Do Standard Error Corrections Exacerbate Publication Bias?

Patrick Vu[†]

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 credibility of published studies. I show that adopting improved but enlarged standard errors in 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 difference-in-differences studies published between 2000 and 2009. Clustering is associated with a near doubling in the magnitude of 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 welfare effects, I estimate a decision-theoretic model of an audience that uses evidence from published studies to inform decisions and overestimates the precision of estimates when standard errors are unclustered. I find that clustering improves welfare, as the benefits from more accurate belief updating outweigh the cost of increased publication bias.

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

[†] *This version:* June 25, 2025. University of New South Wales. Email: patrick.vu@unsw.edu.au. 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, Lewis McLean, Susanne Schennach, Jesse Shapiro and Aleksey Tetenov for helpful comments, as well as seminar participants at Brown University, Caltech, the University of Canterbury, the University of New South Wales, the University of Queensland, the University of Western Australia, 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 methods for estimating standard errors in a wide range of settings (White, 1980; Moulton, 1986; Newey and West, 1987; Staiger and Stock, 1997; Calonico et al., 2014). 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.

Scant attention has been paid to the close connection between standard error corrections and publication bias, despite the fact that both literatures are guided by the shared aim of improving the credibility of empirical research. This paper's main contribution is to demonstrate – both theoretically and empirically – how their interaction can have important implications for bias and coverage, and evidence-based decision-making.

In the first part of the paper, I use newly collected data to document the adoption of clustered standard errors in the empirical DiD literature in the 2000's. I show that the adoption of clustering increased from around one in four studies in the earlier part of the decade to near-universal adoption by the end of it. This increase coincided with the publication of Bertrand et al. (2004), an influential study that documented infrequent use of standard error corrections for serial correlation in applied DiD research, despite earlier emphasis in the econometrics literature (e.g. Moulton (1986)).

The adoption of clustering led to dramatic changes in the distribution of published estimates. Most strikingly, effect sizes in clustered studies are almost twice the magnitude compared to unclustered studies, even after controlling for differences in research topics, sample size, and including year and journal fixed effects. At the same time, clustering had essentially no impact on the magnitude of t-ratios, as higher standard errors from clustering were met with commensurate increases in effect sizes. These patterns are consistent with publication bias favoring statistically significant findings, and complemented by meta-regression results which show a strong positive association between standard errors and effect sizes for both clustered and unclustered studies.

This is the central dynamic explored in this paper: clustering increases standard errors, which raises the bar for statistical significance, inadvertently exacerbating publication bias. What this means for the statistical credibility of published DiD studies, however, is far from

obvious. Larger effect sizes in clustered studies could reflect increased bias or, alternatively, a shift in publication toward studies addressing questions with larger true effects. Moreover, given the possibility of higher bias, it is not even clear whether clustering actually meets its primary objective of improving coverage conditional on publication.

To address these questions, the second part of the paper establishes general theoretical results on the impact of standard error corrections on bias and coverage. The theoretical results apply not only to clustering, but also more generally to any corrections for standard errors that tend to enlarge them (e.g. heteroscedasticity-robust standard errors, heteroscedasticity and autocorrelation consistent standard errors, corrections for weak instruments, or robust standard errors in regression-discontinuity designs).

The theoretical framework builds on the model of selective publication in Andrews and Kasy (2019) to incorporate the possibility that reported standard errors are mismeasured. Using this framework, I show that average bias in published studies can either increase or decrease following standard error corrections, but that increases are inevitable when corrections are sufficiently large. The case of large corrections is empirically relevant because uncorrected standard errors have been shown in many instances to be severely downward biased.[1] Despite the possibility of higher bias, I show that standard error corrections unambiguously increase average coverage in published confidence intervals. This holds under quite general conditions. In particular, it holds for any sized standard error 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. Whether these changes are actually large enough to warrant attention, however, is an empirical question.

Thus, in the final part of the paper, I turn to estimating the model in the empirical DiD setting. Estimation uses the meta-study approach in Andrews and Kasy (2019) on clustered studies. A key aspect of estimation is to calibrate the degree of downward bias in unclustered standard errors. I use two methods. The first method makes the simplifying assumption that all unclustered standard errors are downward biased by a constant factor $r \in (0, 1)$, and then calibrates r using the method of simulated moments (McFadden, 1989). The second method estimates the empirical distribution of the downward bias of unclustered standard errors from a sample of DiD papers in Brodeur et al. (2020). Both provide similar results.

[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.

Results from the estimated model show that clustering led to large improvements in coverage. In the unclustered regime, the coverage probability of published confidence intervals was only 0.36. This implies severe mismeasurement in the calculation of confidence intervals with unclustered standard errors, with only around one in three published confidence intervals containing the true parameter value. By contrast, coverage doubles to 0.72 in the clustered regime, a dramatic improvement but still below nominal coverage of 0.95 due to publication bias.

Despite substantial increases in coverage, clustering also led to average bias in published studies increasing by around 60%, from 1.47 percentage points to 2.34 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 78% 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.

Given both higher bias and coverage, the welfare effects of clustering are unclear. To examine welfare, I use the decision-theoretic model in Frankel and Kasy (2022) to estimate the impact of clustering on the utility of an audience (e.g. policymakers, practitioners, the scientific community) that uses published studies to inform decisions. In the model, if no study is published, then the audience relies on their prior belief. On the other hand, if a study is published, then the audience updates their beliefs about the true effect using Bayes Rule. In contrast to the standard model, I assume that standard errors can be mismeasured, and that the audience naively updates their beliefs based on the reported value.

I operationalize the model using estimates from the empirical DiD literature. Results show that quadratic loss in the clustered regime is 36.5% lower than in the unclustered regime. The key mechanism is that the audience in the unclustered regime drastically overestimates the precision of the published estimate due to downwardly biased standard errors. This causes them to place too much weight on the published estimates when updating their beliefs, which ultimately leads to suboptimal decisions. Empirically, this outweighs the fact that studies in the clustered regime are more likely to be censored by publication bias.

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; 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 aim to improve the credibility of applied research, 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. It is important to emphasize that this paper does not recommend that researchers should not use robust standard errors. Instead, it aims to examine an underappreciated cost of publication bias and quantify its impact.

This paper also contributes to the literature on statistical decision theory and treatment choice (Wald, 1950; Savage, 1951; Stoye, 2009; Tetenov, 2012; Frankel and Kasy, 2022). In the existing literature, treatment choice models typically assume that standard errors are correctly measured. This paper extends the Bayesian framework in Frankel and Kasy (2022) to incorporate broader concerns in the econometrics literature that statistical inference is impaired by mismeasured standard errors.

This paper proceeds as follows. Section 2 describes the empirical setting and presents the descriptive statistics. Section 3 develops the theoretical framework and presents the main propositions. Section 4 presents the results from the empirical model and welfare analysis. Section 5 concludes.

2. Clustering in Difference-in-Difference Studies

This section documents the adoption of clustered standard errors in the empirical DiD literature in the 2000's and its impact on the distribution of published results. This is a compelling setting for at least two reasons. First, DiD is an extremely popular research design in the quantitative social sciences. In economics, it is the most widely referenced quasi-experimental method and its popularity has increased dramatically over time (Currie et al., 2020). Second, failing to cluster frequently results in large downward bias in standard errors, which can lead to exaggerated statistical support for the effectiveness of an intervention (Moulton, 1986, 1990; Bertrand et al., 2004).

2.1. Data

For the empirical analysis, I constructed a 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 a

'main DiD estimate' in each study, excluding placebo tests and tests of alternative hypotheses. The 'main DiD estimate' was chosen as each paper's first full-control DiD specification. 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[2]; the number of observations; and JEL classification codes from *EconLit*. I did not collect the number of clustering groups (e.g. states) because it was frequently unreported.

To ensure meaningful comparisons of effect sizes across studies, I include 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.[3] As an example, consider a study estimating the impact of an educational program on the drop-out rate. In this case, I would 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.[4] 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). The final sample includes 88 studies, 62 of which are clustered. For descriptive statistics comparing clustered and unclustered studies, see Appendix B.

While the main type of serial correlation correction for standard errors in the sample is clustering, a small number of studies implement other corrections e.g. block-bootstrapped standard errors, two-period aggregation.[5] For brevity, I use the term 'clustering' in this

[2]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.

[3]If the paper reported multiple DiD effects, some in levels and others in log units, I selected the log unit regression. Note also that the normalized ATE is a different parameter to the ATE in log differences (Roth and Chen, 2023).

[4]In a small number of cases, normalizing by the mean led to very large percent effects due to low base effect. Four outliers whose effect sizes were above 100% were removed for this reason – two clustered studies and two unclustered studies. This has little impact on the distribution of effect sizes for clustered and unclustered studies (Figure B2). Alternatively, the analysis can be done on the restricted sample of studies that report effect sizes in log units, which does not contain outliers. This approach yields similar results, though with a smaller sample size.

[5]Since GLS corrections do not perform well in Monte Carlo simulations (Bertrand et al., 2004), I exclude them from sample. Six studies with regressions at the state-cohort level (or region-cohort level, or

article to refer to any correction which accounts for the correlation of errors within groups across time. A small number of studies cluster standard errors at a level that does not account for serial correlation (e.g. clustering at the state-year level). These studies are categorized as not having corrected standard errors for serial correlation.

While the 'correct' level of clustering is an active topic of research (e.g. Abadie et al. (2023)), there is little disagreement over whether standard errors should allow for serial correlation in DiD settings. For descriptive statistics in this section, I simply present the reported standard errors for clustered and unclustered studies. For the empirical model in Section 4, I make a stronger assumption that reported clustered standard errors reflect the true standard error.

2.2. Descriptive Statistics

Figure 1 plots the time series for four key variables. Each panel reports a five-year centered moving average to smooth annual variability. The top panel shows 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 is likely in part 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 (7.7%).

The second panel shows that standard error corrections increased over the decade. This is consistent with expectations form the econometrics literature, which emphasizes downward bias in the absence of clustering (Moulton, 1986, 1990; Bertrand et al., 2004; Abadie et al., 2023).

Most strikingly, the third panel shows that average effect sizes almost doubled over the period, from around 10% in the early part of the decade to almost 20% by the end of it. Corresponding regressions are shown in columns 1–4 in Table 1, which report results for regressions of the effect size on an indicator for clustering, adding additional controls with each successive column. The final specification in the fourth column 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

municipality-cohort level etc.) are included. For these studies, clustering at the state level (or similar) is counted as having implemented an appropriate standard error correction.

