard errors are correctly measured, and hence larger, a greater share of estimates will be insignificant and censored through the publication process, resulting in higher bias (Ioannidis, 2008; Andrews and Kasy, 2019; Frankel and Kasy, 2022). However, with small corrections, it is possible to construct examples where bias decreases. For instance, corrections can shift the distribution of published studies to those with larger true effects. Such studies tend to generate larger estimates which are less likely to be censored by selective publication. This can lead to lower bias overall.

Despite the possibility of higher bias, I show that standard error corrections unambiguously increase average coverage in published confidence intervals. This holds under very general conditions. In particular, it holds for any degree of selective publication against null results, any sized correction, and for arbitrary distributions of true treatment effects. In practical terms,

[1] For example, Abadie et al. (2023) find using US Census Data that standard errors clustered at the state level are more than 20 times larger than robust standard errors.

this means that we can extend the common intuition that standard error corrections increase coverage in individual studies to the more realistic case where publication favors statistical significance. Overall, the theoretical results highlight a striking tension: in the presence of publication bias, standard error corrections enhance the credibility of published confidence intervals, but can also inadvertently deteriorate the credibility of published point estimates.

I turn next to studying these issues empirically, using a new dataset I constructed from DiD studies published between 2000–2009. Over this period, clustered standard errors to account for serial correlation became common practice, in part because of an influential study by Bertrand et al. (2004) that demonstrated their practical importance. My data are drawn from the same six economics journals analyzed in that study, but for a later period.[2] The DiD studies in the sample consist primarily of policy evaluations (e.g. health care, tax, education). This is a compelling setting for applying the theoretical results for two reasons. First, DiD is an extremely popular research design in the quantitative social sciences. In economics, it is the most widely referenced quasi-experimental method and its popularity has increased dramatically over time (Currie et al., 2020). Second, failing to cluster frequently results in large downward bias in standard errors, which can lead to exaggerated statistical support for the effectiveness of an intervention (Moulton, 1986, 1990; Bertrand et al., 2004).

Descriptive statistics reveal two striking patterns that are consistent with clustering interacting with publication bias to change the distribution of published estimates. First, the adoption of clustered standard errors in the empirical DiD literature over the 2000's was associated with a near doubling in the magnitude of estimated treatment effects. This large gap remains even after controlling for differences in research topics, sample size, and including year and journal fixed effects. Second, the data exhibit strong evidence for publication bias favoring statistical significance. Following the metaregression approach in Card and Krueger (1995), I find, for both unclustered and clustered studies, a strong positive association between stan-

[2] The journals are: *American Economic Review*, the *Industrial and Labor Relations Review,* the *Journal of Labor Economics*, the *Journal of Political Economy*, the *Journal of Public Economics*, and the *Quarterly Journal of Economics.*

dard errors and effect sizes, such that the overwhelming majority of published studies report statistically significant results. Following Brodeur et al. (2016), I also plot the distributions of test statistics for unclustered and clustered studies. Both distributions are strikingly similar and show substantial bunching around the 5% significance threshold, which is suggestive of publication bias and p-hacking.

The theory emphasizes that we cannot make inferences about the sign of the change in bias or the magnitude of the increase in coverage from these reduced-form facts alone. To learn about the impact of clustering on bias and coverage, I therefore estimate an augmented version of the Andrews and Kasy (2019) model using data from clustered studies.[3] Consistent with estimates in alternative settings, I find a high degree of publication bias in the empirical DiD literature: significant findings at the 5% level over 60 times more likely to be published than insignificant findings.

Next, I use the estimated model to calculate what would have happened if clustered studies had instead reported unclustered standard errors. To do this, I make the simplifying assumption that unclustered standard errors are downward biased by a constant factor r. I then calibrate r such that the model prediction matches differences in key moments between the clustered and unclustered studies, assuming the same underlying distribution of latent (published and unpublished) studies. This gives $\hat{r} = 0.51$, meaning that clustered standard errors tend to be around twice the size of unclustered standard errors.

Model estimates show that clustering led to large improvements in coverage. In the unclustered regime, the coverage probability of published confidence intervals was only 0.28. This implies severe mismeasurement in the calculation of confidence intervals prior to the adoption of clustering, with fewer than one in three published confidence intervals containing the true parameter value. By contrast, coverage increased to 0.70 in the clustered regime, a large improvement but still below nominal coverage of 0.95 due to publication bias.

Despite substantial improvements in coverage, clustering also led to average bias in published

[3] The augmented empirical model follows Vu (2023), which extends the empirical model in Andrews and Kasy (2019) to estimate the latent distribution of standard errors.

studies doubling, from 1.23 percentage points to 2.44 percentage points. This is equivalent to the increase in bias that would occur when moving from a regime with no selective publication (where bias is zero) to one that censors 85% of statistically insignificant results at the 5% level with clustered standard errors. That is, the impact of clustering on bias is comparable to a fairly severe degree of publication bias. The model estimates also show that clustering led to the selection of studies for publication with larger true and estimated treatment effects, since these studies are, all else equal, more likely to produce statistically significant results.

Given the trade-offs between bias and coverage, the welfare implications of clustering are unclear. To understand the implications of clustering on evidence-based policy, I develop a model where policymakers use evidence from published studies to inform a policy decision, but where reported standard errors may be unclustered. In the model, a policymaker chooses a treatment rule which maps findings from published studies to policy choices, with the aim of minimizing maximum regret i.e. the expected welfare loss due to making the inferior decision (Savage, 1951; Manski, 2004; Stoye, 2009; Tetenov, 2012). Following Frankel and Kasy (2022) and Kitagawa and Vu (2023), I consider the case where selective publication can censor studies from being observed by policymakers.

My treatment choice model extends existing frameworks by analyzing treatment choice under the mistaken belief that unclustered standard errors reflect the true standard error. This operationalizes the costs and benefits of clustering in a policy setting. On the one hand, clustered standard errors allow policymakers to more accurately gauge the statistical precision of the evidence contained in published studies, resulting in better informed decisions. On the other hand, studies with larger standard errors are more likely to be insignificant and censored, leaving policymakers to act without evidence.

Calibrating the treatment choice model to the DiD setting, I find that clustering lowers minimax regret when policymakers weigh welfare losses from implementing an ineffective or harmful treatment (Type I error) at least 63 times more than welfare losses from failing to implement a beneficial treatment (Type II error). As a benchmark, note that Type I error would need to be weighed around 100 times more than Type II error for a decision rule that minimizes maximum regret to rationalize hypothesis testing with a 5% statistical significance threshold (Tetenov, 2012). Thus, the model suggests that clustering improves treatment choice if we use the benchmark implicitly implied by conventional hypothesis testing. The intuition behind this result is that decision-makers in the unclustered regime overestimate the precision of published parameter estimates, which leads to a suboptimal decision rule that is too lenient with respect to the evidence required for implementing the policy. This leniency is especially costly when policymakers exhibit a high degree of loss aversion for mistakenly implementing an ineffective or harmful policy (i.e. Type I error).

Related Literature. This paper contributes to, and connects, two large literatures: the metascience literature on publication bias (Card and Krueger, 1995; Ioannidis, 2005, 2008; Franco et al., 2014; Gelman and Carlin, 2014; Ioannidis et al., 2017; Miguel and Christensen, 2018; Amrhein et al., 2019b; Andrews and Kasy, 2019; Frankel and Kasy, 2022; DellaVigna and Linos, 2022) and the econometrics literature on robust measures of uncertainty (Anderson and Rubin, 1949; White, 1980; Moulton, 1986, 1990; Bertrand et al., 2004; Lee et al., 2022; Abadie et al., 2023). While both literatures are guided by the overarching goal of improving the credibility of empirical analysis, little attention has been paid to how they interact. This paper builds on existing publication selection models to provide general theoretical results on how standard error corrections can affect estimated treatment effects, true treatment effects, bias and coverage. Empirically, it uses newly collected data from the DiD literature to show that clustering led to substantial improvements in coverage but also large increases in bias.

This paper also contributes to the literature on statistical decision theory and treatment choice (Wald, 1950; Savage, 1951; Stoye, 2009, 2012; Tetenov, 2012; Kitagawa and Tetenov, 2018; Frankel and Kasy, 2022). In the existing literature, treatment choice models typically assume that standard errors are correctly measured. This paper extends existing minimax regret models to incorporate concerns in the econometrics literature that statistical inference is impaired by mismeasured standard errors. It develops a treatment choice model where policymakers overestimate the precision of published estimates when reported standard errors are unclustered.

This paper proceeds as follows. Section 2.2 develops the theoretical framework and presents the main propositions. Section 1.3 describes the empirical setting and presents the descriptive statistics. Section 1.4 shows the results from the empirical model. Section 1.5 develops the treatment choice model and presents the main welfare results. Section 1.6 concludes.

1.2 Theory

1.2.1 Model of Publication Bias and Standard Error Corrections

I begin by introducing a model of how studies are generated and published in an empirical literature of interest. This could be a literature addressing many different research questions (e.g. the DiD literature). Alternatively, it could be a meta-analysis focused on a single question (e.g. the impact of job training programs on employment outcomes). The model builds on the selective publication model in Andrews and Kasy (2019) to incorporate the possibility that reported standard errors are downward biased. While much of the discussion is framed around clustering to match the empirical application, the same model applies more generally to any method correcting for downward bias in standard errors. For proofs of the propositions, see Appendix 3A.

Suppose we observe estimated treatment effects, standard errors, and an indicator for whether or not standard errors are corrected for a sample of published studies indexed by j. The model of the DGP has five steps:

1. **Draw latent true treatment effect and standard error:** Draw a research question with true treatment effect (β_j) and standard error (σ_j):

$$(\beta_j, \sigma_j) \sim \mu_{\beta,\sigma}$$

where $\mu_{\beta,\sigma}$ is the joint distribution of latent true effects and latent standard errors.

2. **Estimate the treatment effect:** Draw an estimated treatment effect from a normal distribution with parameters from Stage 1:

$$\hat{\beta}_j|\beta_j, \sigma_j \sim N(\beta_j, \sigma_j^2)$$

3. **Report standard errors based on 'standard error regime' r:**

$$\widetilde{\sigma}_j = r \cdot \sigma_j$$

where the corrected regime ($C_j = 1$) has $r = 1$ and the uncorrected regime ($C_j = 0$) has $r \in (0, 1)$.

4. **Publication selection:** Selective publication is modelled by the function $p(\cdot)$, which returns the probability of publication for any given t-ratio using the reported standard error. Let D_j be a Bernoulli random variable equal to one if the study is published and zero otherwise:

$$\mathbb{Pr}(D_j = 1|\hat{\beta}_j, \widetilde{\sigma}_j) = p\left(\frac{\hat{\beta}_j}{\widetilde{\sigma}_j}\right) \tag{1.1}$$

We observe i.i.d. draws from the conditional distribution of $(\hat{\beta}_j, \widetilde{\sigma}_j, C_j)$ given $D_j = 1$. In the corrected regime, standard errors are accurately measured with $r = 1$ and the model coincides with the Andrews and Kasy (2019) model. However, the model differs in the uncorrected regime, since reported standard errors are downward biased with $r \in (0, 1)$. This implies that reported t-ratios are upward biased since $|\hat{\beta}_j|/\widetilde{\sigma}_j > |\hat{\beta}_j|/\sigma_j$. Imposing a constant downward bias factor of r permits a simple exposition of the model.[4] In the empirical application, I perform a robustness exercise where r is drawn from a distribution.

[4] Note however that all theoretical results can be generalized to the case where r is a random variable with support on $(0, 1)$, provided that $r \perp\!\!\!\perp (\hat{\beta}_j, \beta_j, \sigma_j)$.

I impose a number of regularity conditions and assumptions. First, I normalize true treatment effects to be positive and assume a finite first moment:

Assumption 1 (True Treatment Effect Normalization)**.** *Let β_j have support on a subset of the non-negative real line, not be degenerate at zero, and have a finite first moment.*

For empirical literatures examining different questions and outcomes, normalizing true effects to be positive is justified because relative signs across studies are arbitrary. The requirement that β_j not be degenerate at zero is to avoid the special case where coverage probabilities always equal zero when all insignificant results are censored by the publication process.

Second, I assume that true effects are statistically independent of standard errors:

Assumption 2 (Independence of True Effects and Standard Errors)**.** *Let $\beta_j \perp\!\!\!\perp \sigma_j$.*

This is commonly assumed in meta-analyses and is also assumed in the 'meta-study' estimation approach proposed in Andrews and Kasy (2019), which I implement in the empirical section. It is unlikely to hold when experimental researchers choose sample sizes based on predicted effect sizes in power analyses (e.g. Camerer et al. (2016)) or when target parameters are mechanically correlated with standard errors through measurement.[5] However, it may be more likely to hold in experimental settings where exogenous budget constraints are the main determinant of sample sizes, or in observational settings where available datasets are the primary determinant of the sample size.

Finally, I impose the assumption that publication bias depends only on statistical significance:

Assumption 3 (Publication Selection Function)**.** *Let $p(\hat{\beta}_j/\widetilde{\sigma}_j) = 1 - (1-\gamma) \cdot \mathbb{1}[|\hat{\beta}_j|/\widetilde{\sigma}_j < 1.96]$ with $\gamma \in [0, 1)$.*

[5] For example, Chen (2023) considers estimates of tract-level economic mobility in the Opportunity Atlas (Raj et al., 2020). Census tracts with more low-income household have (i) lower true economic mobility and (ii) more precise estimates of economic mobility due to larger sample sizes. This generates a positive correlation between true economic mobility and standard error estimates.

That is, significant results (based on the reported standard error) at the 5% level are published with probability one, while insignificant results are published with probability $\gamma \in [0, 1)$. This assumption is used to match the common concern that publication favors statistical significant findings. The 5% significance level is chosen because it is the most commonly used critical threshold. However, the main theoretical results generalize to other critical thresholds.

Illustrative Example

Consider a simple example to illustrate the model and motivate the general theoretical results which follow. Suppose researchers are interested in studying the impact of a health reform on average life expectancy, and that the reform is implemented in some states and not others.

For the first stage of the model, suppose the average treatment effect for treated states (ATT) is equal to a one-year improvement in life expectancy, $\beta = 1$, and that the standard error is $\sigma_j = 1$ for all studies $j = 1, 2, ...J$ (i.e. the joint distribution of true effects and standard errors, $\mu_{\beta,\sigma}$, is degenerate). In the second stage, researchers conduct a large number of independent DiD studies to learn about the (unobserved) ATT, each producing an unbiased DiD estimate $\hat{\beta}_j$ drawn from a $N(1, 1)$ distribution. For the third stage, we consider two regimes for calculating standard errors. In the clustered regime, researchers correctly cluster by state and reported standard errors equal true standard errors ($\widetilde{\sigma}_j = \sigma_j$). However, in the unclustered regime, researchers fail to cluster by state and erroneously report standard errors which are half their true value ($r = \frac{1}{2}$ and $\widetilde{\sigma}_j < \sigma_j$). In the fourth and final stage, only a subset of the latent DiD estimates $\hat{\beta}_j$ are published due to publication bias. In particular, suppose that the publication process censors all insignificant findings at the 5% level (i.e. $\gamma = 0$ in Assumption 3).

While both standard errors regimes are subject to the same degree of publication bias, statistical significance is easier to obtain in the unclustered regime because t-statistics are upward biased by a factor of two. Thus, the studies selected for publication differ across regimes. We are interested in how this affects both bias and coverage in published DiD studies.

First, consider bias and recall that the true ATT is a one-year improvement in life expectancy. In the unclustered regime, reported standard errors are half the true value such that the effective threshold for statistical significance is half of what it should be. Thus, all DiD estimates $\hat{\beta}_j$ whose absolute values are smaller than $1.96 \times \frac{1}{2} = 0.98$ years are censored by selective publication. This clearly leads to upward bias, such that the average DiD estimate conditional on publication is $\mathbb{E}_r[\hat{\beta}_j | D_j = 1] = 1.64$ years (where the subscript indicates the standard error regime $r = \frac{1}{2}$). Clustering makes matters worse because increasing reported standard errors raises the effective threshold for statistical significance. Now, DiD estimates whose absolute values are smaller than 1.96 years are censored such that the average DiD estimate conditional on publication increases to $\mathbb{E}_1[\hat{\beta}_j | D_j = 1] = 2.45$ years.

Overall, clustering increases bias by 0.81 years (or 125%). This is more than twice the magnitude of bias in the unclustered regime and equal to around four-fifths of the true ATT. It is equivalent to the increase in bias that would arise when moving from a regime with no publication bias to a regime where 88% of insignificant results at the 5% level are censored (based on correctly measured standard errors). In other words, clustering has a large impact on bias which is comparable to very severe levels of selective publication.

Higher bias implies that estimates are, on average, further away from the true ATT. This raises the question of whether clustering could potentially fail to meet its primary goal of improving the average coverage of published confidence intervals (in this example, and also more generally). It turns out that coverage conditional on publication does in fact increase in this case, by 19 percentage points (0.65 to 0.84). The proof in Lemma 1A.6 in Appendix 3A shows that higher coverage is equivalent to showing that the hazard function of the normal distribution is increasing.

This example illustrates a key tension emphasized throughout this paper: for the studies selected for publication, improvements in the credibility of confidence intervals through better coverage ($\uparrow$ 19 ppts) can come at the unintended cost of a deterioration in the credibility of point estimates due to increased bias ($\uparrow$ 125%). It also demonstrates that these effects can be large.

This tension has only been shown here for a special case where $(\mu_{\beta,\sigma}, \gamma, r) = \big(\mathbb{P}\mathrm{r}[\beta_j = 1, \sigma_j = 1] = 1, 0, \frac{1}{2}\big)$. In the remainder of this section, I move beyond this special case to answer, in general, what happens to bias and coverage in published studies when standard error corrections for downward bias are applied. In particular, I derive exact conditions under which the tension between increased bias and coverage generalizes to other settings.

1.2.2 Bias

The illustrative example shows that it is possible for standard error corrections to increase bias in published studies. Under what conditions does this conclusion hold more generally? I find that a sufficient condition for increased bias is that corrections are 'sufficiently' large, and present an example where small corrections can lead to a decrease in bias.

Before presenting the main result, I first define the key measures of interest. Throughout, I normalize the true standard error to $\sigma_j = 1$ and omit it from the notation for clarity. Note that the theoretical results apply both to empirical literatures examining a single question of interest (e.g. the impact of a health reform on life expectancy) and to those addressing different research questions (e.g. the empirical DiD literature examining different policy evaluations).

The theoretical results will apply to several measures of bias. The first measure is *internal-validity bias*, which is defined $\mathbb{E}_r[\hat{\beta}_j - \beta_j | D_j = 1]$ and where the subscript r in the expectation denotes the standard error regime. The publication regime, γ, is implicit in the notation, since the main focus is on standard error corrections. Internal-validity bias asks how far, on average, published estimates are from the questions they answer. The second measure is *study-selection bias*, which is defined as $\mathbb{E}_r[\beta_j | D_j = 1] - \mathbb{E}[\beta_j]$.[6] This measures how far, on average, published true effects are from the average that would occur if there were no publication bias. In certain

[6] In general, study-selection bias is non-zero because true treatment effects β_j follow a distribution. This applies both when the empirical literature of interest is concerned with different questions and when it examines a single question. Variation in true treatment effects may arise in the latter case because of heterogeneity across studies in populations, research design, policies etc.

contexts, this is referred to as 'site-selection bias' (Allcott, 2015).

The relevant measure of bias can depend on context. To illustrate, consider the previous example of the impact of a health reform on life expectancy. Suppose that the true ATT of a one-year improvement in life expectancy is in fact a weighted average of heterogeneous treatment effects across treated states. Moreover, assume that different studies examine different subsets of treated states.[7] First, consider a scenario where study-selection bias is the primary object of interest. Suppose continued federal funding for this health program depends on the average treatment effect in treated states, $\mathbb{E}[\beta_j]$. However, due to publication bias for positive results, studies examining states where the program is most effective are most likely to be published, leading to positive study-selection bias. This exaggerates the average effectiveness of the policy and may lead to a less informed decision with respect to federal funding. Next, consider a scenario where internal-validity bias is the primary concern. Suppose that heterogeneous effects across treated states reflect variation in program features e.g. the cost structure. Policymakers are interested in rolling out the health reform in a new, untreated state and want to know which cost structure will be most effective in producing positive health outcomes. In this scenario, policymakers may be relatively unconcerned if study-selection bias skews toward published studies examining states where the policy is most effective, since this happens to align with their objectives. Instead, their primary concern is internal-study bias conditional on cost structure, so as to correctly gauge the likely impact of the policy in the new, untreated state.[8]

Additionally, in the case where the empirical literature of interest examines many different questions (e.g. the DiD literature analyzed in the empirical section), the primary concern may also be internal-validity bias. In this context, study-selection bias reflects different research questions being addressed in the published literature compared to the case without publication bias. Since different studies are examining different questions, this kind of selection has less clear implications for statistical credibility.

[7] This could arise, for example, due to idiosyncratic data constraints faced by individual researchers.

[8] Selecting policies based on those with the largest estimates is known to induce upward bias in estimated policy impact. Procedures for correcting inference for this 'winner's curse' are studied in Andrews et al. (2023).

Finally, consider *total bias*, which is defined as $\mathbb{E}_r[\hat{\beta}_j|D_j = 1] - \mathbb{E}[\beta_j]$. It asks how far published estimates are from the average true effect across all latent studies, and is equal to the sum of internal-validity bias and study-selection bias. This relationship gives rise to the following decomposition, which provides useful intuition for examining how standard error corrections can affect each type of bias:

$$\begin{aligned} &\underbrace{\mathbb{E}_1[\hat{\beta}_j|D_j = 1] - \mathbb{E}_r[\hat{\beta}_j|D_j = 1]}_{\Delta\text{Estimated Treatment Effects} = \Delta\text{Total Bias}} \\ &= \underbrace{\mathbb{E}_1[\hat{\beta}_j - \beta_j|D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta_j|D_j = 1]}_{\Delta\text{Internal-Validity Bias}} + \underbrace{\mathbb{E}_1[\beta_j|D_j = 1] - \mathbb{E}_r[\beta_j|D_j = 1]}_{\Delta\text{Study-Selection Bias}} \end{aligned} \tag{1.2}$$

That is, the change in total bias is equal to the sum of the change in internal-validity bias and study-selection bias. The main result of this subsection provides a sufficient condition under which all three changes are positive:

Proposition 1 (Large Corrections Increase Bias)**.** *Under Assumptions 1, 2, and 3, there exists an $r^* \in (0, 1]$ such that for any $r \in (0, r^*)$, internal-validity bias, study-selection bias, and total bias all increase with standard error corrections.*[9]

Proposition 1 states that sufficiently large standard error corrections inevitably lead to increases in each of the three types of bias discussed. This is important for two reasons. First, it implies that corrections are most likely to increase bias in published studies in the cases where they are most needed. Second, prior evidence suggests relatively severe downward bias in uncorrected standard errors in practice (Moulton, 1986, 1990; Bertrand et al., 2004). Thus, large downward bias in uncorrected standard errors may be the empirically relevant case, although a definitive answer requires knowledge of the underlying model parameters, which we estimate in the empirical section for DiD studies.

For intuition underlying Proposition 1, consider internal-validity bias (other measures share similar intuition). When standard errors are severely downwardly biased, almost all results are

[9] All inequalities are strict except for study-selection bias, which is a weak inequality. If the latent distribution of true treatment is non-degenerate, then the inequality for study-selection bias is also strict.

reported as significant. Consequently, there is very little selective publication and estimates have relatively small internal-validity bias. However, corrections increase standard errors, which leads to more studies with small effect sizes being censored by the publication process and hence higher bias. It follows that moving from the uncorrected regime with little bias to the corrected regime must necessarily increase bias.

To see why the sufficient condition of large corrections is required, consider an example where small standard error corrections lead to a *decrease* in internal-validity bias.[10] Consider a literature addressing two research questions, one with a small true effect and one with a large true effect. Specifically, let the latent distribution of true effects β_j take on two possible values $(\beta_1, \beta_2) = (1, 4)$ with probabilities $\frac{4}{5}$ and $\frac{1}{5}$, respectively. Assume only one in twenty insignificant studies are published ($\gamma = \frac{1}{20}$) and unclustered standard are 80% of their true value ($r = \frac{4}{5}$).

In the clustered regime, a higher share of studies addressing the question with the larger effect ($\beta_2 = 4$) are published relative to the unclustered regime. This is because studies addressing the question with the smaller true effect ($\beta_2 = 1$) are more likely to be insignificant with clustering and hence censored by selective publication. This decreases average internal-validity bias overall because studies addressing questions with very large effect sizes have bias close to zero.[11] The intuition behind this is that when true effects are large, the probability of obtaining an insignificant result, and thus being subject to publication bias, is low. Overall, then, clustering shifts the distribution of published studies toward those with larger true effects and hence smaller bias.

This example highlights a second important point: it is possible for estimated treatment effects to increase with clustering, despite the fact that internal-validity bias decreases. To see why, consider again the decomposition in equation (1.2). Clustering in this example leads to an overall increase in estimated treatment effects (0.30) that reflects an increase in true

[10] See Appendix 1B for examples where study-selection bias and total bias can decrease with small standard error corrections.

[11] This is shown graphically in Figure 1C.1 in Appendix 1C.

treatment effects (0.31) which outweighs a decrease in internal-validity bias (-0.01). Thus, by observing higher effect sizes in clustered studies, it is not possible, in general, to infer the sign of the change in bias. This underscores the limitations of what we can learn about bias from reduced-form statistics calculated on observed effect sizes. Proposition 1, of course, guarantees that bias must increase if corrections are sufficiently large. Figure 1.1 illustrates this by tracing out the change in internal-validity bias from adopting different sized standard error corrections (r). In this example, we have that $r^* = 0.77$, meaning that corrections that enlarge standard errors by more than 30% will lead to an increase in bias.

In summary, internal-validity bias, study-selection bias, and total bias can in general increase or decrease with corrections, but must always increases when corrections are sufficiently large.

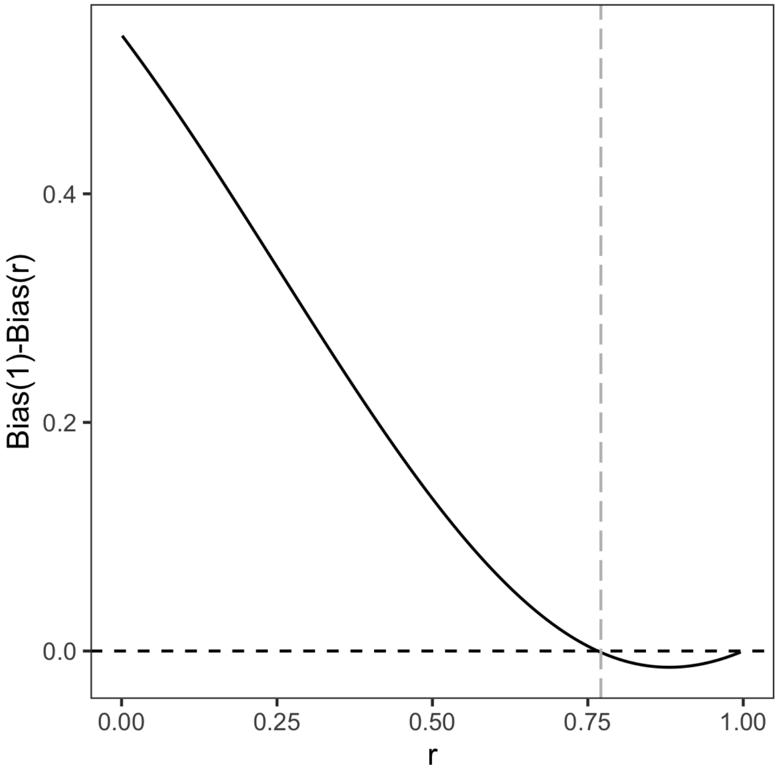

Figure 1.1: Change in Internal-Validity Bias

Notes: Change in Internal-Validity Bias from adopting standard error corrections for different degrees of downward bias r: $\mathbb{E}_1[\hat{\beta}_j - \beta_j | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta_j | D_j = 1]$, with $\gamma = \frac{1}{20}$. The dashed vertical line at $r^* = 0.77$ denotes the value of below which bias always increases with standard error corrections.

1.2.3 Coverage

We turn next to how standard error corrections impact coverage probabilities in the presence of publication bias. First, define *expected coverage conditional on publication* in standard error regime $r \in (0,1]$ as $\text{Coverage}(r) = \mathbb{Pr}_r[\beta_j \in (\hat{\beta}_j - 1.96r, \hat{\beta}_j + 1.96r)|D_j = 1]$ i.e. the probability that published 95% confidence intervals based on reported standard errors contain the true effect.[12] Compare this to expected coverage in a standard econometric analysis without publication bias: $\mathbb{Pr}_r[\beta_j \in (\hat{\beta}_j - 1.96r, \hat{\beta}_j + 1.96r)]$. In the case without publication bias, it is clear that standard error corrections for downward bias will increase coverage.

The presence of publication bias, however, introduces several complications. In the definition of $\text{Coverage}(r)$, see that the degree of downward bias affects not only the width of reported confidence intervals, but also the studies $(\hat{\beta}_j, \beta_j)$ that end up making it into the published literature, since uncorrected standard errors are more likely to lead to statistically significant findings. This can complicate comparisons between uncorrected and corrected regimes. To illustrate, consider Figure 1.2, which depicts three possible realizations of the estimated treatment effect $\hat{\beta}$ (black points) for a fixed true effect β. Each realization would be treated differently under corrected and uncorrected regimes. Confidence intervals with corrections (purple) are twice the width of those without corrections (yellow). Consider each case:

1. **Expand CIs to include β:** an interval that did not cover β or zero in the uncorrected regime now expands to cover β while still not covering zero in the corrected regime.

2. **Expand CI of a covered study to include zero:** an interval that covered β but not zero in the uncorrected regime now expands to cover zero and is therefore censored with some positive probability in the corrected regime.

3. **Expand CI for an uncovered study to include zero:** an interval that did not cover β or zero in the uncorrected regime now covers zero and is censored with some positive

[12] This definition is similar to the coverage concept discussed in Armstrong et al. (2022) in relation to *empirical Bayes confidence intervals*, although here I condition on publication.

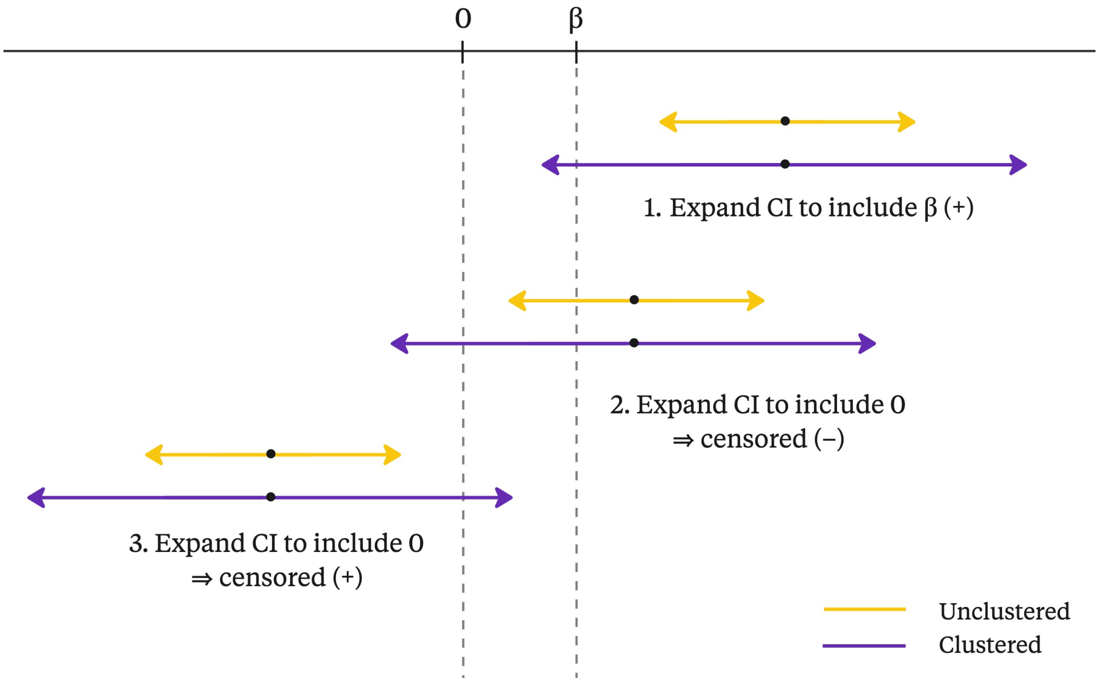
0
β
1. Expand CI to include β (+)
2. Expand CI to include 0
⇒ censored (–)
3. Expand CI to include 0
⇒ censored (+)
Unclustered
Clustered

Figure 1.2: Three Potential Effects of Clustering on Coverage Conditional on Publication

probability in the corrected regime.

In standard analyses that do not account for publication bias, the first effect is the only relevant case and hence corrections clearly improve coverage. The second and third effects occur due to publication bias, since corrections can now censor studies that would otherwise be published. The second effect decreases coverage and the third increases it.

In general, it is not clear a priori which effects dominate or even whether any of them do dominate in all cases. A key reason for this difficulty lies in the fact that different true effects end up in the published literature for the corrected and uncorrected regimes owing to selective publication. Thus, the relative share of published estimates in each of the three cases listed above varies across regimes and ultimately depends on the underlying model parameters. Given that I allow for arbitrary distributions of latent true effects, μ_β, this opens up a large set of possible comparisons, including those which would in principle most favor corrections worsening coverage.

Despite these complications, the next result states, in general, that expected coverage in published studies unambiguously increases:

Proposition 2 (Standard Error Corrections Increase Coverage)**.** *Under Assumptions 2 and 3, $Coverage(1) - Coverage(r) > 0$ for any $r \in (0,1)$.*

In practical terms, Proposition 2 means that we can extend the common intuition that coverage increases with standard error corrections in individual studies to the more realistic case where there is publication bias. It also rules out the possibility that both bias and coverage might worsen with standard error corrections. In conjunction with Proposition 1, this implies that standard error corrections always improve the average quality of variance estimates in published studies, but can worsen bias when corrections are large.

The proof of Proposition 2 builds on the special case where the distribution of true effects β_j is degenerate and $\gamma = 0$.[13] The proof shows that this conclusion holds more generally, in particular, for (i) arbitrary levels of selective publication against null results, $\gamma \in (0,1)$; and for (ii) arbitrary distributions of latent studies μ_β. Both generalizations are non-trivial extensions of the special degenerate case. This is because the distribution of published studies, $\hat{\beta}_j, \beta_j | D_j = 1$, on which expected coverage is calculated, depends jointly on the degree of selective publication γ, the extent to which standard errors are downward biased by r, and the latent distribution of true effects μ_β.

The generalization to any level of selective publication makes use of a result which shows that any publication regime $\gamma \in [0,1]$ can be expressed as a mixture of a publication regime which publishes all insignificant results ($\gamma = 1$) and one that censor all insignificant results ($\gamma = 0$). Loosely speaking, since coverage trivially improves in the former regime, we only need to focus on the latter case where $\gamma = 0$. Generalizing the result to non-degenerate distributions of β_j uses the shape of the coverage probability curve as a function of β_j and the fact that when $\gamma = 0$, the distribution of published true treatment effects $\beta_j | D_j = 1$ in the corrected regime

[13] Coverage is shown to increase in this special case in Lemma 1A.6 in Appendix 3A. The proof shows there are two cases to consider, one where the degenerate value for β is relatively 'large' and another where it is relatively 'small'. For large true effect, only effects one and three in Figure 1.2 occur and thus coverage must increase with corrections. For 'small' true effects, the proof shows that increased coverage is equivalent to showing that the hazard function for normal distribution is increasing.

with $r = 1$ first-order stochastically dominates the corresponding distribution in the uncorrected regime with $r < 1$. Finally, note that the proof is not specific to the 5% significance threshold and thus generalizes to other critical thresholds. For more details, see Appendix 3A.

Remark 1 (Improvements in Coverage)**.** *A common concern with publication bias is that published confidence intervals under-cover the true parameter. However, it is also possible that they over-cover the true parameter, even when standard errors are uncorrected and downward biased. In this case, Proposition 2 implies that corrections would increase coverage further, making them, on average, overly conservative. Lemma 1A.9 in Appendix 3A shows that a sufficient condition for undercoverage in the uncorrected regime when nominal coverage is 0.95 is* $r < 0.8512$. *Thus, corrections that are sufficiently large will either decrease the distance to nominal coverage or achieve coverage that is weakly higher than the nominal target. In the empirical application to the DiD literature, the average coverage of published confidence intervals in uncorrected regime is estimated to be far below nominal coverage.*

1.3 Setting and Data

I turn now to analyzing the implications of the theoretical results in a particular setting: the adoption of clustered standard errors in the empirical DiD literature. There are several motivations for the empirical analysis. First, the theoretical results show that the impact of standard error corrections on bias is ambiguous in general and depends on the distribution of latent studies, the degree of selective publication, and the size of the standard error correction. Second, the magnitude of the change in bias (irrespective of the sign) and coverage is an empirical question. A third motivation is that DiD is an extremely popular research design in economics and the quantitative social sciences more broadly, with growing use over time (Currie et al., 2020). Below, I describe the setting and present descriptive statistics. The following section estimates an empirical model and presents the main results.

1.3.1 Data

The empirical analysis uses a newly constructed dataset of DiD articles published in six journals over 2000–2009: the *American Economic Review*, the *Industrial and Labor Relations Review,* the *Journal of Labor Economics*, the *Journal of Political Economy*, the *Journal of Public Economics*, and the *Quarterly Journal of Economics.* These journals were chosen to match those analyzed in Bertrand et al. (2004) for the previous decade, 1990–2000. Following Currie et al. (2020), I identified DiD articles using a string-search algorithm. I collected data on the 'main' DiD estimate in each study, and excluded placebo tests and tests of alternative hypotheses. The 'main' estimate was chosen from the first DiD table in the paper. When there were multiple estimates, I chose the one emphasized in the discussion of the results or the abstract. When there were several specifications, I selected the one with full controls. For DiD articles that fit the inclusion criteria described below, I manually collected data on the estimated DiD treatment effect; the reported standard error; an indicator for whether a correction for serial correlation is implemented; an indicator for policy evaluations[14]; and the number of observations. I also obtained JEL classification codes from *EconLit.*

While the main type of standard error correction in the sample is clustering, a small number of studies implement other corrections e.g. block-bootstrapped standard errors or two-period aggregation. For brevity, I use the term 'clustering' in this article to refer to any correction which accounts for the correlation of errors within groups across time. While the 'correct' level of clustering is an active topic of research (e.g. Abadie et al. (2023)), there is little disagreement over whether standard errors should allow for serial correlation in DiD settings. For descriptive statistics in this section, I simply present the reported standard errors for clustered and unclustered studies. In the empirical model in the following section, I make a stronger assumption that reported clustered standard errors reflect the true standard error.

[14] This denotes studies that evaluate a specific policy (e.g. by a government or firm) and does not refer to studies which simply have policy relevance. For example, consider a study on the causal effect on the peer effects of boys' schooling outcomes on girls', which is estimated by exploiting the impact of an earthquake on compulsory military service for males. While this may have policy relevance, it is not considered here to be a policy evaluation.

To ensure meaningful comparisons of effect sizes across studies, I included studies where the dependent variable is in percent or log units, or otherwise convertable to percent units. For dependent variables in non-percentage units, the effect is recorded relative to the sample mean of the treatment group prior to the treatment.[15] Consider, for example, a study estimating the impact of an educational program on the drop-out rate. I convert the estimated treatment effect into percent units by dividing it by the mean drop-out rate of the treated group before the intervention. When the mean of the treatment group prior to treatment is unavailable, I instead normalize by the mean of the dependent variable for the whole sample. Two studies did not report an average for the dependent variable and were excluded. For effect size conversions, standard errors are rescaled such that the t-ratio is unchanged. I restrict attention to DiD estimates with an indicator for the treatment variable, and exclude, for example, estimated treatment effects based on changing the rate of a continuous treatment variable (e.g. 10 percentage point change in the share of those eligible for medicare).

Figure 1.3 shows a time series of the fraction of DiD articles implementing a correction for serial correlation between 2000 and 2009. This period saw a dramatic rise in the adoption of clustered standard errors, from around one in four at the beginning of the decade to near universal adoption by the end of it. This could in part be due to the publication of Bertrand et al. (2004), which was highly influential and released as a working paper in the early 2000's. Despite earlier emphasis in the econometrics literature on the importance of accounting for correlation in errors within groups (e.g. Moulton (1986)), Bertrand et al. (2004) showed in a survey of DiD studies that the use of corrections in the empirical literature was very rare between 1990 and 2000. Specifically, Bertrand et al. (2004) identified 65 DiD papers with a potential serial correlation problem and found that only five (7.7%) implemented some form of standard error correction.[16]

Table 1.1 presents summary statistics. The sample consists of 96 DiD studies, 66 of which report clustered standard errors. Clustered studies have, on average, larger standard errors

[15] Note that the normalized ATE is a different parameter to the ATE in log differences (Roth and Chen, 2023).

[16] Four of these five studies used GLS for corrections, which they argue is relative ineffective.

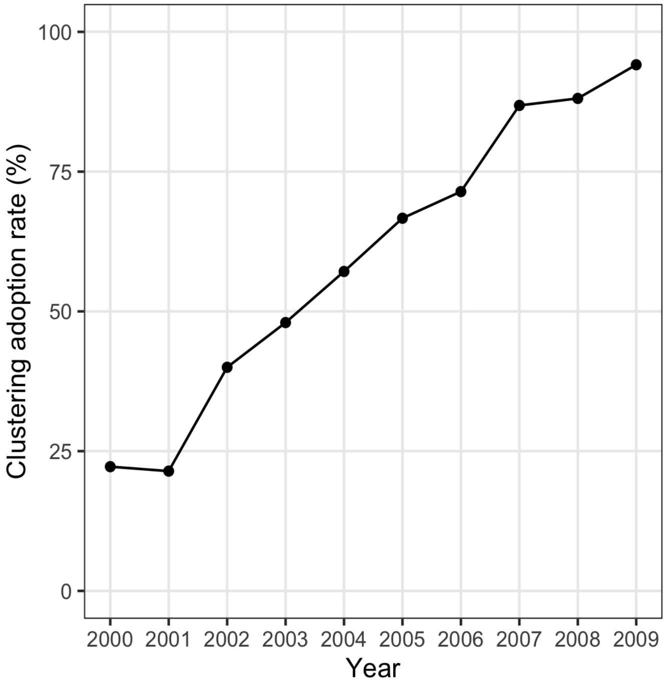

Figure 1.3: Three-Year Centered Moving Average of the Clustering Adoption Rate

than unclustered studies. This is consistent with the econometrics literature that emphasizes downward bias in the absence of corrections (Moulton, 1986, 1990; Bertrand et al., 2004; Abadie et al., 2023). The ratio of the average reported standard errors in unclustered studies to clustered studies is 4.250/6.497 = 0.654 i.e. published clustered standard errors are on average 53% larger than published unclustered standard errors. It is important to note that 0.654 is not an estimate of the degree of downward bias in unclustered standard errors (r), which would be equal to the ratio of unclustered to clustered standard errors in latent studies, not published studies.[17]

Clustered studies are also associated with much larger effect sizes than unclustered studies (19.5% vs. 12.2%). Here, the effect size is defined as the absolute value of the estimated treatment effect. That larger standard errors are accompanied by higher effect sizes is consistent with the main mechanism emphasized in the theory in Section 2.2, namely, that clustering

[17] In fact, this ratio is likely to be an upwardly biased estimate of r. This is because clustering increases reported standard errors which makes publication more difficult. Clustered studies with smaller standard errors are therefore more likely to be statistically significant and published, which would make this ratio larger.

Table 1.1: Summary Statistics: Unclustered and Clustered Studies using Difference-in-Differences

	Unclustered	Clustered	Difference (2)-(1)
Reported standard error (%)	4.253	6.500	2.247
	(4.341)	(6.723)	(1.144)
Effect size (%)	12.182	19.529	7.347
	(14.554)	(18.481)	(3.489)
#JEL codes	3.033	3.333	0.300
	(1.245)	(1.34)	(0.28)
JEL:H (Public)	0.233	0.242	0.009
	(0.430)	(0.432)	(0.095)
JEL:I (Health, Education, & Welfare)	0.433	0.333	-0.100
	(0.504)	(0.475)	(0.109)
JEL:J (Labor and Demographics)	0.667	0.545	-0.121
	(0.479)	(0.502)	(0.107)
JEL:Other	0.533	0.667	0.133
	(0.507)	(0.475)	(0.109)
Policy evaluation	0.867	0.803	-0.064
	(0.346)	(0.401)	(0.080)
log(observations)	9.964	9.849	-0.115
	(2.111)	(2.073)	(0.461)
Number of studies	30	66	36

Notes: The sample is DiD literature over 2000-2009 based on inclusion criteria described in the main text. The first two columns report means and standard deviations below in parentheses. In the final column, robust standard errors are reported from a regression of the row variable on an indicator for clustering. JEL codes H, I and J are presented because they are the most commonly listed codes. JEL:H is an indicator which equals one if at least one of the JEL codes is H; JEL:I and JEL:J are defined similarly. The variable JEL:Other equals one if the study lists at least one code that is not H, I or J.

raises the bar for statistical significance and results in the selection of larger effect sizes due to publication bias. More detailed descriptive statistics consistent with this interpretation are presented further below.

The remaining rows of Table 1.1 show summary statistics on study characteristics. The number of primary JEL categories is around around three for both clustered and unclustered studies.[18] The most common categories are H (Public Economics), I (Health, Education, and Welfare), and J (Labor and Demographic Economics). While a high share of both unclustered

[18] There are 26 primary JEL categories (A to Z) corresponding to different fields of economic research. For the full distribution of JEL codes in unclustered and clustered studies, see Appendix 1D.

and clustered studies belong to these categories, clustered studies are somewhat less likely to report categories I and J. Similarly, while the majority of all studies are policy evaluations, the fraction for clustered studies (0.80) is somewhat lower than in unclustered studies (0.87). These statistics are consistent with DiD research designs being used in a wider variety of settings over time.

1.3.2 Two Stylized Facts

In this subsection, I present descriptive statistics on two stylized facts:

1. Clustering was associated with the magnitude of published estimates almost doubling in size after controlling for differences in research topics, sample size, and including year and journal fixed effects; and

2. There is strong evidence of publication bias favoring statistically significant results.

Effect Size Gap

As shown in Table 1.1, there is a large difference in the magnitude of estimated treatment effects between unclustered and clustered studies. Differences in observable study characteristics cannot explain this gap. Table 1.2 reports results from a regression of the effect size on an indicator for clustering, adding additional controls with each successive column. The final specification includes year and journal fixed effects and controls for sample size, research topic (JEL categories), and an indicator for policy evaluations. The estimated coefficient in the specification with full controls implies that effect sizes in clustered studies are larger than those in unclustered studies by a factor of 1.84 (22.36% vs. 12.18%).

This is a striking gap and consistent with a substantial shift in the distribution of published studies. However, it is important to emphasize that the theoretical results in Subsection 1.2.2 show that observing larger estimated treatment effects in clustered studies does not, in and of itself, tell us whether bias has actually increased. The example presented there shows that

Table 1.2: Impact of Clustering on Effect Sizes

	(1)	(2)	(3)	(4)
Clustered	7.347	8.265	9.464	10.182
	(3.489)	(3.977)	(4.315)	(4.778)
Unclustered mean	12.18	12.18	12.18	12.18
Observations	96	96	96	96
Adjusted-R^2	0.028	0.067	0.056	0.053
Year FE		X	X	X
Journal FE			X	X
Study controls				X

Notes: OLS regressions of estimated treatment effects on an indicator for clustering. The dependent variable is in percent units (or log points for studies where the dependent variable in in logs). The estimated coefficient on the clustering indicator is in percentage point units. Study controls include a quadratic on the log of the number of observations, an indicator for policy evaluations, and a three-way interaction between the three most common JEL primary categories: H (Public Economics), I (Health, Education, and Welfare), and J (Labor and Demographic Economics). Robust standard errors are in parentheses.

higher effect sizes can also be consistent with a decrease in bias.[19] To make inferences about changes in bias, it is therefore necessary to estimate the latent distribution of studies, which we do in the following section.

