Do Standard Error Corrections Exacerbate Publication Bias?

Patrick Vu[†] *(Job market paper)*

Please click here for latest version

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.

Keywords: Standard error corrections, publication bias, difference-in-differences, meta-analysis, statistical decision theory

[†] *This version:* December 27, 2023. Brown University. Email: patrick_vu@brown.edu. I am especially grateful for invaluable advice and encouragement from Jonathan Roth, Peter Hull, and Toru Kitagawa. I would also like to thank Daniel Björkegren, Kenneth Chay, Soonwoo Kwon, Susanne Schennach, Jesse Shapiro and Aleksey Tetenov for helpful comments, as well as seminar participants at Brown University, the University of Canterbury, the Econometrics Society North American Summer Meeting 2023, the 2023 MAER-Net Colloquium, and the AYEW at Monash University. I gratefully acknowledge financial support from the Orlando Bravo Center for Economic Research.

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 correc-

tions 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, 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

[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.

[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.*

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 standard 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

[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.

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 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., 2019; 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 develops the theoretical framework and presents the main propositions. Section 3 describes the empirical setting and presents the descriptive statistics. Section 4 shows the results from the empirical model. Section 5 develops the treatment choice model and presents the main welfare results. Section 6 concludes.

2. Theory

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 A.

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:

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

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.

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 $\gamma \in [0,1)$.*

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.

[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 $r \perp\!\!\!\perp (\hat{\beta}_j, \beta_j, \sigma_j)$.

[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.

2.1.1. 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 A.6 in Appendix A 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 (↑ 19 ppts) can come at the unintended cost of a deterioration in the credibility of point estimates due to increased bias (↑ 125%). It also demonstrates that these effects can be large.

This tension has only been shown here for a special case where $(\mu_{\beta,\sigma}, \gamma, r) = \left(\mathbb{Pr}[\beta_j = 1, \sigma_j = 1] = 1, 0, \frac{1}{2}\right)$. 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.

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 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.

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

[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.

[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).

corrections can affect each type of bias:

$$\underbrace{\mathbb{E}_1[\hat{\beta}_j|D_j=1]-\mathbb{E}_r[\hat{\beta}_j|D_j=1]}_{\Delta\text{Estimated Treatment Effects}\,=\,\Delta\text{Total Bias}}$$

$$=\underbrace{\mathbb{E}_1[\hat{\beta}_j-\beta_j|D_j=1]-\mathbb{E}_r[\hat{\beta}_j-\beta_j|D_j=1]}_{\Delta\text{Internal-Validity Bias}}+\underbrace{\mathbb{E}_1[\beta_j|D_j=1]-\mathbb{E}_r[\beta_j|D_j=1]}_{\Delta\text{Study-Selection Bias}} \tag{2}$$

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 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

[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.

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

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 (2). Clustering in this example leads to an overall increase in estimated treatment effects (0.30) that reflects an increase in true 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 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.

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 $\text{Coverage}(r) = \mathbb{P}\mathbb{r}_r[\beta_j \in (\hat{\beta}_j - 1.96r, \hat{\beta}_j + 1.96r) | D_j = 1]$ 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: $\mathbb{P}\mathbb{r}_r[\beta_j \in (\hat{\beta}_j - 1.96r, \hat{\beta}_j + 1.96r)]$. In the case without publication bias, it is

[11] This is shown graphically in Figure C1 in Appendix C.

[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.

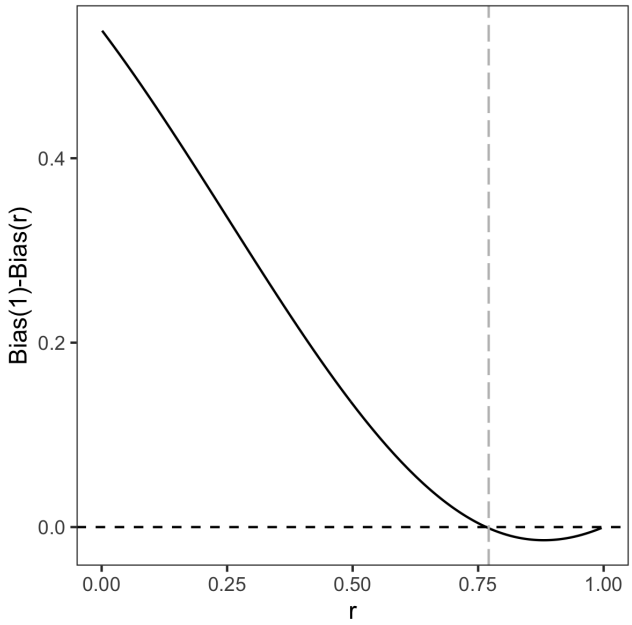

FIGURE 1. Change in internal-study bias from adopting standard error corrections for different degrees of downward bias r: $\mathbb{E}_1[\hat{\beta}_j - \beta_j | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta_j | D_j = 1]$, with $\gamma = \frac{1}{20}$. The dashed vertical line at $r^* = 0.77$ denotes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

clear that standard error corrections for downward bias will increase coverage.

The presence of publication bias, however, introduces several complications. In the definition of 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 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 probability in the corrected regime.