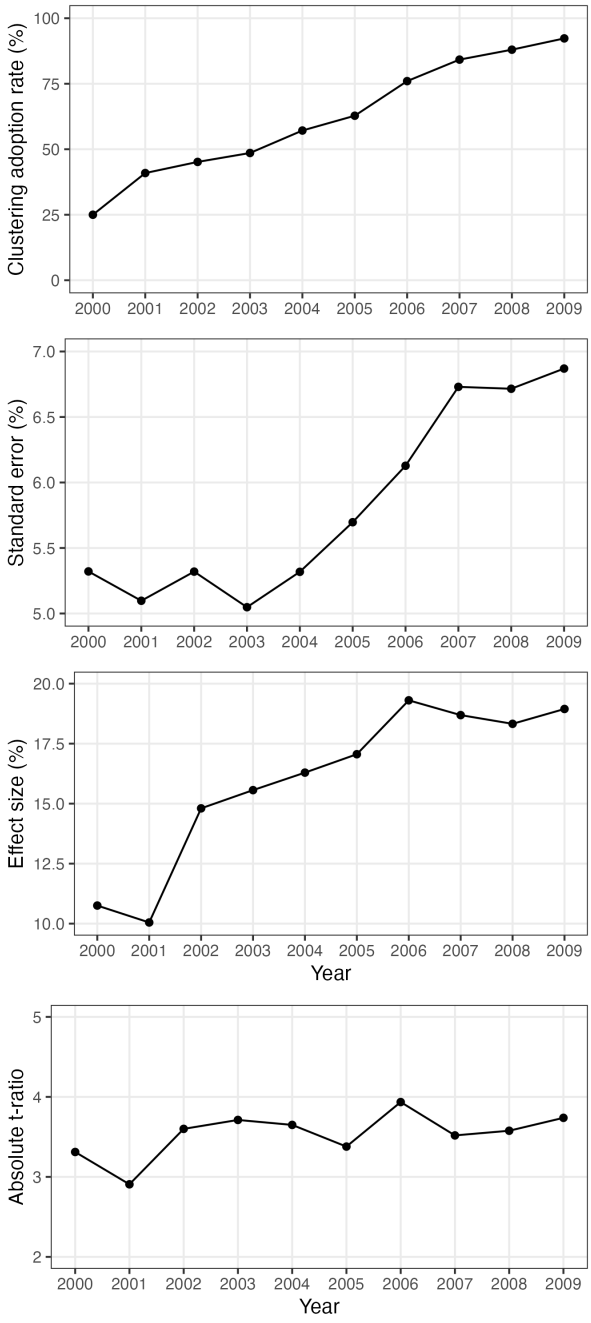

FIGURE 1. Five-Year Centered Moving Average of Adoption and Published Statistics

of 1.8 (19.9% vs. 11.2%). This suggests that increased effect sizes over the decade are driven by the adoption of clustering, rather than differences in observable study characteristics.

Despite rising effect sizes over the decade, the last panel in Figure 1 shows that the average t-ratio is relatively stable over the period. That is, larger standard errors from clustering were accompanied by similar increases in effect sizes, leaving the average magnitude of the t-ratio essentially unchanged at around 3.5. This remains true even after controlling for ob-

TABLE 1 – Impact of Clustering on Effect Sizes and t-Ratios

	Effect Sizes				Absolute t-ratio			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Clustered	8.642	7.882	10.243	10.287	0.023	-0.180	0.149	-0.336
	(3.536)	(4.178)	(4.913)	(5.764)	(0.851)	(0.745)	(0.746)	(0.865)
Unclustered mean	11.23	11.23	11.23	11.23	3.53	3.53	3.53	3.53
Observations	88	88	88	88	88	88	88	88
Adjusted-R^2	0.035	0.012	0.008	0.009	-0.012	-0.002	-0.03	-0.035
Year FE		X	X	X		X	X	X
Journal FE			X	X			X	X
Study controls				X				X

Notes: OLS regressions of estimated treatment effects (columns 1–4) and absolute t-ratios (columns 5–8) on an indicator for clustering. In columns 1–4, the dependent variable is in percent units (or log points for studies where the dependent variable is in logs); and 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.

servable study characteristics (Columns 5–8 in Table 1). In each specification, the estimated coefficient on the clustering indicator is small in magnitude and statistically indistinguishable from zero.

2.3. Publication Bias

The above results are consistent with the following mechanism: clustering enlarges standard errors, which raises the bar on statistical significance, such that larger effect sizes are now required to publish results. To provide further evidence in support of this explanation, this subsection explores two common approaches used in the meta-science literature for detecting publication bias.

The first is the metaregression approach proposed in Card and Krueger (1995). Figure 2 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 relatively 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 1.6–2.2 percentage point increase in standard errors is associated with a little over a two percentage point increase in estimated effect sizes – this is, close to 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

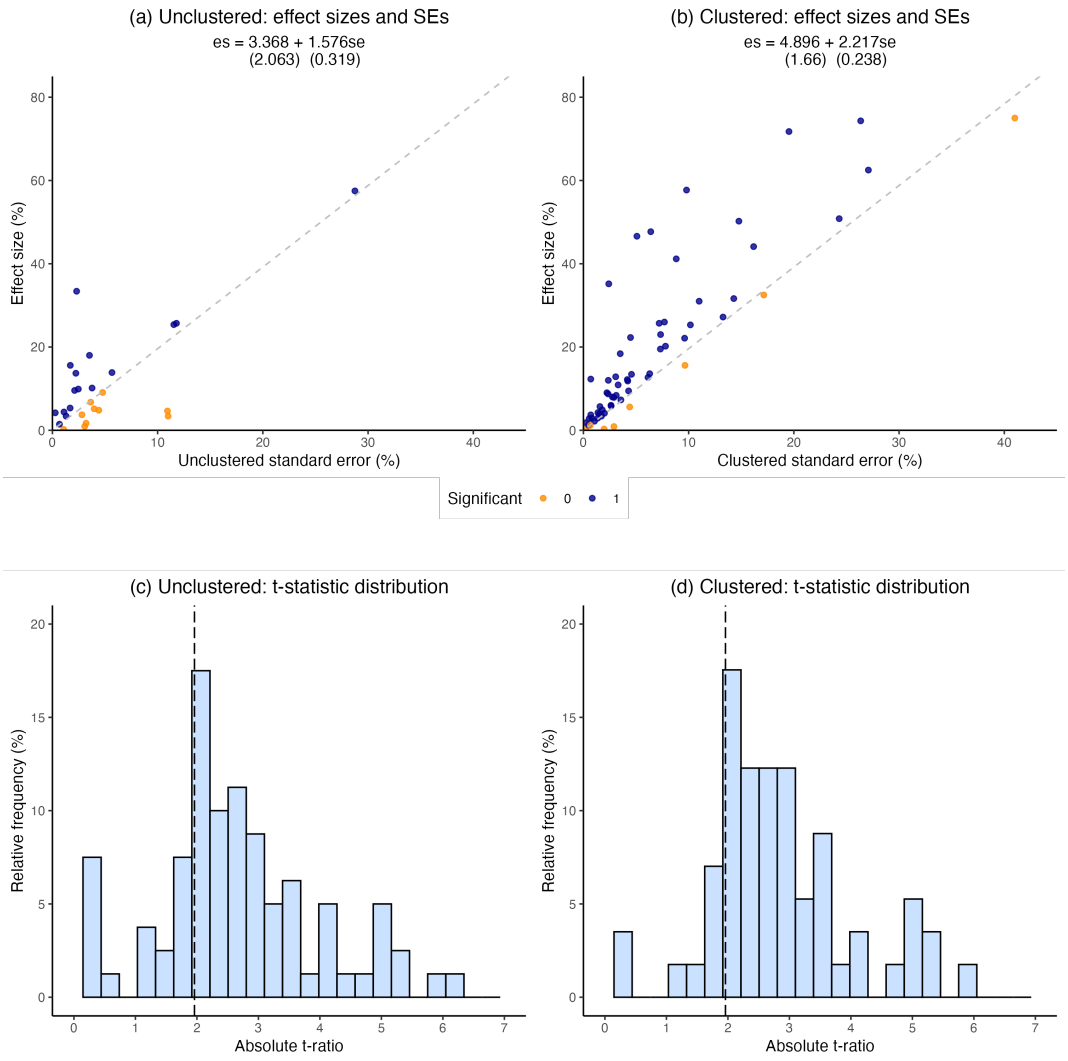

FIGURE 2. SELECTIVE PUBLICATION AND p-HACKING

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.

the sample size in observational studies is not typically chosen but instead predetermined by available datasets.[6] Finally, 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

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

statistically indistinguishable across regimes, and the point estimate in the clustered regime is actually larger than in the unclustered regime.

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. The distributions are also strikingly similar, despite the fact that effect sizes are larger in clustered studies. Thus, it is not only the average t-ratio that is similar across clustered and unclustered studies, as shown earlier in Table 1, but rather the entire distribution.

2.4. Non-Strategic vs. Strategic Clustering

This subsection explores two alternative mechanisms through which the interaction between clustering and publication bias can generate larger published effect sizes. Distinguishing between these two mechanisms is important for understanding the source of the observed effect size gap and will inform the theory in the next section.

The first proposed mechanism is that the choice to cluster is *non-strategic* (exogenous). That is, clustering could be unrelated to whether or not it makes reported estimates more likely to be published. Alternatively, clustering could be *strategic* (endogenous). For example, researchers could strategically choose not to cluster if doing so would overturn a statistically significant result. Both mechanisms generate a positive association between clustering and effect sizes, although each might distort bias and coverage in different ways.

To test which mechanism is driving the results, I examine effect sizes of unclustered studies from the same set of journals in the previous decade, 1990–1999. During this period, the overwhelming majority of studies reported unclustered standard errors (Bertrand et al., 2004) and hence strategic clustering is unlikely to be affecting the distribution of effect sizes. If strategic clustering was absent in the 1990–1999 period, but present during the 2000–2009 period, then, all else equal, we might expect effect sizes for unclustered studies 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 of unclustered studies in the 2000–2009 period is almost exactly the same as in the 1990–1999 period (11.23% and 11.54%). The difference is statistically indistinguishable from zero, although statistical power is somewhat limited. Controlling for differences in observable study characteristics does not change this conclusion. This supports the idea

that strategic clustering of the form discussed here is not driving observed differences in effect sizes across clustered and unclustered regimes. For more details, see Appendix D. Given these results, the theory in the following section assumes non-strategic clustering. Nevertheless, for robustness, results from the empirical model in Section 4 also include an alternative estimation approach that is robust to strategic clustering.[7]

3. Theory

The descriptive statistics in Section 2 document a dramatic increase in reported effect sizes in the DiD literature, in addition to strong evidence for publication bias against null results. While it may seem intuitively reasonable to infer from larger effect sizes that that bias has also increased, the answer is not in fact obvious. This is because larger effect sizes among clustered studies could also reflect a shift toward the publication of studies targeting larger true effects. Moreover, if bias has in fact increased with clustering, then confidence intervals would be shifted away from the true effect, placing downward pressure on coverage. Hence, the implications for coverage conditional on publication are also unclear.

This section addresses these questions in a general theoretical framework and derives the exact conditions under which both bias and coverage increase with clustering. Proofs are in Appendix A. Note that the theoretical framework applies to the DiD setting with clustered standard errors and also, more generally, to any standard error corrections that tend to enlarge reported standard errors (White, 1980; Moulton, 1986; Newey and West, 1987; Staiger and Stock, 1997; Calonico et al., 2014).

3.1. Model of Publication Bias and Standard Error Corrections

The theoretical framework builds on the selective publication model in Andrews and Kasy (2019) to incorporate downwardly biased standard errors. Consider an empirical literature of interest. This could be a literature addressing many different research questions (e.g. the DiD literature) or, alternatively, it could be a meta-analysis focused on a single question (e.g. the impact of job training programs on employment outcomes). From this literature, suppose we observe a sample of published studies, indexed by j. For each study, we observe an estimated treatment effect, standard error, and an indicator for whether or not standard errors are corrected. The model of the data generating process has four steps:

[7]This approach also provides an additional test for the presence of strategic clustering, by comparing robust model estimates to those in the baseline model. Using this approach, I cannot reject the null hypothesis of non-strategic clustering, providing further evidence that this is the relevant channel in the DiD setting. See Subsection 4.1 for further discussion.

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 step 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 $r \in (0, 1)$. Thus, in the uncorrected regime, the reported standard error underestimates the true standard error.

4. **Publication selection:** Selective publication is modeled 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. However, the theoretical results can also be generalized to the case where r is a random variable with support on $(0, 1)$, provided that $r \perp\!\!\!\perp (\hat{\beta}_j, \beta_j, \sigma_j)$. The empirical application in Section 4 also presents results where r is drawn from a distribution. Finally, note that the theoretical results do not assume that β_j and σ_j are independent, and hence extend to settings where sample sizes are chosen based on expected effect sizes in power analyses.

3.2. Illustrative Example

Consider a simple example to illustrate the model. 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. 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 $j = 1, 2, ...J$ (step 1). 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 (step 2). Next, consider two regimes for calculating standard errors (step 3). 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 $\widetilde{\sigma}_j < \sigma_j$). Finally, only a subset of the latent DiD estimates $\hat{\beta}_j$ are published due to publication bias (step 4). In particular, suppose that the publication process censors all insignificant findings at the 5% level.