An alternative explanation for the observed gap is that it is driven by strategic clustering. This is a particular form of endogeneity where researchers p-hack their standard errors to increase the chances of publication. In particular, suppose that researchers strategically choose not to cluster if doing so would overturn a statistically significant result. This behavior would also generate a positive correlation between clustering and estimated treatment effects. Thus, the effect size gap in Table 1.2 might reflect the impact of clustering on estimated treatment effect via selective publication process *and* strategic clustering by researchers.

To test whether strategic clustering is driving this result, I examine effect sizes of unclus-

[19] Strictly speaking, the example shows that the *unnormalizaed* difference in effect sizes, $\mathbb{E}[\hat{\beta}_j|D_j = 1, C_j = 1] - \mathbb{E}[\hat{\beta}_j|D_j = 1, C_j = 0]$, is positive. However, it is also true in this example that the difference in the magnitude of estimated treatment effects, $\mathbb{E}[|\hat{\beta}_j||D_j = 1, C_j = 1] - \mathbb{E}[|\hat{\beta}_j||D_j = 1, C_j = 0]$ is positive. This section focuses on absolute effect sizes because we do not in fact observe unnormalized effect sizes $\hat{\beta}_j$ conditional on our normalization that β_j is positive (Assumption 1). For a concrete example, consider a study with an observed estimate $\hat{\beta}_j$, and an unobserved true effect β_j, which could be positive or negative. Now normalize the true effect to be positive $|\beta_j|$. Whether or not we switch the sign of $\hat{\beta}_j$ to be consistent with this normalization requires knowledge of the sign of unnormalized β_j, which we do not observe.

tered studies in the 1990–1999 period from the same set of journals. During this period, the overwhelming majority of studies reported unclustered standard errors (Bertrand et al., 2004) and hence strategic clustering is unlikely to be affecting the distribution of effect sizes. If strategic clustering was absent in the 1990–1999 period, but present during the 2000–2009 period, then, all else equal, we might expect effect sizes to be smaller in the 2000–2009 period. This is because strategic clustering would increase the fraction of published studies in the unclustered regime with relatively small effect sizes that would be 'just significant' without clustering, but insignificant with it. Instead, I find that the mean effect size in the 2000–2009 period is close to, and in fact slightly larger than, the mean effect size in the 1990–1999 period (12.18% and 10.57%). The difference is statistically indistinguishable from zero, although statistical power is somewhat limited. Controlling for differences in observable study characteristics, including JEL topics and sample sizes, does not change this conclusion. This supports the idea that strategic clustering of the simple form discussed here is not driving observed differences in effect sizes across clustered and unclustered regimes. This, of course, covers only one form of endogeneity and other forms could in principle be present. For more details, see Appendix 1E.

Ultimately, the primary goal of the empirical analysis is to estimate the changes in bias and coverage that occur due to clustering, not simply changes in effect sizes. To this end, in the following section, I propose an estimation approach for the empirical model that yields unbiased estimates of the model parameters irrespective of whether or not there is strategic clustering of the simple form described here. Moreover, this provides an additional test for strategic clustering, by comparing robust model estimates to those in the baseline model. Using this approach, we cannot reject the null hypothesis of no strategic clustering. See Subsection 1.4.1 for further discussion.

Selective Publication on Statistical Significance

The second stylized fact concerns evidence for publication bias favouring statistically significant results. While publication bias has been documented in a wide variety of settings, it is important to test for it in the DiD setting, for two reasons. First, to establish the applicability of the theoretical results; and second, to justify estimating the selective publication model in the following section. I explore two common approaches used in the meta-science literature for detecting selective publication.

The first is the metaregression approach proposed in Card and Krueger (1995). Figure 1.4 visualizes a regression of effect sizes on reported standard errors. Panels (a) and (b) separate articles using clustered and unclustered standard errors, respectively. The results are consistent with selective publication on the basis of statistical significance, for at least three reasons. First, there are simply very few studies with statistically insignificant results. Second, larger standard errors are associated with larger effect sizes. Metaregression estimates in both regimes give a slope coefficient which implies that a one percentage point increase in standard errors is associated with a little over a two percentage point increase in estimated effect sizes – this is, approximately the increment necessary for maintaining statistical significance. In the absence of selective publication, there may be little reason to expect a systematic relationship between estimated treatment effects and standard errors, because the sample size in observational studies is not typically chosen but instead predetermined by available datasets.[20] Finally, the estimated slope coefficient on reported standard errors is very similar across clustered and unclustered regimes. Given that unclustered standard errors are systematically downward biased, one would expect, under the null hypothesis of no selective publication, that clustering would lead to a decrease in the slope coefficient on standard errors. Instead, the estimated linear relationship between treatment effects and reported standard errors is similar across regimes.

Following Brodeur et al. (2016), a second test examines the distribution of t-statistics to determine if there is a bunching around critical significance thresholds. Panel (c) shows the distribution of test statistics for unclustered studies, while Panel (d) shows the same for clustered studies. The vertical dashed line marks the 5% threshold significance level. In both figures, there is a large mass of t ratio values just above this threshold, and a 'missing' mass just below

[20] This contrasts with experimental studies where larger sample sizes may be chosen by authors performing power calculations to detect small expected effect sizes.

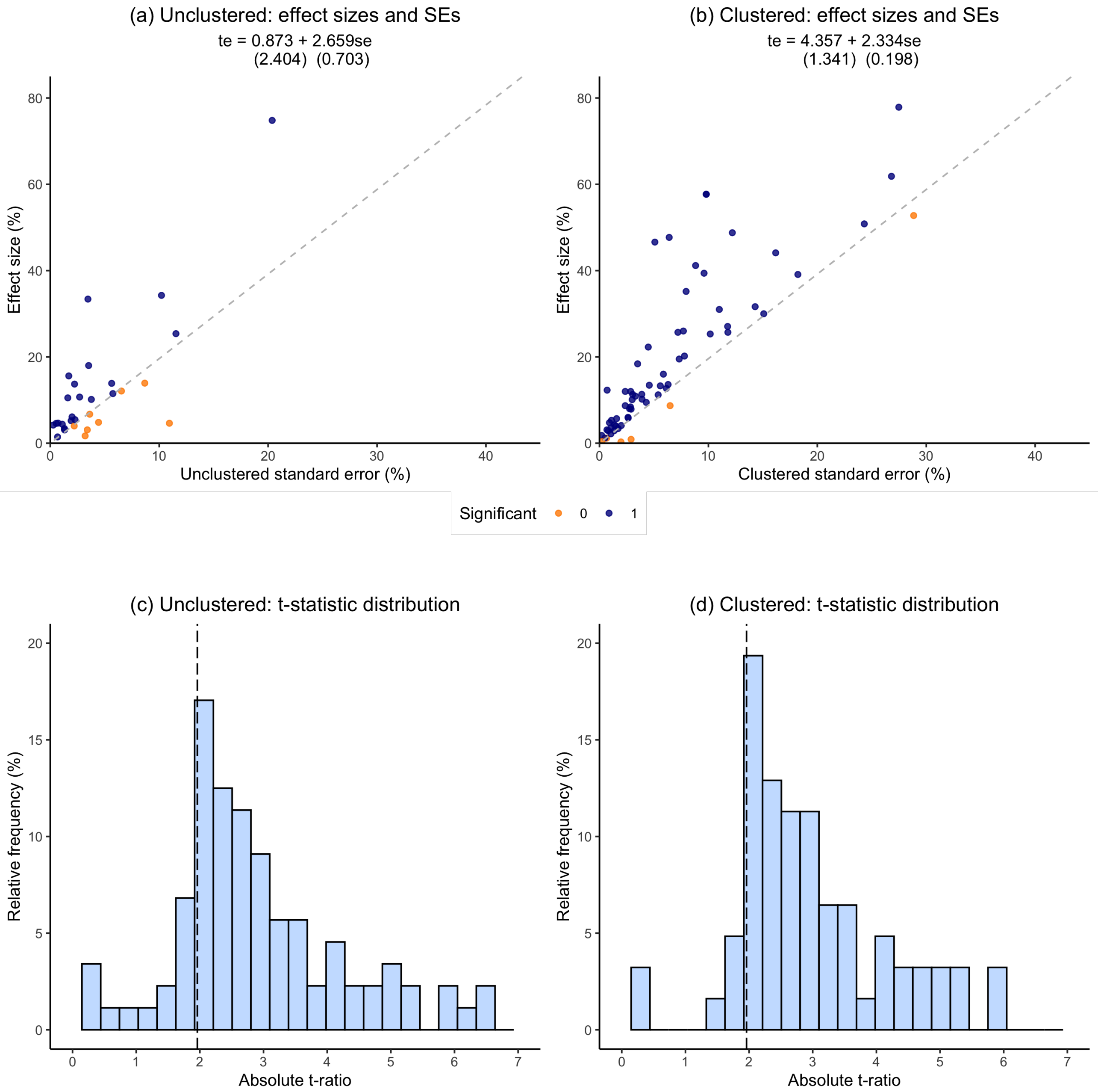

Figure 1.4: Selective Publication and p-Hacking

Notes: These figures present evidence of selective publication and p-hacking in the empirical DiD literature over 2000–2009. Panels (a) and (b) report OLS regressions of estimated treatment effects on standard errors in the unclustered and clustered regime. The dashed line separates statistically significant and insignificant results at the 5% level. Robust standard errors are reported in parentheses. Panels (c) and (d) show the distribution of absolute t-statistics for both regimes; the vertical dashed line is at 1.96, the critical threshold for statistical significance at the 5% level.

it. Despite the fact that standard errors are systematically higher in clustered studies, the distributions appear very similar in both regimes, providing additional evidence of selective publication (or p-hacking).

1.4 Empirical Model

Descriptive statistics provide evidence that clustering led to a change in the distribution of estimated treatment effects via selective publication. However, from these descriptives alone, we cannot make inferences about some of the main quantities of interest, namely, bias and coverage. To do this, I follow an empirical strategy consisting of two steps. In the first, I estimate the model in Section 2.2 using data from clustered DiD studies. This gives parameters governing the latent distribution ($\mu_{\beta,\sigma}$) and selective publication (γ) for clustered studies. With these model estimates, we can analyze counterfactual scenarios of what would have happened had clustered studies instead reported unclustered standard errors which were downward biased by any specified factor r. In the second step, I describe two approaches for calibrating reasonable values for r. I then present the main results.

1.4.1 Estimation

First, I estimate the model of selective publication in Section 2.2 using data from clustered studies. Following Andrews and Kasy (2019), I estimate the latent distribution of true effects assuming that $\beta_j \perp\!\!\!\perp \sigma_j$ (Assumption 2) and $\beta_j | \lambda_\beta, \kappa_\beta \sim \text{Gamma}(\lambda_\beta, \kappa_\beta)$. Following Vu (2023), I augment the baseline model to jointly estimate the distribution of standard errors, assuming this also follows a gamma distribution: $\sigma_j | \lambda_\sigma, \kappa_\sigma \sim \text{Gamma}(\lambda_\sigma, \kappa_\sigma)$. This is necessary for calculating coverage. In line with the theory, I assume publication probabilities follow a step function where the relative probability of publishing a statistically insignificant result at the 5% level is given by γ.[21] Finally, note that clustered standard errors are assumed in estimation

[21] This is similar to Assumption 3 in that selective publication follows a step function at the 5% level. It differs, however, in that it does not impose that $\gamma \in [0, 1)$. In particular, estimation allows the possibility that

to reflect the true variation of estimated treatment effects.

Consistency of the model parameters requires that $C_j \perp\!\!\!\perp \hat{\beta}_j | \beta_j$. This assumption is violated if there is strategic clustering, which I address below in an alternative estimation approach. The assumption is not violated, however, by non-random clustering with respect to study characteristics. For example, there is suggestive evidence in Table 1.1 that DiD studies outside of Health, Education & Welfare (JEL:I) and Labor & Demographics (JEL:J) are more likely to use clustered standard errors. If this were indeed the case, then estimation would still yield consistent estimates of the latent distribution of studies in the clustered regime; however, the latent distribution in the unclustered regime would differ. This has implications for intepreting the main results, which I discuss further below. Finally, note that I restrict attention to clustered studies to avoid imposing strong assumptions about the mapping between unclustered standard errors and (unobserved) clustered standard errors for unclustered studies in the likelihood function.[22]

Table 1.3 presents the maximum likelihood estimates. The estimate $\hat{\gamma} = 0.016$ implies a high degree of selective publication. In particular, it means that statistically significant results are around 60 times more likely to be published than insignificant results. This is broadly similar to estimates of publication bias in Andrews and Kasy (2019) for replication studies in economics ($\hat{\gamma} = 0.038$) and psychology ($\hat{\gamma} = 0.017$).

As mentioned above, the presence of strategic clustering would lead to model misspecification and inconsistent parameter estimates. To address this potential issue, I propose an alternative estimation approach which is robust to the a scenario where researchers choose to cluster if and only if it does not change the significance of their results. For a formal presentation of this augmented model, see Appendix 1F. The main idea in this alternative approach is to estimate

$\gamma \geq 1$ such that the relative probability of publishing insignificant results is the same as, or higher than, for significant results. Note that publication probabilities are only identified up to scale.

[22] This is because publication is based on unclustered standard errors while the true variation of the estimated treatment effect is based on the unobserved clustered standard error. Although we later impose an assumption about this mapping to estimate what would have happened if standard errors were unclustered, conducting estimation without this restrictive assumption means that the consistency of the parameters estimates does not rely on it being correctly specified.

Table 1.3: Maximum Likelihood Estimates

Latent true effects β_j		Latent standard errors σ_j		Selection
κ_β	λ_β	κ_σ	λ_σ	γ
0.154	17.802	1.426	6.475	0.016
(0.035)	(2.692)	(0.167)	(1.282)	(0.007)

Notes: Estimation sample is clustered DiD studies over 2000–2009 ($N = 66$). Robust standard errors are in parentheses. Latent true treatment effects and standard errors are assumed to follow a gamma distribution with shape and scale parameters (κ, λ). The coefficient γ measures the publication probability of insignificant results at the 5% level relative to significant results. For example, $\gamma = 0.016$ implies that significant results are 62.5 times more likely to be published than insignificant results.

the parameters governing the latent distribution of studies on the selected subset of *statistically significant* clustered studies; this entails setting $\gamma = 0$ and not estimating it. The rationale is that the distribution of significant, clustered studies, $\hat{\beta}_j, \sigma_j | D_j = 1, C_j = 1, |\hat{\beta}_j|/\sigma_j \geq 1.96$, is completely invariant to this form of strategic clustering. This is because strategic clustering only affects studies whose results are insignificant when clustered but significant when unclustered. However, none of these studies are included in the subsample of statistically significant clustered studies. Thus, the distribution of studies, and hence the likelihood, is unaffected by whether or not strategic clustering is present. For a formal statement and proof of this claim, see Lemma 1F.1 in Appendix 1F. Robust estimates for the latent distribution of studies are presented in Table 1F.1 and statistically indistinguishable from the baseline estimates in Table 1.3. This suggests that strategic clustering of the form discussed here does not bias baseline parameter estimates.[23] Given these results, I focus on the model estimates in Table 1.3.

1.4.2 Unclustered Counterfactuals

With the model estimates in Table 1.3, we can calculate expected bias, coverage, true treatment effects and estimated treatment effects under the counterfactual scenario where clustered studies report unclustered standard errors that are downward biased by any specified factor

[23] Given similar parameter estimates, the results for bias and coverage using the robust approach are very similar to those presented in the main text. For more details, see Appendix 1F.

$r \in (0, 1)$. We can then compare these statistics across unclustered and clustered regimes. The interpretation of this counterfactual comparison is analogous to an ATT measure of the impact of clustering on the statistical properties of published, clustered studies. If the latent distribution of studies differs across clustered and unclustered regimes, then this ATT measure might differ from an ATE measure which would be the impact of clustering on both unclustered and clustered studies.

This ATT measure can be computed for any specified value of $r \in (0, 1)$ using only the model estimates in Table 1.3. Figure 1G.1 in Appendix 1G shows the results as a function of r over the unit interval. This can be connected directly to Proposition 1, which states that bias must increase for sufficiently large standard error corrections i.e. for any r less than some model-dependent value r^*. Based on the estimates in the DiD setting, I find that $r^* = 0.95$. This implies that any corrections enlarging standard errors by 5.3% or more would lead to an increase in bias in published DiD studies. Since Proposition 2 guarantees increased coverage, it follows that the *qualitative* conclusion of higher coverage but increased bias will exist for all but very small standard error corrections. The *quantitative* results, however, will depend on r, with larger corrections leading to larger changes in both bias and coverage.

1.4.3 Calibrating r

This subsection considers alternative approaches for calibrating r. As a starting point, note that the first-best approach would be to obtain the empirical distribution for r by calculating the ratio of unclustered to clustered standard errors from all studies in the estimation sample of clustered studies. Unfortunately, this is not possible because code and data availability policies were uncommon in the 2000's. Instead, I use two alternative approaches. I focus on the first in the main text and show that the second provides very similar results in Appendix 1H.

In the first approach, I make the simplifying assumption that all unclustered standard errors are downward biased by a constant factor $r \in (0, 1)$. I then calibrate r using the method of simulated moments (McFadden, 1989). Specifically, I select the value of r which minimizes the distance between moments predicted by the model and the actual moments observed in the data. Given that r measures the degree of downward bias in unclustered standard errors, the moment I choose for calibration is the difference in average reported standard errors between clustered and unclustered studies in the published literature. Carrying out this procedure gives $\hat{r} = 0.51$. In other words, clustered standard errors are estimated to be around twice the size of unclustered standard errors.[24] This is a large adjustment. This calibration approach assumes that the distribution of latent studies in clustered studies is the same as in unclustered studies. This would be violated, for example, if there are differences in the datasets which tend to be used in *latent* unclustered and clustered studies, since this would imply differences in the latent distribution of standard errors. Nevertheless, if the assumption is violated, then we still obtain a valid counterfactual for what would have occurred if clustered studies had instead been unclustered and were around half the size of true standard errors.

To address some of the concerns of this first method, I propose an alternative approach which calculates the empirical distribution of r using a sample of DiD studies between 2015–2018. Over this period, code and data availability policies were more common than in the 2000–2009 period. The benefit of this approach is that it does not require the assumption the latent distribution of studies is identical across regimes. Moreover, it is immune to concerns over strategic clustering because unclustered and clustered standard errors are calculated for each individual study. Its main drawback relative to the first approach is external validity, since it is based on data from a later time period.

I consider DiD papers published between 2015–2018 as identified in Brodeur et al. (2020). I collected data on standard errors from six of the 25 journals sampled in that study.[25] While

[24] Lee et al. (2022) propose a standard error adjustment for the single-IV model and apply it to recently published AER papers. In this setting, they find that corrected standard errors are at least 49 percent larger (i.e. $r \leq 0.672$) than conventional 2SLS standard errors at the 5% level.

[25] The journal are *Applied Economic Journal: Applied Economics*, *Applied Economic Journal: Economic Policy*, *American Economic Review*, *Journal of Labor Economics*, *Journal of Political Economy* and the *Quarterly Journal of Economics*. Four overlap with journals from the main analysis. The two excluded journals are the *Industrial and Labor Relations Review*, which is not in the Brodeur et al. (2020) sample; and the *Journal of Public Economics*, which did not require authors to submit data and code over the 2015–2018 period. I included data from *Applied Economic Journal: Applied Economics* and *Applied Economic Journal: Economic Policy* due to a small sample size based on the four overlapping journals alone. The two additional journals were chosen

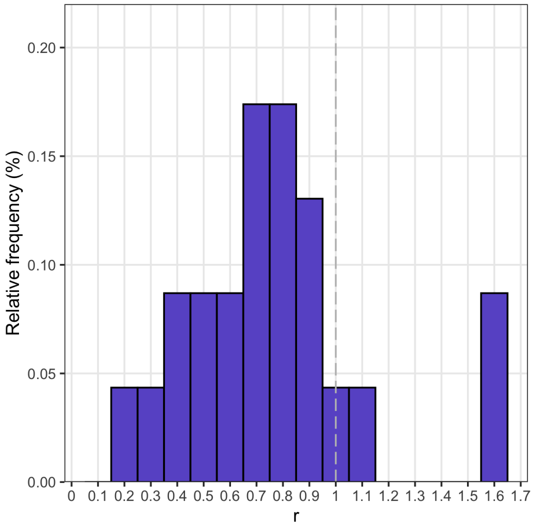

Figure 1.5: Empirical Distribution of r from 2015–2018 DiD Studies

Notes: Calculated from original code, where r equals the ratio of unclustered to clustered standard errors. The sample consists of a subset of DiD studies identified in Brodeur et al. (2022). For more details on sample selection, see the main text.

code is available for almost all studies, not all use publicly available data. Overall, I calculate r in 23 out of 72 DiD studies (31.9%) using non-proprietary data. Figure 1.5 shows the empirical distribution. The mean is 0.76 and a small fraction of studies have clustered standard errors which are *larger* than unclustered standard errors ($r > 1$). For calculating the counterfactual scenario for unclustered studies, we can draw randomly from this distribution to determine the degree bias for each study individually. This is useful because in reality, r varies across studies and depends on the within-cluster correlation of the regressor, the within-cluster correlation of the error, and the number of observations in each cluster (Cameron and Miller, 2015). As mentioned above, both approaches lead to quantitatively similar conclusions. In the main text, I focus on the first approach using the method of simulated moments to calibrate r.

because they: (i) published a high share of DiD studies over this period; and (ii) required replication materials for publication.

1.4.4 Impact of Clustering on Coverage and Bias

Table 1.4 presents the main results. The estimated model shows that clustering increased coverage dramatically, from only 0.28 in the unclustered regime to 0.70 in the clustered regime. This implies severe mismeasurement of standard errors prior to the adoption of clustering, with fewer than one in three published studies reporting confidence intervals covering the true effect. Note that while coverage improves substantially, it still remains, at 0.70, below nominal coverage of 0.95 due to selective publication.

The remaining rows in Table 1.4 show the impact of clustering on various measures of bias. Recall that the change in total bias can be decomposed into the change in internal-validity bias and study-selection bias (equation (1.2)). In this context, the primary measure of interest is internal-validity bias. This is because different studies in the empirical DiD literature address different research questions, and the main concern is therefore each study's internal validity. The model shows that clustering led to internal-validity bias doubling in magnitude, from 1.23 ppts to 2.44 ppts. To gauge the size of this change, we can ask what fraction of insignificant results (with correctly measured standard errors) would need to be censored by publication bias to observe the same increase bias (1.21 ppts)? I find that 85% of null results would need to be censored (i.e. $\gamma = 0.15$). In other words, the increase in internal-validity bias from clustering is comparable to very severe levels of publication bias against null results. Next, see that clustering leads to a large increase in study-selection bias, as studies with larger true treatment effects are more likely to produce statistically significant results and therefore be selected for publication. As mentioned earlier, changes in study-selection bias do not have clear implications for statistical credibility in the DiD context, since different studies address different research questions. Increases in study-selection bias and internal-validity bias mean that total bias rises by 6.48 ppts overall.

Robustness results based on the empirical distribution for r are presented in Table 1H.1 in Appendix 1H. In this alternative approach, the degree of bias of unclustered studies is drawn randomly from the distribution of r (Figure 1.5), such that r varies across unclustered studies.

Table 1.4: Impact of Clustering on Coverage and Bias in Published Studies

	Unclustered ($\hat{r} = 0.51$)	Clustered ($r = 1$)	Change
Coverage	0.28	0.70	0.42
Total Bias ($\mathbb{E}_r[\hat{\beta}_j \mid D_j = 1] - \mathbb{E}_r[\beta_j]$)	3.51 (100%)	10.00 (100%)	6.48 (100%)
Internal-Validity Bias ($\mathbb{E}_r[\hat{\beta}_j - \beta_j \mid D_j = 1]$)	1.23 (34.9%)	2.44 (24.4%)	1.21 (18.7%)
Study-Selection Bias ($\mathbb{E}_r[\beta_j \mid D_j = 1] - \mathbb{E}_r[\beta_j]$)	2.29 (65.1%)	7.56 (75.6%)	5.27 (81.3%)

Notes: These figures are based on the parameter estimates of the empirical model in Table 1.3. Figures are calculated by simulating published studies under unclustered and clustered regimes and assuming that unclustered standard errors are downward biased by a constant factor $\hat{r} = 0.51$.

Results are quantitative similar to those in Table 1.4. In particular, clustering improves coverage from 0.36 to 0.70 and internal-validity bias increases by 1.07 ppts. Alternatively, assuming that unclustered studies are downward biased by a constant factor equal to the mean of the empirical distribution ($\hat{r} = 0.76$) yields qualitatively similar results, but somewhat smaller changes in both coverage and bias. For more details, see Appendix 1H.

Overall, the results underscore the tension from clustering which has been emphasized throughout this paper, namely, that improved credibility of standard errors can come at the unintended cost of declining credibility in point estimates. Quantifying this in the empirical DiD literature shows that both the gains and costs are large.

1.4.5 Non-Selective Publication

A common recommendation to combat distortions arising from publication bias is to implement reforms to publish all results, irrespective of their statistical significance. For example, implementing results-blind peer review (Chambers, 2013; Foster et al., 2019), launching journals dedicated to publishing insignificant findings[26], and even offering cash incentives for publishing null findings (Nature 2020).

To analyze the impact of these reforms in the DiD literature, I perform a counterfactual

[26] Examples include: *Positively Negative (PLOS One); Journal of Negative Results in Biomedicine; Journal of Articles in Support of the Null Hypothesis; Journal of Negative Results - Ecology and Evolutionary Biology.*

analysis where there is no selective publication. In other words, I perform the same empirical exercise as for the main results, but set $\gamma = 1$ such that no insignificant studies are censored. When publication is non-selective, there exists no trade-off between coverage and bias when clustering. Coverage increases from 0.68 to reach nominal coverage of 0.95, and all forms of bias are zero in both standard error regimes. The welfare implications, however, are not clear. In particular, publishing all results is not necessarily without drawbacks. This is because non-selective publication leads to many published studies with small true treatment effects that are very imprecisely measured, and hence relatively uninformative for decision-makers who rely on empirical evidence from published studies to make policy choices. As noted in Frankel and Kasy (2022), if publication comes at a cost (e.g. the opportunity cost of drawing attention away from other studies due to limited journal space), then it is not necessarily the case that the non-selective regime is preferable to the selective regime. To better understand the impact of clustering on welfare, I develop a treatment choice model in the next section to evaluate the impact of clustering on decision-making in a policy context.

1.5 Impact of Clustering on Evidence-Based Policy

The empirical model in Section 1.4 suggests that clustering led to large improvements in coverage but also substantially higher bias. What are the implications of this for evidence-based policy? In this section, I develop a model of a policymaker who chooses whether to implement a policy based on evidence from published studies, but who overestimates the precision of estimates when standard errors are unclustered. I consider a policymaker who aims to minimize maximum regret i.e. the expected welfare loss from making an inferior treatment choice. I derive the minimax decision rule in the clustered and unclustered regimes, and then compare minimax regret across regimes. The main finding is that clustering lowers minimax regret if and only if the policymaker has sufficiently high loss aversion with respect to mistakenly implementing an ineffective or harmful policy i.e. of committing Type I error. Overall, the results suggest that clustering is beneficial if the cost of Type I error is specified in a way that is consistent with hypothesis testing using a 5% significance threshold.

1.5.1 Setup

The basic setup is the same as in Kitagawa and Vu (2023), which extends the model of minimax regret decision-makers in Manski (2004) and Tetenov (2012) to include publication bias. The model presented here makes a further extension to include the possibility that reported standard errors are mismeasured (e.g. from failing to cluster).

The policymaker's problem is to decide whether they should implement a single policy ($a = 1$) or not implement it ($a = 0$).[27] The policy's *unobserved* average treatment effect is denoted by β. All members of the population are assumed to be observationally identical. We normalize utility to be zero when no policy is implemented. Following Tetenov (2012), I consider a policymaker whose utility function may exhibit loss aversion (Kahneman and Tversky, 1979) for implementing a harmful policy ($\beta \leq 0$). The policymaker's utility from an action a with average treatment effect β is given by

$$U(a, \beta | K) = \begin{cases} Ka\beta & \text{if } \beta \leq 0 \\ a\beta & \text{if } \beta > 0 \end{cases} \tag{1.3}$$

where $K \geq 1$ measures the policymaker's loss aversion. As K increases, the policymaker weighs the utility cost of committing Type I error (implementing the policy when $\beta \leq 0$) increasingly high relative to Type II error (not implementing the policy when $\beta > 0$). As a benchmark, note that classical hypothesis testing is consistent with a high degree of loss aversion from Type I error. In particular, regret from committing Type I error would need to be weighed around

[27] A more general formulation of the policymaker's problem is to assign some portion $a \in [0, 1]$ of observationally identical members of a population either a *status quo treatment* or an *innovative treatment*. Assuming $a \in \{0, 1\}$ does not affect the results. This is because in continuous action case for the model in Tetenov (2012), on which this model is based, the policymaker's decision rule for an observational identical population will either treat all or none of the members. For expositional simplicity, I consider the status quo treatment to be not implementing the policy and the innovative treatment to be implementing it.

100 times more than Type II regret for a decision rule that minimizes maximum regret to be consistent hypothesis testing with a 5% statistical significance threshold (Tetenov, 2012).

A study is conducted which provides evidence about true average treatment effect β. However, due to publication bias, it may not be observed by the policymaker. The policymaker's *statistical treatment rule* maps realizations of the publication process to policy decisions. There are two possibilities. First, the case where a study is published and the policymaker uses the evidence contained in it to inform their policy choice. Second, the case where no study is published and the policymaker must rely on a default action.

Let $D = 1$ denote the event when a study is published and $D = 0$ the event where it is not. Consider first the case where $D = 1$. When the study is published, the policymaker observes $(\hat{\beta}, \widetilde{\sigma})$, that is, the estimated treatment effect $\hat{\beta}$ and the *reported* standard error $\widetilde{\sigma}$. If standard errors are clustered, then $\widetilde{\sigma} = \sigma$. If they are unclustered, then $\widetilde{\sigma} = r \cdot \sigma < \sigma$ since $r \in (0, 1)$.

Importantly, the policymaker's statistical decision rule is chosen based on their beliefs about how a study's results, $(\hat{\beta}, \widetilde{\sigma})$, were generated. In the main analysis, I consider a naive policymaker who believes $\hat{\beta}$ is normally distributed on $\mathcal{B} = \mathbb{R}$ according to $N(\beta, \widetilde{\sigma}^2)$, since approximate normality is widely assumed in practice for inference, including in all the DiD papers I examine. This belief can be incorrect on two counts. First, if there is publication bias, then $\hat{\beta}$ is not normally distributed but follows a truncated normal distribution. Thus, in practical terms, the model assumes that policymakers naively take estimates from the published literature at face-value and do not make statistical adjustments to correct for publication bias. Second, beliefs will be wrong about the variance of the estimate $\widetilde{\sigma}^2$ in the case where standard errors are unclustered. In other words, policymakers take reported standard errors in published studies to be accurate measures of the estimate's uncertainty, irrespective of whether they are clustered or not.

We turn next to see how these beliefs affect the policymaker's decision rule. Let $\delta_1 : \mathcal{B} \to [0, 1]$ be the statistical decision rule in the event that a study is published, which maps observed estimates to the probability of implementation. Following Tetenov (2012), it is sufficient to restrict our attention to smaller class of threshold decision rules where a policy is implemented if and only if the published estimate $\hat{\beta}$ is above some chosen threshold T i.e. $\delta_1^T(\hat{\beta}) = \mathbb{1}\{\hat{\beta} > T\}$.[28] Thus the expected welfare of the threshold rule δ_1^T under the misspecified belief that $\hat{\beta}$ is normal and the observed, but potentially mismeasured, standard error $\widetilde{\sigma}$, is equal to

$$\widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) = \begin{cases} K\beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta \leq 0 \\ \beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta > 0 \end{cases} \tag{1.4}$$

To derive a decision rule, it is first necessary to adopt a framework for dealing with the uncertainty of β. Two common approaches are the Bayesian framework and minimax regret framework. For example, in the Bayesian approach, the policymaker sets a prior belief distribution π over the average treatment effect β and chooses a threshold T to maximize (misspecified) expected welfare: $\int \widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma}\big)\pi(\beta)d\beta$.

However, in many situations, policymakers may have insufficient information to form a reasonable prior or priors may conflict when decisions are made by members of a group. In this situation, a common alternative is to introduce ambiguity on the treatment outcomes and pursue robust decisions. Specifically, I consider a policymaker that aims to minimize maximum regret (Manski, 2004; Stoye, 2009; Tetenov, 2012), where regret for a threshold rule δ_1^T equals the difference between the highest possible expected welfare outcome given full knowledge of the true impact of all treatments and the expected welfare attained by the statistical decision rule:

[28] This is because the policymaker believes X to follow a normal distribution, which satisfies the monotone likelihood ratio property. It follows from Karlin and Rubin (1956a) that the class of *threshold decision rules* is essentially complete and consideration of other rules is not necessary.

$$\begin{aligned}\widetilde{R}_1\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) &= W\big(\mathbb{1}\{\beta > 0\}\big) - \widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) \\ &= \begin{cases} -K\beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta \leq 0 \\ \beta\Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big) & \text{if } \beta > 0 \end{cases}\end{aligned} \tag{1.5}$$

In words, regret is equal to the probability of making a mistake multiplied by the magnitude of that mistake $|\beta|$ (and weighted according to K). Thus, the policymaker chooses their minimax regret threshold decision rule based on misspecifed beliefs to minimize regret in the worst-case scenario:

$$T^* = \underset{T \in \mathbb{R}}{\arg\min} \max_{\beta \in \beta} \widetilde{R}_1\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) \tag{1.6}$$

Next, consider the event where no study is published. The no-data decision rule is denoted by $\delta_0 \in [0, 1]$, which denotes the probability of implementing the policy when no evidence is available. Using a similar derivation as above, we arrive at the following expression for regret

$$\widetilde{R}_0\big(\delta_0, \beta|K\big) = \begin{cases} -K\beta\delta_0 & \text{if } \beta \leq 0 \\ \beta(1 - \delta_0) & \text{if } \beta > 0 \end{cases} \tag{1.7}$$

Note that this expression is also misspecified, in that the policymaker makes no inferences about the fact that a study might have been censored. Similar to the event where a study is published, the no-data decision rule is obtained by the following optimization

$$\delta_0^* = \underset{\delta_0 \in [0,1]}{\arg\min} \max_{\beta \in \beta} \widetilde{R}_0\big(\delta_0, \beta|K\big) \tag{1.8}$$

For the no-data decision problem to be well-defined, we impose the following bounds on the support of β:

Assumption 4 (Symmetric Bounds on Average Treatment Effect)**.** *Let the support of β be $[-B, B]$ for some $B > \beta^* > 0$, where $\beta^* = \arg\max_{\beta > 0} \big\{\beta \cdot \Phi(0 - \beta)\big\}$.*

The technical condition requiring that the bound be sufficiently large ensures that the minimax regret problem in the event that a study is published is not constrained by the bound.

Overall, the policymaker's minimax decision rule (T^*, δ_0^*) covers both realizations of the publication process and is chosen according to (1.6) and (1.8).

1.5.2 Minimax Regret Decision Rule

The follow result gives the minimax decision rule under misspecified regret, covering both the clustered regime ($\widetilde{\sigma} = \sigma$) and unclustered regime ($\widetilde{\sigma} < \sigma$):

Lemma 1 (Minimax Regret Decision Rule)**.** *Under Assumptions 3 and 4, the minimax regret decision rule for a publication-bias naive policymaker given reported standard error $\widetilde{\sigma}$ and Type I error loss aversion parameter K is given by*

$$(T^*, \delta_0^*) = \left(g(K) \cdot \widetilde{\sigma},\ \frac{1}{1+K} \right) \tag{1.9}$$

where $g(K)$ is a strictly increasing function of K and $g(1) = 0$

Figure 1.6 illustrates Lemma 3.3.1 calibrating to the level of publication bias ($\hat{\gamma} = 0.016$) and downward bias in standard errors ($\hat{r} = 0.51$) in the empirical DiD literature. In the first panel, observe that the threshold rule in both regimes is increasing in the Type I error loss aversion parameter K, but that in the unclustered regime it is strictly below the clustered regime's threshold rule when $K > 1$.[29] For intuition, see that the threshold rule in equation (1.9) is decreasing in reported precision. That is, higher reported precision means that the policymaker believes the estimate to convey more information about the true treatment effect and hence a less conservative threshold rule is implemented. Thus, in the unclustered regime, the policymaker overestimates the precision of evidence from published studies and is therefore too lenient with their threshold rule for implementing the policy. The absolute size of the

[29] Note that the threshold rule in the clustered regime coincides exactly with the threshold rule in the model with normal signals in Tetenov (2012), although in this setting signals are not in fact normally distributed.

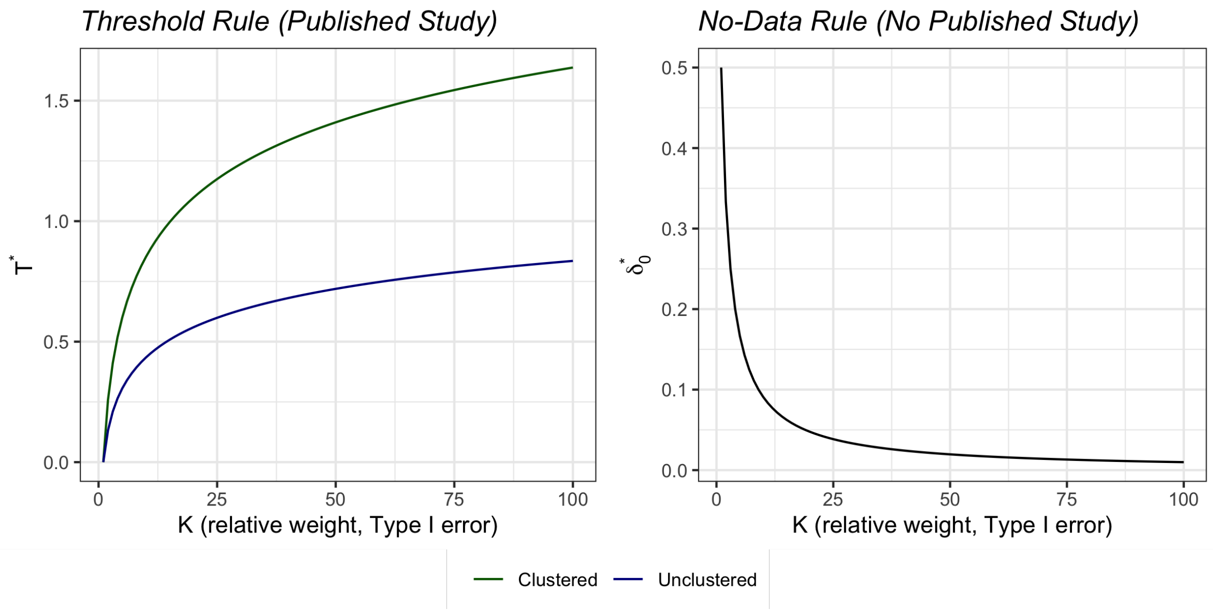

Figure 1.6: Minimax Regret Decision Rule in Clustered and Unclustered Regimes

Notes: The first panel shows the threshold rule in the event that a study is published and given by equation (1.6). The second panel shows the no-data rule in even that a study is not published. The level of publication bias $\hat{\gamma} = 0.016$ and the extent of downward bias $\hat{r} = 0.51$ are based on the empirical model estimated on studies in the DiD literature in Section 1.4.

difference increases with Type I error loss aversion. This is because Lemma 3.3.1 implies that the threshold rule in the unclustered regime is downward biased by a constant factor r, since $T^*_{C=0}/T^*_{C=1} = g(K) \cdot \widetilde{\sigma}/g(K) \cdot \sigma = r$.

In the second panel, we can see that the probability of implementing the policy decreases as K increases (and equals $\frac{1}{2}$ when $K = 1$). This is because the welfare cost of implementing an ineffective or harmful policy increases with K, which leads the policymaker to be more conservative with respect to implementation. Note that the no-data rule is unaffected by whether or not standard errors are clustered since no study is actually observed by the policymaker.

1.5.3 Comparing Regimes Based on True Regret

While the minimax regret decision rule in Lemma 3.3.1 is based on misspecified regret, I evaluate any given decision rule (T, δ_0) based on its *true regret*. True regret is derived from accurate beliefs about β, namely, that it follows a truncated normal distribution with (clustered) standard error σ, and where truncation down-weights the insignificant region of the density (based on γ). The utility of action a_1 when a study is published and action a_0 when it is not, is given by

$$U\big(a_1, a_0, \beta | K\big) = \begin{cases} K\beta D a_1 + \beta(1-D)a_0 & \text{if } \beta \leq 0 \\ \beta D a_1 + \beta(1-D)a_0 & \text{if } \beta > 0 \end{cases} \tag{1.10}$$

and the expected welfare of the decision rule (T, δ_0) is given by

$$W\big(\delta_1^T, \delta_0, \beta, \sigma, \widetilde{\sigma} | K\big) = \begin{cases} K\Big(\beta \cdot \mathbb{Pr}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + \beta \cdot \big(1 - \mathbb{Pr}[D=1|\beta,\widetilde{\sigma}]\big)\delta_0\Big) & \text{if } \beta \leq 0 \\ \beta \cdot \mathbb{Pr}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + \beta \cdot \big(1 - \mathbb{Pr}[D=1|\beta,\widetilde{\sigma}]\big)\delta_0 & \text{if } \beta > 0 \end{cases} \tag{1.11}$$

where $\mathbb{Pr}[D = 1|\beta, \widetilde{\sigma}]$ is the ex-ante publication probability conditional on $(\beta, \widetilde{\sigma})$; and $F(\cdot|\beta, \sigma, \widetilde{\sigma}, D = 1)$ is the cdf of a truncated normal distribution.[30] See that the probability of publication is based on the *reported* standard error and thus the effective significance threshold will differ across regimes. This also shows up in the cdf, where publication probabilities are based on $\widetilde{\sigma}$ but the true variation in the estimated treatment effect is governed by σ.

Finally, for a given average treatment effect β, true (i.e. clustered) standard error σ, and the Type I error loss aversion parameter K, regret is given by the following expression:

$$R\big(\delta_1^T, \delta_0, \beta, \sigma, \widetilde{\sigma} | K\big) = \begin{cases} -K \cdot \beta\Big(\mathbb{Pr}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + (1 - \mathbb{Pr}[D=1|\beta,\widetilde{\sigma}])\delta_0\Big) & \text{if } \beta \leq 0 \\ \beta\Big(\mathbb{Pr}[D=1|\beta,\widetilde{\sigma}] \cdot F(T|\beta,\sigma,\widetilde{\sigma},D=1) + (1 - \mathbb{Pr}[D=1|\beta,\widetilde{\sigma}]) \cdot (1-\delta_0)\Big) & \text{if } \beta > 0 \end{cases} \tag{1.12}$$

Thus, true regret is equal to the ex-ante probability of making an the incorrect treatment choice multiplied by the cost of the mistake $|\beta|$, and then weighted according to the planner's

[30] Specifically, the cdf is given by

$$F(t|\beta, \sigma, \widetilde{\sigma}, D = 1) \equiv \frac{\int_{-\infty}^{t} p\big(\frac{x}{\widetilde{\sigma}}\big)\phi\big(\frac{x-\beta}{\sigma}\big)dx}{\int p\big(\frac{x}{\widetilde{\sigma}}\big)\phi\big(\frac{x-\beta}{\sigma}\big)dx}$$

relative concern over Type I and Type II regret. Another way to interpret this expression is that it is what the policymaker would be using to choose their decision rule in order to minimize maximum regret if they had correct beliefs. The minimax regret of any decision rule (T, δ_0) given σ is given by

$$\text{MMR}(T, \delta_0 | K) = \max_{\beta \in [-B, B]} R\big(\delta_1^T, \delta_0, \beta, \sigma | K\big) \tag{1.13}$$

For any $K \geq 1$, let $\text{MMR}^*_{C=0}(K)$ denote the value of minimax regret in the unclustered regime based on the (misspecified) decision rule from Lemma 3.3.1 and let $\text{MMR}^*_{C=1}(K)$ denote the corresponding statistic for the clustered regime. Then the percent change in minimax regret from moving from the unclustered regime to the clustered regime is given by

$$100 \cdot \left(\frac{\text{MMR}^*_{C=1}(K)}{\text{MMR}^*_{C=0}(K)} - 1 \right) \tag{1.14}$$

Figure 1.7 plots this quantity for different values of the Type I error loss aversion parameter K. Results show that clustering lowers minimax regret if and only if $K > 63$. Recall that classical hypothesis testing at the 5% level entails a much larger level of loss aversion to Type I error i.e. $K = 102.4$ (Tetenov, 2012). Thus, the model suggests that clustering increased welfare if we use the benchmark cost implicitly implied by 5% hypothesis testing, although this could be overly conservative in certain settings.

To understand the intuition behind this result, note that clustering presents a trade-off for the policymaker. On the one hand, it improves the statistical precision of the evidence which leads to a superior threshold rule. On the other hand, clustering increases the probability of censoring studies, which increases the chances that policymakers are forced to make decisions without evidence. Suppose that $K = 1$. In this unique case, the threshold rule is identical across regimes ($T^* = 0$) and thus clustering provides no advantage. However, the probability of publication is lower in the clustered regime such that minimax regret is substantially larger than in the unclustered regime. However, as K increases the trade-off described above gradually favors clustering. This is because the threshold rule in the unclustered regime becomes increasingly

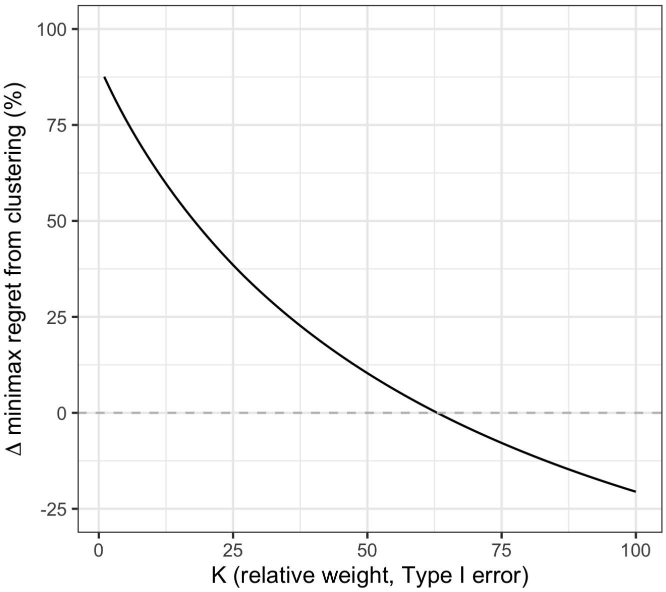

Figure 1.7: Percent Change in Minimax Regret from Clustering

Notes: The percent change in minimax regret moving from the unclustered regime to the clustered regime is calculated according to equation (1.14). The level of publication bias $\hat{\gamma} = 0.016$ and the extent of downward bias $\hat{r} = 0.51$ are based on the empirical model estimated on studies in the DiD literature in Section 1.4.

miscalibrated as K increases, which leads to larger costs in terms of minimax regret. When K is above 63, minimax regret in the clustered regime is lower than in the unclustered regime.

1.6 Conclusion

The econometrics literature on standard error corrections and the meta-science literature on publication bias share the common goal of improving credibility in empirical research. However, they are most often considered in isolation and the interaction between them has received little attention. This paper studies how their interaction affects the statistically credibility of published studies and decision-making among policymakers when treatment choice is informed by published evidence.

A central tension highlighted in the theory is that standard error corrections increase coverage but can also, unintendedly, worsen bias. Empirically, this is the case in the DiD literature, where clustering leads to large improvements in coverage but also sizable increases in the bias of estimated treatment effects. Incorporating this trade-off in a policymaking model with publication bias shows that clustering lowers minimax regret when loss aversion to Type I error is sufficiently high.