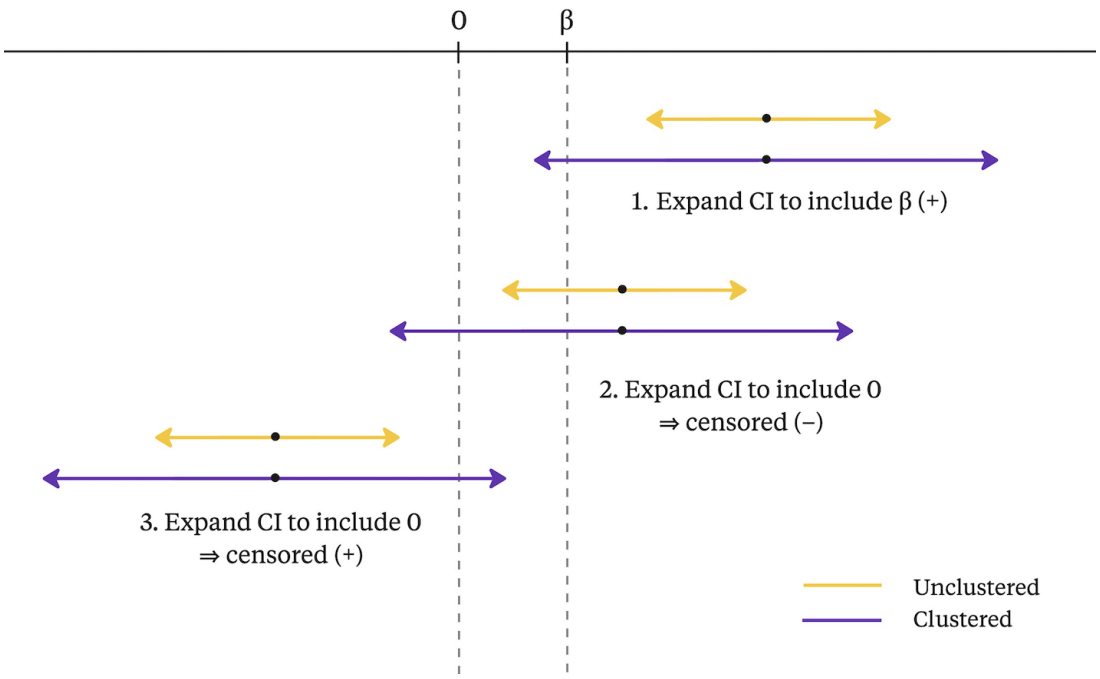

FIGURE 2. Three Potential Effects of Clustering on Coverage Conditional on Publication

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 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 A.

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 A.9 in Appendix A 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.*

[13] Coverage is shown to increase in this special case in Lemma A.6 in Appendix A. 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 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.

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.

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'

[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.

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.

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 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 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 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

[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.

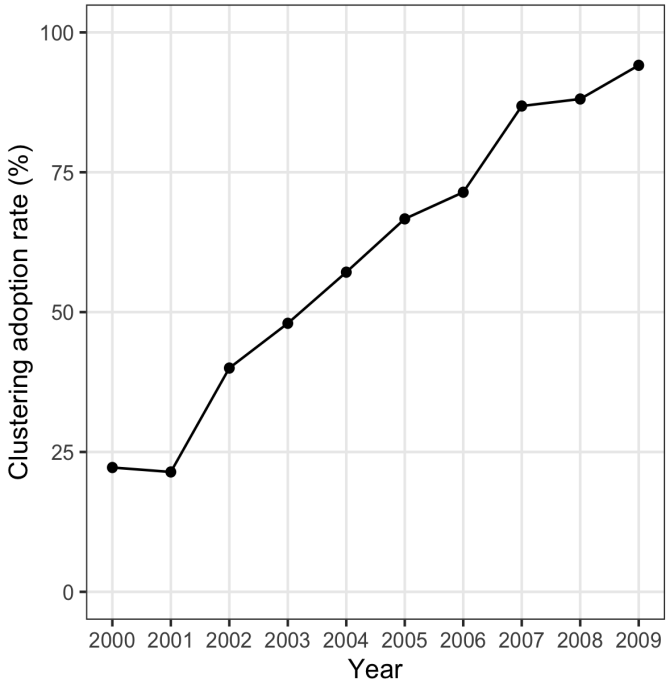

FIGURE 3. Three-Year Centered Moving Average of the Clustering Adoption Rate

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, namely, that clustering 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 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

[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.

[18]There are 26 primary JEL categories (A to Z) corresponding to different fields of economic research. For

TABLE 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.

Welfare), and J (Labor and Demographic Economics). While a high share of both unclustered 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.

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

the full distribution of JEL codes in unclustered and clustered studies, see Appendix D.

journal fixed effects; and

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

3.2.1. Effect Size Gap

As shown in Table 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 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 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 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 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 unclustered 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)

[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.

TABLE 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.

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 E.

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 4.1 for further discussion.

3.2.2. 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 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 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).