Table 2 compares bias and coverage conditional on publication across the unclustered and clustered regime. The results highlight 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%).

Higher bias occurs because clustering widens confidence intervals, which exacerbates publication bias. The magnitude of the change (0.81) is large by several benchmarks. First, it is greater than the level of bias in the unclustered regime, and around four-fifths of the true ATT. Alternatively, the change 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. In other words, it is comparable to very severe levels of publication bias.

Higher bias implies that estimates are, on average, further away from the true ATT. Given this, could clustering potentially fail to meet its primary goal of improving the average coverage of published confidence intervals? It turns out that coverage conditional on publication does in fact increase in this case, by almost 20 percentage points. A discussion of why, and whether it holds in more general settings, is deferred to Subsection 3.4

This illustrative example represents only a particular case. In the following subsections, I move beyond this specific case and derive the exact conditions under which the tension between increased bias and coverage generalizes to other settings.

TABLE 2 – Illustrative Example: Impact of Clustering on Bias and Coverage

	Unclustered	Clustered	Difference
Bias	0.64	1.45	0.81
Coverage	0.648	0.844	0.196

3.3. Bias

The main result in this subsection shows that a sufficient condition for increased bias is that corrections are 'sufficiently' large. To begin, I define three key measures of bias. The first 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. Internal-validity bias corresponds to the most common notion of bias in estimators – $\mathbb{E}_r[\hat{\beta}_j | D_j = 1, \beta_j] - \beta_j$ – averaged across published studies. In other words, it asks how far, on average, published estimates are from the questions they answer. The second measure is *study-selection bias*, which is defined as $\mathbb{E}_r[\beta_j | D_j = 1] - \mathbb{E}[\beta_j]$.[8] This measures how far, on average, published true effects are from the average that would occur if there were no publication bias. In specific contexts, this is referred to as 'site-selection bias' (Allcott, 2015).

The relevant measure of bias is context-dependent. For example, consider a slight variant of the example on the impact of a health reform on life expectancy. Suppose that the ATT of the health reform on life expectancy is in fact a weighted average of heterogeneous treatment effects across treated states, and that different studies may examine different subsets of treated states. If a team of researchers is only concerned with accurately evaluating the impact of the policy in the subset of states in their sample, then internal-validity bias is the relevant measure. However, if researchers are instead interested in the nation-wide impact of the program. Then study-selection bias may also be a concern because limiting attention to states where the health reform was particularly effective – that is, where study-selection bias is positive – would lead researchers to overestimate the true average impact of the policy.[9]

Finally, consider *total bias*, which is defined as $\mathbb{E}_r[\hat{\beta}_j | D_j = 1] - \mathbb{E}[\beta_j]$. This measures 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

[8] 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 where variation true treatment effects may arise due to heterogeneity across studies in populations, research design, policies etc.

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

error 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 (or the difference in estimated treatment effects) is equal to the sum of the change in internal-validity bias and study-selection bias. This decomposition is useful for considering the observed difference in effect sizes in the empirical DiD literature discussed in Section 2. It says that the observed increase in effect sizes must reflect either increased internal-validity bias, increased study-selection bias, or perhaps both.

The main result for bias normalizes true treatment effects to be positive and imposes finite moments:

Assumption 1 (Normalization)**.** *Let β_j have support on a subset of the non-negative real line and not be degenerate at zero.*

Assumption 2 (Regularity Conditions)**.** *Let $\mathbb{E}[|\beta_j|]<\infty$, $\mathbb{E}[|\hat{\beta}_j|]<\infty$, and $\mathbb{E}[\sigma^{-1}]<\infty$.*

For empirical literatures examining different questions and outcomes, normalizing true effects to be positive is justified because relative signs across studies are arbitrary.[10] The requirement that β_j not be degenerate at zero is to rule out the boundary case where coverage probabilities always equal zero when all insignificant results are censored by the publication process.

Next, we assume that publication bias exists and is a weakly increasing step function in the absolute value of the reported t-ratio. Intuitively, this means that studies that are 'more significant' have a (weakly) higher probability of being selected for publication.

Assumption 3 (Monotonic Publication Selection Step Function)**.** *Consider a step function with $K-1$ cutoffs: $0=c_0<c_1<c_2,...,c_{K-1}<c_K=\infty$. Let the corresponding publication probabilities satisfy $0\leq\gamma_1\leq\gamma_2\leq...\leq\gamma_{K-1}\leq\gamma_K=1$ and assume there exists at least one $k=1,2,...K-1$ such that $\gamma_k\in[0,1)$. Then the publication selection function takes the following form:*

$$p(\hat{\beta}_j/\widetilde{\sigma}_j)=\sum_{k=1}^{K}\gamma_k\cdot\mathbb{1}\Big\{c_{k-1}<|\hat{\beta}_j|/\widetilde{\sigma}_j<c_k\Big\}$$

This allows for very general forms of publication bias, and in particular makes no restriction on the the number and location of the critical thresholds. That publication bias

[10] This assumption would not be appropriate when analyzing a single question with heterogeneous treatment effects ranging across both negative and positive values.

exists and is weakly decreasing in the t-ratio is a relatively weak assumption given existing empirical evidence (Franco et al., 2014; Brodeur et al., 2016; Andrews and Kasy, 2019; Brodeur et al., 2022). More restrictive is the fact that Assumption 3 imposes symmetry in the publication selection function about zero. For example, assuming symmetry may not be appropriate when examining the impact of the minimum wage on employment, since there are priors about the sign of the effect. However, for the DiD literature, which is the primary focus of this paper, the assumption is more plausible because studies examine different outcomes and hence, the relative signs of effects across studies are arbitrary.

With this, we can state the main result for this subsection:

Proposition 1 (Large Corrections Increase Bias)**.** *Under Assumptions 1–3, 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.*[11]

Proposition 1 states that sufficiently large standard error corrections inevitably lead to increases in each type of bias. This is important for two reasons. First, it implies that corrections are most likely to exaggerate bias in published studies in the cases where they are most necessary. 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.

To understand the 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. This leads to 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.[12] Suppose we are examining a literature addressing two research questions, one with a small true effect ($\beta_1 = 1$) which occurs with probability $\frac{4}{5}$; and one with a large true effect ($\beta_2 = 6$), which

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

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

occurs with probability $\frac{1}{5}$. Assume only one in twenty insignificant studies at the 5% level are published ($\gamma = \frac{1}{20}$) and unclustered standard are 80% of their true value ($r = \frac{4}{5}$).

In this example, clustering leads to an overall increase in estimated treatment effects (0.51) that reflects an increase in true treatment effects (0.54) which outweighs a decrease in internal-validity bias (-0.03). The reason that internal-validity bias decreases is that clustering shifts the distribution of published studies toward those with larger true effects; and larger true effects tend to have smaller internal-validity bias because they are less likely to be insignificant, and hence subject to publication bias.[13]

This example also illustrates that observing higher effect sizes is not sufficient for determining the sign of the change in internal-validity bias. This underscores the limitations of what we can learn from descriptive statistics calculated on observed effect sizes in the DiD setting in Section 2, and motivates estimating the empirical model in Section 4.

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

3.4. Coverage

We turn next to how standard error corrections impact coverage probabilities in the presence of publication bias. This is of particular importance because improved coverage is typically the primary aim of implementing standard error corrections in the first place. However, the possibility of increased bias, as shown in the previous subsection, raises questions about whether this aim is in fact always met in the presence of publication bias.

For the coverage result, we impose additional restrictions on the shape of the selection function. Specifically, we assume that the selection function has a single critical threshold (e.g. the 5% significance threshold):

Assumption 4 (Publication Selection Function)**.** *Let* $p\left(\frac{\hat{\beta}_j}{\sigma_j \cdot r}\right) = 1 - (1-\gamma) \cdot \mathbb{1}\left\{\frac{|\hat{\beta}_j|}{\sigma_j \cdot r} < c\right\}$ *with* $c > 0$*,* $r \in (0, 1]$*, and* $\gamma \in [0, 1)$*.*

This assumption is more restrictive than Assumption 3 and is made for analytical tractability. However, I conjecture that the same conclusions hold for the more general selection function in Assumption 3.

First, define the main object of interest. Let the *expected coverage conditional on publication* in standard error regime $r \in (0, 1]$ be denoted by $\text{Coverage}(r) = \mathbb{P}\mathbb{r}_r[\beta_j \in$

[13]This is shown graphically in Figure C1 in Appendix C. Note also that Proposition 1 guarantees that bias must increase if corrections are sufficiently large. In this example, we have that $r^* = 0.69$, meaning that corrections that enlarge standard errors by more than 45% will lead to an increase in internal-validity bias.

$(\hat{\beta}_j - 1.96 \cdot \sigma_j r, \hat{\beta}_j + 1.96 \cdot \sigma_j r) | D_j = 1]$ (i.e. the probability that published 95% confidence intervals based on reported standard errors contain the true effect).[14] This can be compared 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 absence of publication bias, it is clear that standard error corrections for downward bias will increase coverage. The presence of publication bias, however, introduces several complications. In the definition of Coverage(r), see that the degree of downward bias r affects not only the width of reported confidence intervals, but also the studies $(\beta_j, \sigma_j, \hat{\beta}_j)$ that end up making it into the published literature through the conditioning $D_j = 1$. This is because statistical significance – and therefore publication probabilities – may depend on the reported standard error.

In particular, consider a case where the confidence interval with unclustered standard errors covers the true effect, but not zero, such that the estimate is statistically significant and published. Now suppose that clustering widens the confidence interval to cover zero, so that now the study is censored by publication bias. All else equal, coverage falls. On the other hand, widening confidence intervals can obviously increase coverage, provided that studies are not censored by publication bias.

In general, it is not clear a priori which effect dominates, or if any effect would dominate in all cases. Moreover, allowing for arbitrary distributions of latent true effects, μ_β opens up a large set of possible comparisons, including those which would in principle most favor corrections worsening coverage.

With this, the main result in this subsection states that expected coverage in published studies unambiguously increases:

Proposition 2 (Standard Error Corrections Increase Coverage)**.** *Under Assumptions 1 and 4, 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 with 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 μ_β is degenerate and all null results are censored, $\gamma = 0$ (Lemma A.6). In this special case, there are two relevant cases. In the first, the degenerate value for β is relatively 'large'

[14] 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.

($\beta \geq 2 \times c \cdot r$), so that a study already covering the true effect without clustering is never censored by clustering. The second case deals with relatively 'small' true effects ($\beta < 2 \times c \cdot r$), where increased coverage is shown to be equivalent to demonstrating that the hazard function for normal distribution is increasing. The general result extends this special case to allow 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. For further 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 theoretically 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 shows that a sufficient condition for undercoverage in the uncorrected regime when nominal coverage is 0.95 is* $r < 0.8512$*. Thus, when this condition is met, applying standard error corrections will either decrease the distance to nominal coverage target or achieve coverage that is weakly higher than the nominal target. In the empirical application to the DiD literature in Section 4, the average coverage of published confidence intervals in the uncorrected regime is estimated to be far below nominal coverage of 0.95.*

4. Empirical Model

We turn next to estimating the empirical model in the previous section using the DiD data from Section 2. A primary motivation for estimating the model is the limitations of reduced-form analysis highlighted by the theoretical results. Specifically, Proposition 1 states that the impact of clustering 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. Moreover, whether the magnitude of the change in bias (irrespective of the sign) and coverage is large enough to warrant serious attention is an empirical question.

The empirical strategy consists of three steps. First, estimate the model in Section 3 using data from clustered DiD studies. Second, estimate the degree of bias in unclustered standard errors. The final step combines the first two steps to examine counterfactual scenarios of what would have happened had clustered studies instead reported unclustered standard errors. The final part of this section explores the welfare effects of clustering.