Appendix

This appendix contain proofs and supplementary materials. Section 1A contains proofs for the Propositions and Lemmas in the main text. Section 1B provides examples showing that bias can decrease when standard error corrections are small. Section 1D provides additional graphs illustrating the data. Section 1E shows descriptive statistics for unclustered studies in the 1990–1999 period. Section 1F introduces an augmented model with strategic clustering and proposes an estimation approach which is robust to certain forms of strategic clustering. It presents results from this alternative approach and compares them to the main results for robustness. Section 1G shows counterfactual comparisons between the clustered regime and the unclustered regime for all values of r on the unit interval. Finally, Section 1H shows robustness of the main results from using the empirical distribution of r calculated from 2015–2018 DiD studies.

1A Proofs

Proof of Proposition 1: The main result follows from two Lemmas which I prove below. First, Lemma 1A.2 shows that there exists an $r_1 \in (0, 1]$ such that for any $r \in (0, r_1)$ internal-validity bias increases:

$$\mathbb{E}_1[\hat{\beta}_j - \beta_j | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta_j | D_j = 1] > 0$$

Next, Lemma 1A.3 claims that there exists an $r_2 \in (0, 1]$ such that for any $r \in (0, r_2)$ study-selection bias weakly increases:

$$\mathbb{E}_1[\beta_j | D_j = 1] - \mathbb{E}_r[\beta_j | D_j = 1] \geq 0$$

Define $r^* = \min\{r_1, r_2\}$. It follows that for any $r \in (0, r^*)$, internal-validity bias and study-selection bias both increase. This immediately implies that the change in total bias (and estimated treatment effects), $\mathbb{E}_1[\hat{\beta}_j|D_j = 1] - \mathbb{E}_r[\hat{\beta}_j|D_j = 1]$, is positive since it is equal to the sum of the change in internal-validity bias and study-selection bias. Below, I present Lemmas 1A.2 and 1A.3 on which this argument is based. Before that, I present Lemma 1A.1, which is used in Lemma 1A.2.

Lemma 1A.1 (Expression for Bias Conditional on Publication)**.** *For a given $\beta \in [0, \infty)$, $\gamma \in [0, 1)$ and $r \in (0, 1]$,*

$$Bias(\beta, \gamma, r) = \frac{(1-\gamma)\big[\phi(1.96r - \beta) - \phi(\beta + 1.96r)\big]}{\Phi(-1.96r - \beta) + \gamma\big[\Phi(1.96r - \beta) - \Phi(-1.96r - \beta)\big] + 1 - \Phi(1.96r - \beta)} \tag{15}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the normal pdf and cdf, respectively.

Proof. Define $Z_j = \hat{\beta}_j - \beta$ so that $Z_j \sim N(0, 1)$ and bias conditional on publication is equal to $\mathbb{E}_r[Z_j|D_j = 1] = \mathbb{E}_r[\hat{\beta}_j|D_j = 1] - \beta$. We can write bias as the weighted sum of conditional expectations of the standard normal distribution:

$$\mathbb{E}_r[Z_j|D_j = 1] = \mathbb{P}\mathrm{r}_r[Z_j \leq -1.96r - \beta|D_j = 1] \cdot \mathbb{E}[Z_j|Z_j \leq -1.96r - \beta]$$

$$+\mathbb{P}\mathrm{r}_r[-1.96r - \beta < Z_j \leq 1.96r - \beta|D_j = 1] \cdot \mathbb{E}[Z_j| - 1.96r - \beta < Z_j \leq 1.96r - \beta]$$

$$+\mathbb{P}\mathrm{r}_r[Z_j > 1.96r - \beta|D_j = 1] \cdot \mathbb{E}[Z_j|Z_j > 1.96r - \beta]$$

$$= \bigg(\frac{\mathbb{P}\mathrm{r}_r[D_j = 1|Z_j \leq -1.96r - \beta]\Phi(-1.96r - \beta)}{\mathbb{P}\mathrm{r}_r[D_j = 1]}\bigg)\bigg(-\frac{\phi(-1.96r - \beta)}{\Phi(-1.96r - \beta)}\bigg)$$

$$+\bigg(\frac{\mathbb{P}\mathrm{r}_r[D_j = 1| - 1.96r - \beta \leq Z_j \leq 1.96r - \beta]\big[\Phi(1.96r - \beta) - \Phi(-1.96r - \beta)\big]}{\mathbb{P}\mathrm{r}_r[D_j = 1]}\bigg)$$

$$\times\bigg(\frac{\phi(-1.96r - \beta) - \phi(1.96r - \beta)}{\Phi(1.96r - \beta) - \Phi(-1.96r - \beta)}\bigg)$$

$$+\bigg(\frac{\mathbb{P}\mathrm{r}_r[D_j = 1|Z_j \geq 1.96r - \beta]\big[1 - \Phi(1.96r - \beta)\big]}{\mathbb{P}\mathrm{r}_r[D_j = 1]}\bigg)\bigg(\frac{\phi(1.96r - \beta)}{1 - \Phi(1.96r - \beta)}\bigg)$$

$$= -\frac{\phi(-1.96r - \beta)}{\mathbb{Pr}_r[D_j = 1]} + \frac{\gamma\big[\phi(-1.96r - \beta) - \phi(1.96r - \beta)\big]}{\mathbb{Pr}_r[D_j = 1]} + \frac{\phi(1.96r - \beta)}{\mathbb{Pr}_r[D_j = 1]}$$

The second equality uses Bayes' Rule on the probability terms and the formula for the expectation of a truncated standard normal on the expectation terms (i.e. for any $a < b$, we have that $\mathbb{E}[Z_j | Z_j \in (a, b)] = [\phi(a) - \phi(b)]/[\Phi(b) - \Phi(a)]$). The final equality uses Assumption 3, which states that the relative publication probabilities are one for significant results and γ for insignificant results. Simplifying the numerator and expanding the denominator gives the desired result. □

Lemma 1A.2 (Sufficient Condition for Increase in Internal-Validity Bias)**.** *Under Assumptions 1, 2, and 3, there exists an* $r_1 \in (0, 1]$ *such that for any* $r \in (0, r_1)$ *internal-validity bias increases with standard error corrections.*

Proof. First, I show that $\mathbb{E}_r[\hat{\beta}_j | D_j = 1] \to \mathbb{E}[\beta_j]$ as $r \to 0$. Using Bayes Rule, we have

$$\mathbb{E}_r[\hat{\beta}_j | D_j = 1] = \int \hat{\beta} f_{\hat{\beta}|D}(\hat{\beta} | D_j = 1; \gamma, r) d\hat{\beta} = \int \hat{\beta} \left(\frac{\mathbb{Pr}_r[D_j = 1 | \hat{\beta}_j] f_{\hat{\beta}}(\hat{\beta})}{\mathbb{Pr}_r[D_j = 1]} \right) d\hat{\beta}$$

$$= \int \left(\frac{\hat{\beta} \cdot p\big(\frac{\hat{\beta}}{r}\big) \int_\beta \phi(\hat{\beta} - \beta) f_\beta(\beta) d\beta}{\int_\beta \mathbb{Pr}_r[D_j = 1 | \beta] f_\beta(\beta) d\beta} \right) d\hat{\beta} \tag{16}$$

Note in the second equality that the distribution of latent studies $f_{\hat{\beta}}(\cdot)$ does not depend on either γ or r. Consider the integrand in (16). First, see that the numerator approaches $\hat{\beta} \int_\beta \phi(\hat{\beta} - \beta) f_\beta(\beta) d\beta$ as $r \to 0$. Next, see that the denominator satisfies

$$\lim_{r \to 0} \int_\beta \mathbb{Pr}_r[D_j = 1 | \beta] f_\beta(\beta) d\beta = 1$$

This equality uses the dominated convergence theorem to move the limit inside the integral and the fact that the probability of publication for any fixed β approaches one as $r \to 0$ (since all results are significant, and hence not censored, in the limit). To see that the conditions for the dominated convergence theorem are met, first see that the integrand converges pointwise to $f_\beta(\beta)$ as $r \to 0$. Second, see that for any $r \in (0, 1]$ and $\beta \geq 0$, the integrand is bounded above by $f_\beta(\beta)$ since $\mathbb{P}\mathrm{r}_r[D_j = 1|\beta] \leq 1$.

Thus, returning to the full expression for the integrand in equation (16), we can see that it converges pointwise to $\hat{\beta}\int_\beta \phi(\hat{\beta}-\beta)f_\beta(\beta)d\beta$ as $r \to 0$. Next, see that for any $r \in (0,1]$ and $\hat{\beta} \in \mathbb{R}$, the absolute value of the integrand satisfies

$$\frac{|\hat{\beta}| \cdot p\big(\frac{\hat{\beta}}{r}\big)\int_\beta \phi(\hat{\beta}-\beta)f_\beta(\beta)d\beta}{\int_\beta \mathbb{P}\mathrm{r}_r[D_j = 1|\beta]f_\beta(\beta)d\beta} \leq \frac{|\hat{\beta}| \cdot \phi(0)}{\int_\beta \mathbb{P}\mathrm{r}_1[D_j = 1|\beta]f_\beta(\beta)d\beta}$$

where the bound follows from the fact that $p\big(\frac{x}{r}\big) \leq 1$ and $\int_\beta \phi(x-\beta)f_\beta(\beta)d\beta \leq \phi(0)$ in the numerator, and $\mathbb{P}\mathrm{r}_r[D_j = 1|\beta]$ is strictly decreasing in r in the denominator.

Since the integrand in equation (16) (i) converges pointwise to $\hat{\beta}\int_\beta \phi(\hat{\beta}-\beta)f_\beta(\beta)d\beta$ and (ii) is dominated by an integrable function, we can apply the dominated convergence theorem to get

$$\lim_{r\to 0}\mathbb{E}_r[\hat{\beta}_j|D_j = 1] = \int_{\hat{\beta}} \hat{\beta}\int_\beta \phi(\hat{\beta}-\beta)f_\beta(\beta)d\beta d\hat{\beta}$$

$$= \int_\beta\bigg(\int_{\hat{\beta}} \hat{\beta}\phi(\hat{\beta}-\beta)d\hat{\beta}\bigg)f_\beta(\beta)d\beta = \int_\beta \mathbb{E}[\hat{\beta}_j|\beta]f_\beta(\beta)d\beta = \mathbb{E}[\beta_j] \tag{17}$$

which is what we wanted to show.

In the next step of the proof, I use similar arguments to also show that $\mathbb{E}_r[\beta_j|D_j = 1] \to \mathbb{E}[\beta_j]$ as $r \to 0$. Using Bayes' Rule, we can write

$$\mathbb{E}_r[\beta_j|D_j = 1] = \int \beta f_{\beta|D}(\beta|D_j = 1;\gamma,r)d\beta$$

$$= \int_\beta\left(\frac{\beta \cdot \mathbb{P}\mathrm{r}_r[D_j = 1|\beta]f_\beta(\beta)}{\int_\beta \mathbb{P}\mathrm{r}_r[D_j = 1|\beta f_\beta(\beta)d\beta}\right)d\beta$$

Note that the latent distribution of true effects, $f_\beta(\beta)$, does not depend on either γ or r. Now see that the integrand converges pointwise to $\beta f_\beta(\beta)$ as $r \to 0$. This follows because $\lim_{r\to 0}\mathbb{P}\mathrm{r}_r[D_j = 1|\beta] = 1$ in the numerator and because the denominator converges to one, as shown earlier.

Next, see that for any $r \in (0, 1]$ and $\beta \geq 0$, we have

$$\frac{\beta \cdot \mathbb{Pr}_r[D_j = 1|\beta] f_\beta(\beta)}{\int_{\beta'} \mathbb{Pr}_r[D_j = 1|\beta'] f_\beta(\beta') d\beta'} \leq \frac{\beta f_\beta(\beta)}{\int_{\beta'} \mathbb{Pr}_1[D_j = 1|\beta'] f_\beta(\beta') d\beta'}$$

where the inequality follows from the fact that $\mathbb{Pr}_r[D_j = 1|\beta]$ is weakly less than one (numerator) and decreasing in r (denominator). Note that the upper bound is integrable since Assumption 1 requires β_j to have a finite first moment. Thus, appealing again to the dominated convergence theorem, we have

$$\lim_{r \to 0} \mathbb{E}_r[\beta_j | D_j = 1] = \int_\beta \beta f_\beta(\beta) d\beta = \mathbb{E}[\beta_j] \tag{18}$$

Using the convergence in mean results in equations (17) and (18) and the linearity of expectations, it follows that

$$\Delta\text{Bias}(r) \equiv \mathbb{E}_1[\hat{\beta}_j - \beta_j | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta_j | D_j = 1]$$

$$\to \mathbb{E}_1[\hat{\beta}_j - \beta_j | D_j = 1] = \int_\beta \text{Bias}(\beta, \gamma, 1) f_\beta(\beta) d\beta > 0 \tag{19}$$

as $r \to 0$. The final inequality follows because it is clear from Lemma 1A.1 that $\text{Bias}(\beta, \gamma, 1) \geq 0$ when $\gamma \in [0, 1)$ (Assumption 3) and $\beta \geq 0$, and with strict inequality when $\beta > 0$. Assumption 1 requires that there exists some $\beta > 0$ on the support of β_j, giving the strict inequality.

Now we can prove the main claim. Consider the following set: $\{r | r \in (0, 1], \Delta\text{Bias}(r) = 0\}$. We know it is non-empty because $\Delta\text{Bias}(1) = 0$. Label the minimum of this set r_1. The claim is that for all $r \in (0, r_1)$, $\Delta\text{Bias}(r) > 0$. We will prove this by contradiction. Suppose instead that there exists an $\bar{r} \in (0, r_1)$ where

$$\Delta\text{Bias}(\bar{r}) \leq 0 < \lim_{r \to 0} \Delta\text{Bias}(r)$$

where the second inequality follows from equation (19). Note that $\Delta\text{Bias}(r)$ is continuous in r over $(0, 1)$ and well-defined for all $r \in (0, 1]$. Thus, there must exist some $\epsilon \in (0, \bar{r})$ such that $\Delta\text{Bias}(\bar{r}) \leq 0 < \Delta\text{Bias}(\epsilon)$. It follows from the intermediate value theorem that there exists an $r' \in (\epsilon, \bar{r})$ such that $\Delta\text{Bias}(r') = 0$ with $r' < \bar{r} < r_1$. But this contradicts the premise that r_1 is the smallest number satisfying this equality. □

Lemma 1A.3 (Sufficient Condition for Increase in Study-Selection Bias)**.** *Under Assumptions 1, 2, and 3, there exists an $r_2 \in (0, 1]$ such that for any $r \in (0, r_2)$ study-selection bias weakly increases with standard error corrections.*

Proof. Consider two cases. The first is the trivial case where the distribution of β_j is degenerate at some $\beta > 0$. Then for any $r \in (0, 1]$, $\Delta\text{SSB}(r) \equiv \mathbb{E}_1[\beta_j | D_j = 1] - \mathbb{E}_r[\beta_j | D_j = 1] = 0$. Let $r_2 = 1$. Then for any $r \in (0, r_2)$ there is no change in study-selection bias with standard error corrections: $\Delta\text{SSB}(r) = 0$.

Next, consider the case where the distribution of β_j is non-degenerate. See that

$$
\begin{aligned}
\lim_{r \to 0} \Delta\text{SSB}(r) &= \mathbb{E}_1[\beta_j | D_j = 1] - \lim_{r \to 0} \mathbb{E}_r[\beta_j | D_j = 1] \\
&= \mathbb{E}_1[\beta_j | D_j = 1] - \mathbb{E}[\beta_j] \\
&= \int_0^\infty [1 - F_{\beta|D}(t | D_j = 1; \gamma, 1)] dt - \int_0^\infty [1 - F_\beta(t)] dt \\
&= \int_0^\infty [F_\beta(t) - F_{\beta|D}(t | D_j = 1; \gamma, 1)] dt \qquad (20)
\end{aligned}
$$

The second equality uses the convergence in expectation result in equation (18) from Lemma 1A.2. The third equality uses the fact that for any non-negative random variable X with cdf F_X, we can write $\mathbb{E}[X] = \int_0^\infty [1 - F_X(t)] dt$. Equation (20) is positive if the distribution of published true treatment effects in the corrected regime, $F_{\beta|D}(\cdot | D_j = 1; \gamma, 1)$, first-order stochastically dominates the latent distribution of true treatment effects $F_\beta(\cdot)$. To show this holds, fix $t \in [0, \infty)$ and see that

$$\int_0^t f_\beta(\beta)d\beta - \int_0^t f_{\beta|D}(\beta|D_j = 1; \gamma, 1)d\beta$$
$$= \frac{1}{\mathbb{P}\mathrm{r}_1(D_j = 1)}\left(\mathbb{P}\mathrm{r}_1(D_j = 1)\int_0^t f_\beta(\beta)d\beta - \int_0^t \mathbb{P}\mathrm{r}_1(D_j = 1|\beta)f_\beta(\beta)d\beta\right)$$
$$= \frac{F_\beta(t)}{\mathbb{P}\mathrm{r}_1(D_j = 1)}\left(\mathbb{E}_\beta\Big[\mathbb{P}\mathrm{r}_1(D_j = 1|\beta)\Big] - \mathbb{E}_\beta\Big[\mathbb{P}\mathrm{r}_1(D_j = 1|\beta)\Big|\beta \leq t)\Big]\right) \geq 0$$

where the first equality uses Bayes' Rule for the second term. The second equality uses the fact that for any function $g(\cdot)$ and $t > 0$ we can write $\int^t g(\beta)f_\beta(\beta)d\beta = \mathbb{E}_\beta[g(\beta)|\beta \leq t; \gamma, 1] \cdot F_\beta(t)$. The final inequality follows from the fact that $\mathbb{P}\mathrm{r}_1(D_j = 1|\beta)$ is an increasing function of β.[31] Since β_j is non-degenerate, there exists some $t \in [0, \infty)$ for which this inequality is strict. This implies that equation (20) is strictly positive, which is what we wanted to show.

With this result, we can prove the main claim for the case where β_j is non-degenerate, namely, that for sufficiently small r, expected true treatment effects will increase following standard error corrections. First, consider the set $\{r|r \in (0, 1], \Delta\text{SSB}(r) = 0\}$. We know it is non-empty because $\Delta\text{SSB}(1) = 0$. Label the minimum of this set r_2. The claim is that for all $r \in (0, r_2)$, $\Delta\text{SSB}(r) > 0$. Suppose in contradiction of the claim that there exists an $\bar{r} \in (0, r_2)$ where

$$\Delta\text{SSB}(\bar{r}) \leq 0 < \lim_{r\to 0} \Delta\text{SSB}(r)$$

where the second inequality follows from the arguments above. Note that $\Delta\text{SSB}(r)$ is continuous in r over $(0, 1)$ and well-defined for all $r \in (0, 1]$. Thus, there must exist some $\epsilon \in (0, \bar{r})$ such that $\Delta\text{SSB}(\bar{r}) \leq 0 < \Delta\text{SSB}(\epsilon)$. It follows from the intermediate value theorem that there exists an $r' \in (\epsilon, \bar{r})$ such that $\Delta\text{SSB}(r') = 0$ with $r' < \bar{r} < r_2$. But this contradicts the premise that r_2 is the smallest number satisfying this equality. □

[31] The derivative is given by:

$$\frac{\partial}{\partial\beta}\Big[\mathbb{P}\mathrm{r}(D_j = 1|\beta; \gamma, 1)\Big] = (1 - \gamma)\Big(\phi(1.96 - \beta) - \phi(1.96 + \beta)\Big) \geq 0$$

which is strictly positive when $\beta > 0$.

Proof of Proposition 2: With a slight abuse of notation, let $f_\beta(\cdot)$ denote the distribution of $|\beta_j|$. This normalization is for notational convenience and is not necessary for proving the result. Next, note that the proof is based on selective publication against insignificant results at the 5% level, in line with Assumption 3; however, all arguments generalize straightforwardly to other critical thresholds.

As a starting point, the following Lemma provides an expression for average coverage in published studies for a fixed true effect, which will be used throughout the proof.

Lemma 1A.4 (Expression for Coverage with Degenerate β_j)**.** *For any $\beta \in [0, \infty)$, $r \in (0, 1]$ and $\gamma \in [0, 1]$, expected coverage in published studies is equal to*

$$Coverage(\beta, r) = \begin{cases} \frac{\gamma[\Phi(1.96r-\beta)-\Phi(-1.96r)]+\Phi(1.96r)-\Phi(1.96r-\beta)}{\Phi(-1.96r-\beta)+1-\Phi(1.96r-\beta)+\gamma[\Phi(1.96r-\beta)-\Phi(-1.96r-\beta)]} & \textit{if } \beta \leq 2 \times 1.96r \\ \frac{\Phi(1.96r)-\Phi(-1.96r)}{\Phi(-1.96r-\beta)+1-\Phi(1.96r-\beta)+\gamma[\Phi(1.96r-\beta)-\Phi(-1.96r-\beta)]} & \textit{if } \beta > 2 \times 1.96r \end{cases} \tag{21}$$

Proof. Fix $\beta \in [0, \infty)$. See that

$$\begin{aligned} \text{Coverage}(\beta, r) &= \mathbb{Pr}_r[\hat{\beta}_j - 1.96r \leq \beta \leq \hat{\beta}_j + 1.96r | D_j = 1] \\ &= \int_{\beta-1.96r}^{\beta+1.96r} f_{\hat{\beta}|D,\beta}(\hat{\beta}|D_j = 1, \beta; \gamma, r) d\hat{\beta} \\ &= \frac{\int_{\beta-1.96r}^{\beta+1.96r} \mathbb{Pr}_r(D_j = 1|\hat{\beta})\phi(\hat{\beta} - \beta) d\hat{\beta}}{\mathbb{Pr}_r(D_j = 1|\beta)} \end{aligned}$$

using Bayes Rule in the last equality and the fact that the probability of publication does not depend on the true effect β after conditioning on the estimate $\hat{\beta}$. Recall that statistically significant results are published with probability one and insignificant results with probability $\gamma \in [0, 1)$ (Assumption 3). Evaluating the integral in the numerator and expanding the denominator gives the desired expression. □

To begin, recall that the publication regime is uniquely characterized by $\gamma \in [0, 1)$, the relative probability of publishing insignificant results (Assumption 3). In the Lemma below, I show that the distribution of published studies in any publication regime $\gamma \in [0,1)$ is isomorphic to a mixture of a publication regime with $\gamma = 0$ (i.e. all insignificant results are censored) and publication regime with $\gamma = 1$ (i.e. all insignificant results are published).

Lemma 1A.5 (Publication Regime as Mixed Distribution)**.** *The density of published studies in publication regime $\gamma \in [0,1]$ and standard error regime $r \in (0,1)$, $f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1;\gamma,r)$, is equivalent to the following mixture of densities:*

$$f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1;\gamma,r) = \omega(r) \cdot f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1;1,r) + \big[1-\omega(r)\big] \cdot f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1;0,r)$$

with

$$\omega(r) = \frac{\gamma}{\mathbb{Pr}_r(D_j = 1)} \in [0,1] \tag{22}$$

Proof. For this proof, I express the probability of publication in publication regime γ and standard error regime r explicitly as $\mathbb{Pr}(D_j = 1;\gamma,r)$ (rather than subscripting the probability). The claim is trivially true in the case where $\gamma = 0$ or $\gamma = 1$. Let $\gamma \in (0,1)$. With Bayes Rule and Assumption 3 which assumes a step-wise publication selection function, we have that

$$\begin{aligned} f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1;\gamma,r) &= \frac{\mathbb{Pr}(D_j = 1|\hat{\beta};\gamma,r)\phi(\hat{\beta}-\beta)f_\beta(\beta)}{\mathbb{Pr}(D_j = 1;\gamma,r)} \\ &= \frac{\mathbb{1}\{|\hat{\beta}| \geq 1.96r\}\phi(\hat{\beta}-\beta)f_\beta(\beta) + \gamma\mathbb{1}\{|\hat{\beta}| < 1.96r\}\phi(\hat{\beta}-\beta)f_\beta(\beta)}{\mathbb{Pr}(D_j = 1;\gamma,r)} \end{aligned} \tag{23}$$

Note in the first equality that the probability of publication does not depend on the true effect β after conditioning on the estimate $\hat{\beta}$.

Now consider the mixture of two publication regimes: (i) a regime where all results are published ($\gamma = 1$) with weight $\omega(r)$ as defined in equation (22); and (ii) a regime where all insignificant results are censored ($\gamma = 0$) with weight $1-\omega(r)$. I show that the density of this mixture is equivalent to the density of published studies for publication regime $\gamma \in (0,1)$ in equation (23). Substituting the weights and densities in the mixture gives

$$\omega(r) \cdot f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1; 1, r) + \big[1 - \omega(r)\big] \cdot f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j = 1; 0, r)$$

$$\begin{aligned}
&= \left(\frac{\gamma}{\mathbb{Pr}(D_j = 1; \gamma, r)}\right)\Big(\mathbb{1}\{|\hat{\beta}| \geq 1.96r\}\phi(\hat{\beta} - \beta)f_\beta(\beta) + \mathbb{1}\{|\hat{\beta}| < 1.96r\}\phi(\hat{\beta} - \beta)f_\beta(\beta)\Big) \\
&\qquad + \left(\frac{\mathbb{Pr}(D_j = 1; \gamma, r) - \gamma}{\mathbb{Pr}(D_j = 1; \gamma, r)}\right)\left(\frac{\mathbb{1}\{|\hat{\beta}| \geq 1.96r\}\phi(\hat{\beta} - \beta)f_\beta(\beta)}{\mathbb{Pr}(D_j = 1; 0, r)}\right) \\
&= \Bigg(\underbrace{\frac{\mathbb{Pr}(D_j = 1; \gamma, r) - \gamma\big(1 - \mathbb{Pr}(D_j = 1; 0, r)\big)}{\mathbb{Pr}(D_j = 1; 0, r)}}_{\equiv \kappa}\Bigg)\left(\frac{\mathbb{1}\{|\hat{\beta}| \geq 1.96r\}\phi(\hat{\beta} - \beta)f_\beta(\beta)}{\mathbb{Pr}(D_j = 1; \gamma, r)}\right) \\
&\qquad + \left(\frac{\gamma\mathbb{1}\{|\hat{\beta}| < 1.96r\}\phi(\hat{\beta} - \beta)f_\beta(\beta)}{\mathbb{Pr}(D_j = 1; \gamma, r)}\right)
\end{aligned}$$

It is clear that this expression equals the density in the publication regime $\gamma \in (0, 1)$ in equation (23) provided that $\kappa = 1$. This is can be verified by substituting the following identify into the numerator:

$$\begin{aligned}
\mathbb{Pr}(D_j = 1; \gamma, r) = &\int_\beta \Big(\Phi(-1.96r - \beta) + 1 - \Phi(1.96r - \beta)\Big)f_\beta(\beta)d\beta \\
&+ \gamma\int_\beta [\Phi(1.96r - \beta) - \Phi(-1.96r - \beta)]f_\beta(\beta)d\beta \\
&= \mathbb{Pr}(D_j = 1; 0, r) + \gamma(1 - \mathbb{Pr}(D_j = 1; 0, r))
\end{aligned}$$

□

In the next step, I show that Lemma 1A.5 implies we only need to show that coverage increases with standard error corrections in the publication regime where $\gamma = 0$. For clarity, let expected coverage in publication regime $\gamma \in [0, 1]$ and standard error regime $r \in (0, 1]$ be denoted by

$$c_\gamma(r) \equiv \int \text{Coverage}(\beta, r)f_{\beta|D}(\beta|D_j = 1; \gamma, r)d\beta$$

Lemma 1A.5 implies that expected coverage in publication regime γ can be written as a weighted average of coverage in the 'publish all insignificant results' regime and the 'publish no insignificant results' regime: $c_\gamma(r) = \omega(r)c_1(r) + \big(1-\omega(r)\big)c_0(r)$. This implies that the change in expected coverage from standard error corrections in publication regime γ is equal to

$$\begin{aligned}
c_\gamma(1) - c_\gamma(r) &= \big[\omega(1)c_1(1) + \big(1-\omega(1)\big)c_0(1)\big] - \big[\omega(r)c_1(r) + \big(1-\omega(r)\big)c_0(r)\big] \\
&= \Big(1-\omega(r)\Big)\Big(c_0(1)-c_0(r)\Big) + \omega(1)\Big(c_1(1)-c_0(1)\Big) - \omega(r)\Big(c_1(r)-c_0(1)\Big) \\
&> \Big(1-\omega(r)\Big)\Big(c_0(1)-c_0(r)\Big)
\end{aligned}$$

where the inequality uses the fact that $c_1(1) - c_1(r) = [\Phi(1.96) - \Phi(-1.96)] - [\Phi(1.96r) - \Phi(-1.96r)] > 0$, and $\omega(1) > \omega(r)$ because the probability of publication in the denominator for the weight in equation (22) is decreasing in r. These two inequalities imply that the product in the second term is strictly greater than the product in the third term. Thus, we only need to show that coverage increases in the case where $\gamma = 0$ to show that coverage increases overall in publication regime $\gamma \in [0, 1)$.

Fix $\gamma = 0$ for the remainder of the proof. We want to show that expected coverage increases with standard error corrections:

$$c_0(1) - c_0(r)$$

$$\begin{aligned}
&= \int_0^\infty \text{Coverage}(\beta,1) f_{\beta|D}(\beta|D_j=1;0,1)d\beta - \int_0^\infty \text{Coverage}(\beta,r) f_{\beta|D}(\beta|D_j=1;0,r)d\beta \\
&= \Bigg(\int_0^{2\times1.96r} \text{Coverage}(\beta,1) f_{\beta|D}(\beta|D_j=1;0,1)d\beta - \int_0^{2\times1.96r} \text{Coverage}(\beta,r) f_{\beta|D}(\beta|D_j=1;0,r)d\beta\Bigg) \\
&+ \Bigg(\int_{2\times1.96r}^\infty \text{Coverage}(\beta,1) f_{\beta|D}(\beta|D_j=1;0,1)d\beta - \int_{2\times1.96r}^\infty \text{Coverage}(\beta,r) f_{\beta|D}(\beta|D_j=1;0,r)d\beta\Bigg)
\end{aligned} \tag{24}$$

We will show that both differences in the parentheses are weakly positive, and that at least one is strictly positive, which gives the desired result.

Consider the second difference, where the integrals are over $\beta \geq 2 \times 1.96r$. Consider the *integrand* in the second term of the difference (and keep the integral limits fixed). Using the expression for coverage when $\beta \geq 2 \times 1.96r$ from Lemma 1A.4 and Bayes' Rule we have that the integrand is equal to

$$\begin{aligned}\text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, r) =& \left(\frac{\Phi(1.96r) - \Phi(-1.96r)}{\mathbb{Pr}(D_j = 1|\beta; 0, r)} \right) \cdot \left(\frac{\mathbb{Pr}(D_j = 1|\beta; 0, r) f_\beta(\beta)}{\mathbb{Pr}(D_j = 1; 0, r)} \right) \\ =& \left(\frac{\Phi(1.96r) - \Phi(-1.96r)}{\mathbb{Pr}(D_j = 1; 0, r)} \right) \cdot f_\beta(\beta)\end{aligned}$$

Consider the term in parentheses in the final line. The numerator is increasing in r and the denominator is decreasing in r. Since both terms are strictly positive, this implies that the integrand is weakly increasing in r (and strictly increasing when $f_\beta(\beta) > 0$). In equation (24), this implies that the difference in the second parentheses is weakly positive, since the integral limits are the same for both terms, but r takes its maximum value of one in the first term.

Next, I show that the first difference in (24) is weakly positive. To do so, I make use of three Lemmas, which I state and prove below.

Lemma 1A.6 (Coverage Increases for Degenerate β_j)**.** *Let* $\gamma = 0$*. For any* $\beta \in (0, \infty)$ *and* $r \in (0, 1]$*, we have*

$$\frac{\partial}{\partial r} \Big(\textit{Coverage}(\beta, r) \Big) > 0$$

Proof. We will show the more general result that coverage increases with corrections for degenerate β_j for any critical threshold $c > 0$ (note that at the 5% significance threshold we have $c = 1.96r$). For convenience, let the second argument in the Coverage$(\cdot, \cdot)$ function be the critical threshold c rather than the reported standard error r. The case where $\beta \geq 2c$ with $c = 1.96r$ has already been shown in the main text of the proof for the more general case where β_j follows a distribution. That proof clearly generalizes to other thresholds. Next, consider the second case where $\beta \in (0, 2c)$. The expression for coverage (Lemma 1A.4) when $\gamma = 0$ is given by

$$\text{Coverage}(\beta, c) = \frac{\Phi(c) - \Phi(c-\beta)}{\Phi(-c-\beta) + 1 - \Phi(c-\beta)}$$

Taking the derivative with respect to c gives

$$\frac{\partial}{\partial c}\Bigg(\text{Coverage}(\beta, c)\Bigg)$$

$$\propto \frac{\partial}{\partial c}\Bigg(\Phi(c){-}\Phi(c{-}\beta)\Bigg)\Bigg(\Phi(-c{-}\beta){+}1{-}\Phi(c{-}\beta)\Bigg){-}\Bigg(\Phi(c){-}\Phi(c{-}\beta)\Bigg)\frac{\partial}{\partial c}\Bigg(\Phi(-c{-}\beta){+}1{-}\Phi(c{-}\beta)\Bigg)$$

where we ignore the denominator in the quotient rule since it is positive. This derivative is weakly positive if and only if

$$\frac{\phi(c+\beta) + \phi(c-\beta)}{1 - \Phi(c+\beta) + 1 - \Phi(c-\beta)} \geq \frac{\phi(c-\beta) - \phi(c)}{\Phi(c) - \Phi(c-\beta)} \tag{25}$$

Now recall that for $Z \sim N(0,1)$ and $a < b$, we have $\mathbb{E}[Z|Z \in (a,b)] = [\phi(a) - \phi(b)]/[\Phi(b) - \Phi(a)]$. Hence we have

$$\mathbb{E}[Z|Z \in (c+\beta, \infty)] = \frac{\phi(c+\beta)}{1 - \Phi(c+\beta)} \equiv \mu_1$$

$$\mathbb{E}[Z|Z \in (c-\beta, \infty)] = \frac{\phi(c-\beta)}{1 - \Phi(c-\beta)} \equiv \mu_2$$

$$\mathbb{E}[Z|Z \in (c-\beta, c)] = \frac{\phi(c-\beta) - \phi(c)}{\Phi(c) - \Phi(c-\beta)} \equiv \mu_3$$

For $\beta \geq 0$, we have that $\mu_1 \geq \mu_2 \geq \mu_3$. Now let

$$\omega = \frac{1 - \Phi(c+\beta)}{1 - \Phi(c+\beta) + 1 - \Phi(c-\beta)}$$

Since $\omega \in (0,1)$, we have that $\omega\mu_1 + (1-\omega)\mu_2 \geq \mu_3$, which gives the desired inequality in (25). □

Lemma 1A.7 (Derivative of Coverage With Respect to r)**.** *For any* $\beta \in [0, \infty)$*,* $r \in (0, 1]$ *and*

$\gamma \in [0,1]$, *we have*

$$\frac{\partial}{\partial\beta}\Big(\mathit{Coverage}(\beta,r)\Big) = \begin{cases} > 0 & \mathit{if}\ \beta \leq 2\times 1.96r \\ < 0 & \mathit{if}\ \beta > 2\times 1.96r \end{cases}$$

Proof. We will prove the more general result for arbitrary critical threshold $c > 0$ (note that $c = 1.96r$ at the 5% significance threshold). That is, we will show that coverage is increasing in β when $\beta \leq 2c$ and decreasing in β when $\beta > 2c$. As in Lemma 1A.6, let the second argument in the Coverage$(\cdot,\cdot)$ function be the critical threshold c rather than the reported standard error r. Consider the expression for coverage in Lemma 1A.4. Consider first the case where $\beta \leq 2c$. Using the quotient rule gives

$$\frac{\partial}{\partial\beta}\big(\text{Coverage}(\beta,c)\big) \propto \phi(c-\beta)d(\beta,c) - \big(\phi(c-\beta)-\phi(c+\beta)\big)n_1(\beta,c) > 0$$

where we define the denominator as $d(\beta,c) \equiv \Phi(-c-\beta)+1-\Phi(c-\beta)+\gamma[\Phi(c-\beta)-\Phi(-c-\beta)] > 0$ and the numerator as $n_1(\beta,c) \equiv \gamma[\Phi(c-\beta)-\Phi(-c)]+\Phi(c)-\Phi(c-\beta) > 0$. The inequality follows because $d(\beta,c) > n_1(\beta,c)$ and $\phi(c-\beta) > \phi(c-\beta)-\phi(c+\beta) > 0$.

Consider next the case where $\beta > 2c$. Define the numerator as $n_2(\beta,c) \equiv \Phi(c)-\Phi(-c) > 0$. Then

$$\frac{\partial}{\partial\beta}\big(\text{Coverage}(\beta,c)\big) \propto -n_2(\beta,c)\cdot\frac{\partial}{\partial\beta}\Big(d(\beta,c)\Big) = -n_2(\beta,c)\cdot\Big[(1-\gamma)\Big(\phi(c-\beta)-\phi(c+\beta)\Big)\Big] < 0$$

□

Lemma 1A.8 (First Order Stochastic Dominance in Corrected Standard Error Regime)**.** *Let $F_{\beta|D}(\beta|D_j = 1;\gamma,r)$ denote the cdf of published true treatment effects in standard error regime $r \in (0,1]$ and publication regime $\gamma \in [0,1]$. Then $F_{\beta|D}(\beta|D_j = 1;0,1)$ first-order stochastically dominates $F_{\beta|D}(\beta|D_j = 1;0,r)$ for any $r \in (0,1)$.*

Proof. I establish first-order stochastic dominance by showing that the monotone likelihood ratio property holds for the following ratio of densities. By Bayes Rule we have

$$\frac{f_{\beta|D}(\beta|D_j=1;0,1)}{f_{\beta|D}(\beta|D_j=1;0,r)} = \frac{\left(\frac{\mathbb{P}\mathrm{r}[D_j=1|\beta;0,1]f_\beta(\beta)}{\mathbb{P}\mathrm{r}[D_j=1;0,1]}\right)}{\left(\frac{\mathbb{P}\mathrm{r}[D_j=1|\beta;0,r]f_\beta(\beta)}{\mathbb{P}\mathrm{r}[D_j=1;0,r]}\right)}$$

$$= \left(\frac{\Phi(-1.96-\beta)+1-\Phi(1.96-\beta)}{\Phi(-c-\beta)+1-\Phi(c-\beta)}\right)\cdot K$$

where $c \equiv 1.96r$ and $K \equiv \mathbb{P}\mathrm{r}[D_j = 1;0,r]/\mathbb{P}\mathrm{r}[D_j = 1;0,1] > 0$. Thus the derivative with respect to β is given by

$$\frac{\partial}{\partial\beta}\left(\frac{f_{\beta|D}(\beta|D_j=1;0,1)}{f_{\beta|D}(\beta|D_j=1;0,r)}\right) \propto \frac{\partial}{\partial\beta}\Big(\Phi(-1.96-\beta)+1-\Phi(1.96-\beta)\Big)\Big(\Phi(-c-\beta)+1-\Phi(c-\beta)\Big)$$

$$-\Big(\Phi(-1.96-\beta)+1-\Phi(1.96-\beta)\Big)\frac{\partial}{\partial\beta}\Big(\Phi(-c-\beta)+1-\Phi(c-\beta)\Big)$$

We want to show this is positive, which is equivalent to showing the following inequality

$$\frac{\phi(1.96-\beta)-\phi(1.96+\beta)}{1-\Phi(1.96-\beta)+1-\Phi(1.96+\beta)} \geq \frac{\phi(c-\beta)-\phi(c+\beta)}{1-\Phi(c-\beta)+1-\Phi(c+\beta)} \tag{26}$$

Note that $c = 1.96r < 1.96$ since $r \in (0,1)$. Hence it suffices to show that the fraction on the right hand side is increasing in c. To show this, first let $Z \sim N(0,1)$. Then using the formula for the expectation of a truncated normal gives

$$\mathbb{E}[Z|Z \in (c-\beta, c+\beta)] = \frac{\phi(c-\beta)-\phi(c+\beta)}{\Phi(c+\beta)-\Phi(c-\beta)} \equiv \mu_1(\beta, c)$$

Next, define

$$\mu_2(\beta,c) \equiv \frac{\Phi(c+\beta)-\Phi(c-\beta)}{1-\Phi(c-\beta)+1-\Phi(c+\beta)}$$

Now see that $\mu_1(\beta,c)\cdot\mu_2(\beta,c)$ gives the right hand side ratio in equation (26). Thus the derivative using the product rule is equal to

$$\frac{\partial}{\partial c}\Big(\mu_1(\beta,c)\cdot\mu_2(\beta,c)\Big)=\frac{\partial}{\partial c}\Big(\mu_1(\beta,c)\Big)\Big(\mu_2(\beta,c)\Big)+\Big(\mu_1(\beta,c)\Big)\frac{\partial}{\partial c}\Big(\mu_2(\beta,c)\Big)$$

Showing that all four terms in this expression are positive is sufficient for proving the derivative is positive. First, see that $\mu_2(\beta, c)$ is clearly positive. Next, see that $\mu_1(\beta, c)$ is positive because it is the conditional expectation of a standard normal over an even interval centered at $c > 0$. Moreover, the derivative $\partial\mu_1(\beta, c)/\partial c$ is positive because the conditional expectation must increase when the fixed-width interval over which the expectation is taken increases (i.e. shifts to the right). Finally, using the quotient rule, we have

$$\begin{aligned}\frac{\partial}{\partial c}\Big(\mu_2(\beta,c)\Big)&\propto\frac{\partial}{\partial c}\Big(n(\beta,c)\Big)\Big(d(\beta,c)\Big)-\Big(n(\beta,c)\Big)\frac{\partial}{\partial c}\Big(d(\beta,c)\Big)\\&=\Big(\phi(c+\beta)-\phi(c-\beta)\Big)d(\beta,c)+n(\beta,c)\Big(\phi(c+\beta)+\phi(c-\beta)\Big)\end{aligned}$$

where $n(\beta, c) \equiv \Phi(c+\beta)-\Phi(c-\beta)$ denotes the numerator and $d(\beta, c) \equiv 1-\Phi(c-\beta)+1-\Phi(c+\beta)$ the denominator. This derivative being positive is equivalent to

$$\frac{\phi(c+\beta)}{d(\beta,c)-n(\beta,c)}\geq\frac{\phi(c-\beta)}{d(\beta,c)+n(\beta,c)}\iff\frac{\phi(c+\beta)}{1-\Phi(c+\beta)}\geq\frac{\phi(c-\beta)}{1-\Phi(c-\beta)}$$

This inequality holds because the hazard function of the normal distribution is increasing and $c + \beta \geq c - \beta$ when $\beta \geq 0$.

Thus, $f_{\beta|D}(\beta|D_j = 1; 0, 1)/f_{\beta|D}(\beta|D_j = 1; 0, r)$ is increasing in β and therefore satisfies the monotone likelihood ratio property. This implies first-order stochastic dominance, giving the desired result. □

Using these three Lemmas, we have that

$$\int_0^{2\times1.96r}\text{Coverage}(\beta,1)f_{\beta|D}(\beta|D_j=1;0,1)d\beta-\int_0^{2\times1.96r}\text{Coverage}(\beta,r)f_{\beta|D}(\beta|D_j=1;0,r)d\beta$$

$$\geq\int_0^{2\times1.96r}\text{Coverage}(\beta,r)f_{\beta|D}(\beta|D_j=1;0,1)d\beta-\int_0^{2\times1.96r}\text{Coverage}(\beta,r)f_{\beta|D}(\beta|D_j=1;0,r)d\beta\geq0$$

The first inequality uses Lemma 1A.6 to replace Coverage$(\beta, 1)$ with Coverage(β, r) in the first term. The final inequality follows from the fact that Coverage(β, r) is strictly increasing over $(0, 2 \times 1.96r)$ (Lemma 1A.7) and first-order stochastic dominance in the distribution of published true effects in the corrected regime as compared with the uncorrected regime (Lemma 1A.8). Thus, the difference is strictly positive if β_j has support on a subset of $(0, 2 \times 1.96r)$ and zero otherwise.

Finally, note that β_j is assumed to have support on a subset of the non-negative real line and not be degenerate at zero (Assumption 1). This implies that both differences in equation (24) are weakly positive and that at least one is strictly positive, completing the proof. □

Lemma 1A.9 (Sufficient Condition for Improved Coverage)**.** *If nominal coverage equals 0.95 and* $r < 0.8512$*, then Coverage(r)*< 0.95.

Proof. Let nominal coverage equal 0.95. Consider coverage conditional on publication in the uncorrected regime:

$$\text{Coverage}(r) = \int \text{Coverage}(\beta, r) f_{\beta|D}(\beta | D_j = 1; \gamma, r) d\beta \leq \text{Coverage}(2 \times 1.96r, r)$$

$$= \frac{\Phi(1.96r) - \Phi(-1.96r)}{\Phi(-3 \times 1.96r) + 1 - \Phi(-1.96r) + \gamma[\Phi(-1.96r) - \Phi(-3 \times 1.96r)]}$$

$$\leq \frac{\Phi(1.96r) - \Phi(-1.96r)}{\Phi(-3 \times 1.96r) + 1 - \Phi(-1.96r)} \tag{27}$$

The first inequality follows from Lemma 1A.7, which shows that Coverage(β, r) is increasing in β when $\beta \leq 2 \times 1.96r$ and decreasing in β when $\beta > 2 \times 1.96r$; this implies that it is maximized when $\beta = 2 \times 1.96r$. The equality in the second line uses the formula for coverage in Lemma 1A.4. The last inequality uses the fact that the expression in the second line is decreasing in γ.

Denote the final expression in equation (27) as $h(r)$. It is straightforward to show that $dh(r)/dr > 0$. Moreover, see that $h(r)$ is continuous in r, and that $h(0) = 0$ and $h(1) = 0.9744$.

By the intermediate value theorem, it follows that there exists some $\bar{r} \in (0,1)$ such that $h(\bar{r}) = 0.95$. Since $dh(r)/dr > 0$, it follows that this value is unique and that $h(r) < 0.95$ for all $r < \bar{r}$. Finally, we can calculate that $\bar{r} = 0.8512$, completing the proof. □

Proof of Lemma 3.3.1: First, consider the threshold rule. Tetenov (2012) considers the case where the estimated treatment effect $\hat{\beta}$ is normally distributed while I consider the case where the policymaker erroneously believes it is normally distributed. Since the derivation of the statistical decision rule is based on identical beliefs, the results from Tetenov (2012) on page 160 immediately apply, despite the fact that those beliefs happen to be incorrect in this setting. (Note however that regret, which is based on the true distribution of studies, will differ in this setting compared to the setting in Tetenov (2012)).

The no-data rule is identical to the one proved in Kitagawa and Vu (2023). □

1B Ambiguous Impact of Corrections on Bias

Proposition 1 shows that bias increases with standard error corrections when they are sufficiently large. This appendix presents examples where bias can decrease when standard error corrections are small. This is formalized in the following lemma:

Lemma 1B.1 (Ambiguous Impact on Bias)**.** *Under Assumptions 1, 2, and 3, standard error corrections have an ambiguous impact on the individual signs for the change in internal-validity bias, study-selection bias and total bias. That is, there exist distinct combinations of* $(\mu_{\beta,\sigma}, \gamma, r)$ *such that their individual signs can be positive, negative, or zero.*

Proof. The proof consists of presenting numerical examples and contains two steps. In the first, I show ambiguity in the sign of the change in internal-validity bias and total bias. In the second, I do the same for study-selection bias.