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

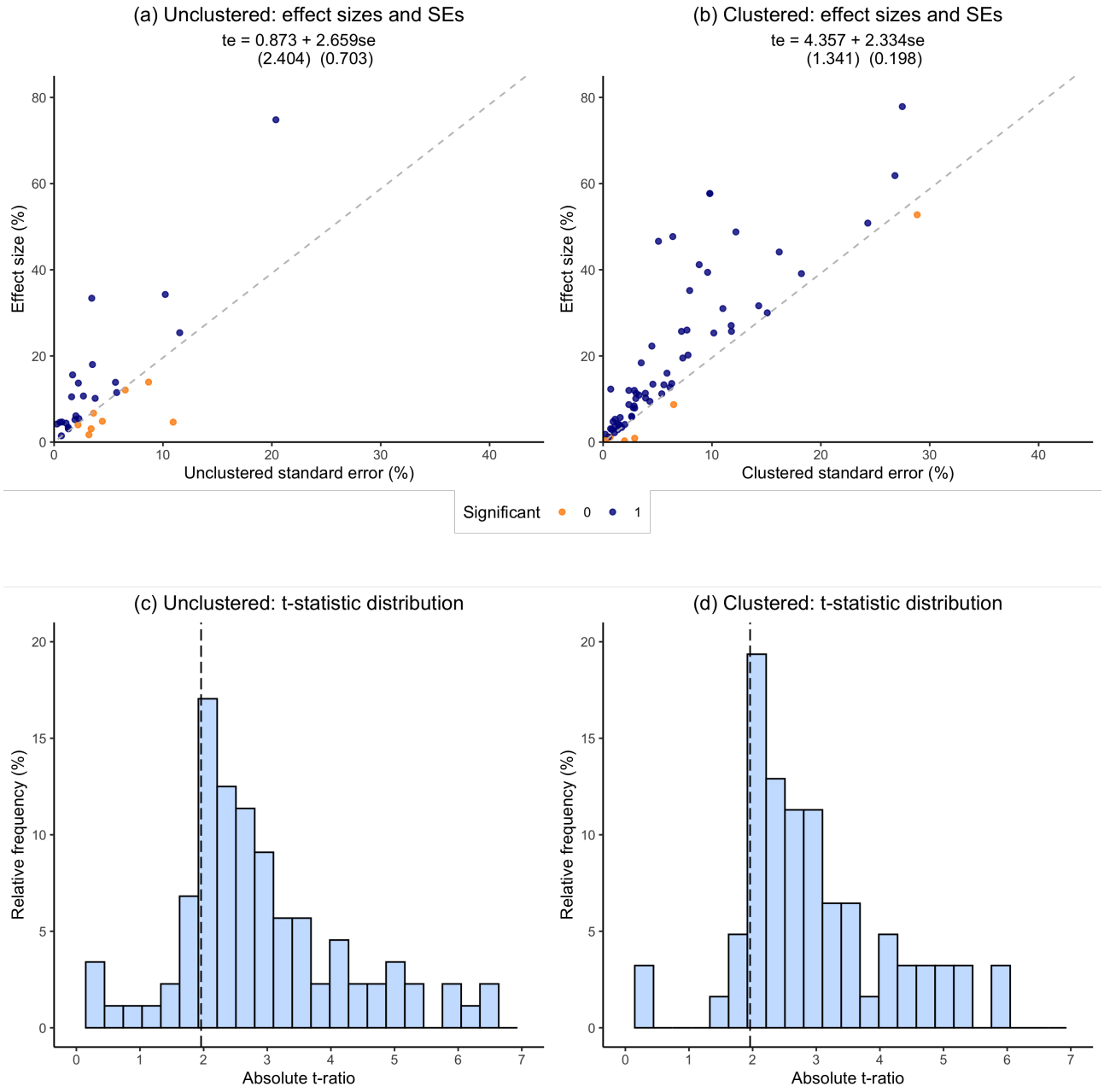

FIGURE 4. SELECTIVE PUBLICATION AND *p*-HACKING

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.

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 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.

4.1. Estimation

First, I estimate the model of selective publication in Section 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 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 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 unclus-

[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 $\gamma \geq 1$ 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.

TABLE 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.

tered standard errors and (unobserved) clustered standard errors for unclustered studies in the likelihood function.[22]

Table 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 F. The main idea in this alternative approach is to estimate 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 F.1 in Appendix F. Robust estimates for the latent distribution of studies are presented in Table F1 and statistically indistinguishable from the baseline estimates in Ta-

[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.

ble 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 3.

4.2. Unclustered Counterfactuals

With the model estimates in Table 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 $r \in (0, 1)$. 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 3. Figure G1 in Appendix G 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.

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 H.

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

[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 F.

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 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 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

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

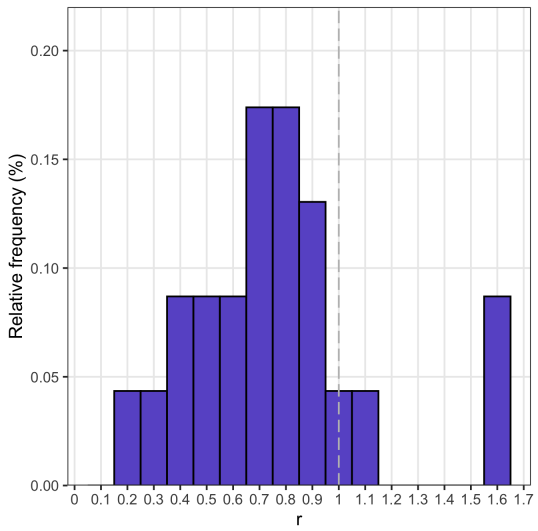

FIGURE 5. Empirical Distribution of r from 2015–2018 DiD Studies

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.

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.

4.4. Impact of Clustering on Coverage and Bias

Table 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 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 (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.

TABLE 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	0.41
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 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 $\hat{r} = 0.51$.

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 H1 in Appendix H. In this alternative approach, the degree of bias of unclustered studies is drawn randomly from the distribution of r (Figure 5), such that r varies across unclustered studies. Results are quantitative similar to those in Table 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 H.

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.

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 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.

5. Impact of Clustering on Evidence-Based Policy

The empirical model in Section 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

[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.*

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.

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

$$U(a, \beta|K) = \begin{cases} Ka\beta & \text{if } \beta \leq 0 \\ a\beta & \text{if } \beta > 0 \end{cases} \tag{3}$$

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 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.

[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.

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

$$\widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma} | K\big) = \begin{cases} K\beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta \leq 0 \\ \beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta > 0 \end{cases} \tag{4}$$

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

[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 (1956) that the class of *threshold decision rules* is essentially complete and consideration of other rules is not necessary.

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:

$$\begin{aligned}\widetilde{R}_1\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) &= W\big(\mathbb{1}\{\beta > 0\}\big) - \widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) \\ &= \begin{cases} -K\beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta \leq 0 \\ \beta\Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big) & \text{if } \beta > 0 \end{cases}\end{aligned} \tag{5}$$

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:

$$T^* = \arg\min_{T \in \mathbb{R}} \max_{\beta \in \beta} \widetilde{R}_1\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) \tag{6}$$

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

$$\widetilde{R}_0\big(\delta_0, \beta|K\big) = \begin{cases} -K\beta\delta_0 & \text{if } \beta \leq 0 \\ \beta(1 - \delta_0) & \text{if } \beta > 0 \end{cases} \tag{7}$$

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

$$\delta_0^* = \arg\min_{\delta_0 \in [0,1]} \max_{\beta \in \beta} \widetilde{R}_0\big(\delta_0, \beta|K\big) \tag{8}$$

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 (6) and (8).

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*

$$(T^*, \delta_0^*) = \left(g(K) \cdot \widetilde{\sigma},\ \frac{1}{1+K} \right) \tag{9}$$

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

Figure 6 illustrates Lemma 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 (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 difference increases with Type I error loss aversion. This is because Lemma 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.