4.1. Estimation

The model is estimated using data from clustered studies. Restricting attention to clustered studies avoids imposing strong assumptions about the mapping between unclustered standard errors and (unobserved) clustered standard errors for unclustered studies in the likelihood function.[15] Following the meta-study approach in Andrews and Kasy (2019), I estimate the latent distribution of true effects assuming that $\beta_j \perp\!\!\!\perp \sigma_j$ and $\beta_j|\lambda_\beta, \kappa_\beta \sim \text{Gamma}(\lambda_\beta, \kappa_\beta)$. Independence is a common, though relatively strong, assumption. It is unlikely to hold in settings where experimental researchers calibrate the sample size according to predicted effect sizes in power analyses, or when target parameters are mechanically correlated with standard errors through measurement (Chen, 2023). However, it may be more likely to hold in the DiD setting because sample sizes are determined primarily by available observational datasets. Following Vu (2024), 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. For the selection function, 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. Finally, note that clustered standard errors are assumed in estimation to reflect the true variation of estimated treatment effects.

Table 3 presents the maximum likelihood estimates. The estimate $\hat{\gamma} = 0.023$ implies a high degree of selective publication. In particular, it means that statistically significant results are around 43 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 ($\hat{\gamma} = 0.017$).

Consistency requires that the choice to cluster is independent of the estimated treatment effect conditional on the true effect: $C_j \perp\!\!\!\perp \hat{\beta}_j|\beta_j$. This assumption is violated if there is strategic clustering to maximize the chances of publication. Subsection 2.4 provides reduced-form evidence that strategic clustering is not taking place. For additional robustness, I also propose an alternative estimation approach that is robust to a particular form of strategic clustering, where researchers choose to cluster if and only if it does not change the significance of their results. The alternative approach focuses exclusively on significant clustered studies, since they are completely invariant to this form of strategic clustering. Robust estimates in Table E1 are statistically indistinguishable from the baseline estimates in Table 3, suggesting that strategic clustering of the form discussed here does not bias baseline parameter estimates. See Appendix E for further details.

[15] This is because for unclustered studies, publication is based on unclustered standard errors while the true variation of the estimated treatment effect is based on the unobserved clustered standard error.

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
0.151	18.202	1.318	7.292	0.023
(0.045)	(6.417)	(0.171)	(1.723)	(0.009)

Notes: Estimation sample is clustered DiD studies over 2000–2009 ($N = 62$). 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, $\gamma = 0.023$ implies that significant results are around 43 times more likely to be published than insignificant results.

Our main aim is to compare bias and coverage in the clustered regime to a counterfactual scenario where clustered studies report unclustered standard errors. The interpretation of this counterfactual comparison is analogous to an ATT measure of the impact of clustering ('treatment') on the statistical properties of published clustered studies.

Model estimates in Table 3 can be used to calculate bias and coverage in the clustered regime. However, to calculate the corresponding figures in the unclustered regime requires an estimate of the degree to which unclustered standard errors are downwardly biased, r.

4.2. Unclustered Standard Errors (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, which end up yielding similar results.

In the first approach, I make the strong, simplifying assumption that all unclustered standard errors are downward biased by a constant factor $r \in (0, 1)$. I then calibrate r using the method of simulated moments (McFadden, 1989). Specifically, I select the value of r which minimizes the distance between moments predicted by the model and the actual moments observed in the data. The moment I choose for calibration is the percent difference in average reported standard errors between clustered and unclustered studies in the published literature. Carrying out this procedure gives $\hat{r} = 0.59$. In other words, clustered standard errors are estimated to be around 1.7 times the size of unclustered standard errors.[16]

[16] 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.

This first 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 so that reported standard errors were 59% the size of true standard errors.

The second approach calculates the empirical distribution of r using a sample of DiD studies from six of the 25 journals examined in Brodeur et al. (2020) over the 2015–2018 period.[17] This approach addresses some of the concerns with the first approach. It does not impose a constant downward bias across studies and does not assume that 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 that it is based on data from a later time period.

Overall, I calculate r in 23 out of 72 DiD studies (31.9%) using non-proprietary data by recalculating both unclustered and clustered standard errors and then taking the ratio. The mean is 0.76, and in 17% of studies, clustered standard errors are more than twice the size of unclustered standard errors ($r < 0.5$). For calculating the counterfactual scenario for unclustered studies, we assume the degree of downward bias is independent and hence can randomly sample from this distribution to determine the degree of bias for each study.

4.3. Impact of Clustering on Coverage and Bias

Table 4 presents the main results. Results are quantitatively similar under both approaches for calibrating r and I therefore focus on the first for discussion. The estimated model shows that clustering increased coverage dramatically, from only 0.36 in the unclustered regime to 0.72 in the clustered regime. This implies severe mismeasurement of standard errors prior to the adoption of clustering, with only a little more than one in three published studies reporting confidence intervals covering the true effect. Note that while coverage improves substantially, it still remains, at 0.72, 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 from equation (2) that the change in total bias can be decomposed into the change

[17] Journals were chosen based on whether they overlapped with the sample in this study, required publication of replication materials, and published a high share of DiD studies. The journals 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.*

TABLE 4 – Impact of Clustering on Coverage and Bias in Published Studies

	Unclustered ($\hat{r} = 0.59$)	Clustered ($r = 1$)	Change
A. Method of Simulated Moments ($\hat{r} = 0.059$)			
Coverage	0.36	0.72	0.36
Total Bias	4.34 (100%)	9.51 (100%)	5.17 (100%)
Internal-Validity Bias	1.47 (33.7%)	2.34 (24.6%)	0.88 (17.0%)
Study-Selection Bias	2.88 (66.3%)	7.17 (75.4%)	4.29 (83.0%)
B. Empirical Distribution of r			
Coverage	0.37	0.72	0.35
Total Bias	4.56 (100%)	9.51 (100%)	4.95 (100%)
Internal-Validity Bias	1.34 (29.5%)	2.34 (24.6%)	1.00 (20.2%)
Study-Selection Bias	3.22 (70.5%)	7.17 (75.4%)	3.95 (79.8%)

Notes: Figures are based on the parameter estimates of the empirical model in Table 3 and calculated by simulating published studies under alternative regimes. Panel A assumes unclustered standard errors are downward biased by a constant factor $\hat{r} = 0.59$. Panel B draws from the empirical distribution of r described in Subsection 4.2. Total bias is defined as $\mathbb{E}_r[\hat{\beta}_j|D_j = 1] - \mathbb{E}_r[\beta_j]$; internal-validity bias as $\mathbb{E}_r[\hat{\beta}_j - \beta_j|D_j = 1]$; and study-selection bias as $\mathbb{E}_r[\beta_j|D_j = 1] - \mathbb{E}_r[\beta_j]$.

in internal-validity bias and study-selection bias. In this context, the primary measure of interest is internal-validity bias ($\mathbb{E}_r[\hat{\beta}_j - \beta_j|D_j = 1]$). 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.

Model results show that clustering led to internal-validity bias increasing by around 60%, from 1.47 ppts to 2.34 ppts. To gauge the size of this change, we can benchmark it against the fraction of insignificant results (with correctly measured standard errors) tjat would need to be censored by publication bias to observe the same increase bias (0.88 ppts). The answer is that 78% of null results would need to be censored (i.e. $\gamma = 0.22$). In other words, the increase in internal-validity bias from clustering is comparable to fairly 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. 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 5.17 ppts overall.

Proposition 1 states that bias must increase for sufficiently large standard error corrections (i.e. for any r less than some model-dependent value r^*). For the estimated model on DiD studies, I find that $r^* = 1$. In other words, *any* correction leading to larger standard errors would lead to an increase in bias in published DiD studies. Thus, the qualitative conclusion of higher bias and coverage would hold for any sized correction (since Proposition

2 guarantees increased coverage). Figure F1 displays the impact of clustering on bias and coverage for r over the unit interval, and shows that larger corrections lead to larger changes in both bias and coverage.

4.4. Welfare Effects of Clustering

Results in Table 4 show that improved credibility of standard errors from clustering comes at the unintended cost of declining credibility in point estimates. The ultimate impact of this on welfare, however, remains unclear. To address this question, this subsection estimates the impact of clustering on welfare in the DiD context, using the Bayesian decision-theoretic framework in Frankel and Kasy (2022).

The decision-theoretic model considers the action of an audience, which could represent the public, practitioners, policymakers, or the scientific community. It can be characterized as an extension of the publication model in Section 3 to include two additional steps.

In the first added step, the audience calculates posterior beliefs based on whether or not a study is published. If no study is published, then the audience maintains their prior belief.[18] The prior, $\mu_{\beta,0}$, is assumed to be equal to the latent distribution of true effects.[19] Alternatively, if the study is published, then the audience updates their prior belief based on the published evidence. Specifically, following Frankel and Kasy (2022), the audience obtains posterior beliefs, $\mu_{\beta,1}$, via Bayesian updating assuming that the estimate is normally distributed. In this paper, I assume that the audience updates beliefs based on the *reported standard error* ($\widetilde{\sigma}$), irrespective of whether it is correctly calculated or not. In practical terms, this means that the audience takes published estimates and standard errors at face value, and does not make sophisticated statistical adjustments for potential biases.

In the second step of the decision-theoretic model, the audience makes an action to maximize expected utility given their posterior beliefs. I assume that the audience has a quadratic loss utility function, so that the chosen action is determined by $a^*(\mu_{\beta,1}) = \arg\max_a \mathbb{E}_{\beta\sim\mu_{\beta,1}}\big[-(a-\beta)^2\big]$, and the utility payoff is equal to $-(a^*(\mu_{\beta,1})-\beta)^2$.

In Frankel and Kasy (2022), welfare is defined as utility net of 'publication costs'. In their framework, publication costs correspond to the opportunity cost of the public's attention arising from limitations in information processing. This paper abstracts away from

[18] This corresponds to 'naive' updating in the Frankel and Kasy (2022) model because a sophisticated audience would understand that not observing a study might be due to the fact that it was censored by publication bias. This is described in their paper as a more realistic description of updating in many settings.

[19] An alternative approach which does not use priors is to assume decision-makers aim to minimize maximum regret. This alternative modeling approach is developed in Appendix G and delivers qualitatively similar results to the Bayesian decision-making model when policymakers exhibit sufficiently high loss aversion for Type I error – that is, for mistakenly implementing an ineffective or harmful policy.

publication costs to avoid making judgments about their magnitude and treats utility differences between the clustered and unclustered regimes as equivalent to welfare differences. Alternatively, if one insists on including publication costs in welfare, then the utility difference can be interpreted as a lower bound on the welfare difference between the clustered and unclustered regime. This is because publication costs in the unclustered regime must be weakly higher than in the clustered regime, as clustering can only decrease the chance of publication.

To implement the model in the DiD setting, I use the estimated model parameters in Table 3 and assume unclustered standard errors are downward biased based on $\hat{r} = 0.59$ to simulate $M = 10^6$ studies across both regimes, $\{\beta_m, \sigma_m, \hat{\beta}_r, D_m^{C=1}, D_m^{C=0}\}_{m=1}^{M}$, where $D_r^{C=c}$ is an indicator for whether the study is published under clustering regime $c = 0, 1$. Under quadratic loss, the optimal Bayesian action corresponds to the posterior mean. Thus, for each simulated study m in standard error regime c, the optimal action is given by

$$a_m^{*,c} = \mathbb{1}\{D_m^{C=c} = 1\} \cdot \bar{\beta}_m^{(\hat{\beta}_m, \widetilde{\sigma_m})} + \mathbb{1}\{D_m^{C=c} = 0\} \cdot \bar{\beta}_r^{\,0}$$

where the mean of the prior is $\bar{\beta}_r^{\,0}$; and the posterior mean after having observed a published study is $\bar{\beta}_r^{(\hat{\beta}_r, \widetilde{\sigma_r})}$ and calculated using MCMC simulation. Averaging across studies, expected welfare in standard error regime $C = c$ is given by $\mathbb{E}[U(a^*, \beta) \mid C = c] \approx -\frac{1}{M} \sum_{m=1}^{M} (a_m^{*,c} - \beta_m)^2$.