(1) Internal-Validity Bias and Total Bias

Suppose that β_j follows a degenerate distribution with $\mathbb{Pr}[\beta_j = \beta] = 1$ for some $\beta > 0$. This implies that the change in internal-validity bias following standard error corrections will be equal to the change in total bias (and the change in estimated treatment effects):

$$\underbrace{\mathbb{E}_1[\hat{\beta}_j - \beta | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta | D_j = 1]}_{\Delta\text{Internal-validity bias}} = \underbrace{\mathbb{E}_1[\hat{\beta}_j | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j | D_j = 1]}_{\Delta\text{Total bias}=\Delta\text{Estimated treatment effects}} \tag{28}$$

We can use the expression for $\text{Bias}(\beta, \gamma, r)$ from Lemma 1A.1 to show that the sign of equation (28) from standard error corrections is ambiguous i.e. the sign of $\text{Bias}(\beta, \gamma, 1) - \text{Bias}(\beta, \gamma, r)$ can be positive, negative or zero. Fix $(\gamma, r) = (0.1, 0.75)$. Then for $\beta = 1.5$ and $\beta = 0.25$, we have that

$$\text{Bias}\big(1.5, 0.1, 1\big) - \text{Bias}\big(1.5, 0.1, 0.75\big) = 0.8244 - 0.6307 = 0.1937 > 0$$

$$\text{Bias}\big(0.25, 0.1, 1\big) - \text{Bias}\big(0.25, 0.1, 0.75\big) = 0.34319 - 0.3722 = -0.0290 < 0$$

Finally, by the intermediate value theorem, there exists some $\beta' \in (0.25, 1.5)$ such that $\text{Bias}(\beta', 0.1, 1) - \text{Bias}(\beta', 0.1, 0.75) = 0$.

(2) Study Selection Bias

Consider a two-point distribution for β_j where $\mathbb{Pr}[\beta_j = \beta] = p_1^* \cdot \mathbb{1}\{\beta = \beta_1\} + (1 - p_1^*) \cdot \mathbb{1}\{\beta = \beta_2\}$ for $0 \leq \beta_1 < \beta_2$ and $p_1^* \in (0, 1)$. Then by Bayes' Rule we have

$$\text{TrueTE}(\beta_1, \beta_2, p_1^*, \gamma, r) \equiv \mathbb{E}_r[\beta_j | D_j = 1] = \frac{p_1^* \beta_1 C(\beta_1, \gamma, r) + (1 - p_1^*)\beta_2 C(\beta_2, \gamma, r)}{p_1^* C(\beta_1, \gamma, r) + (1 - p_1^*) C(\beta_2, \gamma, r)}$$

where $C(\beta, \gamma, r) \equiv \int_{z'} p\big(\frac{\beta + z'}{r}\big)\phi(z')dz'$ is the probability of publication conditional on β.

Now suppose $\beta_1 = 0$ and $p_1^* = 0.5$. Then the change in true treatment effects is given by

$$\text{TrueTE}(0, \beta_2, 0.5, \gamma, 1) - \text{TrueTE}(0, \beta_2, 0.5, \gamma, r)$$

$$= \beta_2 \left(\frac{C(\beta_2, \gamma, 1)}{C(0, \gamma, 1) + C(\beta_2, \gamma, 1)} - \frac{C(\beta_2, \gamma, r)}{C(0, \gamma, r) + C(\beta_2, \gamma, r)} \right) \tag{29}$$

which is strictly positive if and only if

$$\frac{C(\beta_2, \gamma, 1)}{C(0, \gamma, 1)} > \frac{C(\beta_2, \gamma, r)}{C(0, \gamma, r)}$$

That is, true treatment effects will increase if the probability of publication conditional on $\beta_2 > 0$ relative to the probability of publication conditional on $\beta_1 = 0$ is higher in the corrected regime relative to the uncorrected regime.

As in the previous section, fix $(\gamma, r) = (0.1, 0.75)$. We can use the expression in equation (29) to calculate the change in true treatment effects from standard error corrections for different values of β_2. For $\beta_2 = 1.5$ and $\beta_2 = 0.75$, we have that

$$\text{TrueTE}(0, 1.5, 0.5, 0.1, 1) - \text{TrueTE}(0, 1.5, 0.5, 0.1, 0.75) = 0.0261 > 0$$

$$\text{TrueTE}(0, 0.75, 0.5, 0.1, 1) - \text{TrueTE}(0, 0.75, 0.5, 0.1, 0.75) = -0.0016 < 0$$

Finally, by the intermediate value theorem, there exists some $\beta' \in (0.75, 1.5)$ such that $\text{TrueTE}(0, \beta', 0.5, 0.1, 1) - \text{TrueTE}(0, \beta', 0.5, 0.1, 0.75) = 0$. □

Practically, Lemma 1B.1 implies that the impact of standard error corrections on either bias, estimated treatment effects, or true treatment effects is fundamentally an empirical question. In particular, to learn how bias has changed in any given setting, it is necessary to have knowledge about the underlying parameters $(\mu_{\beta,\sigma}, \gamma, r)$.

Recall that the main text provides an example where internal-validity bias decreases with corrections. This example relies on the distribution of published true effects changing. By contrast, Proposition 1B.1 shows that bias can decrease with a degenerate, and hence unchanged, distribution of true effects.

For intuition, consider the example in Lemma 1B.1 which examines bias in the case of an empirical literature examining a single question of interest with a fixed true effect. With $r = \frac{3}{4}$, clustering increases the effective significance threshold from $1.96 \times \frac{3}{4} \approx 1.5$ to approximately 2. With selective publication ($\gamma = \frac{1}{10}$), the clustered regime will therefore censor a large share of studies between 1.5 and 2. How this impacts bias depends on whether censoring these studies tends to increase or decrease the expected estimated treatment effect in the uncorrected regime. In the examples given in the proof, we have that $\mathbb{E}[\hat{\beta}_j | D_j = 1, \beta = 1.5; \gamma = \frac{1}{10}, r = \frac{3}{4}] = 2.13$ and $\mathbb{E}[\hat{\beta}_j | D_j = 1, \beta = \frac{1}{4}; \gamma = \frac{1}{10}, r = \frac{3}{4}] = 0.62$, where β_j is degenerate in both cases. In the first case, moving to the clustered regime censors studies with effect sizes between 1.5 and 2, which are smaller than the mean in the unclustered regime of 2.13; this leads to an increase in estimated treatment effects and thus bias since β_j is degenerate. In the second case, the opposite occurs.

1C Bias and True Treatment Effects

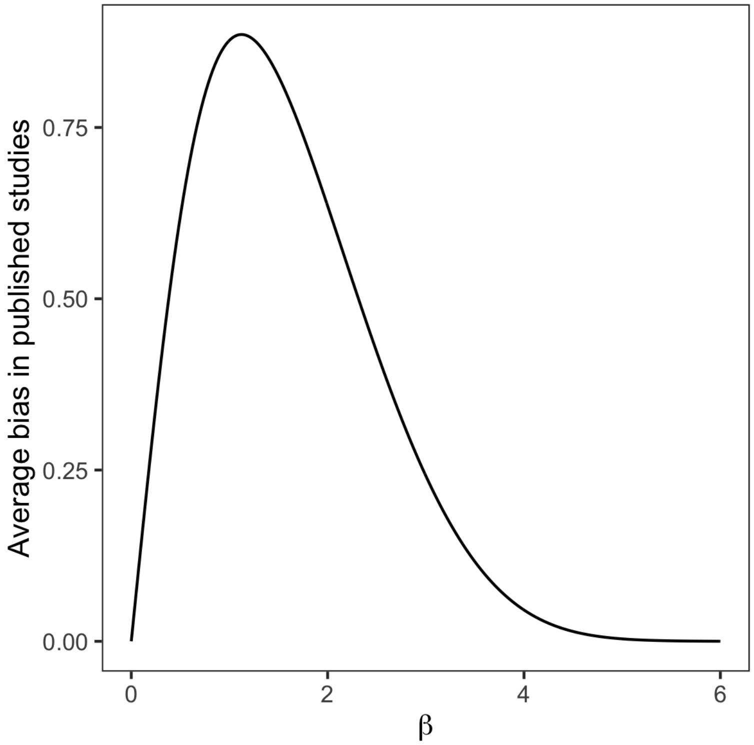

Figure 1C.1: Plot of $\mathbb{E}_1[\hat{\beta}_j - \beta | D_j = 1, \beta]$ for different values of β and $\gamma = 0.1$.

1D Details on Descriptive Statistics

This appendix provides further details on the descriptive statistics in Section 1.3.

Figure 1D.1 shows the distribution of JEL codes. Note that studies typically include multiple JEL codes and Figure 1D.1 plots the distribution at the JEL code level rather than at a study-level e.g. with weighted JEL codes. The results show that clustered articles are less likely to be Health, Education & Welfare (I); and Labor (J), although the difference is not statistically significant. Figure 1D.1 shows that clustered studies are more contain to have JEL codes that are outside the three dominant categories of Public Economics (H); Health, Education & Welfare (I); and Labor (J).

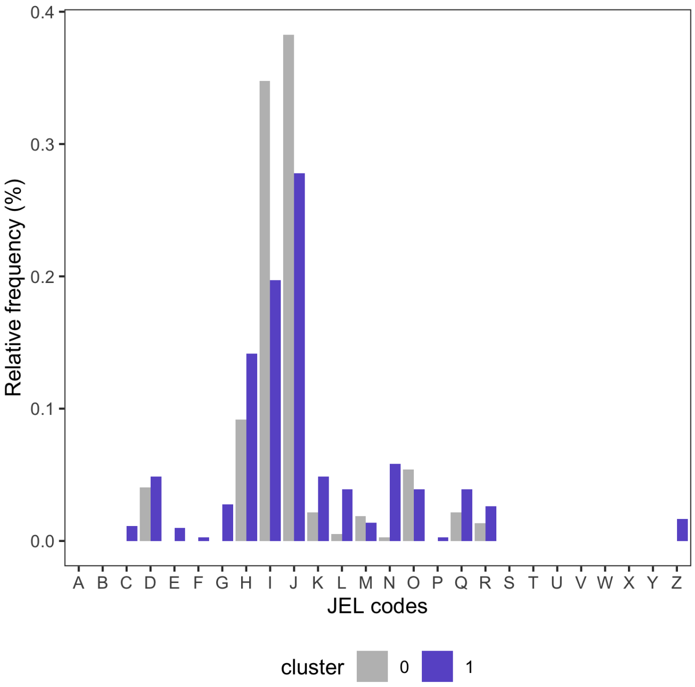

Figure 1D.1: Distribution of JEL codes. The most common JEL codes are: Public Economics (H); Health, Education & Welfare (I); and Labor (J)

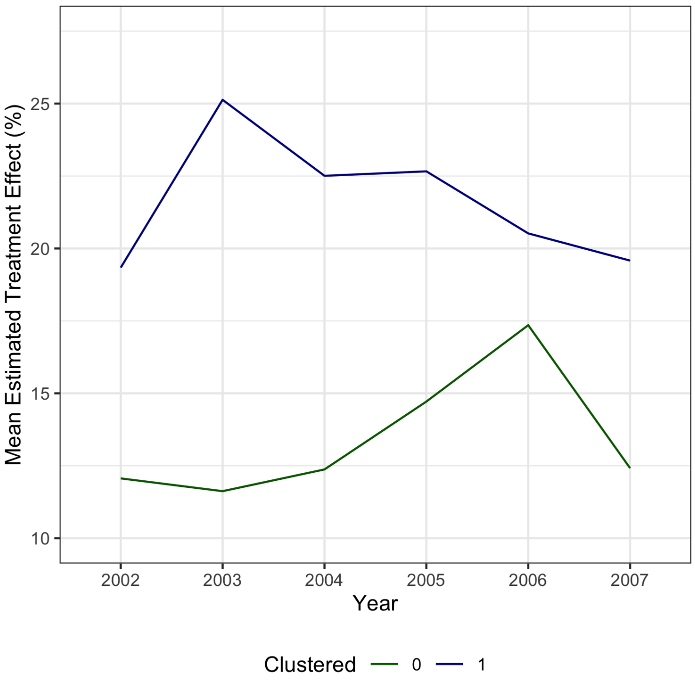

Figure 1D.2: Five-Year Centered Moving Average of the Magnitude Estimated Treatment Effects

Figure 1D.2 shows the five-year centered moving average of estimated treatment effects by clustering regime.[32] Effect sizes are considerably larger for studies reporting clustered standard errors. In particular, the magnitude of estimated treatment effects range approximately between 20–25% in the clustered regime and between 12.5–17.5% in the unclustered regime.

1E Comparative Descriptive Statistics from 1990–1999

This appendix analyzes unclustered studies from the 1990–1999. The main motivation is to examine the extent to which strategic clustering over 2000–2009 (i.e. the time period in the main analysis) might be driving the result that effect sizes in the clustered regime substantially larger than the unclustered regime. Analyzing DiD articles published between 1990 and 1999 is useful because the norm over this period was to report unclustered standard errors (Bertrand et al., 2004). Thus, DiD studies in this period are unlikely to be subject to strategic clustering, providing a useful comparison group.

Table 1E.1 compares effect sizes between unclustered studies published between 2000–2009 to those published between 1990–1999. The average effect size between 2000-2009 is 12.18%. In the earlier 1990-1999 period, effect sizes were only between 1.5–2 ppts smaller. This difference is statistically indistinguishable from zero, although with relatively few observations there is somewhat limited power to reject the null hypothesis. This provides suggestive evidence that the large increase in effect sizes observed over the 2000–2009 period is not driven by strategic clustering of the form discussed here.

There are two reasons for the relatively small sample size. First, the string-search algorithm I use from Currie et al. (2020) which I use is based on searching articles for variations of the term ‘difference-in-differences’ (e.g. DiD, diff-and-diff etc.) Use of this terminology was less consistent in the 1990’s when DiD designs were beginning to be used more frequently in applied work. A second reason for the small sample is that studies must meet the inclusion criteria

[32] A five-year averaging window is used because there are relatively few clustered studies in earlier years of the decade and relatively few unclustered studies in later years of the decade.

described in Section 1.3 which ensure comparability of effect sizes (i.e. estimated treatment effects in percent units from a binary treatment) across studies.

Table 1E.1: Effect Sizes of Unclustered Studies: 1990's vs. 2000's

$\mathbb{1}(1990-1999)$	-1.609	-1.725
	(4.145)	(3.264)
Mean in 2000–2009	12.18	12.18
Observations	43	43
Adjusted-R^2	-0.021	0.054
Study controls		X

Note: The sample is unclustered studies over 1990-2009. Results are from OLS regressions of the magnitude estimated treatment effects on an indicator for whether the study was published between 1990–1999. Study controls include a quadratic on the log of the number of observations, an indicator for policy evaluations, and a three-way interaction between JEL topics H (Public Economics), I (Health, Education, and Welfare), and J (Labor and Demographic Economics). These JEL topics are the most common codes for DiD studies. The dependent variable is in percent units or, for studies where the dependent variable is measured in logs, in log point units. The estimated coefficients are in percentage point units. Robust standard errors are in parentheses.

1F Robust Estimation for Strategic Clustering

The presence of strategic clustering could affect the consistent estimation of parameters of the latent distribution, which could, in turn, affect the main results on the impact of clustering on bias and coverage. This appendix proposes an estimation approach which is robust to the simple form of strategic clustering where researchers choose to cluster only when it does not change the statistical significance of their findings.

First, I extend the model in the main text to include strategic clustering. Second, I present the robust estimation strategy and implement it for the DiD sample. Finally, I compare results from the main text with those using the alternative robust estimation approach. I find very similar results across both approaches, which provides evidence that the form of strategic clustering discussed here is not driving the main conclusions.

1F.1 Model of Strategic Clustering

The model extends the model in Section 2.2 to incorporate strategic clustering:

1. **Draw a latent study:** $(\beta_j, \sigma_j) \sim \mu_{\beta,\sigma}$

2. **Estimate the treatment effect:** $\hat{\beta}_j | \beta_j, \sigma_j \sim N(\beta_j, \sigma_j^2)$

3. **Report standard errors:** This follows a two-stage process. In the first stage, researchers either endogenously cluster with probability $\beta_{c,1} \in [0, 1]$ or otherwise exogeneously cluster with probability $1 - \beta_{c,1}$. In the second stage, researchers choose which standard errors to report depending on the outcome of the first stage.

   (a) Endogenous clustering:

$$\widetilde{\sigma}_j = \begin{cases} r \cdot \sigma_j & \text{if } 1.96r \leq |\hat{\beta}_j|/\Sigma \leq 1.96 \\ \sigma_j & \text{otherwise} \end{cases}$$

   (b) Exogeneous clustering:

$$\widetilde{\sigma}_j = \begin{cases} r \cdot \sigma_j & \text{with probability } 1 - \beta_{c,2} \\ \sigma_j & \text{with probability } \beta_{c,2} \end{cases}$$

   where $r \in (0, 1)$ and $\beta_{c,2} \in (0, 1)$.

4. **Publication selection:**

$$\mathbb{Pr}(D_j = 1 | \hat{\beta}_j, \widetilde{\sigma}_j) = \begin{cases} \gamma & \text{if } |\hat{\beta}_j|/\widetilde{\sigma}_j \geq 1.96 \\ 1 & \text{otherwise} \end{cases} \tag{30}$$

The extension from the baseline model in Section 2.2 is in the third step. There exists some probability $\beta_{c,1}$ that researchers will choose whether or not to cluster strategically. Specifically, researchers strategically choose not to cluster with probability when doing so allows them to obtain statistical significance. Otherwise, they always cluster. When $\beta_{c,1} = 0$ clustering is completely exogenous and the model collapses to the baseline model.

1F.2 Robust Estimation

The follow result provides the basis for an estimation approach which is robust to the form of strategic clustering outlined in the model above:

Lemma 1F.1. *The distribution of statistically significant, published studies in the clustered regime, $\hat{\beta}_j, \sigma_j, \beta_j | D_j = 1, C_j = 1, |\hat{\beta}_j|/\sigma_j \geq 1.96$, does not depend on $(\beta_{c,1}, \beta_{c,2})$.*

Proof. I will show that the density of published *clustered* studies in the endogenous regime is identical to the density in the exogenous regime when $\gamma = 0$. Since the overall density of published clustered studies is simply a mixture of these the endogenous and exogenous regimes, it follows that the overall density must equal to the density in the exogenous regime with $\gamma = 0$, which does not depend on $(\beta_{c,1}, \beta_{c,2})$. Note also that conditioning on statistical significance is equivalent to setting $\gamma = 0$, since doing so censors all insignificant results.

First, consider the endogenous regime, which we denote with $E = 1$. By Bayes Rule we have that the density of published clustered studies is given by

$$f_{\hat{\beta},\sigma,\beta|D}(\hat{\beta}, \sigma, \beta | D_j = 1; \gamma, 1, E = 1) = \frac{\mathbb{Pr}_1[D_j = 1 | \hat{\beta}, \sigma; E = 1] \frac{1}{\sigma} \phi\big(\frac{\hat{\beta} - \beta}{\sigma}\big)}{\mathbb{Pr}_1[D_j = 1; E = 1]}$$

$$\propto \mathbb{1}\{|\hat{\beta}|/\sigma \leq 1.96r\} \cdot \gamma \frac{1}{\sigma} \phi\bigg(\frac{\hat{\beta} - \beta}{\sigma}\bigg) + \mathbb{1}\{|\hat{\beta}|/\sigma > 1.96\} \cdot \frac{1}{\sigma} \phi\bigg(\frac{\hat{\beta} - \beta}{\sigma}\bigg)$$

Note that all studies with $|x|/\sigma \in (1.96r, 1.96)$ are strategically unclustered in the endogenous regime, and hence the density over this region for clustered studies is zero.

Next, consider the density of published clustered studies in the exogenous regime:

$$f_{\hat{\beta},\Sigma,\beta|D,\widetilde{\Sigma}}(\hat{\beta}, \sigma, \beta | D_j = 1; \gamma, 1, E = 0) = \frac{\mathbb{Pr}_1[D_j = 1 | \hat{\beta}, \sigma; E = 0] \frac{1}{\sigma} \phi\big(\frac{\hat{\beta} - \beta}{\sigma}\big)}{\mathbb{Pr}_1[D_j = 1; E = 0]}$$

$$\propto \mathbb{1}\{|\hat{\beta}|/\sigma \leq 1.96\} \cdot \gamma \frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right) + \mathbb{1}\{|\hat{\beta}|/\sigma > 1.96\} \cdot \frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right)$$

When $\gamma = 0$, the densities in these two regimes are clearly identical. □

For intuition, consider the regime where standard errors are chosen strategically. Strategically choosing not to cluster occurs whenever a study is significant without clustering but insignificant with clustering i.e. $|\hat{\beta}|/\sigma \in (1.96r, 1.96)$. But studies with $|\hat{\beta}|/\sigma \in (1.96r, 1.96)$ would never be published in a clustered regime with publication regime $\gamma = 0$, because they are statistically insignificant with clustered standard errors, irrespective of whether there is strategic clustering or not. Thus, strategic clustering has no impact on the distribution of studies once we condition on statistical significance.

This result provides the basis for an approach to obtaining unbiased estimates of the latent distribution in the presence strategic clustering. We do this by estimating the model with the selected sample of statistically significant clustered studies, $\hat{\beta}_j, \sigma_j | D_j = 1, C_j = 1, |\hat{\beta}_j|/\sigma_j \geq 1.96$, and setting $\gamma = 0$ such that we only estimate $\mu_{\beta,\sigma}$. Normally, the selection function $p(\cdot)$ represents selective publication, but now it reflects the joint selection of the publication process and the econometrician who chooses which results to use for estimation. Since we knowingly condition estimation on significant results, we know that $\gamma = 0$ and do not need to estimate it. In other words, once we condition on the selection of the econometrician, conditioning again by selective publication has no impact since it is also based on statistical significance. Thus, we can recover the latent distribution irrespective of whether or not there is strategic clustering.

1F.3 Robust Maximum Likelihood Estimation

Under the null hypothesis of no strategic clustering, the estimated latent distribution using the full sample, $\hat{\beta}_j, \sigma_j | D_j = 1, C_j = 1$, should be similar to the unbiased estimate with the significant sample, $\hat{\beta}_j, \sigma_j | D_j = 1, C_j = 1, |\hat{\beta}_j|/\sigma_j \geq 1.96$. However, if there is strategic clustering, then then the density of the data is different, the model misspecified, and the estimates for the

Table 1F.1: Robust Maximum Likelihood Estimates

	Latent true effects β_j		Latent standard errors σ_j		Selection
	κ_β	λ_β	κ_σ	λ_σ	γ
Restricted (Robust)	0.205	15.126	1.602	6.039	0.000
	(0.102)	(3.220)	(0.260)	(2.006)	–
Standard	0.154	17.802	1.426	6.475	0.016
	(0.0353)	(2.692)	(0.167)	(1.282)	(0.007)

Notes: Estimation sample is clustered DiD studies over 2000–2009. The number of observations is 66 in the standard model and 60 in the restricted model which only uses statistically significant estimates at the 5% level. Robust standard errors are in parentheses. Latent true treatment effects and standard errors are assumed to follow a gamma distribution with shape and scale parameters (κ, λ). The coefficient γ measures the publication probability of insignificant results at the 5% level relative to significant results.

latent distribution should also be different.[33] Thus if the estimates of the latent distribution are sufficiently different, then we can reject the null of no strategic clustering. Otherwise, we do not reject it.

I apply this test to the DiD sample of clustered studies. The full sample has 66 studies and the restricted sample of significant studies consists of 60 studies. Estimates for the latent distribution of studies are similar for both approaches. For each parameter, the 95% confidence interval of the estimated parameters in the restricted model contains the standard model parameter estimate, and vice versa. This implies that we cannot reject the null hypothesis of endogenous clustering.

1F.4 Bias and Coverage Results with Robust Model

Ultimately, we are interested in how differences in parameter estimates from the robust approach could affect our final conclusions about the impact of clustering on bias and coverage. One concern with the statistical test above is that limited power in the above test prevents us from rejecting the null hypothesis despite differences in parameter estimates that have a meaningful impact on the main results examining the impact of clustering on bias and coverage in Section 1.4. To alleviate these concerns, I perform a robustness exercise where I reproduce the main analysis using parameter estimates from the robust model. This allows us to test the sensitivity

[33] Note that the probability of publishing null results γ must be non-zero, since they appear in the sample.

of the main results to the (statistically insignificant) differences in parameter estimates in Table 1F.1.

To estimate the parameters of the latent distribution, the robust model sets $\gamma = 0$ and therefore does not estimate it. Thus, it is necessary to choose the value of γ to calculate the impact of clustering. For robustness, I choose three different values. The first is setting γ to the same value estimated in the standard model for DiD studies (A). The second is to set $\gamma = 0.037$, which is the value estimated by Andrews and Kasy (2019) for replications in experimental economics (B).[34] Finally, to test sensitivity of the results, I set it to $\gamma = 0.1$, a relatively large value which is 6.25 times larger than the value estimated in DiD studies (C).

Table 1F.2 presents the results. Overall, the conclusion from the 'standard model' that clustering increases coverage by a large amount at the expense of increased bias is maintained across all calibrations of the robust model. This suggests that the main results are unlikely to be driven strategic clustering of the form presented in the model above.

[34] This is based on the meta-study estimation approach which is also used in this article.

Table 1F.2: Results for Model Robust to Strategic Clustering

	Unclustered ($\hat{r} = 0.51$)	Clustered ($r = 1$)	Change
Standard Model ($\hat{\gamma} = 0.016$)			
Coverage	0.28	0.70	0.42
Total Bias ($\mathbb{E}_r[\hat{\beta}_j \mid D_j = 1] - \mathbb{E}_r[\beta_j]$)	3.51 (100%)	10.00 (100%)	6.48 (100%)
Internal-Validity Bias ($\mathbb{E}_r[\hat{\beta}_j - \beta_j \mid D_j = 1]$)	1.23 (34.9%)	2.44 (24.4%)	1.21 (18.7%)
Study-Selection Bias ($\mathbb{E}_r[\beta_j \mid D_j = 1] - \mathbb{E}_r[\beta_j]$)	2.29 (65.1%)	7.56 (75.6%)	5.27 (81.3%)
Robust Model			
A DiD Studies ($\gamma = 0.016$)			
Coverage	0.31	0.72	0.41
Total Bias	4.16 (100%)	10.55 (100%)	6.39 (100%)
Internal-Validity Bias	1.52 (36.5%)	2.94 (27.9%)	1.42 (22.3%)
Study-Selection Bias	2.64 (63.5%)	7.60 (72.1%)	4.96 (77.7%)
B Economics Experiments ($\gamma = 0.037$)			
Coverage	0.33	0.75	0.42
Total Bias	3.96 (100%)	9.22 (100%)	5.26 (100%)
Internal-Validity Bias	1.44 (36.4%)	2.56 (27.8%)	1.12 (21.3%)
Study-Selection Bias	2.52 (63.6%)	6.66 (72.2%)	4.14 (78.7%)
C One-in-Ten Censored ($\gamma = 0.1$)			
Coverage	0.38	0.81	0.43
Total Bias	3.46 (100%)	6.70 (100%)	3.24 (100%)
Internal-Validity Bias	1.24 (35.8%)	1.83 (27.3%)	0.59 (18.2%)
Study-Selection Bias	2.22 (64.2%)	4.87 (72.7%)	2.65 (81.8%)

Notes: The 'standard model' results are reprinted from the main text. The remaining results under 'Robust Model' are based on the procedure outlined in Appendix 1F, for different values of γ, which measures the level of publication bias against insignificant results at the 5% level. Figures are calculated by simulating published studies under unclustered and clustered regimes.

1G Impact of Clustering for Different Sized Corrections

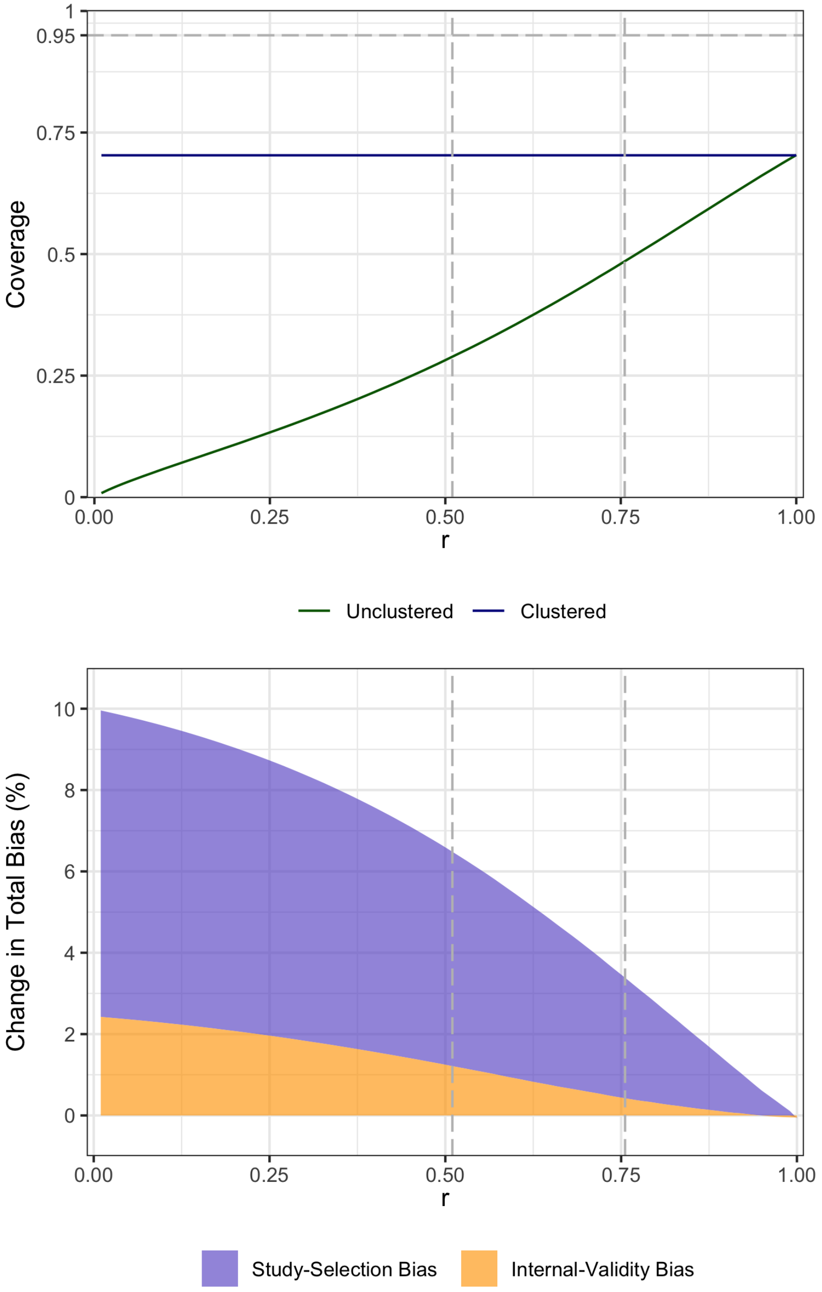

Figure 1G.1: Results on the Impact of Clustering for Different Values of r

Notes: Change in coverage, total bias (and estimated treatment effects), study-selection bias, and internal-validity bias for the estimated model parameters in Table 1.3 as a function of downward bias in unclustered standard errors r. The vertical dashed line at $\hat{r} = 0.51$ represents the calibrated value using the method of simulated moments. The vertical dashed line at $\hat{r} = 0.76$ represents the mean of the empirical distribution of r from 2015–2018 DiD studies.

1H Impact of Clustering on Bias and Coverage Using the 2015–2018 Empirical Distribution of r

Table 1H.1: Impact of Clustering Based on 2015–2018 Empirical Distribution of r

	Unclustered	Clustered ($r = 1$)	Change
Random draws of r			
Coverage	0.36	0.70	0.34
Total Bias	4.67 (100%)	10.00 (100%)	5.32 (100%)
Internal-Validity Bias	1.38 (29.5%)	2.44 (24.5%)	1.07 (20%)
Study-Selection Bias	3.29 (70.5%)	7.55 (75.5%)	4.26 (80%)
Mean: $\hat{r} = 0.76$			
Coverage	0.49	0.70	0.21
Total Bias	6.67 (100%)	10.00 (100%)	3.32 (100%)
Internal-Validity Bias	2.03 (30.4%)	2.44 (24.4%)	0.41 (12.3%)
Study-Selection Bias	4.64 (69.6%)	7.56 (75.6%)	2.91 (87.7%)

Notes: These figures are based on the parameter estimates of the empirical model in Table 1.3. Figures are calculated by simulating published studies under unclustered and clustered regimes. In the unclustered regime, the degree of bias in unclustered studies is based on the empirical distribution of r from 2015–2018 studies. Panel A shows results based on drawing different values of r from the empirical distribution for unclustered studies. Panel B assumes that all unclustered studies are downward biased by a constant factor equal to the mean of the empirical distribution ($\hat{r} = 0.76$).

Chapter 2

Why Are Replication Rates So Low?

Abstract. Many explanations have been offered for why replication rates are low in the social sciences, including selective publication, p-hacking, and treatment effect heterogeneity. This article emphasizes that issues with common power calculations in replication studies may also play an important role. Theoretically, I show in a simple model of the publication process that issues with the way that replication power is commonly calculated imply we should always expect replication rates to fall below their intended power targets, even when original studies are unbiased and there is no p-hacking or treatment effect heterogeneity. Empirically, I find that a parsimonious model accounting only for issues with power calculations can fully explain observed replication rates in experimental economics and social science, and two-thirds of the replication gap in psychology.

2.1 Introduction

In a 2016 survey conducted by *Nature*, 90% of researchers across various fields agreed that the scientific community faces a 'reproducibility crisis' (Baker, 2016). Growing consensus has been supported by high-profile replication projects which find that the replication rate – i.e. the fraction of replications that are significant with the same sign as the original study – is just 36% in psychology, 61% in experimental economics, and 62% in experimental social science (Open Science Collaboration, 2015; Camerer et al., 2016, 2018).

Understanding the underlying cause of low replication rates is important for researchers and reformers aiming to improve the credibility of published research. There is a large literature examining a wide range of explanations, including selective publication against null results (Franco et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018); p-hacking and other questionable research practices (Ioannidis, 2005, 2008; Simonsohn et al., 2014; Brodeur et al., 2016, 2020, 2022; Elliott et al., 2022); and heterogeneity across original studies and replications in research design and experimental subjects (Higgins and Thompson, 2002; Cesario, 2014; Simons, 2014; Stanley et al., 2018; Bryan et al., 2019).

In this article, the main theoretical result shows that we should expect the replication rate to fall short of its intended target, owing to issues with the common approach of setting power in replications. This is true regardless of whether or not there is selective publication, and even in 'ideal' conditions with no p-hacking, no heterogeneity, and relatively high statistical power in original studies (e.g. 80%). Let $RP(x, \sigma_r|\theta)$ be the probability of successfully replicating a study with observed original effect size x and replication standard error σ_r conditional on unobserved true effect θ. Replicators commonly set the replication standard error (or equivalently the replication sample size) as a function of the observed effect size x, such that $RP(x, \sigma_r(x)|\theta)$ equals a pre-specified intended power target $1 - \beta$ when $x = \theta$ (e.g. $1 - \beta = 0.9$ would correspond to 90% intended power target, where β is the target probability of Type II error). This approach was used, for example, in large-scale replication studies in psychology and economics (Open Science Collaboration, 2015; Camerer et al., 2016), and a survey of replications in the psychology literature by Anderson and Maxwell (2017) shows that it is the most commonly implemented approach. In practice, replication rates consistently fall below the intended power target $1-\beta$, which is commonly interpreted as an indicator that original effects are biased due to factors such as selective publication, p-hacking, or treatment effect heterogeneity. However, this article highlights that the replication function $RP(\cdot|\theta)$ is a non-linear, locally concave function. Thus, even if original estimates were unbiased, with $\mathbb{E}_{X|\Theta}[X|\theta] = \theta$, by Jensen's inequality we have that $\mathbb{E}_{X|\Theta}[RP(X, \sigma_r(X)|\theta)|\theta] < RP(\theta, \sigma_r(\theta)|\theta) = 1 - \beta$. That is, stated replication rate targets in large-scale replication studies using the approach described above do not provide an attainable benchmark against which to judge replication rates observed in practice; even if original studies were unbiased, such targets are not in fact reachable in expectation. I also show that the gap between the expected replication rate and its intended power target is larger when the original published studies have low power, a problem that we expect to be severe in practice given evidence of low power in various empirical literatures from (Button et al., 2013; Ioannidis et al., 2017; Stanley et al., 2018; Arel-Bundock et al., 2023).

The main theoretical result applies to studies using what I refer to as the common power rule, which sets replication power to detect the original estimated effect size. More recently, some studies have begun to use a higher-power variant which I refer to as the fractional power rule, wherein replication power is set to detect some fraction of the estimated effect size. Building on results in Andrews and Kasy (2019), I show that the expected replication rate using the fractional power rule can be either above or below the stated power target.

To what extent can these theoretical insights explain the low replication rates actually observed in large-scale replication studies? Although the theory predicts that the actual replication rate will always fall below the target when using the common power rule, the magnitude of this gap is an empirical question. Likewise, for replication studies using the fractional power rule, both the sign and the magnitude of the gap is an empirical question.

To evaluate the importance of power issues in practice, I therefore empirically investigate the results of three replication studies, two of which use the common power rule (Open Science Collaboration, 2015; Camerer et al., 2016) and one of which uses the fractional power rule (Camerer et al., 2018). In each application, I estimate the empirical model in Andrews and Kasy (2019) using a 'metastudy approach' that corrects for publication bias to obtain the underlying distribution of latent studies prior to screening by the publication process.[35] I then use the estimated latent distribution of studies to simulate what we should expect the replication

[35] It is necessary to model publication bias to estimate the latent distribution of studies. However, for a given latent distribution of studies, the replication rate itself does not depend on the degree to which selective publication suppresses insignificant results (Andrews and Kasy, 2019; Kasy, 2021). This is for the simple reason that the replication rate only includes significant results in its definition. See Section I.B below for additional discussion.

rate to be based on the power calculations actually implemented in replications. Importantly, the model and its predictions are based only on data from original studies and assume away researcher manipulation and heterogeneous treatment effects. The empirical exercise asks, in effect, whether observed replication rates could have been predicted by issues with power alone, before the replication studies themselves were actually undertaken and in a parsimonious model without treatment effect heterogeneity or p-hacking.

I find that the predicted replication rate is almost identical to observed replication rates in experimental economics (60% vs. 61%) and experimental social science (54% vs. 57%). Replications in experimental economics implemented the common power rule, while those in experimental social science used a fractional power rule.[36] These empirical results are consistent with the null hypothesis that observed replication rates in these studies are driven entirely by issues with power calculations, rather than other issues such as p-hacking or treatment effect heterogeneity. Of course, failure to reject a hypothesis does not mean that it is true, and thus we should not necessarily conclude that these other factors are not present in these settings. Nevertheless, other evidence has also suggested a relatively limited role for p-hacking in the context of lab experiments studied here (Brodeur et al., 2016, 2020; Imai et al., 2020).

In psychology, the predicted replication rate is 55%, whereas the observed replication rate is 35%. Since the intended power target was 92%, issues with power calculations explain only two-thirds of the gap in psychology. In the case of psychology, we can therefore reject the null that the replication gap is entirely explained by issues with power calculations. This provides strong evidence that some other factors are important in psychology. Some possibilities discussed in the literature include heterogeneous treatment effects, p-hacking, and differences across subfields.

In an extension, I examine the relative effect size (defined as the mean of the ratio of the replication effect size and the original effect size), a common complementary continuous

[36] In the experimental social science replications (Camerer et al., 2018), replicators used a fractional power rule in the first stage of replications predicted here, where replication power was set to detect 75% of the original effect size with 90% intended power.

measure of replication. I generate relative effect size predictions in each field using a similar method as for the replication rate. I once again find that the predictions are quite similar to observed outcomes in economics (0.70 vs. 0.66). The model is somewhat farther off for social sciences (0.53 vs. 0.44), perhaps suggesting some role for other factors, although the difference is not statistically distinguishable from zero. In psychology, predictions are quite far off (0.64 vs. 0.37), again providing strong evidence for alternative factors.

This article contributes to the large metascience literature and the growing literature on predicting research outcomes (Ioannidis, 2005; Franco et al., 2014; Gelman and Carlin, 2014; Dreber et al., 2015; Maxwell et al., 2015; Anderson and Maxwell, 2017; Stanley et al., 2018; Miguel and Christensen, 2018; Altmejd et al., 2019; Amrhein et al., 2019a; DellaVigna et al., 2020; Gordon et al., 2020; Frankel and Kasy, 2022; DellaVigna and Linos, 2022; Nosek et al., 2022). Andrews and Kasy (2019) and Kasy (2021) provide stylized examples showing that the replication rate can vary widely depending on the latent distribution of studies (i.e. the joint distribution of true effects and standard errors for published and unpublished studies). Theoretically, this article builds on this observation by establishing that the expected replication rate is bounded above by its nominal target owing to issues with common power calculations in replication studies. This result holds for any distribution of latent studies. Empirically, I provide evidence that among the profusion of explanations for low replication rates, a parsimonious model accounting only for issues with replication power calculations and low power in original studies can adequately account for observed replication rates in experimental economics and social science.

2.2 Theory

2.2.1 Model of Large-Scale Replication Studies

I consider the model in Andrews and Kasy (2019). Suppose a large-scale replication study is conducted in an empirical literature of interest and we observe the estimated effect sizes and standard errors for original studies and their replications. Let upper case letters denote random variables, lower case letters realizations. Latent studies (published or unpublished) have a superscript * and published studies have no superscript. The model of the DGP has five steps:

1. **Draw a population parameter and standard error:** Draw a research question with population parameter (Θ^*) and standard error (Σ^*):

   $$(\Theta^*, \Sigma^*) \sim \mu_{\Theta,\Sigma}$$

   where $\mu_{\Theta,\Sigma}$ is the joint distribution of latent true effects and latent standard errors.

2. **Estimate the effect:** Draw an estimated effect from a normal distribution with parameters from Stage 1:

   $$X^*|\Theta^*, \Sigma^* \sim N(\Theta^*, \Sigma^{*2})$$

3. **Publication selection:** Selective publication is modelled by the function $p(\cdot)$, which returns the probability of publication for any given t-ratio. Let D be a Bernoulli random variable equal to 1 if the study is published and 0 otherwise:

   $$\mathbb{P}(D = 1|X^*/\Sigma^*) = p\left(\frac{X^*}{\Sigma^*}\right) \tag{31}$$

4. **Replication selection:** Replications are sampled from published studies (X, Σ, Θ) (i.e. latent studies $(X^*, \Sigma^*, \Theta^*)$ conditional on publication $(D = 1)$). Replication selection is modelled by the function $r(\cdot)$, which returns the probability of being chosen for replication for any given t-ratio. Let R be a Bernoulli random variable equal to 1 if the study is chosen for replication and 0 otherwise:

   $$\mathbb{P}(R = 1|X/\Sigma) = r\left(\frac{X}{\Sigma}\right) \tag{32}$$

5. **Replication:** A replication draw is made with:

$$X_r|\Theta, X, \Sigma, \Sigma_r, D = 1, R = 1 \sim N\Big(\Theta, \Sigma_r^2\Big)$$

We observe i.i.d draws of $\big(X, \Sigma, X_r, \Sigma_r\big)$ from the conditional distribution of $\big(X^*, \Sigma^*, X_r, \Sigma_r\big)$ given $D = 1$ and $R = 1$. I consider what happens in the Andrews and Kasy (2019) model outlined above when the replication standard error, Σ_r, is set to detect the original estimate X with a pre-specified power level $1 - \beta$, where β is the target probability of Type II error. This approach is implemented, for example, in Open Science Collaboration (2015) and Camerer et al. (2016), and a survey of replications the psychology literature by Anderson and Maxwell (2017) shows that it is the most commonly implemented approach. I refer to this as the common power rule, which is formalized as follows:

DEFINITION 1 (Common power rule). *The common power rule to detect original effect size x with intended power $1 - \beta$ sets the replication standard error to*

$$\sigma_r(x, \beta) = \frac{|x|}{1.96 - \Phi^{-1}(\beta)} \tag{33}$$

This is equivalent to setting the replication sample size to $N \times \left[\frac{\sigma}{|x|}\big(1.96 - \Phi^{-1}(\beta)\big)\right]^2$, where N and σ are the original study's sample size and standard deviation, respectively.

The justification for the common power rule is that the power in any given replication study will equal its intended power target of $1 - \beta$ when $x = \theta$.[37] In practice, replication rates consistently fall below this benchmark, which is typically taken as evidence that original estimates are biased because of selective publication or p-hacking. While this argument has intuitive appeal, it does not account for the fact that replication power is a non-linear function of the random original estimate X; thus, even if $\mathbb{E}[X|\Theta = \theta] = \theta$, the replication probability evaluated at the expectation (which equals the intended target) will not, in general, be equal

[37] For a formal statement and proof, see Lemma B1.

to the expected replication rate.

This argument is developed more formally in the following section, under a number of regularity conditions and assumptions imposed on the DGP. First, following Andrews and Kasy (2019), we impose the normalization that true effects are positive:[38]

ASSUMPTION 1 (True effect normalization). *The support of* Θ *is a subset of the non-negative real line.*

Second, we impose that the publication probability $p(\cdot)$ is weakly increasing in the t-ratio and symmetric around zero:

ASSUMPTION 2 (Publication selection function). *Let* $p(t) \neq 0$ *for all* $|t| \geq 1.96$*,* $p(t)$ *be weakly increasing for all* $t \geq 1.96$*, and* $p(t) = p(-t)$ *for all* $t \geq 1.96$*. Allow* $p(\cdot)$ *to take any form when* $t \in (-1.96, 1.96)$*.*

This allows for very general forms of publication bias (or lack thereof). Third, in step 4, which models the replication selection mechanism, we assume that the set of significant results chosen for replication is a random sample from published, significant results:

ASSUMPTION 3 (Replication selection function). *For all* $|t| \geq 1.96$*, let* $r(t) = c \in (0, 1]$ *and allow* $r(\cdot)$ *to take any form when* $t \in (-1.96, 1.96)$*.*

Finally, note that the article uses three distinct concepts of statistical power. First, power in an original study is defined as the probability of obtaining a statistically significant estimate in the same direction as the true effect.[39] Second, power in a replication study (or the 'replication probability') is defined as the probability of obtaining a significant effect with the same sign as the original study (Definition 2 below), and will depend on the rule for setting replication power (e.g. the common power rule). Finally, the intended power target of a given rule for setting replication power, which we denote by $1 - \beta$.

[38] Large-scale replications include studies that examine different questions and outcomes. Normalizing true effects to be positive is justified because relative signs across studies are arbitrary.

[39] The arguments made throughout are essentially unchanged if we consider the alternative definition of obtaining a statistically significant estimate irrespective of the sign.

2.2.2 Common Power Calculations and Low Replication Rates

This section defines the replication rate and then discusses the main result. First, we define the replication probability of a single study and then use this to define the expected replication rate over multiple studies.

DEFINITION 2 (Replication probability of a single study). *The replication probability of a published study* (X, Σ, Θ) *chosen for replication* $(R = 1)$ *is*

$$RP\Big(X, \Theta, \sigma_r(X, \beta)\Big) = \mathbb{P}\left(\frac{|X_r|}{\sigma_r(X, \Sigma, \beta)} \geq 1.96, \text{sign}(X_r) = \text{sign}(X)\Big| X, \Theta, \beta, R = 1\right) \tag{34}$$

This definition captures the dual requirement that the replication estimate is statistically significant and has the same sign as the original study.

DEFINITION 3 (Expected replication rate). *The expected replication probability is defined over published studies* (X, Σ, Θ) *which are chosen for replication* $(R = 1)$ *and statistically significant* $(S_X = 1)$*. It is equal to*

$$\mathbb{E}\Big[RP\big(X, \Theta, \sigma_r(X, \beta)\big)\big|R = 1, S_X = 1\Big] \tag{35}$$

Substituting the common power rule in Definition 1 for the replication standard error gives the expected replication rate under the common power rule. Note that while insignificant results may be replicated, they are not included in the replication rate in Definition 3, in line with the main definition reported in most large-scale replication studies (Klein et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Klein et al., 2018).[40] With this, we can state the main theoretical result:

[40] Replication power calculations themselves are typically designed with this definition in mind. Complementary replication measures include: the relative effect size; whether the 95% confidence interval of the replication covers the original estimate; replication based on meta-analytic estimates; the 95% prediction interval approach (Patil et al., 2016); the 'small telescopes' approach (Simonsohn, 2015); and the one-sided default Bayes factor (Wagenmakers et al., 2016).