[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.

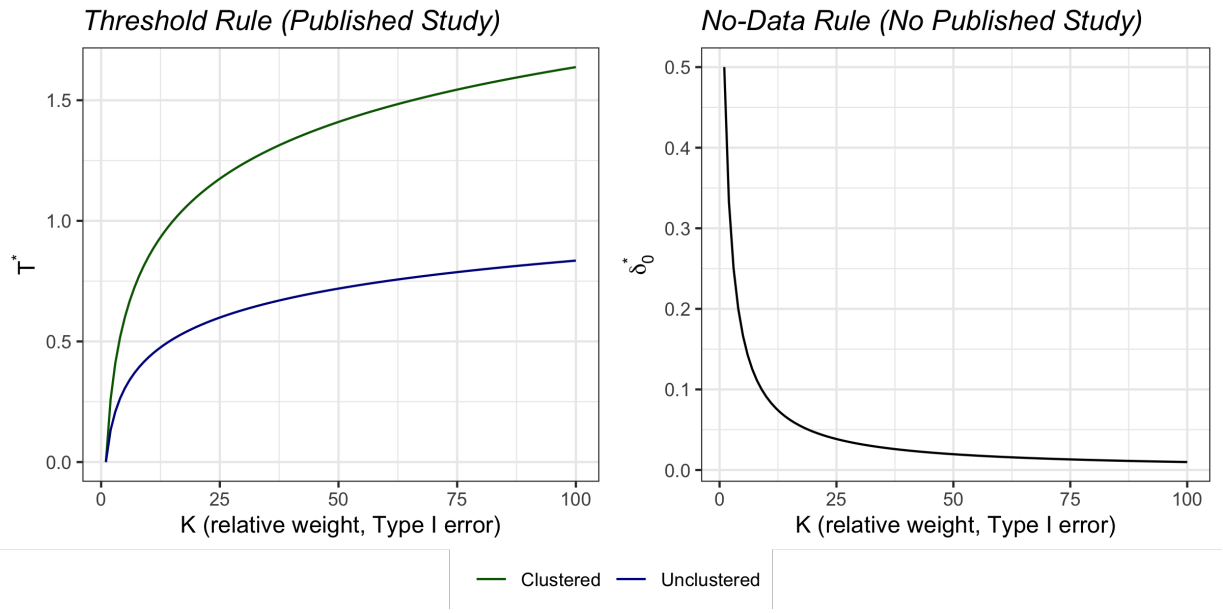

FIGURE 6. Minimax Regret Decision Rule in Clustered and Unclustered Regimes

Notes: The first panel shows the threshold rule in the event that a study is published and given by equation (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 4.

5.3. Comparing Regimes Based on True Regret

While the minimax regret decision rule in Lemma 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

$$U\big(a_1, a_0, \beta | K\big) = \begin{cases} K\beta D a_1 + \beta(1-D)a_0 & \text{if } \beta \leq 0 \\ \beta D a_1 + \beta(1-D)a_0 & \text{if } \beta > 0 \end{cases} \tag{10}$$

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

$$W\big(\delta_1^T, \delta_0, \beta, \sigma, \widetilde{\sigma} | K\big) = \begin{cases} K\Big(\beta \cdot \mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + \beta \cdot \big(1 - \mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}]\big)\delta_0\Big) & \text{if } \beta \leq 0 \\ \beta \cdot \mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + \beta \cdot \big(1 - \mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}]\big)\delta_0 & \text{if } \beta > 0 \end{cases} \tag{11}$$

where $\mathbb{P}\mathbb{r}[D = 1|\beta, \widetilde{\sigma}]$ 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:

$$R\big(\delta_1^T, \delta_0, \beta, \sigma, \widetilde{\sigma}|K\big) = \begin{cases} -K \cdot \beta\Big(\mathbb{P}\mathrm{r}[D = 1|\beta, \widetilde{\sigma}] \cdot [1 - F(T|\beta, \sigma, \widetilde{\sigma}, D = 1)] + (1 - \mathbb{P}\mathrm{r}[D = 1|\beta, \widetilde{\sigma}])\delta_0\Big) & \text{if } \beta \leq 0 \\ \beta\Big(\mathbb{P}\mathrm{r}[D = 1|\beta, \widetilde{\sigma}] \cdot F(T|\beta, \sigma, \widetilde{\sigma}, D = 1) + (1 - \mathbb{P}\mathrm{r}[D = 1|\beta, \widetilde{\sigma}]) \cdot (1 - \delta_0)\Big) & \text{if } \beta > 0 \end{cases} \tag{12}$$

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 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

$$\mathrm{MMR}(T, \delta_0|K) = \max_{\beta \in [-B, B]} R\big(\delta_1^T, \delta_0, \beta, \sigma|K\big) \tag{13}$$

For any $K \geq 1$, let $\mathrm{MMR}^*_{C=0}(K)$ denote the value of minimax regret in the unclustered regime based on the (misspecified) decision rule from Lemma 1 and let $\mathrm{MMR}^*_{C=1}(K)$ 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

$$100 \cdot \left(\frac{\mathrm{MMR}^*_{C=1}(K)}{\mathrm{MMR}^*_{C=0}(K)} - 1\right) \tag{14}$$

Figure 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.