Table 5 presents the results and shows that expected utility in the clustered regime (-18.56) is 36.5% higher than in the unclustered regime (-29.25). This is a substantial gain when compared against the 13.25% expected utility difference between the utility-maximizing scenario with clustered standard errors and no publication bias (-16.10) and the clustered regime with publication bias (-18.56). In other words, the model finds that the audience's welfare gain from implementing clustered standard errors exceeds the gain that would occur from eliminating publication bias, provided that standard errors are correctly calculated.

To understand what is driving this result, the bottom panel decomposes the welfare difference which, for convenience, we can write as $\Delta\mathbb{E}[U(a^*, \beta)] \equiv \mathbb{E}[U(a, \beta) \mid C = 1] - \mathbb{E}[U(a, \beta) \mid C = 0]$. The bottom panel shows that the welfare difference between the clustered and the unclustered regime is equal to the weighted sum of three different 'types' of studies, which we can consider in turn.

First, consider studies whose estimates are published and statistically significant irrespective of the standard error regime: $|\hat{\beta}| > 1.96\sigma$. For these studies, clustered standard errors provide more accurate updating from published evidence, specifically, by correctly weighting the observed estimate and the prior when calculating the posterior mean. By contrast, in the

TABLE 5 – Expected Welfare: Clustered vs. Unclustered

	Expected Utility	
Clustered regime ($\gamma = 0.023$)	-18.56	
Unclustered regime ($\gamma = 0.023$)	-29.25	
	Difference	Share
$\Delta\mathbb{E}[U(a,\beta)]$	10.69	(100%)
(i) $\Delta\mathbb{E}[U(a,\beta) \mid \lvert\hat{\beta}\rvert > 1.96\sigma] \cdot \mathbb{P}[\lvert\hat{\beta}\rvert > 1.96\sigma]$	8.08	(75.6%)
(ii) $\Delta\mathbb{E}[U(a,\beta) \mid 1.96\widetilde{\sigma} < \lvert\hat{\beta}\rvert < 1.96\sigma] \cdot \mathbb{P}[\widetilde{\sigma} < \lvert\hat{\beta}\rvert < 1.96\sigma]$	2.60	(24.3%)
(iii) $\Delta\mathbb{E}[U(a,\beta) \mid \lvert\hat{\beta}\rvert \leq 1.96\widetilde{\sigma}] \cdot \mathbb{P}[\lvert\hat{\beta}\rvert \leq 1.96\widetilde{\sigma}]$	0.01	(0.1%)

Notes: In the top panel, expected utility estimates are based the decision model outlined in Section 4.4 and the model estimates in Table 3. The second panel decompose the difference in expected utility between clustered and unclustered regimes. Results are based on 10^6 simulation draws

unclustered regime, downward biased standard errors cause the audience to overestimate the precision of the parameter estimate, and consequently place too much weight on the published estimate. Results show that over-updating in always-published studies accounts for around three-quarters of the average welfare difference.

The second type of study is defined by estimates that are statistically significant in the unclustered regime, but statistically insignificant in the clustered regime: $1.96\widetilde{\sigma} < |\hat{\beta}| < 1.96\sigma$. In this case, studies in the unclustered regime are more likely to be published, allowing the audience to update their beliefs about the target parameter β. However, as before, unclustered standard errors can lead to over-updating. The overall impact depends on the net outcome of these two effects. Table 5 shows that the welfare cost of over-updating outweighs the benefit of observing more studies. Hence, for this set of studies, observing evidence with incorrectly calculated standard errors is, on average, *worse* for decision-making than simply relying on prior beliefs.

The third type of study contains estimates which are statistically insignificant irrespective of whether or not standard errors are clustered: $|\hat{\beta}| \leq 1.96$. This makes a negligible contribution to the average welfare difference because most studies are censored by publication bias regardless of the regime, so that the audience relies on the prior in either case.

Overall, the results show that average welfare in the clustered regime is substantially larger than in the unclustered regime. Including publication costs would only expand this difference. The main mechanism driving the welfare difference is that downward biased standard errors in the unclustered regime lead to over-updating, which results in suboptimal decisions.

5. 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 statistical credibility of published studies and welfare in a decision-theoretic framework.

A central tension highlighted in the theory is that standard error corrections increase coverage but can also, unintendedly, worsen bias. Empirically, this tension is present in the DiD literature, where the adoption of clustering led to large improvements in coverage but also sizable increases in the bias of estimated treatment effects. Nevertheless, incorporating this trade-off in a decision-theoretic model shows that clustering substantially improves welfare due to more accurate updating.

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, *138* (1), 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.

Calonico, Sebastian, Matias D. Cattaneo, and Rocio Titiunik, "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," *Econometrica*, 2014, *82* (6), 2295–2326.

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.

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.

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.

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 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.

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.

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?," *Journal of Econometrics*, 2024, *245* (1-2).

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

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. To start, however, I present Lemma A.1, which is used in Lemma A.2.

Lemma A.1 (Bias Conditional on Publication)**.** *Fix $\beta \in [0, \infty)$ and $r \in (0, 1]$. Under Assumption 3, internal-validity bias is given by*

$$Bias(\beta, \sigma, \boldsymbol{\gamma}, r) = \frac{\sum_{k=1}^{K-1}(1 - \gamma_k)[d_k - d_{k-1}]}{1 - \sum_{k=1}^{K-1}(1 - \gamma_k)[q_k - q_{k-1}]} \geq 0 \tag{3}$$

$$d_k \equiv \sigma\left[\phi\left(c_k r - \frac{\beta}{\sigma}\right) - \phi\left(-c_k r - \frac{\beta}{\sigma}\right)\right]$$

$$q_k \equiv \Phi\left(c_k r - \frac{\beta}{\sigma}\right) - \Phi\left(-c_k r - \frac{\beta}{\sigma}\right)$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the normal pdf and cdf, respectively, and the inequality is strict when $\beta > 0$.

Proof. We first derive the expression for bias.

$$\begin{aligned}
\text{Bias}(\beta,\sigma,\boldsymbol{\gamma},r) =& \int \hat{\beta} f_{\hat{\beta}|D,\beta,\sigma}(\hat{\beta}|D_j=1,\beta,\sigma;\boldsymbol{\gamma},r)d\hat{\beta} - \beta \\
=& \int \left(\frac{\hat{\beta}\cdot \sum_{k=1}^{K} \gamma_k \mathbb{1}\{c_{k-1}\sigma r < |\hat{\beta}| < c_k \sigma r\} \frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right)}{\sum_{k=1}^{K}\gamma_k \mathbb{P}\mathrm{r}_r\left[c_{k-1}\sigma r < |\hat{\beta}| < c_k \sigma r\right]} \right) d\hat{\beta} - \beta \\
=& \frac{\sum_{k=1}^{K}\gamma_k \overbrace{\mathbb{E}\left[\hat{\beta}\mathbb{1}\{c_{k-1}\sigma r < |\hat{\beta}| < c_k\sigma r\}\right]}^{\equiv m_k}}{\sum_{k=1}^{K}\gamma_k \underbrace{\mathbb{P}\mathrm{r}_r\left[c_{k-1}\sigma r < |\hat{\beta}| < c_k \sigma r\right]}_{\equiv \pi_k}} - \beta \\
=& \frac{\sum_{k=1}^{K-1}(m_k + (\gamma_k - 1)m_k) + m_K}{\sum_{k=1}^{K-1}(\pi_k + (\gamma_k-1)\pi_k) + \pi_K} - \beta \\
=& \frac{\beta - \sum_{k=1}^{K-1}(1-\gamma_k)m_k}{1 - \sum_{k=1}^{K-1}(1-\gamma_k)\pi_k} - \beta \\
=& \frac{\sum_{k=1}^{K-1}(1-\gamma_k)(\beta\pi_k - m_k)}{1-\sum_{k=1}^{K-1}(1-\gamma_k)\pi_k}
\end{aligned}$$

where the second equality uses Bayes Rule and the expression for the selection function under Assumption 3; and the second last equality uses the fact that $\sum_{k=1}^{K} m_k = \beta$ and $\sum_{k=1}^{K}\pi_k = 1$.

It can be straightforwardly verified from this expression that $\pi_k = q_k - q_{k-1}$. Thus, it remains to show that $\beta\pi_k - m_k = d_k - d_{k-1}$.

$$\begin{aligned}
\beta\pi_k - m_k =& \int_{c_{k-1}\sigma r < |\hat{\beta}| < c_k \sigma r} (\beta - \hat{\beta}) \frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right) d\hat{\beta} \\
=& -\sigma\left[\int_{c_{k-1}\sigma r}^{c_k \sigma r} \left(\frac{\hat{\beta}-\beta}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right)d\hat{\beta} + \int_{-c_k\sigma r}^{-c_{k-1}\sigma r} \left(\frac{\hat{\beta}-\beta}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right)d\hat{\beta}\right] \\
=& -\sigma\left[\int_{c_{k-1}r - \frac{\beta}{\sigma}}^{c_k r - \frac{\beta}{\sigma}} z\phi(z)dz + \int_{-c_k r - \frac{\beta}{\sigma}}^{-c_{k-1}r - \frac{\beta}{\sigma}} z\phi(z)dz \right] \\
=& \sigma\left[\phi\left(c_k r - \frac{\beta}{\sigma}\right) - \phi\left(c_{k-1}r - \frac{\beta}{\sigma}\right) + \phi\left(-c_{k-1}r - \frac{\beta}{\sigma}\right) - \phi\left(-c_k r - \frac{\beta}{\sigma}\right)\right] \\
=& d_k - d_{k-1}
\end{aligned}$$

where the third line uses a change in variables, and the fourth line uses the fact that for $Z \sim N(0,1)$ and $a < b$, we have $\mathbb{E}[Z|Z\in(a,b)]\cdot[\Phi(b)-\Phi(a)] = -[\phi(b)-\phi(a)]$.

Next, I show that $\text{Bias}(\beta,\sigma,\boldsymbol{\gamma},r) \geq 0$. First, see that the denominator is equal to the

normalization constant, $\sum_{k=1}^{K}\gamma_k\pi_k$, which is a weighted sum of probabilities and therefore strictly positive. The sign of bias is therefore determined by the numerator, which can be rewritten as follows

$$\begin{aligned}
&\sum_{k=1}^{K-1}(1-\gamma_k)[d_k - d_{k-1}] \\
=&\sum_{k=1}^{K-1}(1-\gamma_k)d_k - \sum_{k=0}^{K-2}(1-\gamma_{k+1})d_k \\
=&\sum_{k=1}^{K-1}(\gamma_{k+1}-\gamma_k)d_k
\end{aligned}$$

where the last line uses the fact that $d_0 = 0$ and $\gamma_K = 1$. First, see that $\beta = 0$ implies $d_k = 0$ for all k, which in turn implies $\text{Bias}(\beta, \sigma, \boldsymbol{\gamma}, r) = 0$.