PROPOSITION 1 (The common power rule implies the expected replication rate is below its target.) *Consider the model in I.A. Under assumptions 1, 2, and 3, if replication standard errors are set by the common power rule to detect original estimates with intended power* $1-\beta \geq 0.8314$*, then*

$$\mathbb{E}\Big[RP\big(X, \Theta, \sigma_r(X, \beta)\big)\big|R = 1, S_X = 1\Big] < 1 - \beta \tag{36}$$

From a practical perspective, Proposition 1 means that replicators who set the replication sample size to detect original effect sizes should not expect the replication rate to reach its intended target, regardless of whether or not there is selective publication, and even under 'ideal' conditions with no researcher manipulation, replications with identical designs and comparable samples (i.e. no heterogeneity in true effects), no measurement error, random sampling in replication selection, and high-powered original studies. That the intended target is not in fact attainable in expectation underscores fundamental difficulties in interpreting replication rate gaps observed in large-scale replication studies.

Figure 2.1 provides intuition for this result. It plots the replication probability of a single study in Definition 2 as a function of the original effect X, for a fixed true effect θ and assuming that the common power rule is applied with an intended power target of $1 - \beta = 0.9$. Denote this conditional replication probability function as $RP\big(X, \sigma_r(X, \beta)\big|\theta)$. It is clear that $RP\big(X, \sigma_r(X, \beta)\big|\theta)$ is non-linear in X, which implies that $\mathbb{E}_{X|\Theta}\big[RP\big(X, \sigma_r(X, \beta)|\theta\big)\big] \neq RP\big(\mathbb{E}_{X|\Theta}[X|\theta], \sigma_r(\mathbb{E}_{X|\Theta}[X|\theta], \beta)|\theta\big)$, even if X is unbiased. If $RP(\cdot|\theta)$ were globally concave, Proposition 1 would immediately follow from Jensen's inequality. However, it is only locally concave around the true effect θ. The proof of Proposition 1 shows that when $1 - \beta > 0.8314$, local concavity is sufficient to arrive at the same result for any distribution of latent studies.

The difference between the expected replication rate and its intended target is larger when power in original studies is low. This is because the concavity of $RP(\cdot|\theta)$ is more pronounced when power in original studies is low. As an illustration, Figure 2.2 plots the relationship

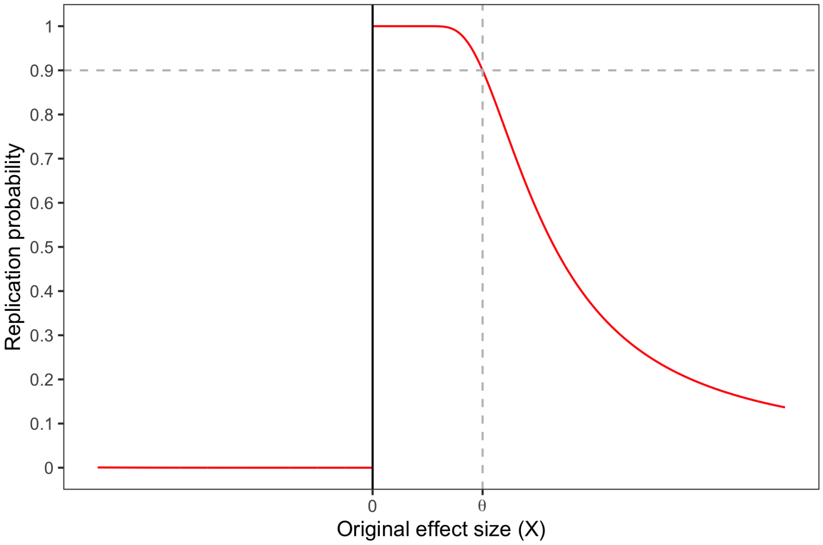

Figure 2.1: Replication Probability Function Conditional on Θ

Notes: Replication probability function in Definition 2 conditional on a fixed θ. The replication standard error is calculated using the common power rule in Definition 1 to detect original effect sizes with 90% power (i.e. $\sigma_r(X, \beta) = |X|/3.242$).

between the expected replication rate and power in original studies, again assuming the intended power target in replications is set to 90%, close to mean reported intended replication power in Open Science Collaboration (2015) and Camerer et al. (2016). To highlight the impact of power in original studies, the relationship is derived assuming no p-hacking, no selective publication, and no heterogeneity (i.e. assuming exact replications). The plot shows that the expected replication rate is bounded above by its intended target of 90%, in line with Proposition 1, and is especially low when power in original studies is low. For instance, the expected probability of replicating an original study with 33% power is around 50%. With relatively low estimates of power across various empirical literatures, this provides strong theoretical grounds for expecting low replication rates in practice, even in the absence of issues with p-hacking or treatment effect heterogeneity. For intuition, note that if the true effect is zero, the replication probability is 0.025 (regardless of the how the replication standard error is chosen). Continuity implies that when original studies have true effects close to zero (and therefore power in original studies is low), replication probabilities will also be very low.

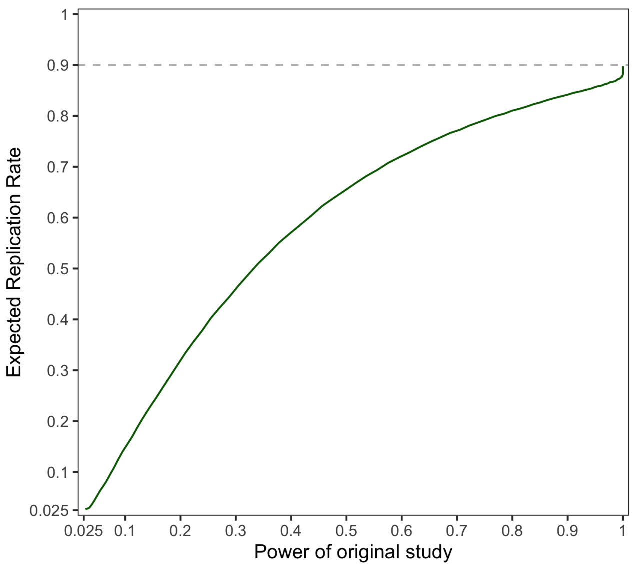

Figure 2.2: Original Power and the Expected Replication Rate Under the Common Power Rule

Notes: Power of original studies and the expected replication rate under the common power rule are both functions of $\omega = \theta/\sigma$ (normalized to be positive). Power is original studies to obtain a significant effect with the same sign as the true effect is equal to $1 - \Phi(1.96 - \omega)$. The expected replication rate is calculated by taking 10^6 draws of Z from $N(\omega, 1)$ and then calculating $10^{-6}\sum_{i=1}^{10^6}\left[1 - \Phi\big(1.96 - \text{sign}(z_i)\frac{\omega}{\sigma_r(z_i,\beta)}\big)\right]$, with intended power equal to $1 - \beta = 0.9$ and depicted by the horizontal dashed line. The replication standard error is calculated using the common power rule (Definition 1) to detect original effect sizes with 90% power, which is given by $\sigma_r(z_i, \beta) = |z_i|/3.242$. This figure assumes no p-hacking, no heterogeneity in true effects, no selective publication and random replication selection. Further details are provided in Section 2.2.

Two other factors affect the replication rate, although empirically their impact turns out to be relatively small. First, as can be seen in Figure 2.1, when original estimates are significant but with the 'wrong' sign, the probability of replication is very low (below 0.025) because it requires the highly unlikely event that the replication estimate also has the wrong sign and is statistically significant. Second, the replication rate induces upward bias in original estimates because it is, by definition, calculated on a selected sample of significant findings. Replication estimates will regress to the mean (Galton, 1886; Hotelling, 1933; Barnett et al., 2004; Kahneman, 2011)[41], although the ultimate impact on the replication rate is ambiguous because conditioning on significance also tends to select larger true effects, which have higher replication probabilities. Appendix C derives and estimates a decomposition of the replication

[41] For a formal statement and proof, see Proposition B1 in Appendix B.

rate gap in economics and psychology, using the empirical methodology described in the next section, and finds that it is almost entirely explained by the concavity of $RP(\cdot)$.

Proposition 1 applies to replications implementing the common power rule. Some more recent replication studies have used a higher-power variant which I refer to as the fractional power rule, wherein replication power is set to detect some fraction ψ of the estimated effect size (Camerer et al., 2018, 2022). In Proposition B2 in Appendix B, I show that the expected replication rate under the fractional power rule can be either above or below the stated power target $1-\beta$. More specifically, the expected replication rate can range anywhere between 0.025 and $\Phi[1.96 - \frac{1}{\psi}\big(1.96 - \Phi^{-1}(\beta)\big)] > 1-\beta$ depending on the statistical power of original studies. For instance, if $\psi = \frac{3}{4}$ and $1-\beta = 0.9$, as in the first-stage in Camerer et al. (2018), then the expected replication rate could range anywhere between 0.025 and 0.99. These results build on those in Andrews and Kasy (2019), who argue that replication rates may vary widely depending on the latent distribution of studies. Finally, note that as with Proposition 1, these conclusions hold whether or not there is selective publication, and even in the absence of p-hacking or treatment effect heterogeneity.

Finally, a common perception is that selective publication favouring significant results – either by authors or journals – produces more 'false-positives' in the published literature, which are in turn harder to replicate. This theory is important to address because it enjoys substantial support: over 90% of researchers cite 'selective reporting' as a contributing factor to irreproducibility, more than any other factor (Baker, 2016). However, Andrews and Kasy (2019) and Kasy (2021) show that the replication rate in fact tells us very little about selective publication. Both provide examples showing that the replication rate can take on almost any value depending on the latent distribution of true effects, irrespective of how selective publication is. In fact, the replication rate in the Andrews and Kasy (2019) model is completely insensitive to selective publication against null results.[42] This follows from the simple fact that the replication rate

[42] For a formal statement, see Proposition B3 in Appendix B, which proves this more generally for measures $g(\cdot)$ that condition on statistical significance. Setting $g(x,\sigma,x_r,\beta) = \mathbb{1}\Big[\frac{|x_r|}{\sigma_r(x,\sigma,\beta)} \geq 1.96, \text{sign}(x_r) = \text{sign}(x)\Big]$ gives the result for the replication rate measure.

definition does not include statistically insignificant results. Thus, even if insignificant results were being widely published, they would not be included in the replication rate.[43][44]

2.3 Empirical Applications

In this section, I test the null hypothesis that observed replication rates can be entirely explained by issues with common power calculations emphasized in Proposition 1, rather than other issues such as p-hacking or heterogeneity. To test this hypothesis, the theory requires that we estimate the latent distribution of studies. This can then be used to generate replication rate predictions which can be compared to observed replication rates. The procedure is as follows:

1. Estimate the latent distribution of studies, $\mu_{\Theta,\Sigma}$ using an augmented version of the Andrews and Kasy (2019) model applied to three large-scale replications.[45] Estimation does not use any data from replications.

2. Use the estimated model to simulate replications and predict what fraction of significant results would replicate, absent any other issues such as p-hacking or heterogeneity.

3. Compare these predictions (which do not use any data from the replications) to actual replication outcomes.

[43] A caveat is that the model assumes a fixed distribution of latent studies, whereas in practice it may be endogenous, for example, if researchers engage in more specification searching when publication bias against null results is high (Simonsohn et al., 2014; Brodeur et al., 2016, 2020, 2022).

[44] Appendix D examines measures of replication which may be more sensitive to changes in selective publication than the replication rate. For evaluating efforts to reduce selective publication, simulation results show that the prediction interval approach (Patil et al., 2016), when calculated over both significant and insignificant results, may provide a useful alternative to the replication rate, the confidence interval measure, and the meta-analysis approach.

[45] Note that estimating the latent distribution of studies requires modelling selective publication, as discussed in the model in Section I.A. However, with estimates of the latent distribution in hand, replication rate predictions in step 2 will not depend on the degree to which null results are suppressed, since the replication rate is defined only over significant results.

2.3.1 Replication Studies

I examine three replication studies. Camerer et al. (2016) replicate results from all 18 between subjects laboratory experiments published in *American Economic Review* and *Quarterly Journal of Economics* between 2011 and 2014. Open Science Collaboration (2015) replicate results from 100 psychology studies in 2008 from *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition.* Following Andrews and Kasy (2019), I consider a subsample of 73 studies with test statistics that are well-approximated by z-statistics. Camerer et al. (2018) replicate 21 experimental studies in the social sciences published between 2010 and 2015 in *Science* and *Nature.*

In Camerer et al. (2016), replicators used the common power rule to detect original effects with at least 90% power. In Open Science Collaboration (2015), replication teams were instructed to achieve at least 80% power using the common power rule, and encouraged to obtain higher power if feasible. Reported mean intended power was 92% in both cases. Camerer et al. (2018) implemented a higher-powered fractional power rule consisting of two stages. In the first stage, replicators aimed to detect 75% of the original effect with 90% power. In the second stage, further data collection was undertaken for insignificant results from the first stage, such that the pooled sample from both stages was calibrated to detect half of the original effect size with 90% power. I predict replication outcomes in the first stage.[46]

Note that the theoretical result in Proposition 1 showing that the expected replication rate is bounded above by its intended target applies to the common power rule and not to the fractional power rule. For the fractional power rule, the expected replication rate can either above or below the stated power target. In both cases, the magnitude of the gap is an empirical question.

[46] Predicting second-stage outcomes is complicated by the fact that one study that was ‘successfully’ replicated in the first stage was erroneously included in the second stage.

2.3.2 Estimation

To calculate the expected replication rate, it is necessary to estimate the latent distribution of studies $\mu_{\Theta,\Sigma}$. To do this, I estimate an augmented version of the empirical model in Andrews and Kasy (2019). Specifically, Andrews and Kasy (2019) develop an empirical model to estimate the marginal distribution of true effects Θ^*, but not of standard errors Σ^*. Since predictions of the replication rate also require knowledge of the distribution of Σ^*, I augment the model to estimate the joint distribution of (Θ^*, Σ^*). Estimation is based on the 'metastudy approach', which only uses data from original studies. Identification requires that true effects are statistically independent of standard errors, a common assumption in meta-analyses. I assume that Σ^* follows a gamma distribution with shape and scale parameters denoted by κ_σ and λ_σ, respectively.

For all other aspects of the model, I implement identical model specifications as Andrews and Kasy (2019), whose focus is on estimating publication bias. Matching their specifications, I assume that $|\Theta^*|$ follows a gamma distribution with shape and scale parameters $(\kappa_\theta, \lambda_\theta)$; and that the joint probability of being published and chosen for replication, $p(X/\Sigma) \times r(X/\Sigma)$, is a step-function parameterized by $\beta_{\mathbf{p}}$. The inclusion of steps at common significance levels $(1.64, 1.96, 2.58)$ varies slightly across applications owing to different approaches for choosing which studies to replicate.[47] Table 2.1 presents the maximum likelihood estimates together with reproduced estimates from Andrews and Kasy (2019) for comparison.[48] For common parameters, estimates are very close.

[47] Details on mechanisms for replication selection are outlined in Appendix E. With $Z = X/\Sigma$, the selection functions in each application are: $r(X/\Sigma) \times p(X/\Sigma) \propto \mathbb{1}\big(1.64 \leq |Z| < 1.96\big)\beta_{p2} + \mathbb{1}\big(|Z| \geq 1.96\big)$ in economics; $r(X/\Sigma) \times p(X/\Sigma) \propto \mathbb{1}\big(|Z| < 1.64\big)\beta_{p1} + \mathbb{1}\big(1.64 \leq |Z| < 1.96\big)\beta_{p2} + \mathbb{1}\big(|Z| \geq 1.96\big)$ in psychology; and $r(X/\Sigma) \times p(X/\Sigma) \propto \mathbb{1}\big(1.96 \leq |Z| < 2.58\big)\beta_{p3} + \mathbb{1}\big(|Z| \geq 2.58\big)$ for social science experiments. Separate identification of the publication probability function, $p()$, requires that we specify the replication selection function $r()$.

[48] Estimates for psychology in this article are slightly different to the meta-study estimates reported in Andrews and Kasy (2019) (their Table 2). The difference is due to a misreported p-value in the raw psychology data for one study, which leads to an erroneous outlier in the distribution of original study standard errors. Table 2.1 in this article reproduces estimates of their model with the corrected data. Excluding this study in the augmented model leads to very similar replication rate predictions.

Table 2.1: Maximum Likelihood Estimates

	Latent true effects Θ^*		Latent standard errors Σ^*		Selection parameters		
	κ_θ	λ_θ	κ_σ	λ_σ	β_{p1}	β_{p2}	β_{p3}
Economics experiments							
Augmented model	1.426	0.148	2.735	0.103	0.000	0.039	–
	(1.282)	(0.072)	(0.536)	(0.031)	(0.000)	(0.05)	–
Andrews and Kasy (2019)	1.343	0.157	–	–	0.000	0.038	–
	(1.285)	(0.075)	–	–	(0.000)	(0.05)	–
Psychology experiments							
Augmented model	0.782	0.179	4.698	0.044	0.012	0.303	–
	(0.423)	(0.055)	(0.605)	(0.008)	(0.007)	(0.134)	–
Andrews and Kasy (2019)	0.734	0.185	–	–	0.012	0.300	–
	(0.405)	(0.056)	–	–	(0.007)	(0.134)	–
Social science experiments							
Augmented model	0.077	0.644	6.249	0.028	0.000	0.000	0.611
	(0.106)	(0.333)	(1.762)	(0.009)	(0.000)	(0.000)	(0.427)
	(0.091)	(0.326)	(1.754)	(0.009)	(0.000)	(0.000)	(0.419)
Andrews and Kasy (2019)	0.070	0.663	–	–	0.000	0.000	0.583
	(0.091)	(0.327)	–	–	(0.000)	(0.000)	(0.418)

Notes: Maximum likelihood estimates for economics (Camerer et al., 2016), psychology (Open Science Collaboration, 2015) and social sciences (Camerer et al., 2018). Robust standard errors are in parentheses. Latent true effects and standard errors are assumed to follow a gamma distribution; parameters (κ, λ) are the shape and scale parameters, respectively. In economics and psychology, joint publication and replication probability coefficients are measured relative to the omitted category of studies significant at 5 percent level. Parameters β_{p1}, β_{p2} in this case are the relative publication probabilities of studies that are insignificant at the 10% level; and significant at the 10% level but not at the 5% level. For example, in experimental economics, an estimate of $\beta_{p2} = 0.039$ implies that results which are significant at the 5% level are about 26 times more likely to be published and chosen for replication than results that are significant at the 5% level. Note that in economics, results which were insignificant at thew 10% level were not selected for replication and hence $\beta_{p1} = 0$. In social sciences, the omitted category is studies significant at the 1% level. Results below the 5% significance level were not chosen for replication so that $\beta_{p1} = \beta_{p2} = 0$, and β_{p3} measures the publication probability of a result that is significant at the 5% level but not at the 1% level, relative to that of a a significant result at the 1% level. Andrews and Kasy (2019) estimates are reproduced from accessible data and code from their analysis.

2.3.3 The Predicted Replication Rate

Model parameters estimates in Table 2.1 can be used to generate replication rate predictions by simulating replications using the following procedure:

1. Draw 10^6 latent (published or unpublished) research questions and standard errors $(\theta^{*sim}, \sigma^{*sim})$ from the estimated joint distribution $\hat{\mu}_{\Theta,\Sigma}(\hat{\kappa}_\theta, \hat{\lambda}_\theta, \hat{\kappa}_\sigma, \hat{\lambda}_\sigma)$.

2. Draw estimated effects $x^{*sim}|\theta^{*sim}, \sigma^{*sim} \sim N(\theta^{*sim}, \sigma^{*sim2})$ for each latent study.

3. Use the estimated selection parameters $\hat{\beta}_{\mathbf{p}}$ to determine the subset of studies that are published and chosen for replication.

4. For studies chosen for replication, calculate the replication standard error σ_r^{sim} according to the following rule

$$\sigma_r^{sim}(x^{sim}, \beta, \psi) = \frac{\psi \cdot |x^{sim}|}{1.96 - \Phi^{-1}(\beta)} \tag{37}$$

   where $\psi = 1$ and $1 - \beta = 0.92$ in economics and psychology, which corresponds to the common power rule; and $\psi = \frac{3}{4}$ and $1 - \beta = 0.9$ in social science experiments, which corresponds to a fractional power rule.[49]

5. Simulate replications by drawing replication effect sizes $x_r^{sim}|\theta^{sim}, \sigma_r^{sim} \sim N(\theta^{sim}, \sigma_r^{sim2})$

Let $\{x_i, \sigma_i, x_{r,i}, \sigma_{r,i}\}_{i=1}^{M_{sig}}$ be the (simulated) set of published, replicated original studies that are significant at the 5% level, and their corresponding replication results.[50] M_{sig} is the number of replicated originally-significant studies. The predicted replication rate is equal to

$$\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \mathbb{1}\Big(|x_{r,i}| \geq 1.96\sigma_{r,i}, \text{sign}(x_{r,i}) = \text{sign}(x_i)\Big) \tag{38}$$

2.3.4 Results

In experimental economics, the predicted replication rate is 60%, which is very close to the observed rate of 61.1% (Table 2.2). This is an "out-of-sample" prediction in the sense that the

[49] This assumes all simulated replications set intended power equal to the mean of reported intended power. In practice, there was some variation in the application of the power rule around the mean. Appendix F reports predicted replication rates allowing for variation in intended power across studies that matches the empirical variation in each application. Results are very similar and in fact slightly more accurate in all three applications (61.5% in economics; 52.2% in psychology; and 55.5% in social science).

[50] In both experimental economics and psychology, a small number of original results whose p-values were slightly above 0.05 were treated as 'positive' results and included in the replication rate calculation. To match this, I set the cutoff for significant findings for the purposes of replication equal to the smallest z-statistic that was treated as a 'positive' result for replication. Predictions are almost identical with a strict 0.05 significance threshold.

model is estimated only using information from the original studies, and does not incorporate any information from the replications. The accuracy of this prediction is consistent with the null hypothesis that the observed replication rate in economics can be explained entirely by a parsimonious model accounting only for issues with power calculations, and not other issues such as p-hacking or treatment effect heterogeneity. Failure to reject the null hypothesis does not, of course, imply that it is true, and thus we should not necessarily conclude that these other factors are not present. Nonetheless, other evidence points to a relatively limited role for p-hacking in the context of lab experiments studied here, perhaps due to fewer researcher degrees of freedom as compared with observational settings (Brodeur et al., 2016, 2020; Imai et al., 2020). Note that despite the very accurate point estimate, standard errors are relatively large, which implies limited power to reject the model's prediction (perhaps owing to the fact that there are only 18 replicated studies).

In psychology, the model predicts a replication rate of 54.5%. This is well below mean intended power of 92%, but higher than the observed replication rate of 34.8%. In this case, the model accounts for around two-thirds of the replication rate gap, and we can reject the null hypothesis that the replication gap is entirely explained by issues with common power calculations. The unexplained portion of the gap in psychology provides evidence that other factors discussed in the literature and not incorporated in the model may be important, including heterogeneity in true effects, p-hacking, and measurement error. Another possibility is that the model should account for differences in replicating main effects and interaction effects, and differences across subfields (Open Science Collaboration, 2015; Altmejd et al., 2019).

A popular variant for the common power rule is the fractional power rule, where replication power is set to detect some fraction of the original effect size with a given level of statistical power (e.g. Camerer et al. (2018) and Camerer et al. (2022)). Theoretically, under the specific rule applied in Camerer et al. (2018), the expected replication rate can range anywhere between 0.025 and 0.99 depending on the power in original studies.[51] Empirically, the predicted repli-

[51] Proposition B2 shows that the expected replication rate can range between 0.025 and $1 - \Phi[1.96 - \frac{1}{\psi}(1.96 -$

cation rate for the experimental social sciences is 54.3%, which is very close to the observed rate of 57.1%. The difference is statistically indistinguishable from zero, although the standard error of the prediction is quite large. Similarly to experimental economics, the accuracy of the point estimate of the prediction implies that we cannot reject the null hypothesis that the observed replication rate can be explained by a parsimonious model accounting only for issues with power calculations.

Table 2.2: Replication Rate Predictions

	Economics experiments	Psychology	Social sciences
Nominal target (intended power)	0.92	0.92	–
Observed replication rate	0.611	0.348	0.571
Predicted replication rate	0.600	0.545	0.543
	(0.122)	(0.054)	(0.134)

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social sciences to Camerer et al. (2018). The replication rate is defined as the share of original estimate whose replications have statistically significant findings of the same sign. Figures in the first row report the mean intended power reported in both applications. The second row shows observed replication rates. The third row reports the predicted replication rate in equation (38) calculated using parameter estimates Table 2.1. The fourth row shows standard errors for the predicted replication rate which are calculated using the delta method. In social sciences, power is set to detect three-quarters of the original effect size with 90% power. This approach does not have a fixed nominal target for the replication rate.

Extensions

I examine three extensions. In Appendix G, I use the empirical models estimated in Table 2.1 to generate predicted average relative effect sizes, using a similar procedure to the replication rate predictions. I find that the predicted relative effect size is quite similar to the observed value in economics (0.70 vs. 0.66). In the social sciences, the model is somewhat farther off (0.53 vs. 0.44), which may suggest a role for other factors, although the difference is not

$\Phi^{-1}(\beta))]$. With the fraction of original effect size to detect equal to $\psi = 3/4$, and intended power set to $1 - \beta = 0.9$, the upper range equals 0.99.

statistically distinguishable from zero. Finally, in psychology, the prediction is quite far off (0.64 vs. 0.37), again providing strong evidence for alternative factors. Note that relative effect sizes are affected both by selection of significant results for replication and the level of statistical power in original studies.[52]

A second extension considers the proposed rule of setting replication power equal to original power in Appendix F. In a review of 108 psychology replications by Anderson and Maxwell (2017), 19 (17.6%) implemented this approach. In all three applications, this approach leads to lower predicted replication rates than under the common power rule.

Given the issues that stem from conditioning on statistical significance, the third extension in Appendix H examines the suggestion of extending the replication rate definition to include null results that are 'replicated' if their replications are also insignificant. For empirical models in economics and psychology, this 'extended' replication rate remains below intended power under the common power rule.

2.4 Conclusion

The prominence of the replication rate stems in part from its apparent transparency and ease of interpretation. However, caution should be applied when interpreting the replication rate from large-scale replication studies using the common power rule for setting replication power. In general, intended replication targets are not attainable in expectation. Moreover, the replication rate gap will be particularly large when original power is low. Empirical evidence supports the importance of these theoretical insights. In a parsimonious model with neither heterogeneity nor p-hacking, predicted replication rates in experimental economics and social science are very close to observed values. This is consistent with the null hypothesis that problems with power calculations alone are sufficient to explain observed replication rates in these fields.

[52] Figure G2 in Appendix G shows that the expected relative effect size is an increasing function of power in original studies and approaches one as original power approach 100%.

Appendix

This appendix contain proofs and supplementary materials for "Why Is the Replication Rate So Low?" Section A derives properties of the replication probability function. Section B contains proofs for results in the main text, in addition to other theoretical results. Section C presents an illustrative example of how the replication rate can vary with changes in selective publication above the 1.96 significance threshold. Section D details replication selection mechanisms implemented in the three applications. Section E presents extensions of the empirical results using alternative power calculations. Section F builds intuition for the empirical replication rate decomposition results. Section G examine two further extensions to the empirical results: examining the impact of p-hacking on the replication rate; and an analysis of the relative effect size measure of replication. Appendix H examines a generalization of the replication rate definition to include insignificant results.

2A Properties of the Replication Probability Function

This Appendix derives properties of the replication probability function (Definition 1). The first 'property' simply provides a convenient, compact notation. The remaining properties consider the replication probability function under the common power rule to detect original effect sizes with $1 - \beta$ intended power (Definition 3). Recall that the replication probability for original study (x, σ, θ) is equal to

$$RP\big(x, \theta, \sigma_r(x, \sigma, \beta)\big) = \mathbb{P}\left(\frac{|X_r|}{\sigma_r(x, \beta)} \geq 1.96, \text{sign}(X_r) = \text{sign}(x)\right) \tag{39}$$

To provide intuition of the properties, Figure A1 provides an illustration of the replication probability function for different values of x under the common power rule for $1 - \beta = 0.9$ and a fixed value of θ.

Lemma A1 (Properties of the replication probability function). *The replication probability function satisfies the following properties:*

1. *For any replication standard error $\sigma_r(x, \sigma, \beta)$, the replication probability for an original study (x, σ, θ) can be written compactly as*

$$RP\big(x, \theta, \sigma_r(x, \sigma, \beta)\big) = 1 - \Phi\left(1.96 - \text{sign}(x)\frac{\theta}{\sigma_r(x, \sigma, \beta)}\right) \tag{40}$$

The remaining properties assume the replication standard error $\sigma_r(x, \beta)$ is set using the common power rule in Definition 3 with intended power $1 - \beta$:

2. *If $1 - \beta > 0.025$, then $RP\big(x, \theta, \sigma_r(x, \beta)\big)$ is strictly decreasing in x over $(-\infty, 0)$ and $(0, \infty)$.*

3. *If $(1 - \beta) > 0.6628$, then $RP\big(x, \theta, \sigma_r(x, \beta)\big)$ is strictly concave with respect to x over the open interval* $(\max\{0, [1 - r^*(\beta)]\theta\}, [1 + r^*(\beta)]\theta)$, *where*

$$r^*(\beta) = -\big(2 + 1.96.h(\beta)\big) + \sqrt{\frac{\big(2 + 1.96.h(\beta)\big)^2 - 4 \times (1 + 1.96.h(\beta) - h(\beta)^2)}{2}} > 0 \tag{41}$$

with $h(\beta) = \big(1.96 - \Phi^{-1}(\beta)\big)$.

4. *The limits of the replication probability function with respect to x are*

$$\lim_{x \to \infty} RP\big(x, \theta, \sigma_r(x, \beta)\big) = 0.025 \text{ and } \lim_{x \to -\infty} RP\big(x, \theta, \sigma_r(x, \beta)\big) = 0.025 \tag{42}$$

$$\lim_{x \uparrow 0} RP\big(x, \theta, \sigma_r(x, \beta)\big) = 0 \text{ and } \lim_{x \downarrow 0} RP\big(x, \theta, \sigma_r(x, \beta)\big) = 1 \tag{43}$$

5. *Suppose $X^* \sim N(\theta, \sigma^2)$. Then $\mathbb{E}\big[RP\big(X, \theta, \sigma_r(X, \beta)\big)\big] \to 1 - \beta$ as $\theta \to \infty$ for fixed σ.*

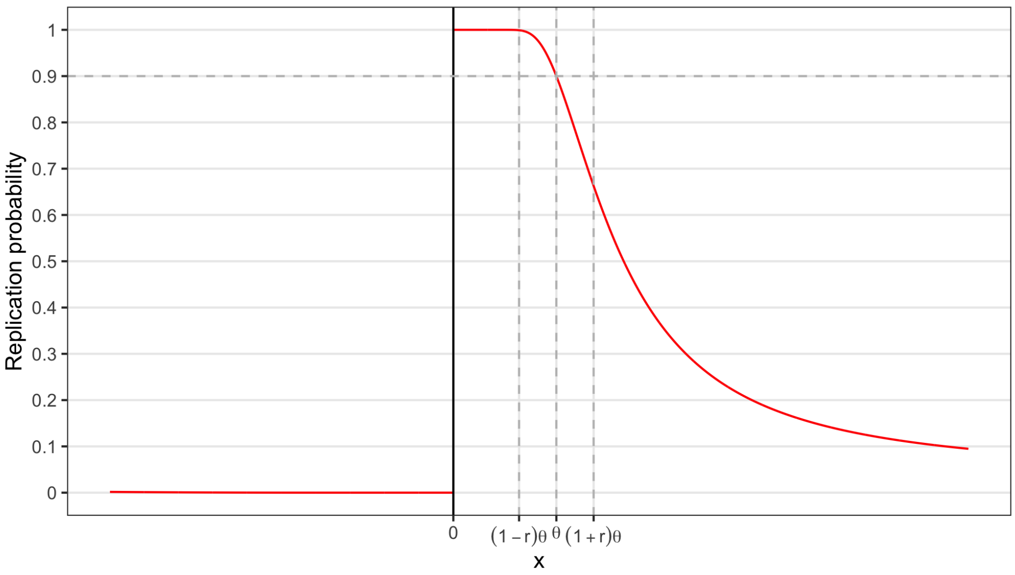

Figure 2A.1: Replication Probability Function Under Common Power Rule

Notes: Example of the replication probability function under the common power rule with intended power $(1-\beta) = 0.9$. The two vertical lines around θ marks the open interval over which the replication probability function is strictly concave, where r^* is given by equation (41).

Proof of 1.

The probability in equation (39) equals $\big[\mathbb{1}(x/\sigma \geq 1.96) \times \big(1 - \Phi\big(1.96 - \frac{\theta}{\sigma_r}\big)\big] + \big[\mathbb{1}(x/\sigma \leq -1.96) \times \Phi\big(-1.96 - \frac{\theta}{\sigma_r}\big)\big]$. This captures the two requirements for 'successful' replication: the replication estimate must attain statistical significance and have the same sign as the original estimate. Equation (40) is obtained using the symmetry of the normal distribution, which implies that $\Phi(t) = 1 - \Phi(-t)$ for any t. □

Proof of 2.

The first derivative of the replication probability function with the common power rule is

$$\frac{\partial RP\big(x,\theta,\sigma_r(x,\beta)\big)}{\partial x} = \begin{cases} -\frac{\theta}{x^2}\big(1.96-\Phi^{-1}(\beta)\big)\times\phi\Big(1.96-\frac{\theta}{x}\big(1.96-\Phi^{-1}(\beta)\big)\Big), & x>0 \\ -\frac{\theta}{x^2}\big(1.96-\Phi^{-1}(\beta)\big)\times\phi\Big(-1.96-\frac{\theta}{|x|}\big(1.96-\Phi^{-1}(\beta)\big)\Big), & x<0 \end{cases} \tag{44}$$

These are strictly negative whenever $\big(1.96 - \Phi^{-1}(\beta)\big) > 0 \iff (1-\beta) > 0.025$. □

Proof of 3.

First, note that for $x > 0$, the second derivative of the replication probability function with the common power rule is

$$\frac{\partial^2 RP\big(x, \theta, \sigma_r(x,\beta)\big)}{\partial x^2} = \left(\frac{h(\beta)\theta}{x^3}\right)\phi\left(1.96 - \frac{h(\beta)\theta}{x}\right)\left[1 + \left(\frac{h(\beta)\theta}{x}\right)\left(1.96 - \frac{h(\beta)\theta}{x}\right)\right] \tag{45}$$

Let $x = (1+r)\theta$. Substituting this into the previous equation and simplifying shows that equation (45) is strictly negative when the following inequality is satisfied

$$r^2 + \big(2 + 1.96h(\beta)\big).r + \big(1 + 1.96h(\beta) - h(\beta)^2\big) < 0 \tag{46}$$

The solution to the quadratic equation has a unique positive solution $r^*(\beta)$ whenever $(1-\beta) > 0.6628$. To see this, note that there exists a unique positive solution when $\big(1+1.96h(\beta) - h(\beta)^2\big) < 0$. This quadratic equation in $h(\beta)$ must have a unique positive and negative solution in turn, since the parabola opens downwards and equals 1 when $h(\beta) = 0$. The positive root can be obtained from the quadratic formula, which gives 2.38014. Since the quadratic function opens downward, this implies that for any $h(\beta) > 2.38014$, we have $\big(1+1.96h(\beta)-h(\beta)^2\big) < 0$. Thus, a unique positive solution to equation (46) exists whenever this condition is satisfied. In particular, a unique positive solution exists whenever

$$\begin{aligned} h(\beta) = 1.96 - \Phi^{-1}(\beta) &> 2.38014 \\ \iff \Phi(1.96 - 2.38014) &> \beta \\ \iff (1-\beta) &> 0.6628 \end{aligned} \tag{47}$$

The unique positive solution for equation (46) can again be obtained by the quadratic formula, which gives equation (41). Note that for any $r > 0$ where the inequality for concavity in equation (46) is satisfied, the same must also be true of $-r$, since it makes the left-hand-side strictly smaller. This implies that the replication probability function is strictly concave (since its second derivative is strict negative) over $(\max\{0, [1 - r^*(\beta)]\theta\}, [1 + r^*(\beta)]\theta)$, where the maximum is taken because the replication probability function is discontinuous at 0. This follows because of the properties of the quadratic function. Specifically, suppose $f(x)$ is a parabola that opens upward and intersects the y-axis at a negative value. Then for any two points (a, b) with $a < b$ and $f(a), f(b) < 0$, it must be that $f(c) < 0$ for any $c \in (a, b)$. □

Proof of 4.

Substituting the common power rule into the replication probability function gives

$$RP\big(x, \theta, \sigma_r(x, \beta)\big) = 1 - \Phi\left(1.96 - \frac{\theta}{x}\big(1.96 - \Phi^{-1}(\beta)\big)\right) \tag{48}$$

The values of the limits can be seen immediately from this expression. □

Proof of 5.

This proof consists of two steps. In the first step, I show that the replication probability function approaches linearity in x in an even interval around θ, as $\theta \to \infty$ for fixed σ. To see this, fix $r \in (0, 1)$. Then the second derivative evaluated at any point $c\theta \in \big(r\theta, (1 + r)\theta\big)$ equals

$$\left.\frac{\partial^2 RP\big(x, \theta, \sigma_r(x, \beta)\big)}{\partial x^2}\right|_{x=c\theta} = \left(\frac{h(\beta)}{c^3\theta^2}\right)\phi\left(1.96 - \frac{h(\beta)}{c}\right)\left[1 + \left(\frac{h(\beta)}{c}\right)\left(1.96 - \frac{h(\beta)}{c}\right)\right] \tag{49}$$

This approaches zero as $\theta \to \infty$, which implies that $RP\big(x, \theta, \sigma_r(x, \beta)\big)$ approaches linearity in x over the interval $\big(r\theta, (1 + r)\theta\big)$ in the limit.

For the second step, see that as $\theta \to \infty$ with fixed σ, we have that

$$\mathbb{P}\big[X^* \in \big(r\theta, (1+r)\theta\big)|\theta, \sigma\big] = \Phi\bigg(\frac{(1+r)\theta - \theta}{\sigma}\bigg) - \Phi\bigg(\frac{r\theta - \theta}{\sigma}\bigg) \to 1 \tag{50}$$

That is, the probability of drawing X^* inside of the range $\big(r\theta, (1+r)\theta\big)$ approaches one in the limit. But from the first step we know that the replication probability function is linear over this range as $\theta \to \infty$ with fixed σ. This implies in the limit that $\mathbb{E}\big[RP\big(X, \theta, \sigma_r(X, \beta)\big)\big] = RP\big(\mathbb{E}[X], \theta, \sigma_r(X, \beta)\big) = RP\big(\theta, \theta, \sigma_r(X, \beta)\big) = 1 - \beta$, as shown in Lemma 1 in the main text.

2B Proofs of Propositions

For convenience, some proofs use notation distinguishing the publication probability function $p(\cdot)$ over significant and insignificant regions:

$$p(X^*/\Sigma^*) = \begin{cases} p_{sig}(X^*/\Sigma^*) & \text{if } S_X^* = 1 \\ p_{insig}(X^*/\Sigma^*) & \text{if } S_X^* = 0 \end{cases}$$

where S_X^* is an indicator variable that equals one if $\big|X^*/\Sigma^*\big| \geq 1.96$ and zero otherwise.

Lemma B1 (Justification of the common power rule). *Consider a published study (x, σ, θ). If $x = \theta$ and a replication uses the common power rule to detect the original effect with intended power $1 - \beta$, then*

$$RP\Big(\theta, \theta, \sigma_r(\theta, \beta)\Big) = 1 - \beta \tag{51}$$

Proof. Substitute the common power rule in the replication probability function derived in Lemma A1.1 in Appendix A. If $x = \theta$, then

$$RP\big(\theta, \theta, \sigma_r(\theta, \beta)\big) = 1 - \Phi\bigg(1.96 - \text{sign}(\theta)\frac{\theta}{\sigma_r(\theta, \beta)}\bigg) = 1 - \Phi\bigg(1.96 - \frac{\theta}{\theta}\big(1.96 - \Phi^{-1}(\beta)\big)\bigg) = 1 - \beta \tag{52}$$

□

Proof of Proposition 1: For notational convenience, let $(X_{sig}, \Sigma_{sig}, \Theta_{sig})$ denote the distribution of latent studies $(X^*, \Sigma^*, \Theta^*)$ conditional on being published $(D = 1)$ and statistically significant $(|X^*/\Sigma^*| \geq 1.96)$. The expected replication probability (Definition 2) under the common power rule (Definition 3) can be written as

$$\begin{aligned}
&\mathbb{E}_{X^*,\Sigma^*,\Theta^*|D,R,S_X^*}\left[RP\Big(X^*, \Theta^*, \sigma_r(X^*, \beta)\Big)\Big|D = 1, R = 1, |X^*/\Sigma^*| \geq 1.96\right] \\
&= \mathbb{E}_{X,\Sigma,\Theta|S_X}\left[RP\big(X, \Theta, \sigma_r(X, \Sigma, \beta)\big)\big||X/\Sigma| \geq 1.96\right] \\
&= \mathbb{E}_{X_{sig},\Sigma_{sig},\Theta_{sig}}\left[RP\Big(X_{sig}, \Theta_{sig}, \sigma_r(X_{sig}, \beta)\Big)\right] \\
&= \mathbb{E}_{\Sigma_{sig},\Theta_{sig}}\left[\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\left[RP\Big(X_{sig}, \Theta_{sig}, \sigma_r(X_{sig}, \beta)\Big)|\Theta_{sig} = \theta, \Sigma_{sig} = \sigma\right]\right] \qquad (53)
\end{aligned}$$

where the second inequality drops the conditioning on being chosen for replication (R) because it is assumed that replication selection on significant results is random; and the last equality uses the Law of Iterated Expectations. The proof shows that the conditional expected replication probability satisfies $\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\big[RP\big(X_{sig}, \Theta_{sig}, \sigma_r(X_{sig}, \beta)\big)|\Theta_{sig} = \theta, \Sigma_{sig} = \sigma\big] < 1 - \beta$ which implies that the expected replication probability is also less than intended power $1 - \beta$. For greater clarity in what follows, let $\mathbb{E}\big[RP\big(X_{sig}|\theta, \sigma, \beta\big)\big]$ be shorthand for $\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\big[RP\big(X_{sig}, \Theta_{sig}, \sigma_r(X_{sig}, \beta)\big)|\Theta_{sig} = \theta, \Sigma_{sig} = \sigma\big]$.

Note that the conditional expected replication probability can be written explicitly as

$$\mathbb{E}\big[RP\big(X_{sig}|\theta, \sigma, \beta\big)\big] = \int \left(1 - \Phi\left(1.96 - \text{sign}(x)\frac{\theta}{|x|}(1.96 - \Phi^{-1}(\beta))\right)\right) \frac{p\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}| \geq 1.96\big)dx}{\int_{x'} p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}| \geq 1.96\big)dx'} \qquad (54)$$

where the integrand in equation (54) is obtained using the compact notation for the replication probability derived in Lemma A1.1 and then substituting the common power rule in Definition 3. This density differs from a normal density in two respects: (1) the publication probability function $p\big(\frac{x}{\sigma}\big)$ reweights the distribution; and (2) conditioning on statistical significance truncates original effects falling in the insignificant region $(-1.96\sigma, 1.96\sigma)$. The denominator is the normalization constant.

First, we introduce some notation. Lemma A1.3 shows that if $(1-\beta) > 0.6628$, then $RP\big(x, |\theta, \sigma, \beta\big)$ is strictly concave over the open interval $(\max\{0, [1-r^*(\beta)]\theta\}, [1+r^*(\beta)]\theta)$, where $r^*(\beta)$ is given by equation (41). This Proposition assumes $(1-\beta) > 0.8314$, so the condition is satisfied. To simplify the notation, define $(l^*, u^*) = \big((1-r^*)\theta, (1+r^*)\big)$ when $r^* \in (0,1)$ and $(l^*, u^*) = \big(0, 2\theta\big)$ when $r^* \geq 1$; in both cases, the replication probability function is strictly concave over an interval with mid-point θ.

Consider first the case where $r^* \geq 1$ so that $(l^*, u^*) = \big(0, 2\theta\big)$. The conditional replication probability can be expressed as a weighted sum

$$
\begin{aligned}
\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big] &= \mathbb{P}\Big(X_{sig} < l^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|X_{sig} < l^*\Big] \\
&+\mathbb{P}\Big(l^* \leq X_{sig} \leq u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leq X_{sig} \leq u^*\Big]+\mathbb{P}\Big(X_{sig} > u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|X_{sig} > u^*\Big] \\
&< \mathbb{P}\Big(X_{sig} < l^*\Big)0.025+\mathbb{P}\Big(l^* \leq X_{sig} \leq u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leq X_{sig} \leq u^*\Big]+\mathbb{P}\Big(X_{sig} > u^*\Big)(1-\beta)
\end{aligned}
\tag{55}
$$

In the last line, the first term in the sum uses the fact that the maximum value of the replication probability when $x < l^* = 0$ is 0.025 (Lemma A1.2 and Lemma A1.4 in Appendix A). The third term follows because $RP\big(2\theta|\theta,\sigma,\beta\big)$ is the maximum value the function takes over $x > u^* = 2\theta$, since the function is strictly decreasing over $x > 0$ (Lemma A1.2); and therefore that $RP\big(2\theta|\theta,\sigma,\beta\big) < RP\big(\theta|\theta,\sigma,\beta\big) = 1-\beta$, where the equality is shown in Lemma 1. From equation (55), we can see that $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta)|l^* \leq X_{sig} \leq u^*\big] < 1-\beta$ is a sufficient condition for $\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\big] < 1-\beta$.

Before showing that this sufficient condition is satisfied, we show that the same sufficient condition holds in the second case, where $r^* \in (0,1)$ so that $(l^*, u^*) = \big((1-r^*)\theta, (1+r^*)\theta\big)$. This requires additional steps. First, express the conditional replication probability as a weighted sum

$$
\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big] = \mathbb{P}\Big(X_{sig} \leq l^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|X_{sig} \leq l^*\Big]
$$

$$+\mathbb{P}\Big(l^* \le X_{sig} \le u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \le X_{sig} \le u^*\Big]+\mathbb{P}\Big(X_{sig} \ge u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|X_{sig} \ge u^*\Big]$$

$$< \mathbb{P}\Big(X_{sig} \le l^*\Big)+\mathbb{P}\Big(l^* \le X_{sig} \le u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \le X_{sig} \le u^*\Big]+\mathbb{P}\Big(X_{sig} \ge u^*\Big)RP\Big(u^*|\theta,\sigma,\beta\Big) \tag{56}$$

The strict inequality follows for two reasons. For the first term in the sum, one is the maximum value the function can take for any x. For the third term, $RP(u^*|\theta,\sigma,\beta)$ is the function's maximum value over $x \ge u^*$, since the integrand is strictly decreasing over positive values (Lemma A1.2). With an additional step, we can write this inequality as

$$\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big] < \frac{1}{2}\Big(1-\mathbb{P}\Big(l^* \le X_{sig} \le u^*\Big)\Big)\Big(1+RP\big(u^*\big|\theta,\sigma,\beta\big)\Big)$$
$$+\mathbb{P}\Big(l^* \le X_{sig} \le u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \le X_{sig} \le u^*\Big] \tag{57}$$

This follows because $\mathbb{P}(X_{sig} \le l^*) \le \mathbb{P}(X_{sig} \ge u^*)$ and $RP(u^*|\theta,\sigma,\beta) < 1$. That is, increasing the relative weight on the maximum value of one, such that both tails are equally weighted, must lead to a (weakly) larger value. The weak inequality $\mathbb{P}(X_{sig} \le l^*) \le \mathbb{P}(X_{sig} \ge u^*)$ required for this simplification is shown below:

Lemma B2. *Suppose $X|\theta,\sigma$ follows the truncated normal pdf in equation (54). Then for any $r^* \in (0,1)$, the following inequality holds: $\mathbb{P}\big(X_{sig} \le (1-r^*)\theta\big) < \mathbb{P}\big(X_{sig} \ge (1+r^*)\theta\big)$.*

Proof. First, note that $\big((1-r^*)\theta,(1+r^*)\theta\big)$ is an interval over the positive real line centered at θ. Consider two cases:

Case 1: Let $(1-r^*)\theta \le 1.96\sigma$. Define the normalization constant $C = \int_{x'} p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}| \ge 1.96\big)dx'$. Then

$$\mathbb{P}\Big(X_{sig} \le (1-r^*)\theta\Big) = \frac{1}{C}\int_{-\infty}^{-1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' \le \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx'$$

$$< \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{\max\{1.96\sigma,(1+r^*)\theta\}}^{2\theta+1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' = \mathbb{P}\Big(X_{sig} \geq (1+r^*)\theta\Big) \tag{58}$$

Consider the weak inequality. Note that the mid-point between -1.96σ and $2\theta+1.96\sigma$ is θ. Thus, with no selective publication (i.e. $p(t) = 1$ for all t), we would have equality owing to the symmetry of the normal distribution. However, recall that $p_{sig}()$ is symmetric about zero and weakly increasing in absolute value. It follows therefore that $|2\theta + 1.96\sigma| > |-1.96\sigma|$ implies $p_{sig}(|2\theta + 1.96\sigma|) \geq p_{sig}(|-1.96\sigma|)$; using this fact and symmetry of the normal distribution about θ gives the weak inequality. The strict inequality follows because the additional term is strictly positive, since $p_{sig}()$ is assumed to be non-zero.