[30] Specifically, the cdf is given by

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

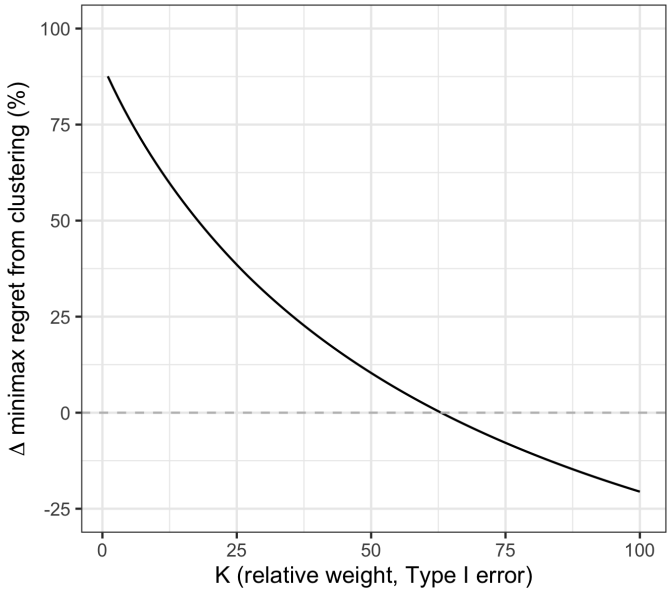

FIGURE 7. Percent Change in Minimax Regret from Clustering

Notes: The percent change in minimax regret moving from the unclustered regime to the clustered regime is calculated according to equation (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 4.

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 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.

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.

References

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).

Amrhein, Valentin, Sander Greenland, and Blake McShane, "Retire Statistical Significance," *Nature*, 2019, *567*, 305–307.

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.

Armstrong, Timothy B., Michal Kolesár, and Mikkel Plagborg-Møller, "Robust Empirical Bayes Confidence Intervals," *Econometrica*, 2022, *90* (6), 2567–2602.

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.

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.

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.

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.

Editorial, "In praise of replication studies and null results," *Nature*, 2020, *578*, 489–490.

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.

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.

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 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.

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.

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.

Manski, Charles F., "Statistical Treatment Rules for Heterogeneous Populations," *Econometrica*, 2004, *72* (4), 1221–1246.

McFadden, Daniel, "A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration," *Econometrica*, 1989, *57* (5), 995–1026.

Miguel, Edward and Garret Christensen, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–980.

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.

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.

Staiger, Douglas and James H. Stock, "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 1997, *65* (3), 557–586.

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.

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.

Appendix

This appendix contain proofs and supplementary materials. Section A contains proofs for the Propositions and Lemmas in the main text. Section B provides examples showing that bias can decrease when standard error corrections are small. Section D provides additional graphs illustrating the data. Section E shows descriptive statistics for unclustered studies in the 1990–1999 period. Section F 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 G shows counterfactual comparisons between the clustered regime and the unclustered regime for all values of r on the unit interval. Finally, Section H shows robustness of the main results from using the empirical distribution of r calculated from 2015–2018 DiD studies.

A. Proofs

Proof of Proposition 1: The main result follows from two Lemmas which I prove below. First, Lemma A.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 A.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 A.2 and A.3 on which this argument is based. Before that, I present Lemma A.1, which is used in Lemma A.2.

Lemma A.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:

$$\begin{aligned}
&\mathbb{E}_r[Z_j | D_j = 1] \\
&= \mathbb{Pr}_r[Z_j \leq -1.96r - \beta | D_j = 1] \cdot \mathbb{E}[Z_j | Z_j \leq -1.96r - \beta] \\
&+ \mathbb{Pr}_r[-1.96r - \beta < Z_j \leq 1.96r - \beta | D_j = 1] \cdot \mathbb{E}[Z_j | -1.96r - \beta < Z_j \leq 1.96r - \beta] \\
&+ \mathbb{Pr}_r[Z_j > 1.96r - \beta | D_j = 1] \cdot \mathbb{E}[Z_j | Z_j > 1.96r - \beta] \\
&= \left(\frac{\mathbb{Pr}_r[D_j = 1 | Z_j \leq -1.96r - \beta]\Phi(-1.96r - \beta)}{\mathbb{Pr}_r[D_j = 1]}\right)\left(-\frac{\phi(-1.96r - \beta)}{\Phi(-1.96r - \beta)}\right) \\
&+ \left(\frac{\mathbb{Pr}_r[D_j = 1 | -1.96r - \beta \leq Z_j \leq 1.96r - \beta]\big[\Phi(1.96r - \beta) - \Phi(-1.96r - \beta)\big]}{\mathbb{Pr}_r[D_j = 1]}\right) \\
&\times \left(\frac{\phi(-1.96r - \beta) - \phi(1.96r - \beta)}{\Phi(1.96r - \beta) - \Phi(-1.96r - \beta)}\right) \\
&+ \left(\frac{\mathbb{Pr}_r[D_j = 1 | Z_j \geq 1.96r - \beta]\big[1 - \Phi(1.96r - \beta)\big]}{\mathbb{Pr}_r[D_j = 1]}\right)\left(\frac{\phi(1.96r - \beta)}{1 - \Phi(1.96r - \beta)}\right) \\
&= -\frac{\phi(-1.96r - \beta)}{\mathbb{Pr}_r[D_j = 1]} + \frac{\gamma\big[\phi(-1.96r - \beta) - \phi(1.96r - \beta)\big]}{\mathbb{Pr}_r[D_j = 1]} + \frac{\phi(1.96r - \beta)}{\mathbb{Pr}_r[D_j = 1]}
\end{aligned}$$

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 A.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

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

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

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

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

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 $\mathbb{P}\mathbb{r}_r[D_j=1|\beta] \leq 1$.

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

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

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

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

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

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

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

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 $\lim_{r\to 0} \mathbb{Pr}_r[D_j=1|\beta] = 1$ 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 A.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 A.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

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

The second equality uses the convergence in expectation result in equation (18) from Lemma A.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

$$\begin{aligned}
&\int_0^t f_\beta(\beta) d\beta - \int_0^t f_{\beta|D}(\beta | D_j = 1; \gamma, 1) d\beta \\
&= \frac{1}{\mathbb{P}\mathbb{r}_1(D_j = 1)} \left(\mathbb{P}\mathbb{r}_1(D_j = 1) \int_0^t f_\beta(\beta) d\beta - \int_0^t \mathbb{P}\mathbb{r}_1(D_j = 1 | \beta) f_\beta(\beta) d\beta \right)
\end{aligned}$$

$$= \frac{F_\beta(t)}{\mathbb{P}\mathbb{r}_1(D_j = 1)}\Big(\mathbb{E}_\beta\Big[\mathbb{P}\mathbb{r}_1(D_j = 1|\beta)\Big] - \mathbb{E}_\beta\Big[\mathbb{P}\mathbb{r}_1(D_j = 1|\beta)\Big|\beta \leq t)\Big]\Big) \geq 0$$

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 $\mathbb{P}\mathbb{r}_1(D_j = 1|\beta)$ 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 $\big\{r|r \in (0, 1], \Delta\text{SSB}(r) = 0\big\}$. 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. □

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.