In the case where $\beta > 0$, we have that $\text{Bias}(\beta, \sigma, \boldsymbol{\gamma}, r) > 0$, which follows from two facts. First, Assumption 3 implies $\gamma_{k+1} - \gamma_k \geq 0$, with strict inequality for at least one k. Second, $d_k > 0$ for all k because $\phi(\cdot)$ is strictly decreasing over $(0, \infty)$ and symmetric, which implies $\phi\Big(c_k r - \frac{\beta}{\sigma}\Big) > \phi\Big(-c_k r - \frac{\beta}{\sigma}\Big) = \phi\Big(c_k r + \frac{\beta}{\sigma}\Big)$. □

Lemma A.2 (Sufficient Condition for Increase in Internal-Validity Bias)**.** *Under Assumptions 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

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

$$= \int \left(\hat{\beta} \cdot \frac{\int_{\beta,\sigma} p\Big(\frac{\hat{\beta}}{\sigma \cdot r}\Big) \frac{1}{\sigma}\phi\Big(\frac{\hat{\beta}-\beta}{\sigma}\Big) f_{\beta,\sigma}(\beta, \sigma) d\beta d\sigma}{\int_{\beta,\sigma} \mathbb{P}\mathbb{r}_r[D_j = 1|\beta, \sigma] f_{\beta,\sigma}(\beta, \sigma) d\beta d\sigma}\right) d\hat{\beta} \tag{4}$$

Note in the second equality that the density $f_{\hat{\beta}}(\cdot)$ does not depend on either $\boldsymbol{\gamma}$ or r. To evaluate the limit we apply the dominated convergence theorem (DCT) separately to the numerator and the denominator. First, in the numerator, see that for any fixed $\hat{\beta}$, the integrand converges pointwise to $\int_{\beta,\sigma} \frac{1}{\sigma}\phi\Big(\frac{\hat{\beta}-\beta}{\sigma}\Big) f_{\beta,\sigma}(\beta, \sigma) d\beta d\sigma$ as $r \to 0$ (since all results are published in the limit). The integrand is bounded by $\frac{1}{\sigma}\phi\Big(\frac{\hat{\beta}-\beta}{\sigma}\Big) f_{\beta,\sigma}(\beta, \sigma)$, which is integrable under Assumption 2. Thus, by the DCT, the numerator converges to the unconditional density of $\hat{\beta}$.

Next, see that the denominator satisfies $\lim_{r\to 0}\int_{\beta,\sigma}\mathbb{Pr}_r[D_j=1|\beta,\sigma]f_{\beta,\sigma}(\beta,\sigma)d\beta d\sigma=1$. This follows by applying the DCT to pass the limit inside the integral. For each fixed (β,σ), the probability of publication satisfies $\mathbb{Pr}_r[D_j=1|\beta,\sigma]\to 1$ as $r\to 0$, so the integrand converges pointwise to $f_{\beta,\sigma}(\beta,\sigma)$, which also serves as the upper bound because $\mathbb{Pr}_r[D_j=1|\beta,\sigma]\leq 1$.

Combining these two results, and using the law of iterated expectations, we have

$$\lim_{r\to 0}\mathbb{E}_r[\hat{\beta}_j|D_j=1]=\int\hat{\beta}f_{\hat{\beta}}(\hat{\beta})d\hat{\beta}=\mathbb{E}[\hat{\beta}_j]=\mathbb{E}[\mathbb{E}[\hat{\beta}_j|\beta_j]]=\mathbb{E}[\beta_j] \tag{5}$$

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,

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

where the latent distribution of true effects, $f_\beta(\beta)$, does not depend on either $\boldsymbol{\gamma}$ or r. From earlier, we know that the denominator converges to one. Expanding the numerator gives

$$\int_{\sigma,\hat{\beta}}\beta\cdot p\left(\frac{\hat{\beta}}{\sigma\cdot r}\right)\frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right)f_{\beta,\sigma}(\beta,\sigma)d\sigma d\hat{\beta}$$

See that the integrand converges pointwise to $\beta\cdot\frac{1}{\sigma}\phi\left(\frac{\hat{\beta}-\beta}{\sigma}\right)f_{\beta,\sigma}(\beta,\sigma)$ and is dominated by the same function, which is integrable under Assumption 2. Hence, by the DCT, and integrating out $\hat{\beta}$, we have

$$\lim_{r\to 0}\mathbb{E}_r[\beta_j|D_j=1]=\int_{\beta,\sigma}\beta f_{\beta,\sigma}(\beta,\sigma)d\beta d\sigma=\mathbb{E}[\beta_j] \tag{6}$$

Using the convergence in mean results in equations (5) and (6) and the linearity of expectations, it follows that

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

$$\to\mathbb{E}_1[\hat{\beta}_j-\beta_j|D_j=1]=\int_{\beta,\sigma}\text{Bias}(\beta,\sigma,\boldsymbol{\gamma},1)f_{\beta,\sigma}(\beta,\sigma)d\beta d\sigma>0 \tag{7}$$

as $r\to 0$. The final inequality follows because Lemma A.1 shows that $\text{Bias}(\beta,\sigma,\boldsymbol{\gamma},1)\geq 0$ when the publication selection function satisfies Assumption 3 and $\beta\geq 0$, and with strict inequality when $\beta>0$. Assumption 2 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 (7). 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 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. First, we show that $\Delta\text{SSB}(r) \equiv \mathbb{E}_1[\beta_j | D_j = 1] - \mathbb{E}_r[\beta_j | D_j = 1] \geq 0$. 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

$$\begin{aligned}
\lim_{r \to 0} \Delta\text{SSB}(r) &= \mathbb{E}_1[\beta_j | D_j = 1] - \lim_{r \to 0} \mathbb{E}_r[\beta_j | D_j = 1] \\
&= \mathbb{E}_1[\beta_j | D_j = 1] - \mathbb{E}[\beta_j] \\
&= \int_0^\infty [1 - F_{\beta|D}(t | D_j = 1; \boldsymbol{\gamma}, 1)] dt - \int_0^\infty [1 - F_\beta(t)] dt \\
&= \int_0^\infty [F_\beta(t) - F_{\beta|D}(t | D_j = 1; \boldsymbol{\gamma}, 1)] dt \qquad (8)
\end{aligned}$$

The second equality uses the convergence in expectation result in equation (6) 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 (8) is positive if the distribution of published true treatment effects in the corrected regime, $F_{\beta|D}(\cdot | D_j = 1; \boldsymbol{\gamma}, 1)$, first-order stochastically dominates the latent distribution of true treatment effects $F_\beta(\cdot)$. To show this holds, fix $t \in [0, \infty)$ and see that

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

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 $\int^t g(\beta)f_\beta(\beta)d\beta = \mathbb{E}_\beta[g(\beta)|\beta \leq t]\cdot F_\beta(t)$. Finally, the last inequality follows from the fact that $\mathbb{P}\mathbb{r}_1(D_j=1|\beta)$ is an increasing function of β, which implies that the unconditional expectation must be larger than the conditional expectation.[20] Since β_j is non-degenerate, there exists some $t \in [0,\infty)$ for which this inequality is strict. This implies that (8) 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: Without loss of generality, and with a slight abuse of notation, let $f_\beta(\cdot)$ denote the distribution of $|\beta_j|$. Note that lemmas in the proof implicitly impose Assumption 4.

[20] The derivative can be derived using very similar arguments to Lemma A.1 and is given by:

$$\frac{\partial}{\partial\beta}\Big[\mathbb{P}\mathbb{r}_r(D_j=1|\beta)\Big] = \int_\sigma \sum_{k=1}^{K-1}(\gamma_{k+1}-\gamma_k)\cdot\frac{1}{\sigma}\left[\phi\left(c_k\cdot r - \frac{\beta}{\sigma}\right) - \phi\left(-c_k\cdot r - \frac{\beta}{\sigma}\right)\right]f_\sigma(\sigma)d\sigma$$

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

As a starting point, the following Lemma provides an expression for expected coverage in published studies for a fixed true effect.

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*

$$Coverage(\beta, r) = \begin{cases} \frac{\gamma[\Phi(c\cdot r-\beta)-\Phi(-c\cdot r)]+\Phi(c\cdot r)-\Phi(c\cdot r-\beta)}{\Phi(-c\cdot r-\beta)+1-\Phi(c\cdot r-\beta)+\gamma[\Phi(c\cdot r-\beta)-\Phi(-c\cdot r-\beta)]} & \text{if } \beta \leq 2 \times c \cdot r \\ \frac{\Phi(c\cdot r)-\Phi(-c\cdot r)}{\Phi(-c\cdot r-\beta)+1-\Phi(c\cdot r-\beta)+\gamma[\Phi(c\cdot r-\beta)-\Phi(-c\cdot r-\beta)]} & \text{if } \beta > 2 \times c \cdot r \end{cases} \tag{9}$$

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

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

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 4). Evaluating the integral in the numerator and expanding the denominator gives equation (9). □

The publication regime is uniquely characterized by $\gamma \in [0, 1)$, the relative probability of publishing insignificant results (Assumption 4). 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

$$\omega(r) = \frac{\gamma}{\mathbb{P}\mathrm{r}_r(D_j = 1)} \in [0, 1] \tag{10}$$

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

bility). The claim is trivially true in the case where $\gamma = 0$ or $\gamma = 1$. Let $\gamma \in (0, 1)$. With Bayes Rule and Assumption 4 which assumes a step-wise publication selection function, we have that

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

$$= \frac{\mathbb{1}\{|\hat{\beta}| \geq c \cdot r\}\phi(\hat{\beta} - \beta)f_\beta(\beta) + \gamma\mathbb{1}\{|\hat{\beta}| < c \cdot r\}\phi(\hat{\beta} - \beta)f_\beta(\beta)}{\mathbb{Pr}(D_j = 1; \gamma, r)} \tag{11}$$

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 (10); 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 (11). Substituting the weights and densities in the mixture gives

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

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

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

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

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

It is clear that this expression equals the density in the publication regime $\gamma \in (0, 1)$ in equation (11) provided that $\kappa = 1$. This can be verified by substituting the following identify into the first term of the numerator:

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

□

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

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

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: $g_\gamma(r) = \omega(r)g_1(r) + \big(1 - \omega(r)\big)g_0(r)$. Hence, the change in expected coverage from standard error corrections in publication regime γ is equal to

$$\begin{aligned} g_\gamma(1) - g_\gamma(r) &= \big[\omega(1)g_1(1) + \big(1 - \omega(1)\big)g_0(1)\big] - \big[\omega(r)g_1(r) + \big(1 - \omega(r)\big)g_0(r)\big] \\ &= \Big(1 - \omega(r)\Big)\Big(g_0(1) - g_0(r)\Big) + \omega(1)\Big(g_1(1) - g_0(1)\Big) - \omega(r)\Big(g_1(r) - g_0(1)\Big) \\ &> \Big(1 - \omega(r)\Big)\Big(g_0(1) - g_0(r)\Big) \end{aligned}$$

where the inequality uses the fact that $g_1(1) - g_1(r) = [\Phi(c) - \Phi(-c)] - [\Phi(c \cdot r) - \Phi(-c \cdot r)] > 0$, and $\omega(1) > \omega(r)$ because the probability of publication in the denominator for the weight in equation (10) 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:

$$\begin{aligned} & g_0(1) - g_0(r) \\ = & \int_0^\infty \text{Coverage}(\beta, 1) f_{\beta|D}(\beta|D_j = 1; 0, 1) d\beta - \int_0^\infty \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, r) d\beta \\ = & \left(\int_0^{2 \times c \cdot r} \text{Coverage}(\beta, 1) f_{\beta|D}(\beta|D_j = 1; 0, 1) d\beta - \int_0^{2 \times c \cdot r} \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, r) d\beta \right) \\ + & \left(\int_{2 \times c \cdot r}^\infty \text{Coverage}(\beta, 1) f_{\beta|D}(\beta|D_j = 1; 0, 1) d\beta - \int_{2 \times c \cdot r}^\infty \text{Coverage}(\beta, r) f_{\beta|D}(\beta|D_j = 1; 0, r) d\beta \right) \end{aligned} \tag{12}$$

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

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

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, the integrand must therefore be weakly increasing in r (and strictly increasing when $f_\beta(\beta) > 0$). Thus, in equation (12), 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 (12) 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)**.** *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. Let the second argument in the Coverage$(\cdot, \cdot)$ be redefined as the critical threshold $t \equiv c \cdot r$ rather than the reported standard error r. Showing the $\frac{\partial}{\partial t}\Big(\text{Coverage}(\beta, t)\Big) > 0$ is clearly equivalent to the main claim.