Case 2: Let $(1-r^*)\theta > 1.96\sigma$. The argument is similar to the first case:

$$\mathbb{P}\Big(X_{sig} \leq (1-r^*)\theta\Big) = \frac{1}{C}\int_{-\infty}^{-1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{1.96\sigma}^{(1-r^*)\theta} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx'$$

$$< \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{(1+r^*)\theta}^{2\theta-1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx'$$

$$+\frac{1}{C}\int_{2\theta-1.96\sigma}^{2\theta+1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' = \mathbb{P}\Big(X_{sig} \geq (1+r^*)\theta\Big) \tag{59}$$

□

The inequality in equation (57) can be further simplified by placing restrictions on intended power. In particular, if intended power satisfies $1-\beta \geq 0.8314$, then

$$\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big] < \Big(1 - \mathbb{P}\big(l^* \leq X_{sig} \leq u^*\big)\Big)\big(1-\beta\big)$$

$$+\mathbb{P}\big(l^* \leq X_{sig} \leq u^*\big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big| l^* \leq X_{sig} \leq u^*\Big] \tag{60}$$

This follows because with $u^* = (1+r^*)\theta$, we have

$$\frac{1}{2}\Big(1 + RP\big(u^*\big|\theta,\sigma,\beta\big)\Big) = \frac{1}{2}\left(1 + \left(1 - \Phi\left(1.96 - \frac{1.96 - \Phi^{-1}(\beta)}{1 + r^*(\beta)}\right)\right)\right)$$
$$\leq 1 - \beta \iff 1 - \beta \geq 0.8314 \tag{61}$$

From equation (60), we can see that $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta)\big|l^* \leq X_{sig} \leq u^*\big] < 1-\beta$ is a sufficient condition for $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta)\big] < 1-\beta$. Thus, in both cases, the sufficient condition for the desired result is the same.

This sufficient condition is shown in two steps. In the first, I show that this inequality holds even in the case where there is no selective publication and all published results are replicated (i.e. when $X \sim N(\Theta, \Sigma^2)$). In the second, I show that this inequality remains true once we allow for selective publication and truncation of the distribution due to conditioning on statistical significance.

Lemma B3 states the first intermediate step. Its implications are of independent interest and discussed in the main text. It shows that even in the optimistic scenario where original estimates are unbiased, there is no selective publication, and all results are published and replicated, that the expected replication probability still falls below intended power.

Lemma B3. *Let published effects be distributed according to $X|\theta,\sigma \sim N(\theta,\sigma^2)$. Suppose $p(t) = 1$ and $r(t) = 1$ for all $t \in \mathbb{R}$. Assume all results are included in the replication rate calculation. Let power in replications is set according to the common power rule with intended power $1-\beta \geq 0.8314$. Then $\mathbb{E}\big[RP(X|\theta,\sigma,\beta)\big] < 1-\beta$.*

Proof. Recall that $RP(x|\theta,\sigma,\beta)$ is strictly concave with respect to x over the interval (l^*, u^*), where $(l^*, u^*) = \big((1-r^*)\theta, (1+r^*)\big)$ when $r^* \in (0,1)$ and $(l^*, u^*) = \big(0, 2\theta\big)$; in both cases, the mid-point of the interval is θ. We have that

$$\mathbb{E}\Big[RP\big(X|\theta,\sigma,\beta\big)\Big|l^* \leq X \leq u^*\Big] = \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{l^*}^{u^*}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} < RP\Big(\theta\Big|\theta,\sigma,\beta\big)\Big) = 1-\beta \tag{62}$$

where the strict inequality follows from Jensen's inequality and the fact that $\mathbb{E}[X|l^* \leq X \leq u^*] = \theta$. The final equality is a property of the replication probability function shown in Lemma 1 in the main text. This is the sufficient condition required for the desired result.

Note that the inequalities in equations (57) (for when $r^* \geq 1$) and (60) (for when $r^* \in (0,1)$) were derived under more general conditions, where the normal distribution may we reweighted by $p()$ and truncated based on significance. This setting is a special case with no selective publication (i.e. $p(t) = 1$ for all t), and no truncation such that all results are included in the replication rate irrespective of statistical significance. □

The same conclusions hold when we introduce selective publication (which reweights the normal distribution) and condition on statistical significance (which truncates the 'insignificant' regions of the density). Consider three cases. First, suppose that $u^* \leq 1.96\sigma$. Then $\mathbb{E}\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leq X_{sig} \leq u^*\big) = 0 < 1-\beta$ because of truncation. Second, suppose that $l^* \geq 1.96\sigma$. Then

$$\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leq X_{sig} \leq u^*\Big] = \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{l^*}^{u^*} p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}$$

$$\leq \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{l^*}^{u^*}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} < RP\Big(\theta\Big|\theta,\sigma,\beta\big)\Big) = 1-\beta \tag{63}$$

Note that the distribution is invariant to the scale of $p_{sig}()$. Consider first the weak inequality. This follows because $p_{sig}()$ is assumed to be weakly increasing over (l^*, u^*). When it is a constant function over the interval, the equality holds. If $p_{sig}(x/\sigma) > 0$ for some $x \in (l^*, u^*)$

then the function redistributes weight to larger values of x. Since $RP(x|\theta,\sigma,\beta)$ is strictly decreasing over positive values of x (Lemma A1.2), placing higher relative weight on lower values implies that the weak inequality becomes strict. As in the proof to Lemma B3, the strict inequality follows from Jensen's inequality, since $RP(x|\theta,\sigma,\beta)$ is strictly concave over (l^*,u^*), and the fact that the expected value of X over this interval is equal to the true value θ. The last equality follows from Lemma 1 in the main text.

Finally, consider the case where $l^* < 1.96\sigma < u^*$. Then

$$\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \le X_{sig} \le u^*\Big] = \int_{1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{1.96\sigma}^{u^*}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}$$

$$= \int_{1.96\sigma}^{2\theta-1.96\sigma} RP\big(x|\theta,\sigma,\beta\big)\frac{p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{1.96\sigma}^{u^*}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} + \int_{2\theta-1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{1.96\sigma}^{u^*}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}$$

$$= \omega\int_{1.96\sigma}^{2\theta-1.96\sigma} RP\big(x|\theta,\sigma,\beta\big)\frac{p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{1.96\sigma}^{2\theta-1.96\sigma}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} + (1-\omega)\int_{2\theta-1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{p_{sig}\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{2\theta-1.96\sigma}^{u^*}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}$$

$$= \omega\int_{1.96\sigma}^{2\theta-1.96\sigma} RP\big(x|\theta,\sigma,\beta\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{1.96\sigma}^{2\theta-1.96\sigma}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} + (1-\omega)\int_{2\theta-1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{2\theta-1.96\sigma}^{u^*}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}$$

$$< \omega RP\Big(\theta\Big|\theta,\sigma,\beta\Big)\Big) + (1-\omega).RP\Big(2\theta-1.96\sigma\Big|\theta,\sigma,\beta\Big)\Big) < 1-\beta \tag{64}$$

with

$$\omega = \frac{\int_{1.96\sigma}^{2\theta-1.96\sigma}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}{\int_{1.96\sigma}^{u^*}p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} \tag{65}$$

The second row simply breaks up the integral. The third row rearranges the sum so that the conditional expectation of the replication probability appears in both terms. The third line follows because, as in the previous case, the p_{sig} function redistributes weight to large values of x and hence lower values of $RP(x|\theta,\sigma,\beta)$. In the last line, the first term uses the concavity of $RP(x|\theta,\sigma,\beta)$ over $(1.96\sigma, 2\theta-1.96\sigma) \subset (l^*, u*)$, Jensen's inequality, and the fact that the expected value of X over this interval is equal to θ. The second term follows because $2\theta - 1.96\sigma$ is the maximum value the function can take because $RP(x|\theta,\sigma,\beta)$ is strictly decreasing in x

over positive values. The final inequality follows because $RP\big(\theta\big|\theta,\sigma,\beta)\big) = 1-\beta$ (Lemma 1) and $RP\big(2\theta - 1.96\sigma\big|\theta,\sigma,\beta)\big) < 1-\beta$ because $2\theta - 1.96\sigma > \theta$ and the function is strictly decreasing over positive values.

This covers all cases, proving the proposition.

Proposition B1 (Regression to the mean in replications). *Suppose $p_{sig}()$ is symmetric about zero, non-zero over all values, differentiable, and weakly increasing in absolute value. Allow $p_{insig}()$ to take any form. Published original estimates X and corresponding replication estimates X_r satisfy*

$$\mathbb{E}\big[X\big|\Theta=\theta, S_X=1\big] > \theta = \mathbb{E}\big[X_r|\Theta=\theta\big] \tag{66}$$

Proof. We have $\mathbb{E}\big(X_r\big|\Theta=\theta\big) = \theta$ by assumption. Next, note that

$$\begin{aligned}\mathbb{E}_{X^*|\Theta^*,S^*_X,D}\Big(X^*|\Theta^*=\theta, |X^*/\Sigma^*| \geq 1.96, D=1\Big) &= \mathbb{E}_{X|\Theta,S_X}\Big(X|\Theta=\theta, |X/\Sigma| \geq 1.96\Big)\\ &= \mathbb{E}_{\Sigma|\Theta,S_X}\left(\mathbb{E}_{X|\Theta,\Sigma,S_X}\Big(X|\Theta=\theta,\Sigma=\sigma,|X/\sigma|\geq 1.96\Big)\right)\end{aligned} \tag{67}$$

where the last line uses the Law of Iterated Expectations. We will prove $\mathbb{E}_{X|\Theta,\Sigma,S^*_X}\big(X|\Theta = \theta, \Sigma=\sigma, |X/\sigma| \geq 1.96\big) > \theta$, which implies that the expression in equation (67) is also greater than θ. Recall that $X|\theta,\sigma$ is the effect size of published studies and follows a truncated normal distribution:

$$\frac{p\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\big(\frac{x-\theta}{\sigma}\big)\mathbb{1}\big(|\frac{x}{\sigma}| \geq 1.96\big)}{\int p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\big(\frac{x'-\theta}{\sigma}\big)\mathbb{1}\big(|\frac{x}{\sigma}| \geq 1.96\big)dx'} \tag{68}$$

Define $X = \theta + \sigma Z$. Then the density for the transformed random variable Z is

$$\frac{p\big(z+\frac{\theta}{\sigma}\big)\phi\big(z\big)\mathbb{1}\big(|z+\frac{\theta}{\sigma}| \geq 1.96\big)}{\int p\big(z'+\frac{\theta}{\sigma}\big)\phi\big(z'\big)\mathbb{1}\big(|z+\frac{\theta}{\sigma}| \geq 1.96\big)dz'} \tag{69}$$

For notational convenience, define the following normalization constants:

$$\bar{\eta} = \mathbb{P}(X \leq -1.96\sigma) + \mathbb{P}(X \geq 1.96\sigma) = \mathbb{P}\left(Z \leq -1.96 - \frac{\theta}{\sigma}\right) + \mathbb{P}\left(Z \geq 1.96 - \frac{\theta}{\sigma}\right) \tag{70}$$

$$\eta_1 = \mathbb{P}(X \leq -1.96\sigma) = \mathbb{P}\left(Z \leq -1.96 - \frac{\theta}{\sigma}\right) \tag{71}$$

$$\eta_2 = \mathbb{P}(X \geq 2\theta + 1.96\sigma) = \mathbb{P}\left(Z \geq \frac{\theta}{\sigma} + 1.96\right) \tag{72}$$

$$\eta_3 = \mathbb{P}(1.96\sigma \leq X \leq 2\theta - 1.96\sigma) = \mathbb{P}\left(1.96 - \frac{\theta}{\sigma} \leq Z \leq \frac{\theta}{\sigma} - 1.96\right) \tag{73}$$

Case 1.

Consider two cases. First, suppose $\theta \in (0, 1.96\sigma)$. Conditional on (θ, σ) (where we suppress the conditional notation on (θ, σ) for clarity), the expected value of a published estimate conditional of statistical significance is

$$\begin{aligned}\mathbb{E}(X|1.96\sigma \leq |X|) = \frac{1}{\bar{\eta}}\Bigg(&\eta_1\mathbb{E}(X|X \leq -1.96\sigma) + \eta_2\mathbb{E}(X|X \geq 2\theta + 1.96\sigma) \\ &+\big(\bar{\eta} - \eta_1 - \eta_2\big)\mathbb{E}(X|1.96\sigma \leq X \leq 2\theta + 1.96\sigma)\Bigg)\end{aligned} \tag{74}$$

First note that $\mathbb{E}(X|1.96\sigma \leq X \leq 2\theta + 1.96\sigma) > \theta$ since we assume that $\theta \in (0, 1.96\sigma)$ and $p_{sig}() > 0$. If $\eta_1\mathbb{E}(X|X \leq -1.96\sigma) + \eta_2\mathbb{E}(X|X \geq 2\theta + 1.96\sigma) \geq \big(\eta_1 + \eta_2\big)\theta$, it follows that $\mathbb{E}(X|1.96\sigma \leq |X|) > \theta$, which is what we want to show. Consider the first expectation in this expression:

$$\mathbb{E}(X|X \leq -1.96\sigma) = \mathbb{E}\Big(\theta + \sigma Z|Z \leq -1.96 - \frac{\theta}{\sigma}\Big) = \theta + \sigma\mathbb{E}\Big(Z|Z \leq -1.96 - \frac{\theta}{\sigma}\Big) \tag{75}$$

Evaluating the expectation in the right-hand-side of equation (75) gives

$$\mathbb{E}\Big(Z|Z \leq -1.96-\frac{\theta}{\sigma}\Big) = \frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} zp_{sig}\Big(z+\frac{\theta}{\sigma}\Big)\phi(z)dz = -\frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p_{sig}\Big(z+\frac{\theta}{\sigma}\Big)\phi'(z)dz$$

$$= -\frac{1}{\eta_1}\Bigg[p_{sig}(-1.96)\phi\Big(-1.96-\frac{\theta}{\sigma}\Big) - p_{sig}(-\infty)\phi(-\infty) - \int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z+\frac{\theta}{\sigma}\Big)\phi(z)dz\Bigg]$$

$$= -\frac{1}{\eta_1}p_{sig}(-1.96)\phi\Big(-1.96-\frac{\theta}{\sigma}\Big) + \frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z+\frac{\theta}{\sigma}\Big)\phi(z)dz \tag{76}$$

where the second equality uses $\phi'(z) = -z\phi(z)$; the third equality uses integration by parts; and the final equality follows because $p_{sig}(-\infty)\phi(-\infty) = 0$ since $p_{sig}()$ is bounded between zero and one. Substituting this into equation (75) gives

$$\mathbb{E}(X|X \leq -1.96\sigma) = \theta - \frac{\sigma}{\eta_1}p_{sig}(-1.96)\phi\Big(-1.96-\frac{\theta}{\sigma}\Big) + \frac{\sigma}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z+\frac{\theta}{\sigma}\Big)\phi(z)dz \tag{77}$$

Next, note that

$$\mathbb{E}(X|X \geq 2\theta + 1.96\sigma) = \theta + \sigma\mathbb{E}\Big(Z|Z \leq \frac{\theta}{\sigma} + 1.96\Big) \tag{78}$$

where

$$\mathbb{E}\Big(Z|Z \leq \frac{\theta}{\sigma} + 1.96\Big) = \frac{1}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} zp_{sig}\Big(z+\frac{\theta}{\sigma}\Big)\phi(z)dz \geq \frac{1}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} zp_{sig}\Big(z-\frac{\theta}{\sigma}\Big)\phi(z)dz \tag{79}$$

since $p_{sig}(z+\theta/\sigma) \geq p_{sig}(z-\theta/\sigma)$ for all $z \in (1.96+\theta/\sigma, \infty)$ because $p_{sig}(t)$ is weakly increasing over $t > 1.96$. For the right-hand-side of this equation, we can apply similar arguments used to derive equation (76). Substituting the result into equation (78) gives

$$\mathbb{E}(X|X \geq 2\theta + 1.96\sigma) \geq \theta + \frac{\sigma}{\eta_2} p_{sig}(1.96)\phi\bigg(1.96 + \frac{\theta}{\sigma}\bigg) + \frac{\sigma}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(z - \frac{\theta}{\sigma}\bigg)\phi(z)dz \quad (80)$$

Equations (77) and (80) imply

$$\begin{aligned} &\eta_1\mathbb{E}(X|X \leq -1.96\sigma) + \eta_2\mathbb{E}(X|X \geq 2\theta + 1.96\sigma) \\ &\geq (\eta_1+\eta_2)\theta + \sigma\bigg[p_{sig}(1.96)\phi\bigg(1.96 + \frac{\theta}{\sigma}\bigg) - p_{sig}(-1.96)\phi\bigg(-1.96 - \frac{\theta}{\sigma}\bigg)\bigg] \\ &+\sigma\bigg[\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz + \int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(z - \frac{\theta}{\sigma}\bigg)\phi(z)dz\bigg] = (\eta_1+\eta_2)\theta \quad (81) \end{aligned}$$

In the second line, the second term in the sum equals zero because symmetry of $p_{sig}()$ and $\phi()$ about zero implies that both terms in the brackets are equal. To see why the third term in the sum equals zero, note that

$$\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz = \int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(-u + \frac{\theta}{\sigma}\bigg)\phi(u)du = -\int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(u - \frac{\theta}{\sigma}\bigg)\phi(u)du \quad (82)$$

The first equality follows from both changing the order of the integral limits and applying the substitution $u = -x$; it also uses the symmetry of $\phi()$. The final equality holds because symmetry of $p_{sig}()$ about zero implies that for any $t > 1.96$, $p'_{sig}(t) = -p'_{sig}(-t)$.

Case 2.

Consider the second case where $\theta \geq 1.96\sigma$. For a given (θ, σ), we have

$$\mathbb{E}(X|1.96\sigma \leq |X|) = \frac{1}{\bar{\eta}}\Bigg(\eta_1 \mathbb{E}(X|X \leq -1.96\sigma) + \eta_2 \mathbb{E}(X|X \geq 2\theta + 1.96\sigma)$$

$$\eta_3 \mathbb{E}(X|1.96\sigma \leq X \leq 2\theta - 1.96\sigma) + \big(\bar{\eta} - \eta_1 - \eta_2 - \eta_3\big)\mathbb{E}(X|2\theta - 1.96\sigma \leq X \leq 2\theta + 1.96\sigma)\Bigg)$$

$$> \frac{1}{\bar{\eta}}\Bigg(\theta(\eta_1 + \eta_2) + \big(\bar{\eta} - \eta_1 - \eta_2 - \eta_3\big)\theta + \eta_3 \mathbb{E}(X|1.96\sigma \leq X \leq 2\theta - 1.96\sigma)\Bigg) \tag{83}$$

The inequality follows from two facts. First, the inequality proved in the first case: $\eta_1 \mathbb{E}(X|X \leq -1.96\sigma) + \eta_2 \mathbb{E}(X|X \geq 2\theta + 1.96\sigma) \geq (\eta_1 + \eta_2)\theta$. Second, the expectation in the third term of the sum satisfies $\mathbb{E}(X|2\theta - 1.96\sigma \leq X \leq 2\theta + 1.96\sigma) > \theta$ because $\theta \geq 1.96\sigma \iff 2\theta - 1.96\sigma \geq \theta$ and we assume that $p_{sig}() > 0$.

It remains to show that $\mathbb{E}(X|1.96\sigma \leq X \leq 2\theta - 1.96\sigma) \geq \theta$. Then it follows that $\mathbb{E}(X|1.96\sigma \leq |X|) > \theta$, which is what we want to show. First, note that

$$\mathbb{E}(X|1.96\sigma \leq X \leq 2\theta - 1.96\sigma) = \theta + \sigma \mathbb{E}\bigg(Z\bigg|1.96 - \frac{\theta}{\sigma} \leq Z \leq -1.96 + \frac{\theta}{\sigma}\bigg) \tag{84}$$

It is therefore sufficient to show that $\mathbb{E}\Big(Z\Big|1.96 - \frac{\theta}{\sigma} \leq Z \leq -1.96 + \frac{\theta}{\sigma}\Big) \geq 0$. Writing out the expectation in full gives

$$\mathbb{E}\bigg(Z\bigg|1.96 - \frac{\theta}{\sigma} \leq Z \leq -1.96 + \frac{\theta}{\sigma}\bigg) = \frac{1}{\eta_3}\Bigg(\int_{1.96-\frac{\theta}{\sigma}}^{0} z p_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz + \int_{0}^{\frac{\theta}{\sigma}-1.96} z p_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz\Bigg)$$

$$= \frac{1}{\eta_3}\Bigg(\int_{0}^{\frac{\theta}{\sigma}-1.96} z\bigg[p_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg) - p_{sig}\bigg(-z + \frac{\theta}{\sigma}\bigg)\bigg]\phi(z)dz\Bigg) \geq 0 \tag{85}$$

The second equality follows because

$$\int_{1.96-\frac{\theta}{\sigma}}^{0} z p_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz = -\int_{0}^{1.96-\frac{\theta}{\sigma}} z p_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz = -\int_{0}^{\frac{\theta}{\sigma}-1.96} u p_{sig}\bigg(-u + \frac{\theta}{\sigma}\bigg)\phi(u)du \tag{86}$$

which uses the substitution $u = -x$ and the symmetry of $\phi()$. The weak inequality in equation (85) follows because $p_{sig}()$ is assumed to be weakly increasing over positive values. Thus, $z - \theta/\sigma > -z + \theta/\sigma$ for all $z \in (0, \theta/\sigma - 1.96)$ implies $p_{sig}\big(z + \theta/\sigma\big) - p_{sig}\big(-z + \theta\sigma\big) \geq 0$.

This covers all cases and proves the proposition. □

Proposition B2 *Under the fractional power rule which sets the replication standard error according to* $\sigma_r(X, \beta, \psi) = \frac{\psi \cdot |X|}{1.96 - \Phi^{-1}(\beta)}$ *with* $\psi < 1$*, the expected replication rate can range between 0.025 and* $1 - \Phi[1.96 - \frac{1}{\psi}\big(1.96 - \Phi^{-1}(\beta)\big)] > 1 - \beta$.

Proof of Proposition B2: Under the fractional power rule, the expected replication rate conditional on fixed (θ, σ) is given by

$$\mathbb{E}[RP(X, \Theta, \sigma_r(X, \beta, \psi)|\Theta = \theta, \Sigma = \sigma]$$

$$= \int \left[1 - \Phi\Big(1.96 - \text{sign}(x)\frac{\theta}{\psi \cdot |x|}\big(1.96 - \Phi^{-1}(\beta)\big)\Big)\right]\frac{1}{\sigma}\phi\Big(\frac{x - \theta}{\sigma}\Big)dx \tag{87}$$

If $\theta = 0$, then this equals 0.025. Next, suppose that $\theta > 0$ and consider the case where $\sigma \to 0$ such that power in original studies approaches one. See that the integrand is bounded above by one and converges pointwise as $\sigma \to 0$ to

$$1 - \Phi\Big(1.96 - \text{sign}(x)\frac{\theta}{\psi \cdot |x|}\big(1.96 - \Phi^{-1}(\beta)\big)\Big)\mathbb{1}\{x = \theta\} \tag{88}$$

since the normal distribution converges to a degenerate distribution when the variance goes to zero. Thus, by the dominated convergence theorem (and the fact that $\theta > 0$), we have that

$$\lim_{\sigma \to 0} \mathbb{E}[RP(X, \Theta, \sigma_r(X, \beta, \psi)|\Theta = \theta, \Sigma = \sigma] = 1 - \Phi\Big(1.96 - \frac{1}{\psi}\big(1.96 - \Phi^{-1}(\beta)\big)\Big) \tag{89}$$

When $\psi = 1$, this equals $1 - \beta$. Since equation (89) is strictly decreasing in ψ, it follows that equation (89) is strictly above $1 - \beta$ when $\psi < 1$.

This shows that the expected replication of an *individual* study can range between 0.025 and $1-\Phi[1.96-\frac{1}{\psi}\big(1.96-\Phi^{-1}(\beta)\big)] > 1-\beta$. Integrating over the distribution of latent studies gives the desired result. □

Proposition B3 *For any function* $g(X,\Sigma,X_r,\beta)$,$\mathbb{E}\big[g(X,\Sigma,X_r,\beta)|D=1,R=1,S_X=1\big]$ *does not depend on* $p_{insig}()$.

Proof of Proposition B3: We can write $\mathbb{E}\big[g(X,\Sigma,X_r,\beta)|D=1,R=1,S_X=1\big]$ as

$$\int g(x,\sigma,x_r,\beta)f_{X^*,\Sigma^*,\Theta^*,X_r|D,R,S_X^*}\Big(x,\sigma,\theta,x_r\big|D=1,R=1,S_{X^*}=1\Big)dxd\sigma d\theta dx_r$$

$$=\int_{x,\sigma,\theta}\bigg(\int_{x_r}g(x,\sigma,x_r,\beta)f_{X_r|X^*,\Sigma^*,\Theta^*}\Big(x_r|\theta,\sigma_r(x,\sigma,\beta)\Big)dx_r\bigg)f_{X^*,\Sigma^*,\Theta^*|D,R,S_X^*}\big(x,\sigma,\theta|D=1,R=1,S_X^*=1\big)dxd\sigma d\theta \quad (90)$$

The equality uses the Law of Iterated Expectations and $f_{X_r|X^*,\Sigma^*,\Theta^*,D,R,S_X^*}\big(x_r|\theta,\sigma_r(x,\sigma,\beta)\big) = f_{X_r|X^*,\Sigma^*,\Theta^*}\big(x_r|\theta,\sigma_r(x,\sigma,\beta)\big)$. Replication estimates are not subject to selective publication, which implies this is a normal density that does not depend on $p()$. Hence, the term in parentheses can only be affected by $p()$ indirectly through $f_{X^*,\Sigma^*,\Theta^*|D,R,S_X^*}$, which is the joint distribution of original studies conditional on being published, chosen for replication, and statistically significant at the 5% level. However, this distribution does not depend on the probability of publishing insignificant findings. To see this, apply Bayes rule twice to get

$$f_{X^*,\Sigma^*,\Theta^*|D,R,S_X^*}\big(x,\sigma,\theta|D=1,R=1,S_X^*=1\big)$$

$$=\frac{\mathbb{P}\Big(D=1\big|X^*=x,\Sigma^*=\sigma,\Theta^*=\theta,R=1,S_X^*=1\Big)}{\mathbb{P}\Big(D=1\big|R=1,S_X^*=1\Big)}\times\frac{\mathbb{P}\Big(R=1\big|X^*=x,\Sigma^*=\sigma,\Theta^*=\theta,S_X^*=1\Big)}{\mathbb{P}\Big(R=1\big|S_X^*=1\Big)}$$

$$\times f_{X^*,\Theta,\Sigma^*|S_X^*}\Big(x,\theta,\sigma\big|S_X^*=1\Big)$$

$$=\frac{p_{sig}(x/\sigma)}{\mathbb{E}\big(p_{sig}(X^*/\Sigma^*)\big|S_X^*=1\big)}\cdot\frac{r_{sig}(x/\sigma)}{\mathbb{E}\big(r_{sig}(X^*/\Sigma^*)\big|S_X^*=1\big)}\cdot f_{X^*,\Sigma^*,\Theta^*|S_X^*}\Big(\theta,x,\sigma\big|S_X^*=1\Big) \quad (91)$$

In the final line, the first factor in the product includes only $p_{sig}()$; the denominator does not condition on R because replication selection is assumed to be random for significant findings. The second factor equals one because replication selection for significant results is assumed to be random. The final factor in the product is the density of latent studies conditional on significance, which is not affected by selective publication. □

2C Replication Rate Gap Decomposition

How can we measure the relative importance of non-linearities as compared to distortions from selection on significance? To answer this question, I derive a decomposition of the replication rate gap, which I implement in the empirical section.

The decomposition is based on two regimes. Regime 1 ($M1$) assumes use of the standard definition of the replication rate: only significant results are included, and replication selection is a random sample of significant results. Regime 2 ($M2$) is based on a counterfactual scenario where all results are published and replication is random. This implies the distribution of published, replicated studies coincides with the distribution of latent studies. Formally, the expectation operators under both regimes are defined by:

$$\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big] = \int RP(x,\theta,\sigma_r(x,\beta)) f_{X^*,\Theta^*|D,R,S_X^*}(x,\theta|D=1,R=1,S_X^*=1)dxd\theta \tag{92}$$

$$\mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big] = \int RP(x,\theta,\sigma_r(x,\beta)) f_{X^*,\Theta^*}(x,\theta)dxd\theta \tag{93}$$

Using these, we have the following decomposition:

$$\begin{aligned}\underbrace{(1-\beta) - \mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big]}_{\text{replication rate gap}} &= \underbrace{(1-\beta) - \mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geq 0\big]}_{\text{(i) concavity gap}} \\ &+ \underbrace{\mathbb{P}_{M1}\big(X<0\big)\Big(\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geq 0\big] - \mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X<0\big]\Big)}_{\text{(ii) wrong-sign gap}} \\ &+ \underbrace{\mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geq 0\big] - \mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geq 0\big]}_{\text{(iii) selection-on-significance gap}}\end{aligned} \tag{94}$$

Proof. Write the expected replication probability under model 1 as

$$\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big] = \mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X \geq 0\big]$$
$$+\mathbb{P}_{M1}\big(X<0\big)\Big(\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X<0\big]\Big) - \mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X \geq 0\big]\Big) \tag{95}$$

To arrive at equation (94), substitute equation (95) into the replication rate gap; add and subtract $\mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X \geq 0\big]$; and rearrange the terms. □

Note that the concavity gap and the selection-on-significance gap condition on estimates with the same sign as the underlying true effect. This allows us to determine their contribution separate from the impact of attempting to replicate original estimates with the 'wrong' sign.

Table C1 presents the results. Panel A reproduces the results in the main text, and Panel B present the decomposition results. The empirical results for the decomposition show that failing to account for the concavity of the replication power function explains the overwhelming majority of the explained replication rate gap in both economics and psychology. The selection-on-significance gap in small, explaining only 3.1% of the gap in economics, while actually *decreasing* the replication rate in psychology. The latter outcome arises because conditioning on statistical significance tends to select larger true effects, which have higher replication probabilities than smaller true effects.

Table 2C.1: Replication Rate Predictions and Decomposition Results

	Economics experiments	Psychology	Social sciences
A. Replication rate predictions			
Nominal target (intended power)	0.92	0.92	–
Observed replication rate	0.611	0.348	0.571
Predicted replication rate	0.600	0.545	0.543
B. Decomposition of explained gap			
Predicted replication rate gap	0.320 (100%)	0.375 (100%)	–
Concavity gap	0.292 (91.16%)	0.364 (97.16%)	–
Wrong-sign gap	0.018 (5.72%)	0.030 (8.03%)	–
Selection-on-significance gap	0.010 (3.12%)	-0.019 (-5.18%)	–

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social sciences to Camerer et al. (2018). The replication rate is defined as the share of original estimate whose replications have statistically significant findings of the same sign. Figures in the first row report the mean intended power reported in both applications. The second row shows observed replication rates. The third row reports the predicted replication rate in equation (38) calculated using parameter estimates Table 2.1. In social sciences, power is set to detect three-quarters of the original effect size with 90% power. This approach does not have a fixed nominal target for the replication rate.

Below I provide details underlying the intuition behind the decomposition results.

Concavity gap.—Figure C1 presents normal simulations showing that the non-linearity gap is largest for standardized true effects $\omega \equiv \theta/\sigma$ which are close to 0, and remains above 0.2 for $\omega \leq 1$. It decreases monotonically as the true effect size ω increases and approaches zero in the limit.[53] It follows that the size of the non-linearity gap depends on the distribution of ω. The first row of graphs in Figure F2 plot the distribution of latent studies that have the 'correct' sign (this corresponds to the expression for the 'non-linearity' gap in equation (**??**)). We see that a high fraction of latent studies have $\omega < 1$, which explains why the non-linearity

[53] See Lemma A1.5 in Appendix A for a proof which shows that the non-linearity issue vanishes as true effect sizes approach infinity.

gap explains such a large role.

Wrong-sign gap.—Random sampling variation means that original estimates will occasionally have the 'wrong' sign. When this occurs, the replication probability is bounded above by 0.025. The extent to which this issue contributes to low replication rates therefore depends on the share of studies that have the wrong sign among significant studies. This share will be higher in settings with small true effects and low statistical power (Gelman and Carlin, 2014; Ioannidis et al., 2017). As power approaches 100%, the 'wrong-sign gap' approaches zero because the probability of drawing an estimate with the 'wrong' sign shrinks to zero.

Table C2 presents figures based on the estimated models, which show that significant results in experimental economics and psychology are relatively low-powered. The share of significant studies with the 'wrong' sign is 3% in economics, and 5% in psychology owing to lower statistical power. As a consequence, the wrong-sign gap is around 1 percentage point higher in psychology compared to economics.

Table 2C.2: Power and Estimates With the Wrong Sign For Statistically Significant Studies

	Experimental economics	Experimental psychology
Mean normalized true effect	2.835	2.251
Mean power	0.550	0.486
Share with wrong sign	0.030	0.054
Wrong-sign gap	0.018	0.030

Notes: Figures are based on simulated draws from the estimated distribution of latent studies in Table 1 in the main text. All statistics are calculated on the subset of statistically significant studies. The normalized true effect is defined as θ/σ. Power is defined as the probability of obtaining a statistically significant effect at the 5% level. The wrong-sign gap is defined in (**??**).

Selection-on-significance gap.—The Selection-on-significance gap is 1% in economics and slightly negative for psychology (i.e. conditioning on statistical significance increases the replication rate compared to when there is no conditioning). The sign of this gap is ambiguous because of two opposing effects from conditioning on statistical significance. To see these two effects, consider the figures in Table C2 which are based on the estimated empirical models. For the first effect, note that conditioning on significant findings increases mean bias in both

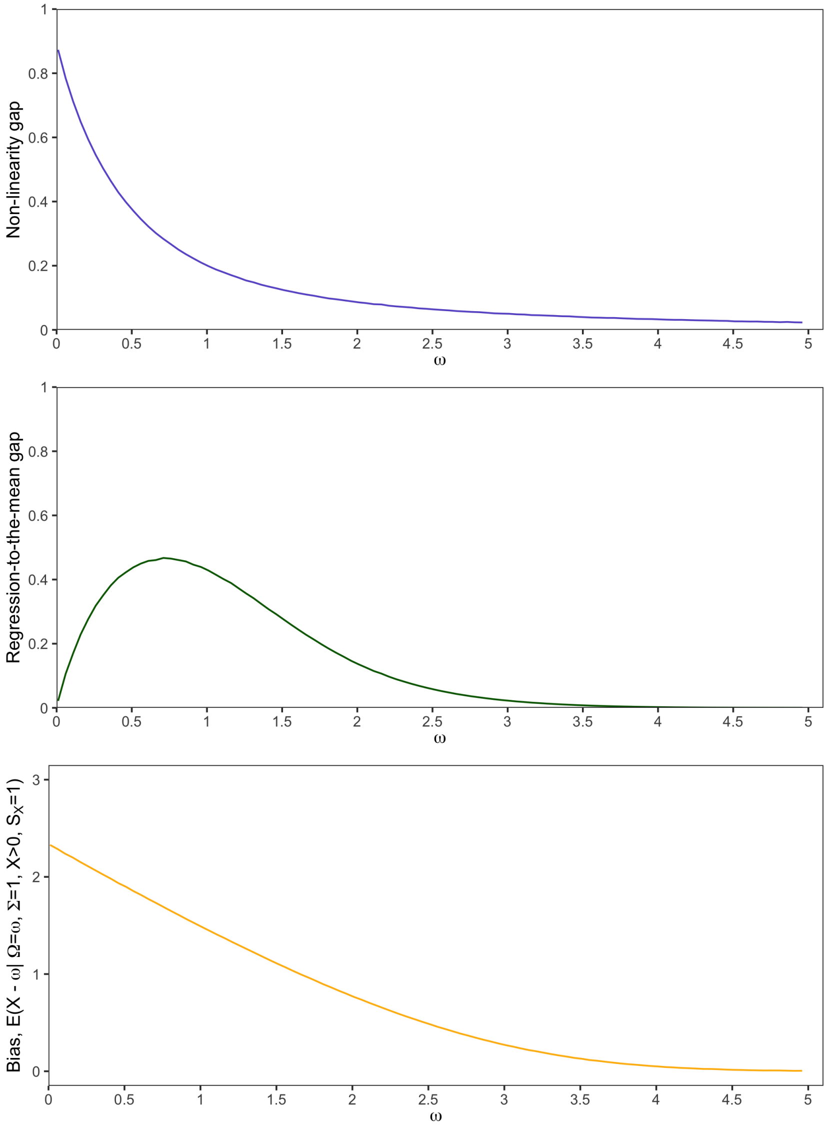

Figure 2C.1: Replication Rate Gap Decomposition: Monte Carlo Simulations

Notes: Plots are based on simulating studies from an $N(\omega, 1)$ distribution, for different values of ω. Replication estimates are drawn from a $N(\omega, \sigma_r(x, \beta)^2)$, where $\sigma_r(x, \beta)$ is set based on the common power rule to detect the original effect x with $1 - \beta = 0.92$ intended power. The non-linearity gap and regression-to-the-mean gap are based on equation (**??**) and calculated using Monte Carlo methods.

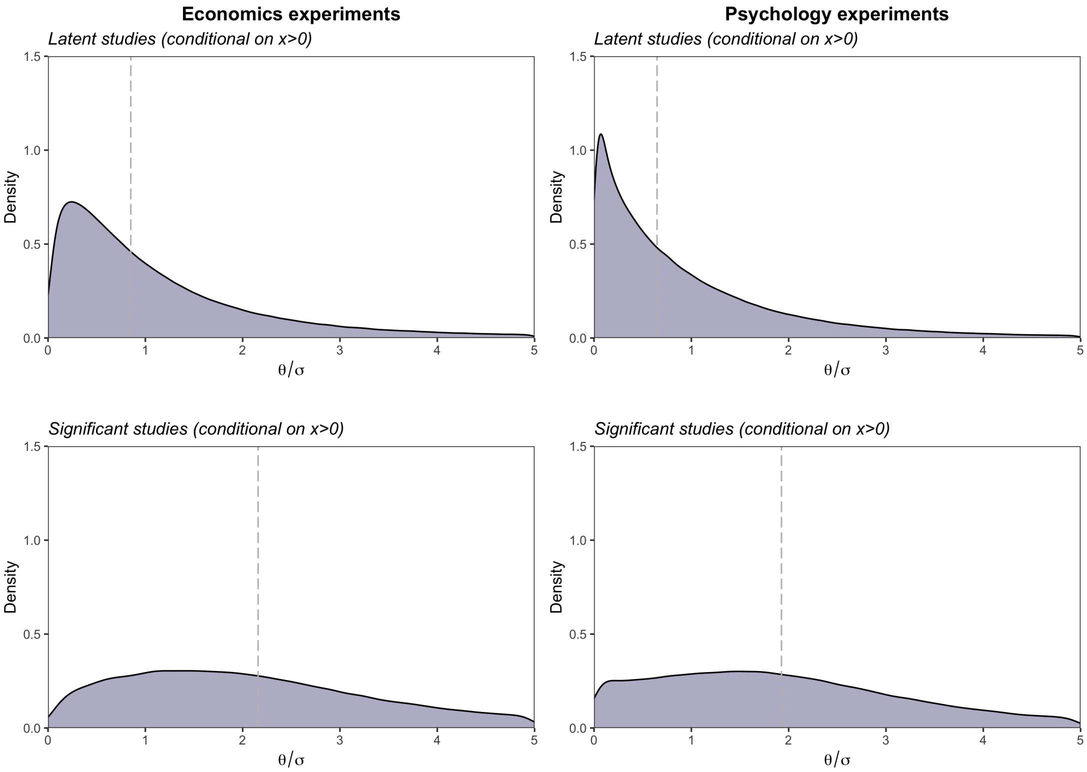

Figure 2C.2: Distribution of Normalized True Effects: Latent Studies and Significant Studies

Notes: Economics experiments refers to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). Densities are based on simulated draws from the estimated distribution of latent studies in Table 1 in the main text. Dashed vertical lines show the median of the distribution.

applications.[54] This makes replication more difficult for any fixed level of ω. For the second effect, note that conditioning also tends to select studies with larger standardized true effects ω, which have higher replication probabilities.[55] Higher replication probabilities arise because (i) bias is lower for larger true effects; and (ii) non-linearity effects are less severe for more highly powered studies.

The bottom panel in Figure C1 present normal simulations which show that mean bias decreases as the standard effect size increases, and approaches zero in the limit. The intuition is that censoring insignificant original estimates has little 'bite' when the true effect is very large, since the probability of drawing an insignificant estimate is very small. Thus, as true

[54] Bias is positive for latent studies because these statistics condition on original estimates X^* to have the same sign as true effects.

[55] The impact of conditioning on the full distribution of ω can be seen in Figure C2.

Table 2C.3: True Effect Sizes and Bias For Studies with the ‘Correct’ Sign

	Economics experiments		**Psychology experiments**	
	Latent	Published & significant	Latent	Published & significant
Mean bias	0.113	0.200	0.091	0.173
Mean standardized true effect	1.415	2.915	1.084	2.367

Notes: Economics experiments refers to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). Figures are based on simulated draws from the estimated distribution of latent studies from Table 1 in the main text. The mean of the standardized true effect is equal to $\mathbb{E}[\Omega^*|S^*_X, X^* > 0, D]$. Mean Bias is equal to $\mathbb{E}[X^* - \Omega^*|S^*_X, X^* > 0, D]$. ‘Latent studies’ allow S^*_X and D to be either 0 or 1. ‘Published & significant studies’ set $S^*_X = 1$ and $D = 1$.

effects become very large, the regression-to-the-mean gap approaches zero because the expected replication probability of statistically significant findings with the ‘correct’ sign converges to the expected replication probability of latent studies with the ‘correct’ sign.

2D Alternative Measures of Selective Publication

Proposition 1 shows that the replication rate is unresponsive to the most salient form of selective publication. For journals and policymakers seeking to change current norms, this highlights the need for more informative measures. In this section, I conduct policy simulations using the estimated model to show how three alternative measures respond to changes in the selective publication of null results:

1. **Replication CI:** This measure counts a replication as ‘successful’ if its 95% confidence interval covers the original estimate: $\mathbb{1}\big[X \in \big(X_r - 1.96\Sigma_r, X_r + 1.96\Sigma_r\big)\big]$.

2. **Meta-analysis:** The standard criterion of replication with the same sign and significance is applied to a fixed-effect meta-analytic estimate combining the original and replication estimate (uncorrected for selective publication): $\mathbb{1}\big[|X_m| \geq 1.96\Sigma_m, \text{sign}(X_m) = \text{sign}(X)\big]$ where X_m and Σ_m are the meta-analytic estimate and standard error, respectively.[56]

[56] The fixed-effects meta-analytic estimate is a weighted average of original and replication estimates: $X_m = \big(\omega_o X + \omega_r X_r\big)/(\omega_o + \omega_r)$, where the weights are equal to the precision of each estimate i.e. $(\omega_o, \omega_r) = (\Sigma^{-2}, \Sigma^{-2})$.

3. **Prediction interval:** Original and replication estimates are counted as 'consistent' under this approach if their difference is not statistically different from zero at the 5% level (Patil et al., 2016). This is equivalent to estimating a 95% 'prediction interval' for the original estimate and then determining if it covers the replication estimate: $\mathbb{1}\big[X_r \in \big(X - 1.96\sqrt{\Sigma^2 + \Sigma_r^2}, X + 1.96\sqrt{\Sigma^2 + \Sigma_r^2}\big)\big]\big)$.[57]

These alternative replication measures are frequently reported in large-scale replication studies (Open Science Collaboration, 2015; Camerer et al., 2016, 2018). In simulations, I calculate these measures over significant and insignificant published results, since conditioning on statistical significance makes them unresponsive to selective publication on null results (Proposition B2).

Simulations assume that all results significant at the 5% level are published, and that results insignificant at the 5% level are published with probability β_p. I then calculate how the various measures change with β_p to see how well they capture changes in selective publication (e.g. because of policy changes that reduce selective publication). Policymakers' successful efforts to increase the probability of publishing null results lead to an increase in the policy variable, β_p. Note that while model estimation assumes multiple cutoffs, policy simulations are performed assuming policymakers influence publication probabilities at a single cutoff (1.96) for simplicity (i.e. in the policy simulations I set $\beta_p = \beta_{p1} = \beta_{p2}$ and $\beta_{p3} = 1$ in social science).

Figure D1 shows the results. In line with Proposition 1, the replication rate is completely unresponsive to changes in the probability of publishing null results, making it a poor measure to evaluate efforts to reduce selective publication. Turning to alternative measures, note that the replication CI and meta-analysis measures actually *worsen* when more null results are published ($\beta_p \to 1$). This is because less selective publication leads to more small effects being selected for replication, which have relatively low replication probabilities under these

These weights minimize the mean-squared error of X_m (Laird and Mosteller, 1990). The variance of this estimator is given by $\Sigma_m^2 = 1/(\omega_o + \omega_r)$.

[57] This approach assumes that original and replication estimates share the same true effect and are statistically independent. For more details, see the Supplementary Materials for Patil et al. (2016).