[31] The derivative is given by:

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

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

Lemma A.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*

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

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

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

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 A.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

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

Proof. For this proof, I express the probability of publication in publication regime γ and standard error regime r explicitly as $\mathbb{P}\mathbb{r}(D_j = 1; \gamma, r)$ (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

$$f_{\hat{\beta},\beta|D}(\hat{\beta},\beta|D_j=1;\gamma,r) = \frac{\mathbb{Pr}(D_j=1|\hat{\beta};\gamma,r)\phi(\hat{\beta}-\beta)f_\beta(\beta)}{\mathbb{Pr}(D_j=1;\gamma,r)}$$

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

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

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

$$= \left(\frac{\gamma}{\mathbb{Pr}(D_j=1;\gamma,r)}\right)\Big(\mathbb{1}\{|\hat{\beta}|\geq 1.96r\}\phi(\hat{\beta}-\beta)f_\beta(\beta)+\mathbb{1}\{|\hat{\beta}|<1.96r\}\phi(\hat{\beta}-\beta)f_\beta(\beta)\Big)$$

$$+\left(\frac{\mathbb{Pr}(D_j=1;\gamma,r)-\gamma}{\mathbb{Pr}(D_j=1;\gamma,r)}\right)\left(\frac{\mathbb{1}\{|\hat{\beta}|\geq 1.96r\}\phi(\hat{\beta}-\beta)f_\beta(\beta)}{\mathbb{Pr}(D_j=1;0,r)}\right)$$

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

$$+\left(\frac{\gamma\mathbb{1}\{|\hat{\beta}|<1.96r\}\phi(\hat{\beta}-\beta)f_\beta(\beta)}{\mathbb{Pr}(D_j=1;\gamma,r)}\right)$$

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:

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

□

In the next step, I show that Lemma A.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 A.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:

$$\begin{aligned} & c_0(1) - c_0(r) \\ = & \int_0^\infty \text{Coverage}(\beta,1) f_{\beta|D}(\beta|D_j=1;0,1)d\beta - \int_0^\infty \text{Coverage}(\beta,r) f_{\beta|D}(\beta|D_j=1;0,r)d\beta \\ = & \left(\int_0^{2\times 1.96r} \text{Coverage}(\beta,1) f_{\beta|D}(\beta|D_j=1;0,1)d\beta - \int_0^{2\times 1.96r} \text{Coverage}(\beta,r) f_{\beta|D}(\beta|D_j=1;0,r)d\beta \right) \\ + & \left(\int_{2\times 1.96r}^\infty \text{Coverage}(\beta,1) f_{\beta|D}(\beta|D_j=1;0,1)d\beta - \int_{2\times 1.96r}^\infty \text{Coverage}(\beta,r) f_{\beta|D}(\beta|D_j=1;0,r)d\beta \right) \end{aligned} \tag{24}$$

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 A.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 A.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 A.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

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

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

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 A.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*

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

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 A.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 A.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 A.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

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

where $c\equiv 1.96r$ and $K\equiv\mathbb{Pr}[D_j=1;0,r]/\mathbb{Pr}[D_j=1;0,1]>0$. 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

$$\int_0^{2\times 1.96r} \text{Coverage}(\beta, 1) f_{\beta|D}(\beta|D_j = 1; 0, 1) d\beta - \int_0^{2\times 1.96r} \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, r) d\beta$$

$$\geq \int_0^{2\times 1.96r} \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, 1) d\beta - \int_0^{2\times 1.96r} \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, r) d\beta \geq 0$$

The first inequality uses Lemma A.6 to replace $\text{Coverage}(\beta, 1)$ with $\text{Coverage}(\beta, r)$ in the first term. The final inequality follows from the fact that $\text{Coverage}(\beta, r)$ is strictly increasing over $(0, 2 \times 1.96r)$ (Lemma A.7) and first-order stochastic dominance in the distribution of

published true effects in the corrected regime as compared with the uncorrected regime (Lemma A.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 A.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:

$$\begin{aligned}\text{Coverage}(r) &= \int \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; \gamma, r) d\beta \leq \text{Coverage}(2 \times 1.96r, r) \\ &= \frac{\Phi(1.96r) - \Phi(-1.96r)}{\Phi(-3 \times 1.96r) + 1 - \Phi(-1.96r) + \gamma[\Phi(-1.96r) - \Phi(-3 \times 1.96r)]} \\ &\leq \frac{\Phi(1.96r) - \Phi(-1.96r)}{\Phi(-3 \times 1.96r) + 1 - \Phi(-1.96r)} \qquad (27)\end{aligned}$$

The first inequality follows from Lemma A.7, which shows that $\text{Coverage}(\beta, r)$ 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 A.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 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). □

B. Ambiguous Impact of Corrections on Bias (and Other Measures)

Proposition 1 shows that when standard error corrections are sufficiently large, bias, estimated treatment effects and true treatment effects must all increase. This appendix presents examples for when these quantities decrease in the case where standard error corrections are small. This is formalized in the following lemma:

Lemma B.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 $\mathbb{P}\mathrm{r}[\beta_j = \beta] = 1$ 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 A.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 $\mathbb{P}\mathrm{r}[\beta_j = \beta] = p_1^* \cdot \mathbb{1}\{\beta = \beta_1\} + (1 - p_1^*) \cdot \mathbb{1}\{\beta = \beta_2\}$ 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 $C(\beta, \gamma, r) \equiv \int_{z'} p\left(\frac{\beta + z'}{r}\right)\phi(z')dz'$ 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

$$\begin{aligned} &\text{TrueTE}(0, \beta_2, 0.5, \gamma, 1) - \text{TrueTE}(0, \beta_2, 0.5, \gamma, r) \\ &= \beta_2 \left(\frac{C(\beta_2, \gamma, 1)}{C(0, \gamma, 1) + C(\beta_2, \gamma, 1)} - \frac{C(\beta_2, \gamma, r)}{C(0, \gamma, r) + C(\beta_2, \gamma, r)} \right) \end{aligned} \tag{29}$$

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 B.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 B.1 shows that bias can decrease with a degenerate, and hence unchanged, distribution of true effects.