Consider two cases for the value of β. First, suppose that $\beta \geq 2t$. This case has already been shown in the main text of the proof for the more general scenario where β follows a distribution. In the second case, suppose that $\beta \in (0, 2t)$. The expression for coverage when $\gamma = 0$ (Lemma A.4) is given by

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

Taking the derivative with respect to t gives

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

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

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

$$\frac{\phi(t+\beta)+\phi(t-\beta)}{1-\Phi(t+\beta)+1-\Phi(t-\beta)} \geq \frac{\phi(t-\beta)-\phi(t)}{\Phi(t)-\Phi(t-\beta)} \tag{13}$$

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

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

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

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

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

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

Since $\omega \in (0,1)$, we have that $\omega\mu_1 + (1-\omega)\mu_2 \geq \mu_3$, which gives the desired inequality in (13). □

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

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

Proof. Let the second argument in the Coverage$(\cdot,\cdot)$ be redefined as the critical threshold $t \equiv c\cdot r$ rather than the reported standard error r.

Consider two cases. First, suppose that $\beta \leq 2t$. Using the quotient rule on the expression for coverage in Lemma A.4 gives

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

where we define the denominator as $d(\beta,t) \equiv \Phi(-t-\beta)+1-\Phi(t-\beta)+\gamma[\Phi(t-\beta)-\Phi(-t-$

$\beta)] > 0$ and the numerator as $n_1(\beta, t) \equiv \gamma[\Phi(t-\beta) - \Phi(-t)] + \Phi(t) - \Phi(t-\beta) > 0$. The inequality follows because $d(\beta, t) > n_1(\beta, t)$ and $\phi(t-\beta) > \phi(t-\beta) - \phi(t+\beta) > 0$.

Next, suppose that $\beta > 2t$. Define the numerator as $n_2(\beta, t) \equiv \Phi(t) - \Phi(-t) > 0$. Then

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

□

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, namely, that $f_{\beta|D}(\beta|D_j = 1; 0, 1)/f_{\beta|D}(\beta|D_j = 1; 0, r)$ is increasing in β. By Bayes Rule we have

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

where $K \equiv \mathbb{Pr}[D_j = 1; 0, r]/\mathbb{Pr}[D_j = 1; 0, 1] > 0$ does not depend on β. Thus the derivative with respect to β is given by

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

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

$$\frac{\phi(c-\beta) - \phi(c+\beta)}{1 - \Phi(c-\beta) + 1 - \Phi(c+\beta)} \geq \frac{\phi(c\cdot r-\beta) - \phi(c\cdot r+\beta)}{1 - \Phi(c\cdot r-\beta) + 1 - \Phi(c\cdot r+\beta)} \tag{14}$$

Thus, it suffices to show that

$$g(t) \equiv \frac{\phi(t-\beta) - \phi(t+\beta)}{1 - \Phi(t-\beta) + 1 - \Phi(t+\beta)}$$

is increasing in t. To show this, first write $g(t) = \mu_1(t) \cdot \mu_2(t)$, where

$$\mu_1(t) \equiv \frac{\phi(t-\beta) - \phi(t+\beta)}{\Phi(t+\beta) - \Phi(t-\beta)}$$

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

The derivative using the product rule gives

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

Showing that all four terms in this expression are positive is sufficient for proving the derivative is positive. First, see that $\mu_1(t) = \mathbb{E}[Z|Z \in (t-\beta, t+\beta)]$ with $Z \sim N(0,1)$, based on the formula for the expectation of a truncated normal. This implies $\mu_1(t) > 0$ for $t, \beta > 0$; and also that $\partial\mu_1(t)/\partial t > 0$, since the conditional expectation must increase when the fixed-width interval over which the expectation is taken shifts to the right.

Next, note that $\mu_2(\beta, c)$ is clearly positive. Finally, using the quotient rule gives

$$\frac{\partial}{\partial t}\Big[\mu_2(t)\Big] \propto \frac{\partial}{\partial t}\Big[n(t)\Big] \cdot d(t) - n(\beta, c) \cdot \frac{\partial}{\partial c}\Big[d(\beta, c)\Big]$$

$$= \Big[\phi(t+\beta) - \phi(t-\beta)\Big] \cdot d(t) + n(t) \cdot \Big[\phi(c+\beta) + \phi(c-\beta)\Big]$$

where $n(t) \equiv \Phi(t+\beta) - \Phi(t-\beta)$ denotes the numerator and $d(t) \equiv 1 - \Phi(t-\beta) + 1 - \Phi(t+\beta)$ the denominator. This derivative is positive if and only if

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

This inequality holds because the hazard function of the normal distribution is increasing and $t + \beta \geq t - \beta$ 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

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

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

The first inequality uses Lemma A.6 to replace Coverage$(\beta, 1)$ with Coverage(β, r) in the first term. The final inequality follows from the fact that Coverage(β, r) is strictly increasing in β over $(0, 2 \times c \cdot r)$ (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 $(0, 2 \times c \cdot r)$ 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 (12) are weakly positive and that at least one is strictly positive, completing the proof. □

Lemma A.9 (Sufficient Condition for Undercoverage in Uncorrected Regime)**.** *If nominal coverage equals 0.95 and* $r < 0.8512$*, then Coverage(r)*< 0.95*.*

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

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

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

$$\leq \frac{\Phi(1.96r)-\Phi(-1.96r)}{\Phi(-3\times 1.96r)+1-\Phi(-1.96r)} \tag{15}$$

The first inequality follows from Lemma A.7, which shows that Coverage(β, r) is increasing in β when $\beta \leq 2 \times 1.96r$ and decreasing in β when $\beta > 2 \times 1.96r$; this implies that it is maximized when $\beta = 2 \times 1.96r$. The equality in the second line uses the formula for coverage in Lemma 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 (15) 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. □

Online Appendix

This online appendix supplementary materials. Section B presents additional descriptive statistics from the DiD data and treatment of outliers. Section C provides examples showing that bias can decrease when standard error corrections are small. Section D shows descriptive statistics for unclustered studies in the 1990–1999 period. Section E introduces an augmented model with strategic clustering and proposes an estimation approach which is robust to certain forms of strategic clustering. Section F shows counterfactual comparisons between the clustered regime and the unclustered regime for all values of r on the unit interval. Finally, Section G develops an alternative decision-theoretic model based on minimax regret.

B. Summary Statistics

Table B1 presents summary statistics. The sample consists of 88 DiD studies, 62 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 et al., 2023). The ratio of the average reported standard errors in unclustered studies to clustered studies is $4.989/6.755 = 0.739$ i.e. published clustered standard errors are on average 35% larger than published unclustered standard errors. It is important to note that 0.739 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 (published and unpublished), not published studies.[21]

Clustered studies are also associated with much larger effect sizes than unclustered studies (19.9% vs. 11.2%). Here, the effect size is defined as the absolute value of the estimated treatment effect.

The remaining rows of Table B1 show summary statistics on study characteristics. The number of primary JEL categories is around around three for both clustered and unclustered studies.[22] The most common categories are H (Public Economics), I (Health, Education, and Welfare), and J (Labor and Demographic Economics). While a high share of both unclustered and clustered studies belong to these categories, clustered studies are somewhat less likely to report category I. Similarly, while the majority of all studies are policy evaluations, the fraction for clustered studies (0.79) is somewhat lower than in unclustered studies (0.92).

[21] 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.

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

TABLE B1 – Summary Statistics: Unclustered and Clustered Studies using Difference-in-Differences

	Unclustered	Clustered	Difference (2)-(1)
Reported standard error (%)	4.989	6.755	1.765
	(5.935)	(7.756)	(1.520)
Effect size (%)	11.232	19.873	8.642
	(12.641)	(19.944)	(3.536)
Number of JEL codes	2.885	3.290	0.406
	(1.211)	(1.260)	(0.285)
JEL:H (Public)	0.231	0.226	-0.005
	(0.430)	(0.422)	(0.099)
JEL:I (Health, Education & Welfare)	0.500	0.306	-0.194
	(0.510)	(0.465)	(0.116)
JEL:J (Labor and Demographics)	0.577	0.548	-0.029
	(0.504)	(0.502)	(0.117)
JEL:Other	0.577	0.661	0.084
	(0.504)	(0.477)	(0.115)
Policy evaluation	0.923	0.790	-0.133
	(0.272)	(0.410)	(0.074)
log(observations)	10.225	9.896	-0.329
	(2.146)	(2.070)	(0.494)
Number.of.studies	26	62	–

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.

These comparisons are consistent with DiD research designs being applied to a wider variety of settings over time.

B.1. JEL Codes

Figure B1 shows the distribution of JEL codes. Note that studies typically include multiple JEL codes and Figure B1 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. Moreover, clustered studies are more likely to have at least one JEL code that is outside the three dominant categories of Public Economics (H); Health, Education & Welfare (I); and Labor (J).

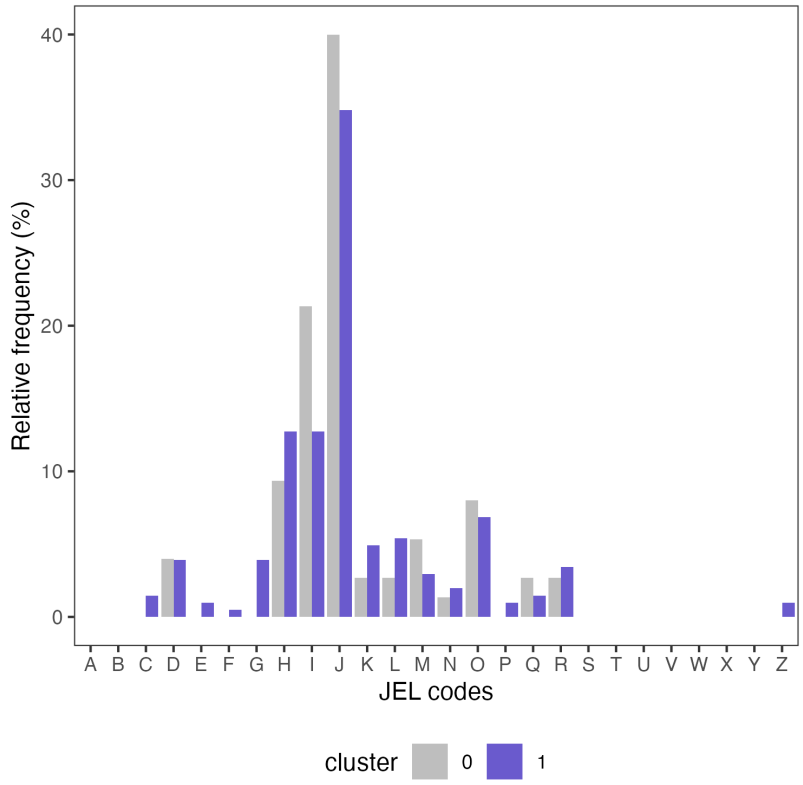

FIGURE B1. Distribution of JEL codes. The most common JEL codes are: Public Economics (H); Health, Education & Welfare (I); and Labor (J)

B.2. Outliers

As discussed in the main text, for dependent variables in non-percentage units, effects are recorded relative to the sample mean of the treatment group prior to the treatment. In four cases, this leads to very large percent effects due to low base effects. For example, one study estimates that an exogenous reallocation of police away from sporting events reduced the average number of violent incidents from 1.03 to 3.41, representing a more than 300% effect. Three other studies whose effect sizes were above 100% were removed for similar reasons – two clustered studies and two unclustered studies. Figure B2 shows the density of normalized effect sizes in the full sample which includes outliers (top panel) and the sample with the outliers removed (bottom panel).