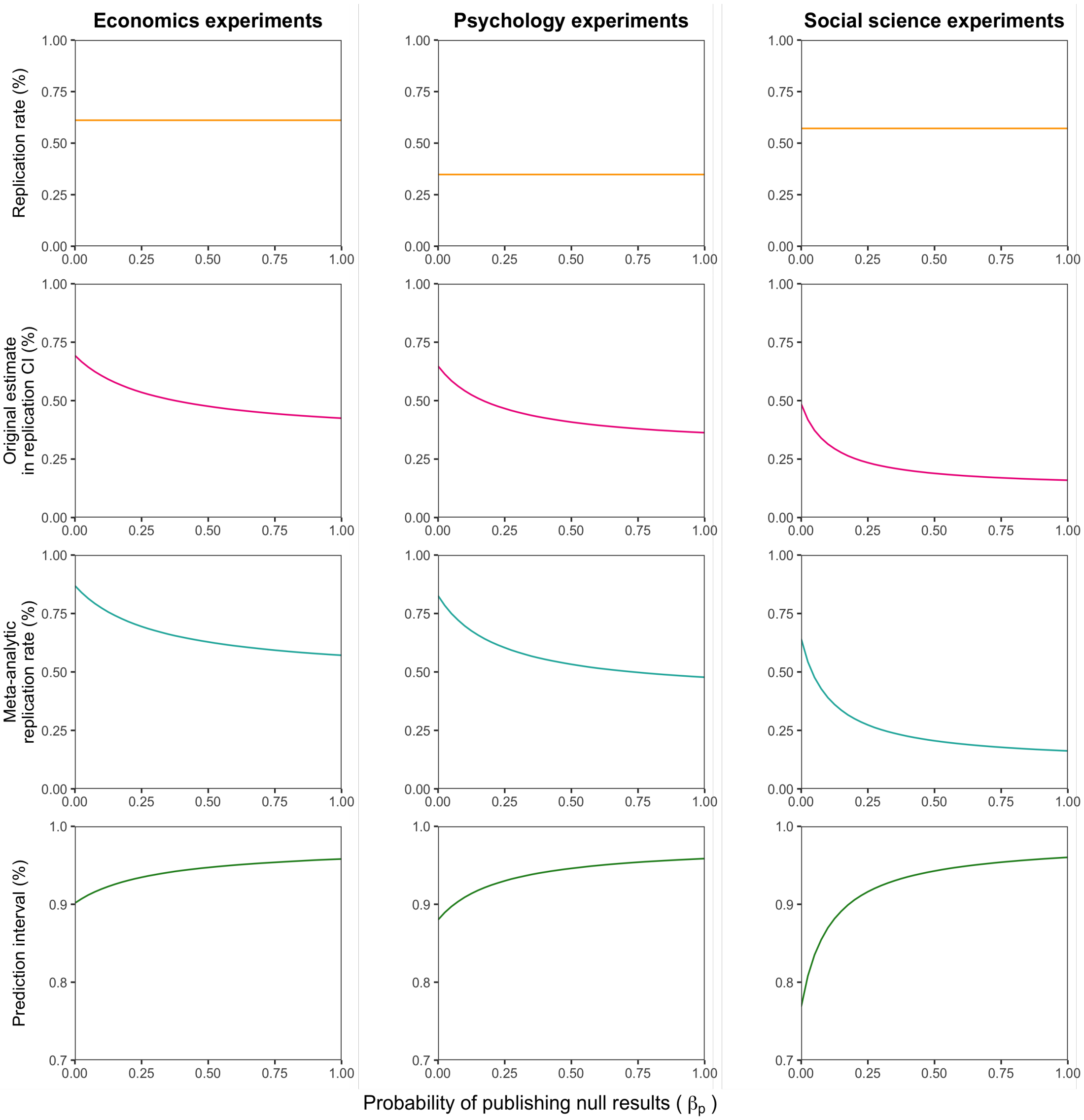

Figure 2D.1: Policy Simulations: Alternative Measures of Replication and Selective Publication

Notes: Details of each measure are provided in the main text. All measures except for the replication rate are calculated over significant and insignificant published results. Simulations use model estimates of the latent distribution of studies from Table 2.1 and set different levels of selective publication β_p. The first column reproduces replication rate predictions in Table 2.2.

approaches. By contrast, the prediction interval measure is low when selective publication is high, and approaches close to 95% as the probability of publishing null results approach one.[58] The prediction interval measure performs well because it explicitly accounts for the

[58] When $\beta_p = 1$, the prediction interval measure is slightly higher than 95% in all applications. This is because it assumes that the original estimate X and the replication estimate X_r are uncorrelated. In practice, the replication standard error is a function of the original estimate via the common power rule, which generates

decline in original power as more small effects are selected for replication. Noisy low-powered original studies contain limited information about true effects, which implies that a large range of replication estimates are statistically consistent with them.

Overall, for the purpose of evaluating efforts to reduce selective publication, these results suggest that calculating the prediction interval measure over a random sample of all published results could provide a useful alternative to the replication rate.

2E Replication Selection in Empirical Applications

Replication selection is a multi-step mechanism that first selects studies, and then selects results within those studies to replicate (since studies typically report multiple results). It consists of three steps:

1. **Eligibility**: define the set of eligible studies (e.g. journals, time-frame, study designs).
2. **Study selection:** on the set of eligible studies, a mechanism that select which studies will be included in the replication study.
3. **Within-study replication selection:** for selected studies, a mechanism for selecting which result(s) to replicate.

These three features of the replication selection mechanism influence the interpretation of the selection parameters $(\beta_{p1}, \beta_{p2}, \beta_{p3})$.

Economics experiments.—Consider these three steps in Camerer et al. (2016):

1. **Eligibility**: Between-study laboratory experiments in *American Economic Review* and *Quarterly Journal of Economics* published between 2011 and 2014.

some correlation between X and X_r.

2. **Study selection:** Camerer et al. (2016) select for publication all eligible studies that had ‘at least one significant between subject treatment effect that was referred to as statistically significant in the paper.’ Andrews and Kasy (2019) review eligible studies and conclude that no studies were excluded by this restriction. Thus, the complete set of eligible studies was selected for replication.

3. **Within-study replication selection:** the most important *statistically significant* result within a study, as emphasized by the authors, was chosen for replication. Further details are in the supplementary materials in Camerer et al. (2016). Of the 18 replication studies, 16 were significant at the 5% level and two had p-values slightly above 0.05 but were treated as ‘positive’ results for replication and included in the replication rate calculation.

I assume replication selection is random with respect to the t-ratio for results whose p-values are below or only slightly above 0.05. This implies that β_{p2} measures the relative probability of being published and chosen for replication for a result whose p-value is slightly above 0.05, compared to if it were strictly below 0.05. Overall, the empirical results are valid for the population of ‘most important’ significant (or ‘almost significant’) results, as emphasized by authors, in experimental economics papers published in top economics journals between 2011 and 2014.

Psychology.—Next, consider replication selection in Open Science Collaboration (2015):

1. **Eligibility**: Studies published in 2008 in one of the following journals: *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

2. **Study selection:** Open Science Collaboration (2015) write: ‘The first replication teams could select from a pool of the first 20 articles from each journal, starting with the first article published in the first 2008 issue. Project coordinators facilitated matching articles

with replication teams by interests and expertise until the remaining articles were difficult to match. If there were still interested teams, then another 10 articles from one or more of the three journals were made available from the sampling frame.' Importantly, the most common reason why an article was not matched was due to feasibility constraints (e.g. time, resources, instrumentation, dependence on historical events, or hard-to-access samples).

3. **Within-study replication selection:** the last experiment reported in each article was chosen for replication. Open Science Collaboration (2015) write that, 'Deviations from selecting the last experiment were made occasionally on the basis of feasibility or recommendations of the original authors.' A small number of results had p-values just above 0.05 but were treated as 'positive' results for replication, as in Camerer et al. (2016).

This selection mechanism implies that the empirical results are valid for the distribution of last experiments in the set of eligible journals. Since neither studies nor results were selected based on statistical significance, it is reasonable to treat the 'last experiment' rule as effectively random. In this case, we can interpret the results are being valid for all results in the eligible set of journals.

Social science experiments.—Finally, consider replication selection in Camerer et al. (2018):

1. **Eligibility**: Experimental studies in the social sciences published in *Nature* or *Science* between 2010 and 2015.

2. **Study selection:** Camerer et al. (2018) include all studies that: '(1) test for an experimental treatment effect between or within subjects, (2) test at least one clear hypothesis with a statistically significant finding, and (3) were performed on students or other accessible subject pools. Twenty-one studies were identified to meet these criteria.'

3. **Within-study replication selection:** Camerer et al. (2018) write, 'We used the following three criteria in descending order to determine which treatment effect to replicate

> within each of these 21 papers: (a) select the first study reporting a significant treatment effect for papers reporting more than one study, (b) from that study, select the statistically significant result identified in the original study as the most important result among all within- and between-subject treatment comparisons, and (c) if there was more than one equally central result, randomly select one of them for replication.' All results selected for replication had p-values strictly below 0.05.

This selection mechanism implies that the empirical results are valid for the population of statistically significant between- or within-subject treatment comparisons in experimental social science, which were identified by authors as the most 'important' and published in *Nature* or *Science* between 2010 and 2015.

2F Predicted Replication Rates Under Alternative Power Calculations

This appendix presents several extensions to the main empirical results on predicting replication rates in experimental economics, psychology and social science. The first extension allows for variation in the application of the common power rule around mean intended power. Results are similar to those in the main text, which assume no variability in the application of the common power rule. The second extension generates replication rate predictions under the rule of setting replication power equal to original power. This delivers lower replication rates than the common power rule.

Alternative power calculation rules.—Consider first the rule used for calculating replication power in the main text, and then two additional approaches. For concreteness, suppose we want to calculate the replication standard error for a simulated original study $(x^{sim}, \sigma^{sim}, \theta^{sim})$.

1. **Common power rule (mean):** This is the rule reported in the results in the main text. It assumes no variability in the application of the common power rule, such that all replications have mean intended power $1 - \beta$. This rule implies

$$\sigma_r^{sim}(x^{sim}, \beta) = \frac{|x^{sim}|}{1.96 - \Phi^{-1}(\beta)} \tag{96}$$

2. **Common power rule (realized):** Intended power for individual replications varied around mean intended power for at least two reasons. First, replication teams were instructed to meet minimum levels of statistical power, and encouraged to obtain higher power if feasible. Second, a number of replication in Open Science Collaboration (2015) did not meet this requirement. Figure F1 shows the distribution of realized intended power in replications for experimental economics and psychology. Realized intended power is right-skewed for psychology. In experimental economics and social science, realized intended power is distributed more tightly around mean.

   To capture variability in the application of the common power rule, take a random draw from the empirical distribution of $|x|/\sigma_r$ and denote it $1.96 - \widehat{\beta}^n$. Then realized intended power for simulated study $(x^{sim}, \sigma^{sim}, \theta^{sim})$ is equal to

$$\sigma_r^{sim}(x^{sim}, \widehat{\beta}^n) = \frac{|x^{sim}|}{1.96 - \Phi^{-1}(\widehat{\beta}^n)} \tag{97}$$

3. **Same power:** Set replication power equal to the power in the original study:

$$\sigma_r^{sim}(\sigma^{sim}) = \sigma^{sim} \tag{98}$$

   This rule has been proposed as a straightforward, intuitive approach for designing replication studies. In a review of replication studies by Anderson and Maxwell (2017), 19 of 108 studies used this approach.

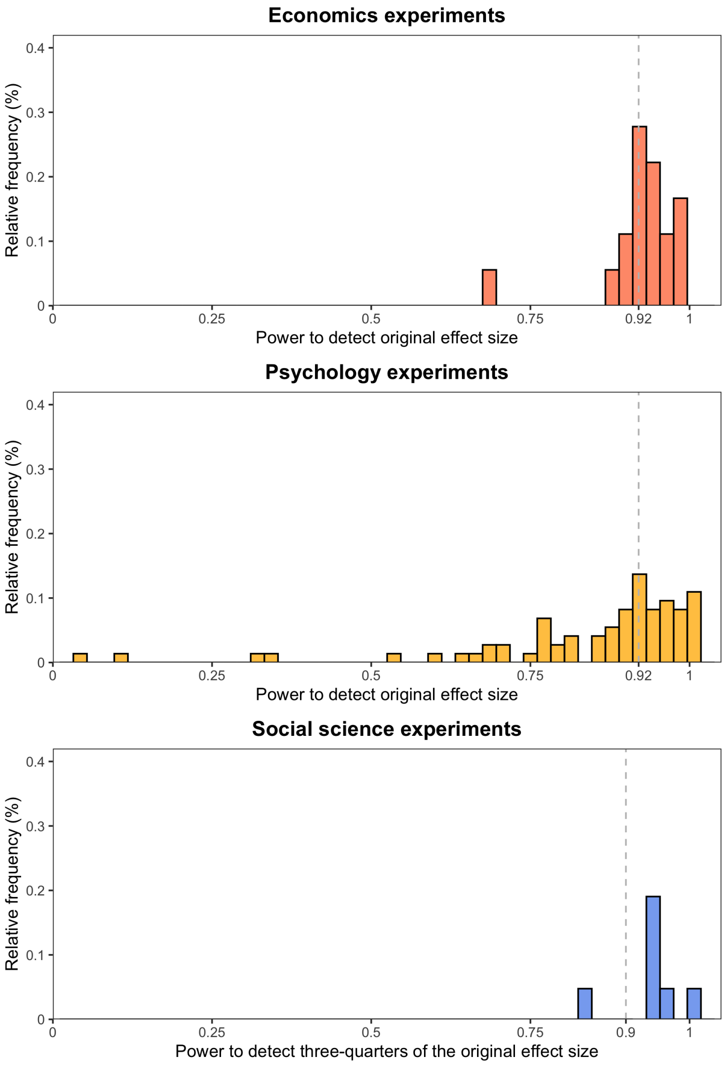

Figure 2F.1: Histograms of Realized Intended Power in Replication Studies

Notes: Data are from Camerer et al. (2016), Open Science Collaboration (2015), and Camerer et al. (2018), respectively. Realized intended power is defined as $1-\Phi(1.96-\psi\cdot\frac{x}{\sigma_r})$ with $\psi=1$ in economics and psychology and $\psi=3/4$ in social science. The horizontal dashed line is reported mean power in each application. In economics and psychology, this is 92% to detect the original effect size. In social science, this is 90% to detect three quarters of the effect size.

Results.—Table F1 presents the results for all three applications. Panel A shows that allowing intended power to vary across replications ('Realized power') yields similar replication rate prediction to assuming all replications have intended power equal to the report mean ('92% on X'). In fact, in all three applications, the accuracy improves very slightly under the realized power rule. The biggest differences is in psychology, because the realized power rule accounts for the fact that the distribution of intended power is right skewed.

Panel B examines the proposed rule of setting replication power equal to original power. In all three cases, the expected replication rate is lower than under the common power rule.

Table 2F.1: Replication Rate Predictions Under Alternative Replication Power Rules

	Economics	Psychology	Social science
A. Replication rate predictions			
Nominal target (intended power)	0.92	0.92	–
Observed replication rate	0.611	0.348	0.571
Mean power	0.600	0.545	0.543
Realized power	0.615	0.522	0.555
B. Alternative rule			
Same power	0.550	0.486	0.494

Notes: Economics experiments refer to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015), and social science experiments to Camerer et al. (2018). The replication rate is defined as the share of original estimate whose replications have statistically significant findings of the same sign. Figures in the first row are observed outcomes from large-scale replication studies. Remaining rows report predicted replication rates using parameter estimates Table 1 in the main text and assuming different rules for calculating replication power.

2G Relative Effect Size Predictions

The main focus of this article is the binary measure of replication based on the statistical significance criterion. This is because of its status as the primary replication indicator in the large-scale replication studies.[59] However, complementary measures are frequently presented alongside the replication rate. Perhaps the most common is the relative effect size, a continuous measure of replication defined as the ratio of replication effect size and original effect size. Relative effect sizes typically range between 0.35 and 0.7. Below, I include a brief theoretical discussion of the relative effect size and then present predictions of this measure using the

[59] Power calculations in replications are themselves typically designed to measure a binary notation of replication ‘success’ or ‘failure’.

estimated models.

Theoretical discussion.—The relative effect size for individual studies may be informative about biases affecting original studies, especially when original studies are well-powered. However, as an *aggregate* measure of reproducibility, the relative effect size measure may be subject to similar issues to the replication rate, at least in the case where it is defined exclusively over significant findings.

First, if the relative effect size is defined over significant original results, then it will be largely uninformative about the 'file-drawer' problem (Proposition B2).[60] Second, non-random sampling of significant results for replication mechanically induces inflationary bias in original estimates and regression to the mean in replication estimates, such that relative effect sizes are below one in expectation. Thus, similar to the replication rate, it has no natural benchmark against which to judge deviations, making it challenging to interpret. Relatedly, the average relative effect size is also very sensitive to power in original studies, which is unobserved. Figure G1 provides an illustration with intended power set to 0.9, which shows that the expected relative effect size for significant results is increasing in the power of original studies, and approaches one only as statistical power approaches 100%.

[60] Defining it over null results may present its own difficulties. For a perfectly measured null effect, the denominator in the statistic is equal to zero and the statistic is not well defined. On the other hand, if it is close but not equal to zero, then the statistic is highly sensitive to the precision of replication estimates; this raises questions about how one should set replication power when replicating a null effect.

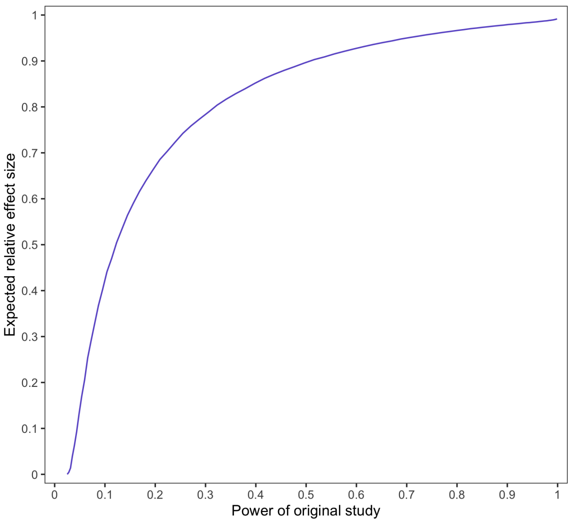

Figure 2G.1: Expected Relative Effect Size of Significant Original Studies and their Statistical Power

Notes: Illustration for the relationship between original power and the expected relative effect size of significant findings under the common power rule are both functions of $\omega = \theta/\sigma$ (normalized to be positive). Original power to obtain a significant effect with the same sign as the true effect is equal to $1 - \Phi(1.96 - \omega)$. The expected relative effect size is calculated by taking 10^6 draws of Z from $N(\omega, 1)$ and then calculating $\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \rho_{i,r}^{sig} / \rho_i^{sig}$, where $\rho = \tanh z$ denotes the Pearson correlation coefficient obtained by transforming the Fisher-transformed correlation coefficient (Fisher, 1915); and M_{sig} is the number of significant latent studies. The superscript *sig* reflects the fact that only statistically significant original results at the 5% level and their replications are included in the calculation. Replication estimates $z_{i,r}$ are drawn from an $N(\omega, \sigma_{r,i}(z_i, \beta)^2)$ distribution. The replication standard error is calculated using the common power rule to detect original effect sizes with 90% power (i.e. $1 - \beta = 0.9$), which is given by $\sigma_r(z_i, \beta) = |z_i|/[1.96 - \Phi^{-1}(\beta)] = |z_i|/3.242$.

Empirical results.—The estimated models in Table 1 in the main text can be used to generate predictions of the average relative effect sizes. To procedure for simulating replications is identical to the procedure outlined in the main text for the replication rate case. Let $\{x_i, \sigma_i, x_{r,i}, \sigma_{r,i}\}_{i=1}^{M_{sig}}$ be the set of simulated original studies that are published and significant, and their corresponding replication results; M_{sig} is the size of the set. The predicted relative effect size is equal to

$$\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \frac{\rho_{i,r}^{sig}}{\rho_i^{sig}} \tag{99}$$

where $\rho = \tanh z$ denotes the Pearson correlation coefficient which is obtained by transforming the Fisher-transformed correlation coefficient (Fisher, 1915). I also present results for the median relative effect size. Results are presented in Table G2. The predicted average relative effect size is relatively close to observed average relative effect size in economics, somewhat further off in social science, and quite far off in psychology. In each case, the predicted average relative effect size is optimistic compared to the observed value. In economics and psychology, the difference in predicted and observed relative effect sizes is not statistically different from zero, while in psychology it is. Predictions for median relative effect sizes show qualitatively similar results.

Table 2G.1: Average Relative Effect Size Predictions

	Economics	Psychology	Social Sciences
Observed relative effect size (mean)	0.657	0.374	0.443
Predicted relative effect size (mean)	0.703	0.637	0.533
	(0.135)	(0.060)	(0.141)
Observed relative effect size (median)	0.691	0.292	0.527
Predicted relative effect size (median)	0.747	0.674	0.595
	(0.129)	(0.063)	(0.240)

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social science experiments to Camerer et al. (2018). Observed relative effect sizes are based on data from large-scale replication studies. Predicted average relative effect sizes are calculated using equation (99) and the procedure outlined in the text. Standard errors are calculated using the delta method.

2H Extending the Replication Rate Definition

This appendix analyzes a generalization of the replication rate definition that extends to insignificant results. It outlines a number of issues with this proposal.

The Generalized Replication Rate.—Suppose we extend the definition of the replication rate such that insignificant original results are counted as 'successfully replicated' if they are also insignificant in replications. Assume replication selection is a random sample of published results. Then we have the following definitions:

Definition H1 (Generalized replication probability of a single study). *The replication probability of a study (X, Σ, Θ) which is published $(D = 1)$ and chosen for replication $(R = 1)$ is*

$$\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Big) = \begin{cases} \mathbb{P}\Big(\frac{|X_r|}{\sigma_r(X,\Sigma,\beta)} \geq 1.96, \text{sign}(X) = \text{sign}(X_r)\Big| X, \Theta, \sigma_r(X, \Sigma, \beta)\Big) & \text{if } 1.96.\Sigma \leq |X| \\ \mathbb{P}\Big(\frac{|X_r|}{\sigma_r(X,\Sigma,\beta)} < 1.96\Big| X, \Theta, \sigma_r(X, \Sigma, \beta)\Big) & \text{if } 1.96.\Sigma > |X| \end{cases} \tag{100}$$

Definition H2 (Expected generalized replication probability). *The expected generalized replication probability equals*

$$\mathbb{E}\Big[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Big)\Big] = \mathbb{P}\Big(1.96.\Sigma \leq |X|\Big)\mathbb{E}\Bigg[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Bigg| X, \Theta, \sigma_r(X, \Sigma, \beta), 1.96.\Sigma \leq |X|\Bigg]$$

$$+\Big(1 - \mathbb{P}\Big(1.96.\Sigma \leq |X|\Big)\Big)\mathbb{E}\Bigg[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Bigg| X, \Theta, \sigma_r(X, \Sigma, \beta), 1.96.\Sigma > |X|\Bigg] \tag{101}$$

First, note that Definition H2 equals the standard replication rate definition when the expectation is taken only over significant studies because, in this case, $\mathbb{P}\Big(|X| \leq 1.96.\Sigma\Big) = 0$. Thus, the degree to which the expected generalized replication probability differs from the standard expected replication probability depends on two factors. First, the share of published results that are insignificant. Second, the expected probability that replications will be insignificant conditional on original estimates being insignificant.[61]

Empirical Results.—To analyze the generalized replication rate, we can apply the empirical approach outlined in the main text, but using the generalized definition in place of the original definition. Recall that the original replication rate is invariant to publication bias against null results. The generalized replication rate, by contrast, does vary as the degree of selective publication against null results changes. Thus, two sets of results are presented for comparison. The first set assumes selective publication using estimated selection parameters in Table 1 in the main text. The second set assumes no selective publication (i.e. that all results are published with equal probability). We examine two rules for calculating replication power: the common power rule and the original power rule (where the replication standard error is set equal to the

[61] Additionally, note that this definition implies that if $\theta = 0$, then $\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)|\Theta = 0\Big) = 0.90375$. That is, the replication probability of null results is constant and independent of power in original studies and replication studies.

original standard error). For more details on different rules for calculation replication power, see Appendix E.

Table H1 reports the results for both applications. Under the common power rule, the simulated generalized replication rate remains below intended power in both publication regimes. Under the original power rule, it is relatively low when there is selective publication and around 80% when there is no selective publication.

These generalized replication rate predictions differs from the standard replication rate predictions for two reasons: (i) the share of insignificant results in the published literature and (ii) the replication probability when results are insignificant, which depends on the power rule used in replication studies. On the first point, moving from the selective publication regime to the no selective publication regime implies a dramatic increase in the share of insignificant published results; in both applications, null results change from a minority of published results to a majority. On the second point, the results show that the replication power rules considered here have some undesirable properties. First, note that the common power rule is designed to detect original estimates with high statistical power. This implies that low-powered, insignificant original results will be high-powered in replications, which increases the probability that they are significant and thus counted as replication 'failures' under the generalized definition. The original power rule has the reverse problem. On the one hand, low-powered, insignificant original studies are likely to be insignificant in replications, which counts as a 'successful' replication under the generalized definition. However, on the other hand, low-powered, significant original studies will have low replication probabilities when the same low-powered design is repeated in replications. The generalized replication rate therefore depends crucially on the share of significant and insignificant findings in the published literature, and the distribution of standard errors. Under the original power rule with no selective publication, the generalized replication rate is around 80% in both applications; however, with greater power in original studies, the replication rate would fall.

While the generalized replication rate changes as selective publication is reduced, the direction of this change depends on which replication power rule is used: with the original power rule the replication rate increases, while with the common power rule it decreases.

Overall, generalizing the replication rate with Definition H2 does not deliver replication rates close to intended power under the common power rule. For the original power rule, it is higher when there is no selective publication because replications repeat low-power designs for low-powered original studies with insignificant results. The generalized replication rate under this original power rule will therefore be sensitive to the distribution of power in original studies.

Table 2H.1: Predicted Generalized Replication Rate Results

	Simulated statistics	
A Economics experiments	92% for X	Original power
Selective publication		
Generalized replication rate	0.600	0.553
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.600	0.551
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.574	0.789
$\mathbb{P}(S_X = 1)$	0.993	0.993
$\mathbb{P}(S_X = 0)$	0.007	0.007
No selective publication		
Generalized replication rate	0.432	0.773
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.582	0.515
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.378	0.867
$\mathbb{P}(S_X = 1)$	0.268	0.268
$\mathbb{P}(S_X = 0)$	0.732	0.732
B Psychology experiments		
Selective Publication		
Generalized replication rate	0.546	0.526
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.544	0.487
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.563	0.839
$\mathbb{P}(S_X = 1)$	0.890	0.890
$\mathbb{P}(S_X = 0)$	0.110	0.110
No selective publication		
Generalized replication rate	0.490	0.798
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.535	0.469
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.478	0.886
$\mathbb{P}(S_X = 1)$	0.209	0.209
$\mathbb{P}(S_X = 0)$	0.791	0.791

Notes: Economics experiments refer to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). The generalized replication rate is defined in the text. The indicator variable S_X equals one for significant results and zero otherwise. Economics experiments refers to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). Simulated statistics are based on parameter estimates in Table 1 in the main text. Different column represent different rules for calculating power in replications.

Chapter 3

Optimal Publication Rules for Evidence-Based Policy

Abstract. Empirical research can inform evidence-based policy choice but may be censored due to publication bias. How does this impact the decisions of policymakers who do not have, or are unwilling to use, prior beliefs about a policy's impact? For minimax regret policymakers, we characterize the optimal treatment rule with selective publication against statistically insignificant results. We then show that the optimal publication rule which minimizes maximum regret is non-selective. This contrasts with the optimal publication rule for Bayesian policymakers studied in the literature, where only 'extreme' results that sufficiently move the prior are published. Thus, in the minimax regret framework, the optimal publication regime for policy choice is consistent with valid statistical inference in scientific research.

3.1 Introduction

Publication bias has been widely-documented across various fields and led to debates in the scientific community about reforming publication norms (Ioannidis, 2005; Franco et al., 2014; Nosek et al., 2015; Miguel and Christensen, 2018; Nosek et al., 2018; Andrews and Kasy, 2019). Proposals to combat publication bias are often aimed at mitigating selective publication of statistically significant findings. For example, launching journals dedicated to publishing null results (e.g. *PLOS One*); promoting preregistered analysis plans which are reviewed and published prior to data collection (e.g. the *Journal of Development Economics*); banning the use of stars to denote significance when presenting estimation results (e.g. the *American Economic Review*); and even abandoning statistical significance altogether (McShane et al., 2019).

However, non-selective publication may not necessarily be optimal from the perspective of a decision-maker who uses evidence from published studies to inform a policy decision. Frankel and Kasy (2022) develop a model of a Bayesian decision-maker who has a prior distribution over possible treatment outcomes and updates their beliefs using evidence from published studies before making a policy decision. When publication entails a cost (e.g. the opportunity cost of drawing attention away from other studies), the optimal rule is to publish only 'extreme' results that sufficiently move prior beliefs. This gives rise to a striking trade-off: selective publication enhances policy relevance while at the same time deteriorating statistical credibility.

While selective publication may be optimal for a Bayesian decision-maker, in many situations, policymakers may be unable or unwilling to base decisions on prior beliefs about treatment outcomes. For example, they may have insufficient information to form a reasonable prior, or if when decisions are made by a group, prior beliefs of different group members may conflict with one another. A common alternative to relying on prior beliefs is to introduce ambiguity on the treatment outcomes and pursue robust decisions.

In this paper, we consider a policymaker that aims to minimize maximum regret (Savage, 1951; Manski, 2004), where regret equals the difference between the highest possible expected welfare outcome given full knowledge of the true impact of all treatments and the expected welfare attained by the statistical decision rule. We first characterize the minimax regret decision rule of the policymaker in the presence of publication bias, and then derive the optimal publication rule that minimizes the value of minimax regret. In contrast to the Bayesian framework, we show that the optimal publication rule for minimax regret decision-makers is completely non-selective i.e. publication decisions do not depend on statistical significance. Importantly, non-selective publication implies valid statistical inference. Thus, in the minimax regret framework, there exists no trade-off between policy relevance and statistical credibility.

Following Manski (2004), Stoye (2009), and Tetenov (2012), our model considers a policymaker whose problem is to assign members of a population one of two treatments: a status quo treatment and an innovative treatment. A study about the relative effectiveness of the treatments is conducted. However, the study is only observed by the policymaker if it is published, which may depend on its statistical significance. We consider the case where t-ratios in a symmetric interval around zero are censored with probability $\beta_p \in [0, 1]$ e.g. statistical significance at the 5% level implies that t-ratios between -1.96 and 1.96 will be published with probability β_p. Additionally, publication may also entail a cost $c \geq 0$. We consider a policymaker who correctly accounts for publication bias when choosing their statistical treatment rule (and later consider a naive policymaker who does not account for it). If a study is published, the policymaker observes it and implements the innovative policy if its relative effect size is greater than a chosen threshold value T. Alternatively, if a study is not published, then the policymaker must act without evidence and implements the innovative treatment with probability δ_0. The policymaker chooses a statistical treatment rule – consisting of the threshold rule T and the default action δ_0 – that minimizes their maximum regret, that is, the expected welfare loss relative to optimal welfare attained with knowledge of the true treatment effect.

We show that the minimax regret decision rule implements the innovative treatment if and only if the published estimate of the relative efficacy of the innovative treatment is positive, and randomizes between treatments with equal probability when no study is published i.e. $(T^*, \delta_0^*) = (0, \frac{1}{2})$. The intuition for this result follows from the two key factors. First, the decision-maker's welfare equally weighs Type I errors (from mistakenly implementing an inferior treatment) and Type II errors (failing to implement the superior treatment). Second, the class of publication rules we consider censors insignificant empirical results symmetrically around zero. The first symmetry implies that the decision-maker will implement the innovative treatment when the published evidence supports the innovative treatment having a positive effect, and remain with the status quo treatment otherwise. Combined with the second symmetry, we can conclude that when the study is published, the sign of the estimate (i.e., $T^* = 0$) is sufficient for the decision-maker to infer the sign of the effect, and when no study is published, the decision-maker has no evidence regarding the sign of the relative treatment effect and will therefore randomize between treatments.

Given the minimax regret rule of the decision-maker, we optimize the value of minimax regret with respect to the publication rule. As the main result, we show that the resulting optimal publication rule is to *publish all results*. This accords well with common intuition: receiving evidence from a published study about the relative effectiveness of treatments allows the policymaker to do better than in the case where no study is published and they must randomize between treatments.

It is notable, however, that the opposite conclusion is reached when considering a Bayesian decision-maker, for whom the optimal publication rule censors relatively uninformative studies that do little to move prior beliefs of the decision-maker (Frankel and Kasy, 2022). Two differences account for this. First, publication costs enter the expected welfare function in the Bayesian framework of Frankel and Kasy (2022), while they do not appear in the expression for regret in our framework. This is because regret equals the probability of making an inferior treatment choice multiplied by the magnitude of the loss from doing so. Neither quantity is affected by publication costs. Put differently, publication costs are constant with respect to the decision rule and therefore have no impact on regret. The second difference is that Frankel and Kasy (2022) define null results in the Bayesian framework as those which do not move prior beliefs. By contrast, there is no notion of prior beliefs in the minimax regret framework. We instead use the common definition of null results as those which are statistically indistinguishable from zero. Accordingly, we consider a class of symmetric publication rules that yield no information about the sign of the true effect when studies are not published. Since published studies will always provide some evidence on the sign of the true effect, the optimal publication regime in terms of the regret criterion is to publish all the results.

Our results highlight that the optimal publication regime can change drastically depending on what optimality criterion the policymaker pursues for policy choice. Which optimality criterion is relevant in practice may depend on the factors such as behavioural axioms of the decision-makers, availability of the prior belief of the policy effect, and/or the form of publication bias relevant to the empirical literature of interest.

We consider three main extensions to the baseline model. Following Tetenov (2012), we first extend the model to incorporate decision criteria that asymmetrically weigh Type I error (from mistakenly implementing the inferior treatment) and Type II error (from mistaking rejecting a superior treatment). We provide numerical evidence consistent with the conjecture the optimal publication rule for minimax regret decision-makers with asymmetric regret criteria is also non-selective.

Second, we consider a naive policymaker who, unlike the sophisticated policymaker in the main analysis, does not account for publication bias when choosing their decision rule. Naive policymakers could in some cases be more realistic than sophisticated policymakers, because sophistication demands both knowledge of the publication rule and the ability to correctly adjust for it. In this model, the naive policymaker's expected welfare (and regret) is misspecified because they believe, erroneously, that there is no publication bias.[62] We evaluate their subsequent decision rule based on the worse case scenario under *correctly* specified regret. We show that minimax regret for the naive policymaker is weakly higher than for the sophisticated policymaker.[63] Thus, in general, the naive policymaker chooses a non-optimal decision rule because they fail account for publication bias.

The optimal publication rule in the main analysis assumes that the policymaker and the journal have the same preferences, namely, to minimize maximum regret. In the third extension, we consider the optimal publication rule under misaligned preferences. In particular, we consider the case where the policymaker chooses their decision rule to minimize maximum regret, but where the journal chooses the publication rule to maximize welfare (under some

[62] This affects: (i) their beliefs about the distribution of the published estimates; and (ii) implies that they make no inferences about the size of the treatment effect when no study is published.

[63] Minimax regret for the naive policymaker is strictly higher when the Type I and Type II error are unequally weighted.

prior distribution for the policy's effect). The main result shows that the journal's optimal action takes the form a simple threshold rule: publish all results if the cost c is sufficiently low; otherwise, censor all null results. Thus, in the case where publication costs are low, the optimal publication rule under misaligned preferences is the same as with aligned preferences. However, when publication costs are high, it is possible that censoring all null results is optimal.

Related Literature. This article contributes to the literature on statistical decision theory (Manski, 2004; Stoye, 2009; Tetenov, 2012). It generalizes the canonical model in the minimax regret framework to incorporate publication bias against null results. We characterize the optimal decision rule that minimizes maximum regret and extend results to the case where Type I and Type II error are weighted asymmetrically. Our model coincides with the canonical model in the special case where there is no publication bias.

This article also contributes to the meta-science literature on publication bias and optimal publication rules (Ioannidis, 2005; Andrews and Kasy, 2019; Miguel and Christensen, 2018; Frankel and Kasy, 2022). It is most closely related to Frankel and Kasy (2022), who examine a similar problem in a Bayesian framework. In contrast to a Bayesian framework, where the optimal publication rule selects only 'extreme' results for publication, we show in a minimax regret framework that the optimal publication rule is completely non-selective.

3.2 Model

3.2.1 Setup

The policymaker's problem is to assign two treatments to a population with observationally identical members: the status quo treatment ($t = 0$) and the innovative treatment ($t = 1$). Following Manski (2004), suppose that each member j in population J has a treatment response function $y_j(\cdot) : \{0, 1\} \to Y$ mapping treatments into outcomes. The population is a probability space (J, Ω, P). The probability distribution $P[y(\cdot)]$ of the random vector $y(\cdot)$ describes treatment response across the population. The population is "large" in the sense that J is uncountable and $P(j) = 0$. Next, let $\mathbb{E}[y(1)] - \mathbb{E}[y(0)] \equiv \theta \in \Theta \subseteq \mathbb{R}$ be the unknown average treatment effect of the innovative treatment relative to the status quo treatment, with the status quo treatment normalized to zero. When $\theta > 0$, the innovative treatment is preferred; otherwise, the status quo treatment is preferred.

Evidence about θ may be observed by the policymaker in the form of a published study. However, not all studies are necessarily published. Consider first a *latent study* (published or unpublished), which is characterized by (X, σ), where X is the *estimated treatment effect* and σ is the known *standard error*. We assume X is normally distributed on $\mathcal{X} = \mathbb{R}$ and normalize $\sigma = 1$, so that $X|\theta \sim N(\theta, 1)$. This assumption is motivated by the fact that study estimates are widely assumed to be approximately normal in practice. The normalization is for notational convenience, since σ is known and fixed. The journal observes the latent study $(X, 1)$ and decides the probability of publication according to their publication rule, $p : \mathcal{X} \to [0, 1]$. Let $D = 1$ denote the event when a study is published and $D = 0$ the event when it is not. We consider the class of publication rules where absolute t-ratios below a critical threshold t_α may be published with a lower probability than those above that threshold:

Assumption 3.2.1 (Publication Selection Function)**.** *Let* $p(X) = 1 - (1 - \beta_p) \cdot \mathbb{1}[|X| < t_\alpha]$ *with* $\beta_p \in [0, 1]$.

The form of publication bias in Assumption 1 implies that published estimated treatment effects follow a mixture of truncated normal densities, where the region below the critical threshold of the density is down-weighted and the region above it is up-weighted. Denote the cdf as

$$F(x|\theta, D = 1) \equiv \frac{\int_{-\infty}^{x} p(y)\phi(y - \theta)dy}{\int p(y)\phi(y - \theta)dy}, \tag{102}$$

where $\phi(x) = (2\pi)^{-1/2} \exp(2^{-1}x^2)$ is the probability density function of the standard normal distribution.

The policymaker's decision rule must cover two possible realizations of the publication process: the event when the study is published ($D = 1$) and event when it is not ($D = 0$). Let $Z = X \cdot D + \{\text{missing}\}(1 - D)$ and the *statistical treatment rule* be $\delta : Z \to [0, 1]$, with

$$\delta(X, D) = \begin{cases} \delta_1(X) & \text{if } D = 1 \\ \delta_0 & \text{if } D = 0 \end{cases} \tag{103}$$

That is, $\delta(X, D)$ maps study outcomes to treatment assignment proportions when the study is published, and assigns a default action $\delta_0 \in [0, 1]$ when it is not.

We first consider a sophisticated policymaker who knows the exact form of publication bias and correctly accounts for it when choosing their optimal decision rule. For example, a sophisticated policymaker could estimate $p(\cdot)$ from a sample of studies in the published literature (e.g. by using the Andrews and Kasy (2019) model). Their utility from treatment rule $\delta(X, D)$ with treatment effect θ and observed data X is given by

$$U\big(\delta, \theta\big) = \theta D \delta_1(X) + \theta(1 - D)\delta_0 - Dc \tag{104}$$

where $c \geq 0$ represents the cost of publishing an article. Following Frankel and Kasy (2022), we interpret this cost as the opportunity cost of directing the public's limited attention away from other studies. Welfare for a statistical decision rule $\delta(X, D)$ corresponds to a shared objective by the policymaker and the journal. Expected welfare is obtained by integrating over possible study outcomes:

$$\begin{aligned} W\big(\delta, \theta\big) = & \int U\big(\delta, \theta\big) f(x'|\theta) dx' \\ = & \theta \cdot \mathbb{P}[D = 1|\theta] \cdot \mathbb{E}[\delta_1(X)|\theta, D = 1] + \theta \cdot \big(1 - \mathbb{P}[D = 1|\theta]\big)\delta_0 - \mathbb{P}[D = 1|\theta] \cdot c \\ = & W_1(\delta_1, \theta) + W_0(\delta_0, \theta) - \mathbb{P}[D = 1|\theta] \cdot c \end{aligned} \tag{105}$$

where $W_1(\delta_1, \theta) = \theta \cdot \mathbb{P}[D = 1|\theta] \cdot \mathbb{E}[\delta_1(X)|\theta, D = 1]$ is the welfare for the case that the study is published, and $W_0(\delta_0, \theta) = \theta \cdot \big(1 - \mathbb{P}[D = 1|\theta]\big)\delta_0$ is the welfare for the case that the study is not published.

Finally, regret is given by the difference between the highest possible expected welfare conditional on θ and the expected welfare under the treatment rule. Let $W^*(\theta)$ be the welfare attained by the oracle rule $\delta_1 = \delta_0 = \mathbb{1}(\theta > 0)$. Then regret is given by

$$R\big(\delta, \theta\big) = W^*(\theta) - W\big(\delta, \theta\big)$$

$$= \begin{cases} -\theta\Big(\mathbb{P}[D = 1|\theta] \cdot \mathbb{E}[\delta_1(X)|\theta, D = 1] + (1 - \mathbb{P}[D = 1|\theta])\delta_0\Big) & \text{if } \theta \leq 0 \\ \theta\Big(\mathbb{P}[D = 1|\theta] \cdot \big(1 - \mathbb{E}[\delta_1(X)|\theta, D = 1]\big) + (1 - \mathbb{P}[D = 1|\theta])(1 - \delta_0)\Big) & \text{if } \theta > 0 \end{cases} \tag{106}$$

That is, regret equals the magnitude of the loss $|\theta|$ multiplied by the expected probability of assigning the inferior treatment choice. The expected probability of assigning the wrong treatment is a weighted average of making the incorrect decision, where weights correspond to different realizations of the publication process. Two points are worth noting. First, the publication cost does not appear in the expression for regret because it is constant with respect to the policymaker's decision rule. Second, this expression for regret reflects a sophisticated policymaker who has complete knowledge of publication bias. In particular, the sophisticated policymaker correctly accounts for publication when considering the distribution of the estimated treatment effect X, and the probability that a study is or is not published. In a later section, we study a naive policymaker who does not account for publication bias.

The expression for minimax regret can be further simplified by restricting the class of decision rules for $\delta_1(X)$ to threshold rules. As in Tetenov (2012) for the Gaussian signal case, this restriction is innocuous since in terms of welfare $W_1(\delta_1, \theta)$ for published case, the class of threshold rules is essentially complete, i.e., for any admissible rule $\delta_1(X)$ in terms of $W_1(\delta_1, \theta)$, its welfare level can be replicated by a threshold rule.

Lemma 3.2.1 (Threshold Rules are Essentially Complete)**.** *Under Assumption 3.2.1, the class* *of threshold decision rules $\delta_1^T(X) = \mathbb{1}[X > T]$ is essentially complete in terms of the welfare of $W_1(\delta_1, \theta)$.*

With Lemma 1, any decision rule δ is fully characterized by a tuple (δ_1^T, δ_0). The first element corresponds to the threshold rule $\mathbb{1}[X > T]$ and is applicable when a study is published. The second element is a default action δ_0 and is applicable when a study is not published. With this simplification, we can rewrite regret for decision rule δ in equation (111) as

$$R\big((\delta_1^T, \delta_0), \theta\big) = \begin{cases} -\theta\Big(\mathbb{P}[D = 1|\theta] \cdot [1 - F(T|\theta, D = 1)] + (1 - \mathbb{P}[D = 1|\theta])\delta_0\Big) & \text{if } \theta \leq 0 \\ \theta\Big(\mathbb{P}[D = 1|\theta] \cdot F(T|\theta, D = 1) + (1 - \mathbb{P}[D = 1|\theta]) \cdot (1 - \delta_0)\Big) & \text{if } \theta > 0 \end{cases} \tag{107}$$

Finally, the optimal decision rule (T^*, δ_0^*) selects the rule which minimizes maximum regret:

$$(T^*, \delta_0^*) = \arg \min_{(T,\delta_0) \in \mathbb{R} \times [0,1]} \max_{\theta \in \mathbb{R}} R\big((\delta_1^T, \delta_0), \theta\big) \tag{108}$$

3.3 Optimal Publication Rule For Minimax Regret

In this section, we first characterize the optimal minimax regret decision rule for the sophisticated policymaker. Given this decision rule, we then show analytically that the optimal publication rule that minimizes the value of minimax regret is non-selective. Finally, we provide numerical evidence that this result generalizes to the case where the policy-maker's concerns over Type I and Type II error are asymmetric. Proofs are in Appendix 3A.

3.3.1 Optimal Minimax Regret Decision Rule

In the presence of publication bias, decision-makers must choose optimal actions for when studies are published and when they are not. The following lemma characterizes the optimal minimax regret decision rule:

Lemma 3.3.1 (Minimax Regret Decision Rule Under Publication Bias)**.** *Under Assumption 3.2.1 for the sophisticated policymaker,* $(T^*, \delta_0^*) = \left(0, \frac{1}{2}\right)$ *for any* $\beta_p \in [0, 1]$.

When a study is published, the optimal minimax decision rule implements the innovative treatment if the published estimate is positive; and when no study is published, the policymaker randomly choose between treatments with equal probability. With symmetric concern of Type I and Type II error, the policymaker will implement the innovative treatment when there is evidence that it is superior to the status quo treatment. When no study is published there exists no such evidence and hence the policymaker randomizes between treatments. The only information available to the policymaker when no study is published is that the difference in the efficacy of treatments is likely to be small, since publication bias censors small effect sizes. However, because publication bias is symmetric around zero, no information is gained about which treatment might be superior. When the study is published, a positive estimate is more likely to come from a positive true treatment effect, while a negative estimate is more likely to come from a negative true treatment effect. Hence, the policymakers' threshold rule implements the innovative policy if and only if the signal is positive.

It is noteworthy that the optimal minimax regret threshold decision rule in the presence of publication bias is identical to the case where there is no publication bias (Tetenov, 2012). This is a consequence of the symmetry of the problem when Type I error and Type II error are equally weighted by the policymaker.

3.3.2 Optimal Non-Selective Publication Rule

Given the minimax decision rule $(T^*, \delta_0^*) = \left(0, \frac{1}{2}\right)$, what publication rule minimizes the value of minimax regret? The following result provides the answer:

Proposition 3 (Non-Selective Optimal Publication Rule)**.** *Under Assumption 3.2.1, the value of minimax regret is minimized for the sophisticated policymaker when the publication rule is non-selective, that is, when* $\beta_p = 1$.

The optimal publication rule for a minimax regret policymaker publishes all results. Thus, under the optimal publication regime, the policymaker's problem collapses into the standard model with no publication bias in Tetenov (2012) where signals are normally distributed. The intuition behind this result is straightforward: publishing a study always provides useful information about the relative effectiveness of the treatments, which allows the policymaker to do better than in the case where no study is published and they randomize between choices.

This conclusion differs starkly from the optimal publication rule in a Bayesian framework. Frankel and Kasy (2022) show that the optimal publication rule in a Bayesian model only publishes extreme results, that is, results that move prior beliefs sufficiently. In this framework, null results are defined as those which do not change the policy-maker's prior. By contrast, the minimax regret framework does not rely on a prior distribution about treatment efficacy and thus this notion of 'null results' is not well-defined. Instead, we model publication bias based on the common definition of results being statistically indistinguishable from zero (Assumption 3.2.1).

A second key difference across these frameworks is the role of publication costs. In the Bayesian setting, publishing relatively uninformative results that do little to move the policymaker's prior belief yields no benefits, while at the same time incurring a cost; it is thus not optimal to publish such results. By contrast, in the minimax regret framework, the cost parameter c does not appear in the expression for regret, as can be seen in equation (107). This is because regret equals the size of the loss from making an inferior treatment choice, $|\theta|$, multiplied by the probability of this occurring. Since the expected cost of publication is the same irrespective of the decision rule, the expression for regret does not include it. Thus, publication costs have no impact on the minimax decision rule and therefore the optimal publication rule.

3.3.3 Type I Error Loss Aversion

Up until now, we have made the implicit assumption of equal weight for Type I error (of mistakenly implementing an inferior policy) and Type II error (of failing to implement the superior policy). However, in practice, policymakers may exhibit loss aversion and weigh the regret from Type I error higher relative to Type II error. In fact, Tetenov (2012) finds that classical hypothesis testing at the 5% level is consistent with a policymaker who weighs the regret from Type I error around 100 times more than the regret arising from Type II error.