For intuition, consider the example in Lemma B.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.

C. Bias and True Treatment Effects

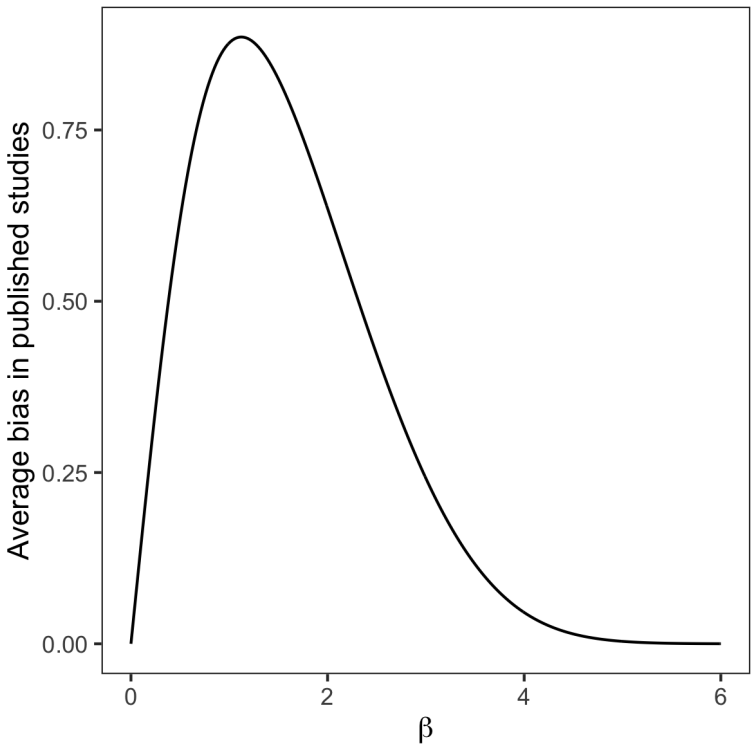

FIGURE C1. Plot of $\mathbb{E}_1[\hat{\beta}_j - \beta | D_j = 1, \beta]$ for different values of β and $\gamma = 0.1$.

D. Details on Descriptive Statistics

This appendix provides further details on the descriptive statistics in Section 3.

Figure C1 shows the distribution of JEL codes. Note that studies typically include multiple JEL codes and Figure C1 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 C1 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).

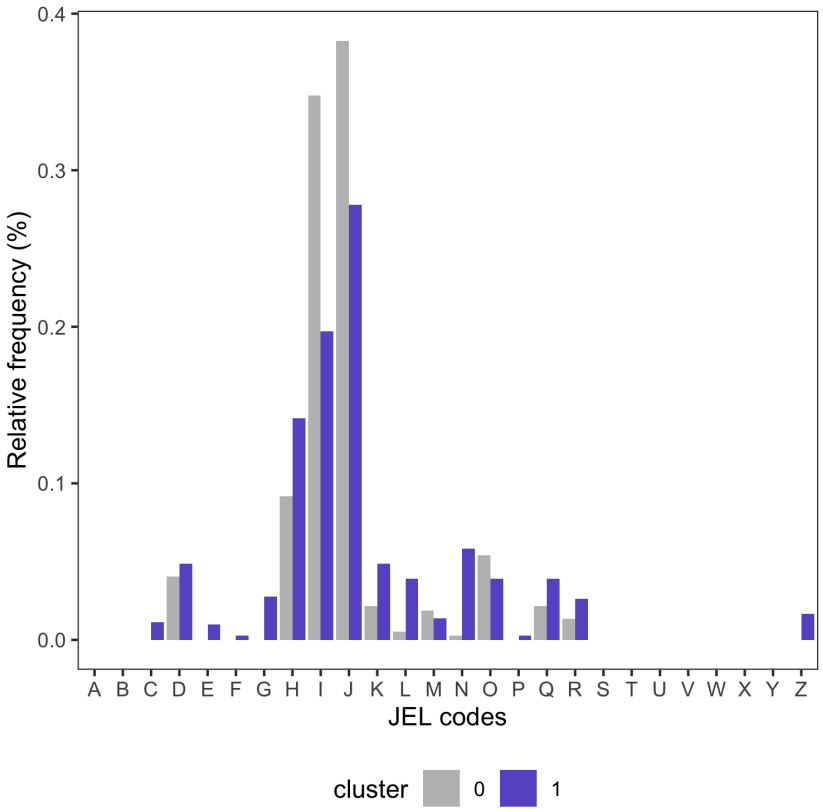

FIGURE C1. Distribution of JEL codes. The most common JEL codes are: Public Economics (H); Health, Education & Welfare (I); and Labor (J)

Figure C2 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.

[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.