To incorporate this asymmetry in the concern over different types of error, we follow Tetenov (2012) and introduce a Type I error loss aversion parameter $K > 0$. With this, the policymakers utility is given by

$$U\big(\delta, \theta, c\big) = \begin{cases} K\Big(\theta D\delta_1(X) + \theta(1-D)\delta_0 - Dc\Big) & \text{if } \theta \leq 0 \\ \theta D\delta_1(X) + \theta(1-D)\delta_0 - Dc & \text{if } \theta > 0 \end{cases} \tag{109}$$

Expected welfare is given by

$$W\big(\delta, \theta, c\big) = \begin{cases} K\Big(\theta \cdot \mathbb{P}[D=1|\theta] \cdot \mathbb{E}[\delta_1(X)|\theta, D=1] + \theta \cdot \big(1 - \mathbb{P}[D=1|\theta]\big)\delta_0 - \mathbb{P}[D=1|\theta] \cdot c\Big) & \text{if } \theta \leq 0 \\ \theta \cdot \mathbb{P}[D=1|\theta] \cdot \mathbb{E}[\delta_1(X)|\theta, D=1] + \theta \cdot \big(1 - \mathbb{P}[D=1|\theta]\big)\delta_0 - \mathbb{P}[D=1|\theta] \cdot c & \text{if } \theta > 0 \end{cases} \tag{110}$$

and regret is equal to

$$R\big(\delta, \theta\big) = W\big(\mathbb{1}(\theta > 0), \theta\big) - W\big(\delta, \theta\big)$$

$$= \begin{cases} -K\theta\Bigg(\mathbb{P}[D=1|\theta] \cdot \mathbb{E}[\delta_1(X)|\theta, D=1] + (1 - \mathbb{P}[D=1|\theta])\delta_0\Bigg) & \text{if } \theta \leq 0 \\ \theta\Bigg(\mathbb{P}[D=1|\theta] \cdot \big(1 - \mathbb{E}[\delta_1(X)|\theta, D=1] + (1 - \mathbb{P}[D=1|\theta])(1 - \delta_0)\big)\Bigg) & \text{if } \theta > 0 \end{cases} \tag{111}$$

What is the optimal publication rule for different levels of loss aversion for Type I error K? Figure 3.1 plots minimax regret as a function of β_p for different values of K, in addition to the optimal minimax decision rule in each case. These figures are computed numerically. As a benchmark, the first column shows the regime where $K = 1$. First, see that minimax regret is decreasing β_p, in line with Proposition 3. Second, see that the optimal minimax regret decision rule is $(T^*, \delta_0^*) = (0, \frac{1}{2})$ for all $\beta_p \in [0, 1]$, in line with Lemma 3.3.1.

These are two particular cases. Numerical results for other values of K show similar patterns, namely, that the value of minimax regret is a decreasing function of β_p. Based on this, we conjecture that the optimal publication rule minimizing maximum regret being non-selective generalizes to any $K \geq 1$, although we do not have at present an analytical proof.

Now consider the case where $K = 3$ i.e. the policymaker weighs the Type I error of imple-

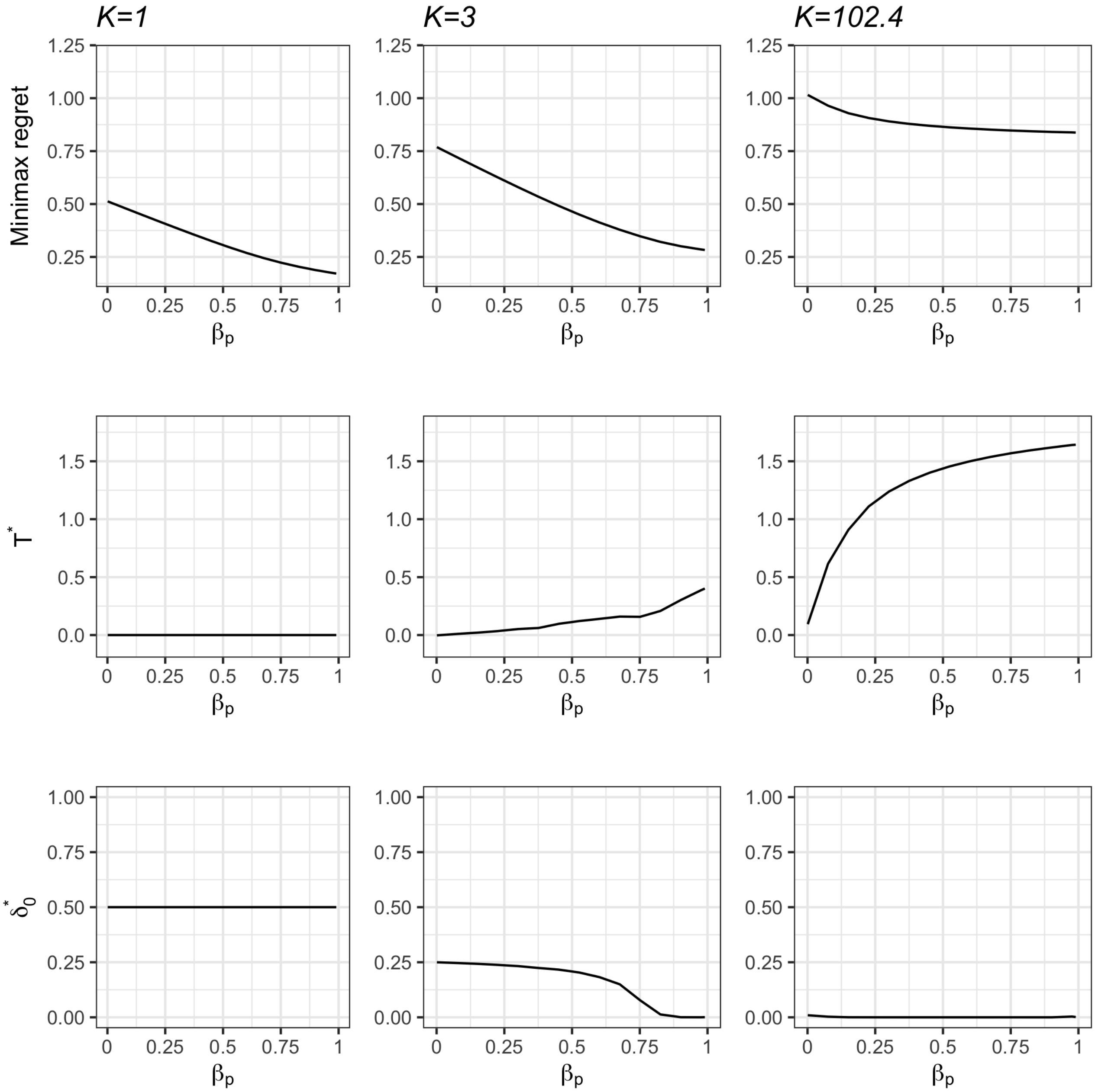

Figure 3.1: Minimax Regret and Optimal Decision Rule for Different Value of K

menting the inferior treatment three times larger than the Type II error of failing to implement the superior treatment. As in the case where $K = 1$, minimax regret is also decreasing in β_p. See that the threshold rule is increasing in β_p. Similarly, the default probability of implementing the innovative policy in the event that a study Just not published, δ_0^*, is decreasing in β_p (and weakly less than $\frac{1}{2}$). That is, as β_p gets larger, the policymakers decision rule becomes more conservative with respect to assigning the innovative treatment. The intuition behind this is that as β_p increases, the possibility of noisier small effect being published increases, which increases the risk of committing Type I error.

Finally, consider the case where $K = 102.4$, which is the value that rationalizes hypothesis testing at the 5% significance level (Tetenov, 2012). Again, the level of minimax regret decreases as a the relative probability of publishing null results increases. Given the very high level of Type I loss aversion, the no-data rule is essentially zero for any value of β_p. Again, the threshold rule is increasing in β_p, and at a faster rate than as for the case where $K = 3$.

3.4 Naive Policymakers

The sophisticated policymaker knows the exact form of publication bias and can accurately account for it. This is a strong assumption. As an alternative, we may consider a policymaker who naively chooses their decision rule without accounting for selective publication. This is perhaps more realistic, in the sense that most published research reports standard errors and assumes (approximately) normally distributed treatment effects for inference. 'Naiveity' impacts both realizations of the publication process. When a study is published, the policymaker erroneously believes it is normally distributed; and in the event that a study is not published, the naive policymaker fails to account for censoring in the publication process when choosing their default action. As in the sophisticated policymaker's problem, a decision rule δ is equivalent to the tuple (T, δ_0). The naive policymaker's *misspecified welfare* is equal to

$$\widetilde{W}\big((T,\delta_0),\theta\big) = \begin{cases} \theta[1-\Phi\big(T-\theta\big)] & \text{if } D=1 \\ \theta \cdot \delta_0 & \text{if } D=0 \end{cases} \tag{112}$$

This gives rise to two misspecified regret expressions. First, in the event that a study is published

$$\widetilde{R}_1\big(T,\theta\big) = \begin{cases} -\theta[1-\Phi\big(T-\theta\big)] & \text{if } \theta \leq 0 \\ \theta\Phi\big(T-\theta\big) & \text{if } \theta > 0 \end{cases} \tag{113}$$

and second, in the event that no study is published,

$$\widetilde{R}_0\big(\delta_0,\theta\big) = \begin{cases} -\theta\delta_0 & \text{if } \theta \leq 0 \\ \theta(1-\delta_0) & \text{if } \theta > 0 \end{cases} \tag{114}$$

Misspecified regret in equation (113) when a study is published is equivalent to the expression for regret in the model in Tetenov (2012) with normally distributed signals. This expression is misspecified because the policymaker does not account for the fact that selective publication distorts the distribution of estimated treatment effects. Misspecified regret when no study arrives, in equation (114), is simply a function of the default action δ_0 and the true effect θ. It is misspecified in that it ignores that possibility that a study was not published because of selective publication. For the minimax problem to be well-defined, we need to impose bounds on θ. For the naive policymaker, we impose the following assumption:

Assumption 3.4.1 (Symmetric Bounds on Average Treatment Effect)**.** *Let the support of* Θ *be* $[-B,B]$ *for some* $B > \theta^* > 0$*, where* $\theta^* = \arg\max_{\theta>0}\big\{\theta \cdot \Phi(0-\theta)\big\}$.

The technical condition that the bound is larger than $\theta^* = \arg\max_{\theta>0}\big\{\theta \cdot \Phi(0-\theta)\big\}$ ensures that the minimax problem when a study is published is not constrained by the bound.[64]

[64] Tetenov (2012) shows that the maximum θ^* is attained on a closed interval $[0,H]$ for some $H>0$.

The naive policymaker has, in effect, two decision problems, one for each realization of the publication process.

$$T^* = \arg\min_{T \in \mathbb{R}} \max_{\theta \in [-B,B]} \widetilde{R}_1(T, \theta) \tag{115}$$

$$\delta_0^* = \arg\min_{\delta_0 \in [0,1]} \max_{\theta \in [-B,B]} \widetilde{R}_0\big(\delta_0, \theta\big) \tag{116}$$

While (T, δ_0) are chosen by the naive policymaker under misspecified beliefs about the DGP, regret of any decision rule is assessed against the 'true' worst-case scenario which accounts for publication bias. That is, regret for any decision rule (T, δ_0) is identical to regret in the sophisticated policymaker's problem in equation (107).

To compare the 'cost' of naivity with respect to publication bias, we make the following calculation for some fixed K and assuming that $t_\alpha = 1.96$:

$$100 \cdot \left(\frac{\text{MMR}^*_{Naive}(K)}{\text{MMR}^*_{Soph}(K)} - 1 \right) \tag{117}$$

where $\text{MMR}^*_{Naive}(K)$ is the value of minimax regret for the naive policymaker and $\text{MMR}^*_{Soph}(K)$ is the value of minimax regret for the sophisticated policymaker.

Figure 3.2 illustrates the cost of naivity when $K = 3$. Results show that the cost of naivity if weakly positive. This is to be expected, since the naive planner chooses their decision rule under misspecified beliefs. Interestingly, the results show that the costs of naivety are highest when publication bias is moderate. When there is no publication bias, the cost of naivety is zero because the naive policymaker belief that there is no publication bias is correct in this special case. More surprisingly, the cost of naivety is also zero when there is extreme publication bias, such that no insignificant results are published. This is because the optimal threshold rule when the study is published is set identified and the solution for the naive policymaker and the sophisticated policymaker both fall within this set. In particular, any threshold rule above which the innovative treatment is implemented in the range of (-1.96σ, 1.96σ) will be effectively identical, because no insignificant studies within this range are ever published.

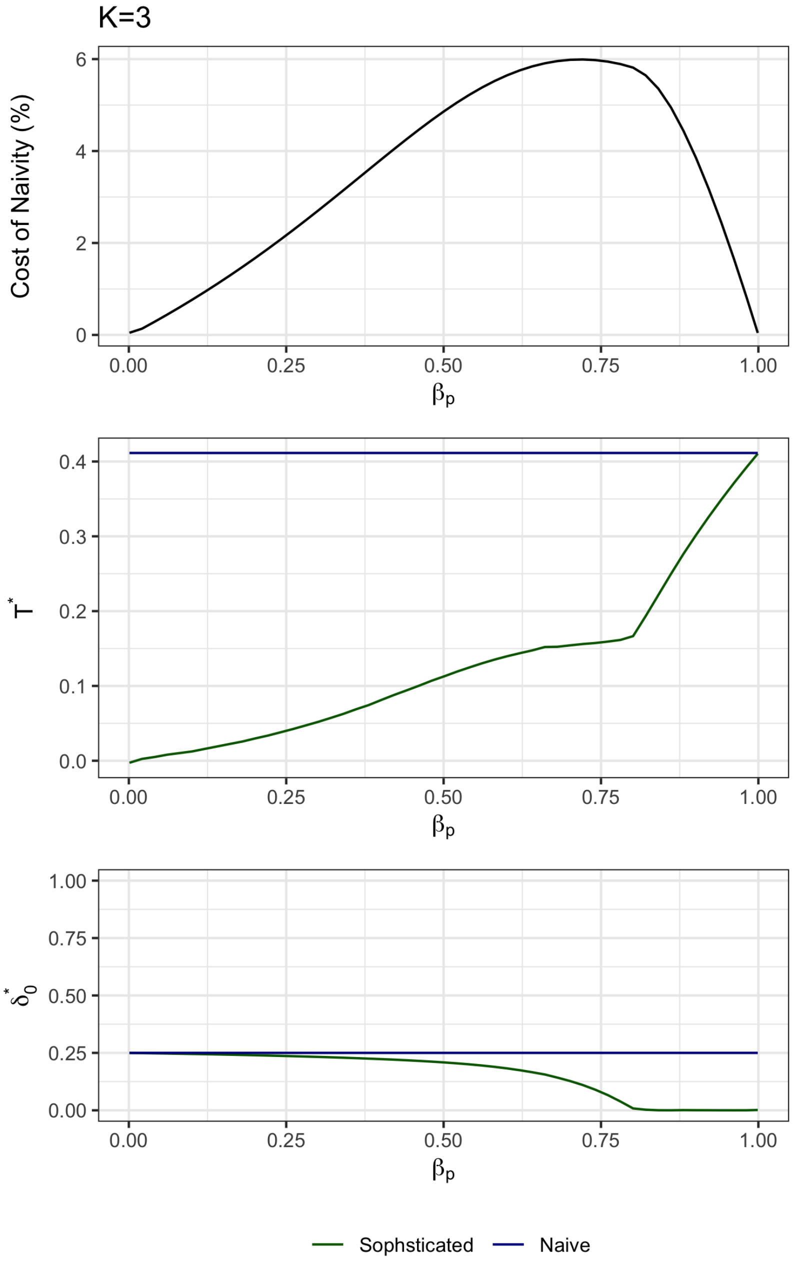

Figure 3.2: Cost of Naivity ($K = 3$)

3.5 Misaligned Preferences

In the main analysis, the policymaker chooses a decision rule to minimize maximum regret, and we consider the optimal publication rule of a journal editor who chooses $\beta_p \in [0,1]$ with the same preferences. In this extension, we consider what happens when the policymaker and the journal editor do not have aligned preferences. In particular, we continue to assume that the policymaker optimizes using minimax regret, but instead consider a journal editor who chooses $\beta_p \in [0,1]$ to maximize welfare under a Bayesian prior. Since the policymakers' decision rule could in theory depend on the journal editor's choice for β_p, we can view the equilibrium outcome as resulting from a game between the editor and the policymaker. However, since Lemma 3.3.1 shows that the minimax decision rule is the same for any value of $\beta_p \in [0,1]$, there are no strategic considerations at play. Throughout, we assume that Type I and Type II error are equally weighted ($K = 1$).

More formally, for the policymaker's decision rule δ and publication cost c, the Bayesian journal editor's problem is given by

$$\max_{\beta_p \in [0,1]} \int W\big(\delta, \theta, c\big) \pi(\theta) d\theta \tag{118}$$

where welfare is given by equation 105 and $\pi(\cdot)$ denotes the prior belief distribution of the journal editor. We assume that the prior satisfies the following regularity conditions:

Assumption 3.5.1 (Support of Journal Editor's Prior)**.** *Let the prior distribution $\pi(\cdot)$ have support on an open subset of the real line.*

Recall the policymaker's optimal minimax rule from Lemma 3.3.1 and that it is identical for under both sophisticated and naive beliefs when $K = 1$. The following Proposition gives the optimal publication rule of the Bayesian journal editor:

Proposition 4 (Optimal Bayesian Publication Rule)**.** *Suppose the policymaker implements the optimal minimax regret decision rule $(T^*, \delta_0^*) = (0, \frac{1}{2})$. Under Assumptions 3.2.1 and 3.5.1, the*

Bayesian journal editor's optimal publication rule for any $c \geq 0$ is given by

$$\beta_p^* = \begin{cases} 1 & \textit{if } c \leq T \\ \\ 0 & \textit{if } c > T \end{cases} \tag{119}$$

where

$$T = \frac{\frac{1}{2}\int \theta \Big(\big[\Phi(t_\alpha - \theta) - \Phi(-\theta)\big] - \big[\Phi(-\theta) - \Phi(-t_\alpha - \theta)\big]\Big)\pi(\theta)d\theta}{\int \big[\Phi(t_\alpha - \theta) - \Phi(-t_\alpha - \theta)\big]\pi(\theta)d\theta} > 0$$

The journal's optimal action takes the form a simple threshold rule: publish all results if publication costs are sufficiently low; otherwise, censor all null results. Thus, when publication costs are low, the optimal publication rule under misaligned preferences is the same as with aligned preferences, namely, it is non-selective. However, when publication costs are high, it will be optimal to censor all null results.

For the Bayesian policymaker in the Frankel and Kasy (2022) model, the optimal publication rule recommends censoring results which do not sufficiently move the prior. In other words, the journal does not publish 'unsurprising' findings close to its prior beliefs on a given research question (which is assumed to be shared by the public). Our result in Proposition 4 differs because we consider the class of publication rules which censor statistically insignificant results. The rationale behind this is that the censoring of null results is the most common form of publication bias highlighted in the literature.

3.6 Conclusion

This paper studies treatment choice in the presence of publication bias in the case where policymakers are unwilling or unable to rely on prior beliefs about relative treatment efficacy. We show that the optimal publication rule which minimizes maximum regret is non-selective. This holds whether or not policymakers account for publication bias in choosing their treatment rule i.e. whether they are sophisticated or naive in their beliefs about the DGP. This contrasts with the Bayesian policymaker studies in the literature, where the optimal publication rule for policy choice censors results close to the decision-maker's prior. Thus, the optimal publication regime – and hence the statistical credibility of published research – can vary drastically depending on the optimality criterion pursued by the policymaker and journals. In the minimax framework, the publication regime which is optimal for treatment choice also delivers valid statistical inference.

Appendix

3A Proofs

Proof of Lemma 3.2.1: We focus on the welfare function $W_1(\delta_1, X)$ for the published case. Karlin and Rubin (1956b) shows that if the distribution of sufficient statistics for θ satisfies the monotone likelihood ratio property, the class of threshold decision rules is essentially complete for a class of loss functions including the current one. Under Assumption 3.2.1, $F(X|\theta, D = 1)$ is an exponentially family distribution with pdf

$$C(\theta)h(x)\exp(x\theta), \tag{120}$$

where $C(\theta) = \frac{\exp(-\theta^2/2)}{\sqrt{2\pi}\int p(t)\phi(t-\theta)dt}$ and $h(x) = p(x)\exp(-x^2/2)$, and X being a sufficient statistics for θ. Since the exponential family distribution satisfies the monotone likelihood ratio property (see, e.g., Section 3.4 in Lehmann and Romano (2005)), the current lemma follows. □

Proof of Lemma 3.3.1: The proof follows two main steps. First, we solve the minimax problem for the sophisticated policymaker. In the second step, we show that the naive policymaker, who optimizes under misspecified beliefs about the DGP, nonetheless arrives at the optimal solution.

Sophisticated policymaker.—First, we show that the optimal decision rule for the sophisticated policymaker is $(T^*, \delta_0^*) = \left(0, \frac{1}{2}\right)$. To do this, we use the following theorem (for reference, see Theorem 1 in section 2.11 (pg 90) in Ferguson (1967)):

Lemma 3A.1. *Suppose δ minimizes Bayes risk under π:*

$$\delta \in arg \min_{\delta' \in \mathcal{D}} \int_\theta R(\delta', \theta) d\pi(\theta)$$

and

$$R(\delta,\theta) \leq \int_\theta R(\delta,\theta) d\pi(\theta)$$

for all $\theta \in \Theta$. Then δ is a minimax rule and π is least favourable.

Using this lemma, we first propose a guess for δ and π and then show that this guess satisfies the sufficient conditions in Theorem A1 which imply that δ is the minimax regret decision rule.

Our guess is that the minimax regret decision rule is $(T^*, \delta_0^*) = (0, \frac{1}{2})$. Regret under this proposed rule for any θ is equal to:

$$R\big((0,0.5),\theta\big) = \begin{cases} -\theta\Big(\mathbb{P}[D=1|\theta]\cdot[1-F(0|\theta,D=1)] + (1-\mathbb{P}[D=1|\theta])\frac{1}{2}\Big) & \text{if } \theta \leq 0 \\ \theta\Big(\mathbb{P}[D=1|\theta]\cdot\big(F(0|\theta,D=1) + (1-\mathbb{P}[D=1|\theta])\frac{1}{2}\big)\Big) & \text{if } \theta > 0 \end{cases} \tag{121}$$

Next, guess that Nature's least favorable prior is equal to

$$\pi = \begin{cases} \theta_+^* & \text{with probability } \frac{1}{2} \\ -\theta_+^* & \text{with probability } \frac{1}{2} \end{cases} \tag{122}$$

where $\theta_+^* = \arg\max_{\theta>0} R\big((0,0.5),\theta\big)$. We know that $\theta_+^* \in (0,\infty)$ because $R\big((0,0.5),0\big) = 0$; $R\big((0,0.5),\theta\big) \to 0$ as $\theta \to \infty$; and $R\big((0,0.5),\theta\big) > 0$ for any $\theta > 0$. The first and third claims can be seen directly from equation (121). To see why the second claim is true see that

$$\lim_{\theta\to\infty}\Big\{\theta\cdot\mathbb{P}[D=1|\theta]\cdot F(0|\theta,D=1)\Big\} + \frac{1}{2}\lim_{\theta\to\infty}\Big\{\theta\cdot(1-\mathbb{P}[D=1|\theta])\Big\} \tag{123}$$

The first term equals zero because

$$\lim_{\theta\to\infty}\Big\{\theta\cdot\mathbb{P}[D=1|\theta]\cdot F(0|\theta,D=1)\Big\} < \lim_{\theta\to\infty}\Big\{\theta\cdot\Phi(0-\theta)\Big\} = \lim_{\theta\to\infty}\Big\{\theta^2\cdot\phi(0-\theta)\Big\} = 0 \tag{124}$$

where the first inequality follows because $F(.|\theta, D = 1)$ is a truncated normal cdf and $\theta > 0$; the second last equality follows from L'Hôpital's rule; and the final equality follows from the fact that $\theta^2 \cdot \phi(0 - \theta)$ has finite moments. The second term also equals zero since we have

$$\theta \cdot (1 - \mathbb{P}[D = 1|\theta])\big) = (1 - \beta_p)(2\pi)^{-1} \int_{-t_\alpha}^{t_\alpha} \theta \exp\left(-\frac{1}{2}(t - \theta)^2\right) dt \tag{125}$$

and $\lim_{\theta \to \infty} \theta \exp\left(-\frac{1}{2}(t - \theta)^2\right) = 0$ at every $t \in [-t_\alpha, t_\alpha]$, and apply the dominated convergence theorem.

Next, we will show that $(T^*, \delta_0^*) = (0, \frac{1}{2})$ minimizes Bayes risk with respect to π. For any decision rule (T, δ_0), Bayes risk equals

$$\int_\theta R\big((T, \delta_0), \theta\big) d\pi(\theta) = \frac{1}{2} \cdot \theta_+^* \Big(\mathbb{P}[D = 1|\theta_+^*] \cdot F(T|\theta_+^*, D = 1) + (1 - \mathbb{P}[D = 1|\theta_+^*])(1 - \delta_0)\Big)$$

$$+\frac{1}{2} \cdot \theta_+^* \Big(\mathbb{P}[D = 1|\theta_+^*] \cdot [1 - F(T|\theta_+^*, D = 1)] + (1 - \mathbb{P}[D = 1|\theta_+^*])\delta_0\Big)$$

$$= \frac{1}{2} \cdot \theta_+^* \Big(1 - \mathbb{P}[D = 1|\theta_+^*]\Big) + \frac{1}{2} \cdot \theta_+^* \mathbb{P}[D = 1|\theta_+^*] \Big(F(T|\theta_+^*, D = 1) + F(-T|\theta_+^*, D = 1)\Big) \tag{126}$$

Note that any δ_0 is optimal, so we can choose $\delta_0^* = \frac{1}{2}$. We will show that $T^* = 0$ minimizes Bayes risk by showing that $F(T|\theta_+^*, D = 1) + F(-T|\theta_+^*, D = 1)$ is minimized when $T = 0$. To do this, we will show that any other choice of T leads to higher regret. Since the Bayes risk under π (126) is symmetric in T, without loss of generality, we assume $T > 0$. Consider first the case where $T > t_\alpha > 0$. We have

$$F(-T|\theta_+^*, D = 1) + F(T|\theta_+^*, D = 1) = \frac{1}{C}\Big(\Phi(-T - \theta_+^*) +$$

$$+\Phi(-T - \theta_+^*) + \big[\Phi(-t_\alpha - \theta_+^*) - \Phi(-T - \theta_+^*)\big]$$

$$+\beta_p\big[\Phi(0 - \theta_+^*) - \Phi(-t_\alpha - \theta_+^*)\big] + \beta_p\big[\Phi(t_\alpha - \theta_+^*) - \Phi(0 - \theta_+^*)\big] + \big[\Phi(T - \theta_+^*) - \Phi(t_\alpha - \theta_+^*)\big]\Big)$$

$$> \frac{2}{C}\bigg(\Phi(-t_\alpha - \theta_+^*) + \beta_p\big[\Phi(0-\theta_+^*) - \Phi(-t_\alpha - \theta_+^*)\big]\bigg) = 2 \cdot F(0|\theta_+^*, D=1) \tag{127}$$

where $C = \int p(z)\phi(z - \theta_+^*)dz$ is the normalization constant of the truncated normal distribution. The case where $t_\alpha > T > 0$ follows a similar argument. Thus, $(T^*, \delta_0^*) = (0, \frac{1}{2})$ minimizes Bayes risk with respect to π.

Finally, see that for any $\theta \in \mathbb{R}$, we have that

$$R\big((0,0.5),\theta\big) \leq R\big((0,0.5),\theta_+^*\big) = \frac{1}{2}R\big((0,0.5),\theta_+^*\big) + \frac{1}{2}R\big((0,0.5),-\theta_+^*\big) = \int_\theta R(\delta,\theta)d\pi(\theta) \tag{128}$$

The first inequality follows from the construction of θ_+^*. The next equality uses the symmetry of the regret function with respect to θ around zero. From Theorem A1, it then follows that the minimax regret decision rule for the sophisticated policymaker is $(T^*, \delta_0^*) = (0, \frac{1}{2})$ and the least favourable prior is π in equation (122).

Naive policymaker.—Next, we show that the naive policymaker arrives at the same decision rule, despite ignoring selective publication. The naive policymaker's optimal decision rule consists of two problems, when a study is published and when it is not. When a study is published, the policymaker (erroneously) believes the signal is normally distributed. This is equivalent to the problem in Tetenov (2012), who proves that the optimal solution in the symmetric case is $T^* = 0$.

Next, consider the case where no study is published. Misspecified regret is equal to

$$\widetilde{R}_0(\delta_0, \theta) = \begin{cases} -\theta\delta_0 & \text{if } \theta \leq 0 \\ \theta(1-\delta_0) & \text{if } \theta > 0 \end{cases} \tag{129}$$

and thus misspecified worse-case regret given bounds in Assumption 3.4.1 is given by $\max_{\theta\in[-B,B]} \widetilde{R}_0(\delta_0,\theta) = \max\{B\theta_0, B(1-\delta_0)\}$. The minimax regret decision rule equalizes the arguments in the max operator, giving $\delta_0^* = \frac{1}{2}$.

Proof of Proposition 3: We have shown that for any $\beta_p \in [0,1]$, both the sophisticated and naive policymakers' optimal minimax decision rule is $(T^*, \delta_0^*) = (0, \frac{1}{2})$. It remains to show that $\beta_p = 1$ is the optimal publication rule, in the sense that it minimizes minimax regret.

Denote the value of minimax regret as a function of parameter β_p:

$$V(\beta_p) \equiv \max_{\theta>0} \left\{ \theta \Big(\mathbb{P}[D=1|\theta] F(0|\theta, D=1) + \big(1 - \mathbb{P}[D=1|\theta]\big)\frac{1}{2} \Big) \right\} \tag{130}$$

$$= \max_{\theta>0} \left\{ \theta \int_{-\infty}^{0} p(y)\phi(y-\theta)dy + \frac{\theta}{2}\int_{-\infty}^{\infty}[1-p(y)]\phi(y-\theta)dy \right\},$$

$$= \max_{\theta>0} f(\theta, \beta), \tag{131}$$

where $f(\theta,\beta) = \theta\int_{-\infty}^{0} p(y)\phi(y-\theta)dy + \frac{\theta}{2}\int_{-\infty}^{\infty}[1-p(y)]\phi(y-\theta)dy$ and its dependence on β_p is only through $p(\cdot)$.

Note that the value function inside the maximum operator is continuously differentiable in β_p with an integrable envelope over the domain of $\beta_p \in [0,1]$. Hence, by the generalized envelope theorem (Theorem 2 in Milgrom and Segal (2002)), $V(\beta_p)$ is absolutely continuous and admits the following integral representation:

$$V(\beta_p) = V(0) + \int_0^{\beta_p} f_{\beta_p}(\theta^*(\beta_p'), \beta_p')d\beta_p', \tag{132}$$

where $f_{\beta_p}(\cdot,\cdot) = \frac{\partial}{\partial\beta} f(\theta,\beta)$ and $\theta^*(\beta_p)$ is a maximizer of $f(\theta,\beta_p)$ in θ given β_p. Note that for $\theta > 0$, we can show

$$f_{\beta_p}(\theta,\beta_p) = \frac{\theta}{2}\left[\int_{-t_\alpha}^{0} \phi(y-\theta)dy - \int_0^{t_\alpha}\phi(y-\theta)dy\right] < 0. \tag{133}$$

To see this inequality holds, consider two cases. First, suppose that $\theta \geq t_\alpha$. Then we immediately get the desired result because $\phi(z-\theta)$ is strictly increasing over $(-t_\alpha, t_\alpha)$.

Next consider the case where $\theta \in (0, t_\alpha)$. Then $\int_0^\theta \phi(y-\theta)dy > \int_{-\theta}^0 \phi(y-\theta)dy$ since $\phi(y-\theta)$ is strictly increasing in y for any $y < \theta$. And we also have that $\int_{\theta}^{t_\alpha} \phi(y-\theta)dy = \int_{2\theta-t_\alpha}^{\theta} \phi(y-\theta)dy > \int_{-t_\alpha}^{-\theta} \phi(y-\theta)dy$, where the first equality uses symmetry of the normal distribution about θ and the second equality again uses the fact that $\phi(y-\theta)$ is strictly increasing in y for any $y < \theta$. Taking these two inequalities together leads to the inequality of (133).

Combining (132) and (133), we conclude that $V(\beta_p)$ is a monotonically decreasing function, and $\beta_p = 1$ minimizes the value of minimax regret.

Proof of Proposition 4: Fix the optimal minimax rule for the policymaker: $\delta^* = (T^*, \delta_0^*) = (0, \frac{1}{2})$. Then

$$\int W\big(\delta^*, \theta, c\big)\pi(\theta)d\theta = \int \theta \cdot \mathbb{P}[D=1|\theta,\beta_p]\big[1 - F(0|D=1,\theta,\beta_p)\big]\pi(\theta)d\theta$$

$$+\frac{1}{2}\int \theta \cdot \big(1 - \mathbb{P}[D=1|\theta,\beta_p]\big)\pi(\theta)\theta - c\int \mathbb{P}[D=1|\theta,\beta_p]\pi(\theta)d\theta$$

Now see that

$$\frac{\partial}{\partial \beta_p}\bigg(\mathbb{P}[D=1|\theta,\beta_p]\bigg) = \Phi(t_\alpha - \theta) - \Phi(-t_\alpha - \theta)$$

$$\frac{\partial}{\partial \beta_p}\bigg(F(0|D=1,\theta,\beta_p)\cdot \mathbb{P}[D=1|\theta,\beta_p]\bigg) = \Phi(-\theta) - \Phi(-t_\alpha - \theta)$$

which implies that

$$\frac{\partial}{\partial \beta_p}\bigg[\int W\big(\delta^*,\theta,c\big)\pi(\theta)d\theta\bigg] = \frac{1}{2}\int \theta\Big(\big[\Phi(t_\alpha - \theta) - \Phi(-\theta)\big] - \big[\Phi(-\theta) - \Phi(-t_\alpha-\theta)\big]\Big)\pi(\theta)d\theta$$

$$-c\int \big[\Phi(t_\alpha - \theta) - \Phi(-t_\alpha - \theta)\big]\pi(\theta)d\theta$$

It is clear that the integral in the second term multiplied by c is positive. If the integral in the first term is strictly positive, then the desired result clearly follows. That is, for sufficiently low c, the derivative will be positive and the optimal rule will be $\beta_p^* = 1$. Conversely, for sufficiently high c, the derivative will be negative and the optimal publication rule will be

$\beta_p^* = 0$.

In the remainder of the proof, we show the integral is indeed positive. For clarity, define the integrand $g(\theta) \equiv \theta\big([\Phi(t_\alpha - \theta) - \Phi(-\theta)] - [\Phi(-\theta) - \Phi(-t_\alpha - \theta)]\big)$. First, see that $g(0) = 0$. However, Assumption 3.5.1 implies that there exists some $\theta \neq 0$ on the support of $\pi(\cdot)$. Thus, to show that the integral is positive, it suffices to show that $g(\theta) > 0$ for all $\theta \neq 0$.

To show this, first note that $g(\cdot)$ is symmetric about zero i.e. $g(\theta) = g(-\theta)$. We can therefore restrict our attention to $\theta > 0$. Consider two cases. First, suppose $t_\alpha - \theta \leq 0$. Then $g(\theta) > 0$ if and only if $[\Phi(t_\alpha - \theta) - \Phi(-\theta)] - [\Phi(-\theta) - \Phi(-t_\alpha - \theta)] > 0$, which clearly holds because the normal density is increasing over $(-\infty, 0)$.

Next, suppose that $t_\alpha - \theta > 0 \iff t_\alpha > \theta > 0$. Then breakup up the integral and using the symmetry of the normal density, we have

$$g(\theta) = [\Phi(t_\alpha - \theta) - \Phi(-\theta)] - [\Phi(-\theta) - \Phi(-t_\alpha - \theta)]$$

$$= \bigg([\Phi(t_\alpha - \theta) - \Phi(0)] + [\Phi(0) - \Phi(-\theta)]\bigg) - \bigg([\Phi(-\theta) - \Phi(-2\cdot\theta)] + [\Phi(-2\cdot\theta) - \Phi(-t_\alpha - \theta)]\bigg)$$

$$= \bigg([\Phi(0) - \Phi(-\theta)] - [\Phi(-\theta) - \Phi(-2\cdot\theta)]\bigg) + \bigg([\Phi(t_\alpha - \theta) - \Phi(0)] - [\Phi(t_\alpha + \theta) - \Phi(2\cdot\theta)]\bigg) > 0$$

where the inequality follows because both differences in the parentheses are strictly positive. The first difference is positive because the normal density if increasing over $(\infty, 0)$. The second difference is positive because the normal density if decreasing over $(0, \infty)$. □

Bibliography

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *The Quarterly Journal of Economics*, 2023, pp. 1–35.

Allcott, Hunt, "Site Selection Bias in Program Evaluation," *Quarterly Journal of Economics*, 2015, *130* (3).

Altmejd, Adam, Anna Dreber, Eskil Forsell et al., "Predicting the Replicability of Social Science Lab Experiments," *PLoS ONE*, 2019, *14* (12).

Amrhein, Valentin, David Trafimow, and Sander Greenland, "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication," *The American Statistician*, 2019, *73* (1), 262–270.

_ , Sander Greenland, and Blake McShane, "Retire Statistical Significance," *Nature*, 2019, *567*, 305–307.

Anderson, Samantha F. and Scott E. Maxwell, "Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power," *Multivariate Behavioral Research*, 2017, *52* (3), 305–324.

Anderson, T. W. and Herman Rubin, "Estimation of the parameters of a single equation in a complete system of stochastic equations," *Annals of Mathematical Statistics*, 1949, *20* (1), 46–63.

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

__ , Toru Kitagawa, and Adam McCloskey, "Inference on Winners," *Quarterly Journal of Economics*, 2023.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos et al., "Quantitative Political Science Research is Greatly Underpowered," *OSF Preprint*, 2023.

Armstrong, Timothy B., Michal Kolesár, and Mikkel Plagborg-Møller, "Robust Empirical Bayes Confidence Intervals," *Econometrica*, 2022, *90* (6), 2567–2602.

Baker, Monya, "1,500 Scientists Lift the Lid on Reproducibility," *Nature*, 2016, *533*, 452–454.

Barnett, Adrian G., Jolieke C. Van Der Pols, and Annette J. Dobson, "Regression To The Mean: What It Is and How To Deal with It," *Journal of Business and Psychology*, 2004, *34* (1), 215–220.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan, "How Much Should We Trust Differences-In-Differences Estimates?," *Quarterly Journal of Economics*, 2004, *110* (1), 249–275.

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

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

__ , __ , and __ , "We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell Us about Publication Bias and p-Hacking in Online Experiments," *IZA Discussion Paper 15478*, 2022.

Bryan, Christopher J., David S. Yeager, and Joseph M. O'Brien, "Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate," *Proceedings of the National Academy of Sciences of the United States of America*, 2019, *116* (51), 25535–25545.

Button, Katherine S., John P.A. Ioannidis, Claire Mokrysz et al., "Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience," *Nature Reviews Neuroscience*, 2013, *14* (5), 365–376.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, and Magnus Johannesson, "Evaluating replicability of laboratory experiments in economics," *Science*, 2016, *351* (6280), 1433–1437.

_ , _ , Felix Holzmeister et al., "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature Human Behaviour*, 2018, *2* (9), 637–644.

Camerer, Colin, Yiling Chen, Anna Dreber et al., "Mechanical Turk Replication Project," 2022.

Cameron, Miller A. and Douglas L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

Card, David and Alan B. Krueger, "Time-Series Minimum-Wage Studies: A Meta-analysis," *American Economic Review: Papers and Proceedings*, 1995, *85* (2), 238–243.

Cesario, Joseph, "Priming, Replication, and the Hardest Science," *Perspectives on Psychological Science*, 2014, *9* (1), 40–48.

Chambers, Christopher D., "Registered Reports: A new publishing initiative at Cortex," *Cortex*, 2013, *49* (3), 609–610.

Chen, Jiafeng, "Empirical Bayes When Estimation Precision Predicts Parameters," *arXiv Working Paer*, 2023.

Currie, Janet, Henrik Kleven, and Esmee Zwiers, "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, 2020, *110*, 42–48.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence From Two Nudge Units," *Econometrica*, 2022, *90* (1), 81–116.

_ , Nicholas Otis, and Eva Vivalt, "Forecasting the Results of Experiments: Piloting an Elicitation Strategy," *AEA Papers and Proceedings*, 2020, *110*, 75–79.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg et al., "Using Prediction Markets to Estimate the Reproducibility of Scientific Research," *Proceedings of the National Academy of Sciences of the United States of America*, 2015, *112* (50), 15343–15347.

Editorial, "In praise of replication studies and null results," *Nature*, 2020, *578*, 489–490.

Elliott, Graham, Nikolay Kudrin, and Kaspar Wüthrich, "Detecting p-Hacking," *Econometrica*, 2022, *90* (2), 887–906.

Ferguson, Thomas S., *Mathematical Statistics: A Decision Theoretic Approach*, New York, Academic Press, 1967.

Fisher, Ronald A., "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, 1915, *10* (4), 507–521.

Foster, Andrew, Dean Karlan, Edward Miguel, and Aleksandar Bogdanoski, "Pre-results Review at the Journal of Development Economics: Lessons Learned So Far," *World Bank Development Impact Blog*, 2019.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication bias in the social sciences: Unlocking the file drawer," *Science*, 2014, *345* (6203), 1502–1505.

Frankel, Alexander and Maximilian Kasy, "Which Findings Should Be Published?," *American Economic Journal: Microeconomics*, 2022, *14* (1), 1–38.

Galton, Francis, "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 1886, *15*, 246–263.

Gelman, Andrew and John Carlin, "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 2014, *9* (6), 641–651.

Gordon, Michael, Domenico Viganola, Michael Bishop et al., "Are Replication Rates the Same Across Academic Fields? Community Forecasts from the DARPA SCORE Programme," *Royal Society Open Science*, 2020, *7*.

Higgins, Julian P.T. and Simon G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Statistics in Medicine*, 2002, *21* (11), 1539–1558.

Hotelling, Harold, "Review: The Triumph of Mediocrity in Business, By Horace Secrist," *Journal of the American Statistical Association*, 1933, *28* (184), 463–465.

Imai, Taisuke, Klavdia Zemlianova, Nikhil Kotecha et al., "How Common are False Positives in Laboratory Economics Experiments? Evidence from the P-Curve Method," *Working Paper*, 2020.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics Research," *The Economic Journal*, 2017, *127* (605), 236–265.

Ioannidis, John P.A., "Why Most Published Research Findings Are False," *PLoS Med*, 2005, *2* (8).

__ , "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 2008, *19* (5), 640–648.

Kahneman, Daniel, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

__ **and Amos Tversky**, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 1979, *47* (2), 263–292.

Karlin, Samuel and Herman Rubin, "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio," *The Annals of Mathematical Statistics*, 1956, *27* (2), 272–299.

__ **and** __ , "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio," *Annals of Mathematical Statistics*, 1956, *27*, 272–299.

Kasy, Maximilian, "Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It," *Journal of Economic Perspectives*, 2021, *35* (3), 175–192.

Kitagawa, Toru and Alex Tetenov, "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 2018, *86* (2), 591–616.

__ **and Patrick Vu**, "Optimal Publication Rules for Evidence-Based Policy," *Working Paper*, 2023.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello et al., "Investigating Variation in Replicability: A "Many Labs" Replication Project," *Social Psychology*, 2014, *45* (3), 142–152.

_ **, Michelangelo Vianello, Fred Hasselman et al.**, "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (4), 443–490.

Laird, Nan M. and Frederick Mosteller, "Some Statistical Methods for Combining Experimental Results," *International Journal of Technology Assessment in Health Care*, 1990, *6* (1), 5–30.

Lee, David S., Justin McCrary, Marcelo J. Moreira, and Jack Porter, "Valid t-Ratio Inference for IV," *American Economic Review*, 2022, *112* (10), 3260–3290.

Lehmann, Erlich L. and Joseph P. Romano, *Testing Statistical Hypotheses*, Springer, 2005.

Manski, Charles F., "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 2004, *72* (4), 1221–1246.

Maxwell, Scott E., Michael Y. Lau, and George S. Howard, "Is Psychology Suffering from a Replication Crisis? What Does "Failure to Replicate" Really Mean? ," *American Psychologist*, 2015, *70* (6), 487–498.

McFadden, Daniel, "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 1989, *57* (5), 995–1026.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett, "Abandon Statistical Significance," *The American Statistician*, 2019, *73* (1), 235–245.

Miguel, Edward and Garret Christensen, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–980.

Milgrom, Paul and Ilya Segal, "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 2002, *70* (2), 583–601.

Moulton, Brent R., "Random group effects and the precision of regression estimates ," *Journal of Econometrics*, 1986, *32* (3), 385–397.

—, "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units ," *The Review of Economics and Statistics*, 1990, *72* (2), 334–338.

Newey, Whitney K. and Kenneth D. West, "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 1987, *55* (3), 703–708.

Nosek, Brian A., Charles R. Ebersole, Alexander C. DeHaven et al., "The Preregistration Revolution," *Proceedings of the National Academy of Sciences*, 2018, *115* (11), 2600–2606.

—, George Alter, George C. Banks et al., "Promoting an Open Research Culture," *Science*, 2015, *348* (6242), 1422–1425.

—, Tom E. Hardwicke, Hannah Moshontz et al., "Replicability, Robustness, and Reproducibility in Psychological Science," *Annual Review of Psychology*, 2022, *73*, 719–748.

Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, 2015, *349* (6251).

Patil, Prasad, Roger D. Peng, and Jeffrey T. Leek, "What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science," *Perspectives on Psychological Science*, 2016, *11* (4), 539–544.

Raj, Chetty, John Friedman, Nathaniel Hendren et al., "The Opportunity Atlas: Mapping the Childhood Roots of Social Mobility," *Working Paer*, 2020.

Roth, Jonathan and Jiafeng Chen, "Logs With Zeros? Some Problems and Solutions," *Working paper*, 2023.

Savage, Leonard J., "The Theory of Statistical Decision," *Journal of the American Statistical Association*, 1951, *46* (253), 55–67.

Simons, Daniel J., "The Value of Direct Replication," *Perspectives on Psychological Science*, 2014, *9* (1), 76–80.

Simonsohn, Uri, "Small Telescopes: Detectability and the Evaluation of Replication Results," *Psychological Science*, 2015, *26* (5), 559–69.

_ , **Leif D. Nelson, and Joseph P. Simmons**, "P-Curve: A Key to the File-Drawer," *Journal of Experimental Psychology: General*, 2014, *143* (2), 534–547.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65* (3), 557–586.

Stanley, T. D., Evan C. Carter, and Hristos Doucouliagos, "What Meta-Analyses Reveal About the Replicability of Psychological Research," *Psychological Bulletin*, 2018, *144* (12), 1325–1346.

Stoye, Jörg, "Minimax Regret Treatment Choice With Finite Samples," *Journal of Econometrics*, 2009, *151* (1), 70–81.

_ , "New Perspectives on Statistical Decisions Under Ambiguity," *Annual Review of Economics*, 2012, *4*, 257–282.

Tetenov, Aleksey, "Statistical treatment choice based on asymmetric minimax regret criteria," *Journal of Econometrics*, 2012, *166*, 157–165.

Vu, Patrick, "Why Are Replication Rates So Low?," *Working Paper*, 2023.

Wagenmakers, Eric-Jan, Josine Verhagen, and Alexander Ly, "How to Quantify the Evidence for the Absence of a Correlation," *Behavior Research Methods*, 2016, *48* (2), 413–26.

Wald, Abraham, *Statistical Decision Functions*, New York: John Wiley & Sons, 1950.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 1980, *48* (4), 817–838.