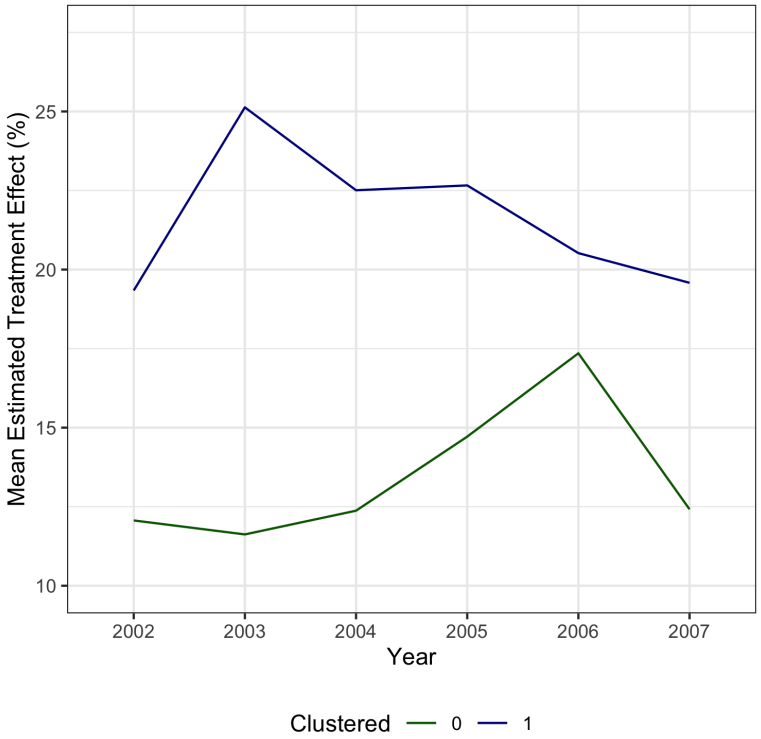

FIGURE C2. Five-Year Centered Moving Average of the Magnitude Estimated Treatment Effects

E. 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 D1 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 described in Section 3 which ensure comparability of effect sizes (i.e. estimated treatment effects in percent units from a binary treatment) across studies.

TABLE D1 – 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.

F. 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.

F.1. Model of Strategic Clustering

The model extends the model in Section 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:**

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

The extension from the baseline model in Section 2 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.

F.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 F.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

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

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

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:

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

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

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.

F.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 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.

TABLE F1 – 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.

F.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

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

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 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 of the main results to the (statistically insignificant) differences in parameter estimates in Table F1.

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 F2 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 F2 – 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	0.41
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 F, 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.

G. Impact of Clustering for Different Sized Corrections

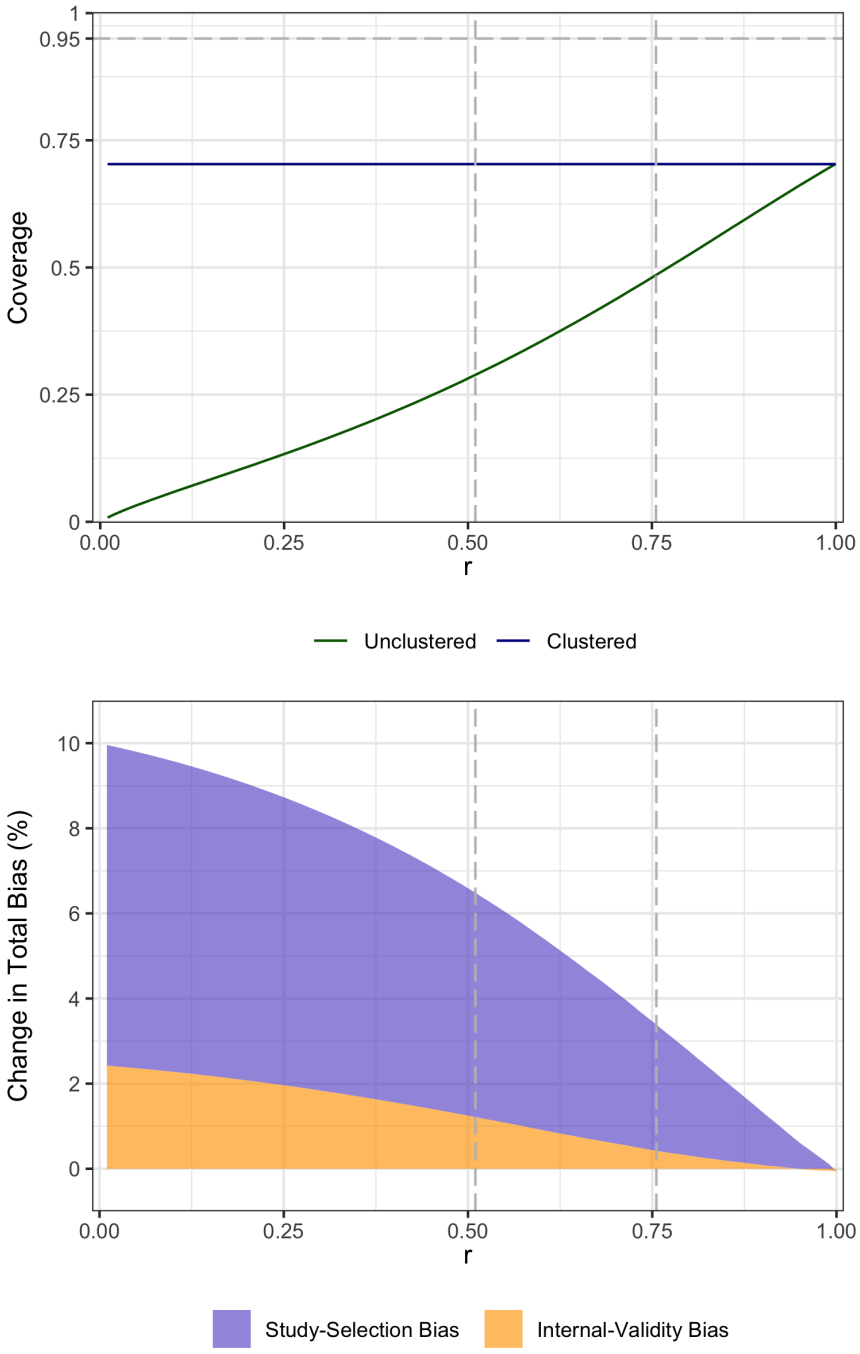

FIGURE G1. Results on the Impact of Clustering for Different Values of r

Notes: Change in coverage, total bias (and estimated treatment effects), study-selection bias, and internal-validity bias for the estimated model parameters in Table 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.

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

TABLE H1 – 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 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 ($\hat{r} = 0.76$).