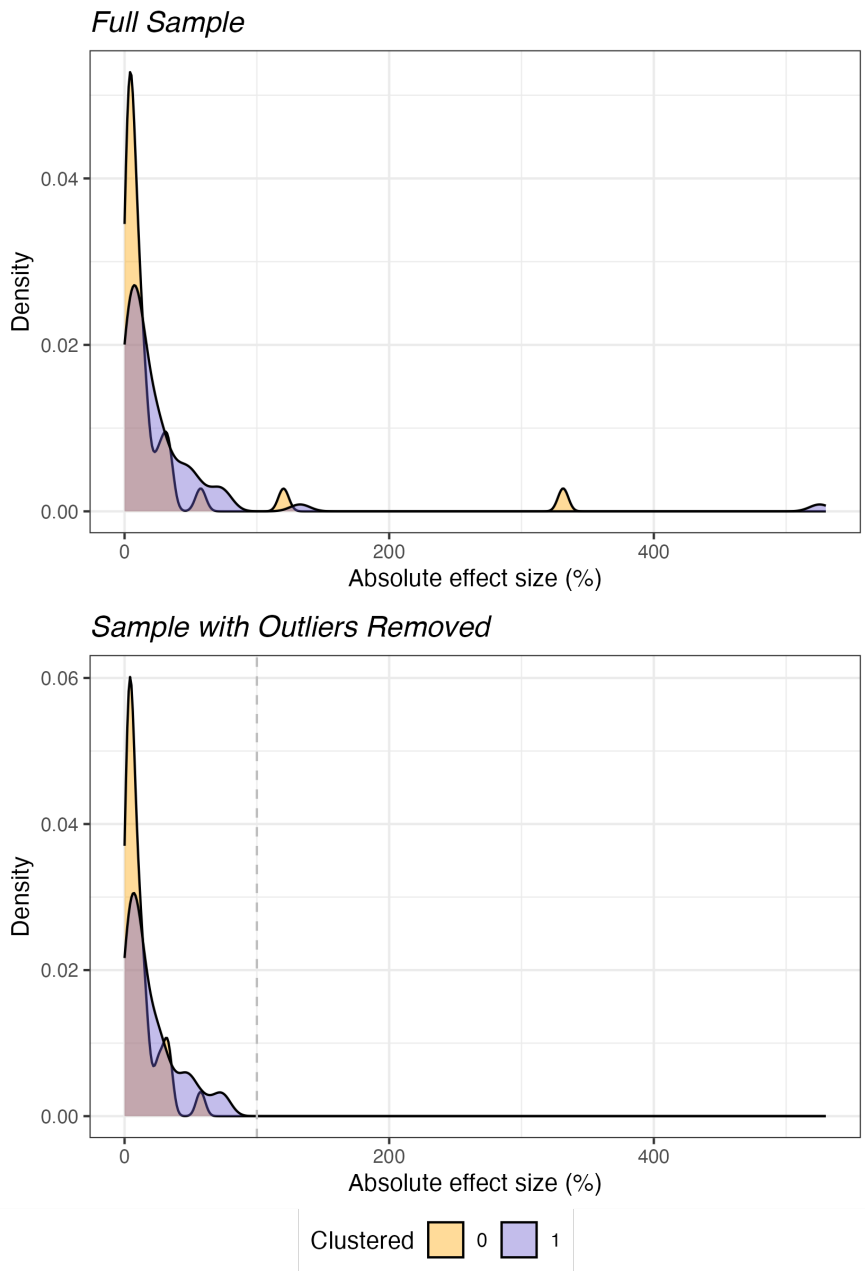

FIGURE B2. Density of Normalized Effect Sizes With and Without Outliers, for Clustered and Unclustered Studies

C. Ambiguous Impact of Corrections on Bias

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

Lemma C.1 (Ambiguous Impact on Bias)**.** *Under Assumptions 2,* **??***, and* **??***, 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):

$$\underbrace{\mathbb{E}_1[\hat{\beta}_j - \beta | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j - \beta | D_j = 1]}_{\Delta\text{Internal-validity bias}} = \underbrace{\mathbb{E}_1[\hat{\beta}_j | D_j = 1] - \mathbb{E}_r[\hat{\beta}_j | D_j = 1]}_{\Delta\text{Total bias}=\Delta\text{Estimated treatment effects}} \tag{16}$$

We can use the expression for $\text{Bias}(\beta, \gamma, r)$ from Lemma A.1 to show that the sign of equation (16) 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 $C(\beta, \gamma, r) \equiv \int_{z'} p\big(\frac{\beta + z'}{r}\big)\phi(z')dz'$ is the probability of publication conditional on β.

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

$$\begin{aligned} &\text{TrueTE}(0, \beta_2, 0.5, \gamma, 1) - \text{TrueTE}(0, \beta_2, 0.5, \gamma, r) \\ &= \beta_2 \left(\frac{C(\beta_2, \gamma, 1)}{C(0, \gamma, 1) + C(\beta_2, \gamma, 1)} - \frac{C(\beta_2, \gamma, r)}{C(0, \gamma, r) + C(\beta_2, \gamma, r)} \right) \end{aligned} \tag{17}$$

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 (17) 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 C.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 and uses the fact that studies with very large true effects have low bias (Figure C1). By contrast, Proposition C.1 shows that bias can decrease with a degenerate, and hence unchanged, distribution of true effects.

For intuition, consider the example in Lemma C.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 there-

fore 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.

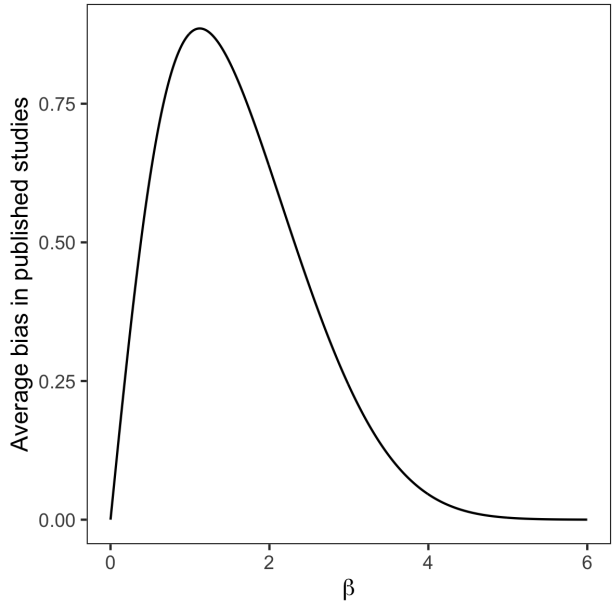

FIGURE C1. Plot of $\mathbb{E}_1[\hat{\beta}_j - \beta|D_j = 1, \beta]$ for different values of β, assuming $\gamma = 0.1$.

D. 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 observed effect size gap between clustered and unclustered studies in between 2000 and 2009. 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.

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 pub-

lished studies in the unclustered regime with relatively small effect sizes that would be 'just significant' without clustering, but insignificant with it.

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 11.2%. In the earlier 1990-1999 period, effect sizes were almost identical, at 11.5%. This difference is statistically indistinguishable from zero, although one should be cautious given the relatively small sample size. Adding controls for observable study characteristics implies that average effect sizes are slightly larger in the 2000-2009 period, which is the opposite of what we would expect if there were strategic clustering present, although the point estimate is small and statistically insignificant. Overall, 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 specific 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 2 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

$\mathbb{1}(1990-1999)$	0.313	-1.652
	(4.703)	(4.138)
Mean in 2000–2009	11.23	11.23
Observations	35	35
Adjusted-R^2	-0.03	0.152
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.

E. 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.

To begin, I extend the model in the main text to include strategic clustering. Then 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.

E.1. Model of Strategic Clustering

The model extends the model in Section 3 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:

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

   (b) Exogeneous clustering:

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

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

4. **Publication selection:**

$$\mathbb{Pr}(D_j = 1|\hat{\beta}_j, \widetilde{\sigma}_j) = \begin{cases} \gamma & \text{if } |\hat{\beta}_j|/\widetilde{\sigma}_j \geq 1.96 \\ 1 & \text{otherwise} \end{cases} \tag{18}$$

The extension from the baseline model in Section 3 is in the third step. There exists some probability $\beta_{c,1}$ that researchers will choose whether or not to cluster strategically. Specifically, researchers may strategically choose not to cluster 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.

E.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 E.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 of published clustered studies in the exogenous regime once we set $\gamma = 0$ in both regimes (this is equivalent to conditioning on statistical significance). 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 $(\beta_{c,1}, \beta_{c,2})$.

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

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

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

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

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

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

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

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. $|\hat{\beta}| \in (1.96r\sigma, 1.96\sigma)$. But studies with $|\hat{\beta}| \in (1.96r\sigma, 1.96\sigma)$ 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, which is equivalent to setting $\gamma = 0$.

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.

E.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.[23] 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. Results are in Table E1, with the robust model in the first row ($n = 54$), and the standard model from the main text in the second row ($n = 62$). Note that in the robust model, the selection parameter γ is not estimated but set to zero. Estimates for the latent distribution of studies are relatively

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

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 E1 – 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)	0.167	16.442	1.508	6.212	0.000
	(0.059)	(7.234)	(0.193)	(1.405)	–
Standard (Main Text)	0.151	18.202	1.318	7.292	0.023
	(0.045)	(6.417)	(0.171)	(1.723)	(0.009)

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.

E.4. Bias and Coverage Results with Robust Model

Ultimately, we are interested in how differences in parameter estimates from the robust approach could affect our final conclusions about the impact of clustering on bias and coverage. One concern with the statistical test above is that limited power 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 E1.

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).[24] Finally, to test sensitivity of the results, I set it to $\gamma = 0.1$, a relatively large value which is 4.35 times larger than the value estimated in DiD studies (C).

Table E2 presents the results. Overall, the conclusion from the 'standard model' that clustering increases coverage by a large amount at the expense of increased internal-validity

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

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.

TABLE E2 – Results for Model Robust to Strategic Clustering

	Unclustered ($\hat{r} = 0.51$)	Clustered ($r = 1$)	Change
Standard Model ($\hat{\gamma} = 0.023$)			
Coverage	0.36	0.72	0.36
Total Bias ($\mathbb{E}_r[\hat{\beta}_j \vert D_j = 1] - \mathbb{E}_r[\beta_j]$)	4.34 (100%)	9.51 (100%)	5.17 (100%)
Internal-Validity Bias ($\mathbb{E}_r[\hat{\beta}_j - \beta_j \vert D_j = 1]$)	1.47 (33.7%)	2.34 (24.6%)	0.88 (17.0%)
Study-Selection Bias ($\mathbb{E}_r[\beta_j \vert D_j = 1] - \mathbb{E}_r[\beta_j]$)	2.88 (66.3%)	7.17 (75.4%)	4.29 (83.0%)
Robust Model			
A DiD Studies ($\gamma = 0.023$)			
Coverage	0.36	0.71	0.35
Total Bias	4.25 (100%)	9.37 (100%)	5.12 (100%)
Internal-Validity Bias	1.51 (35.6%)	2.47 (26.4%)	0.96 (18.7%)
Study-Selection Bias	2.73 (64.4%)	6.90 (73.6%)	4.16 (81.3%)
B Economics Experiments ($\gamma = 0.037$)			
Coverage	0.37	0.74	0.37
Total Bias	4.07 (100%)	8.53 (100%)	4.46 (100%)
Internal-Validity Bias	1.45 (35.6%)	2.25 (26.4%)	0.80 (18.0%)
Study-Selection Bias	2.62 (64.4%)	6.28 (73.6%)	3.66 (82.0%)
C One-in-Ten Censored ($\gamma = 0.1$)			
Coverage	0.44	0.80	0.36
Total Bias	3.41 (100%)	6.03 (100%)	2.62 (100%)
Internal-Validity Bias	1.21 (35.6%)	1.59 (26.4%)	0.38 (14.5%)
Study-Selection Bias	2.20 (64.4%)	4.44 (73.6%)	2.24 (85.5%)

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 E, 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.

F. Impact of Clustering for Different Sized Corrections

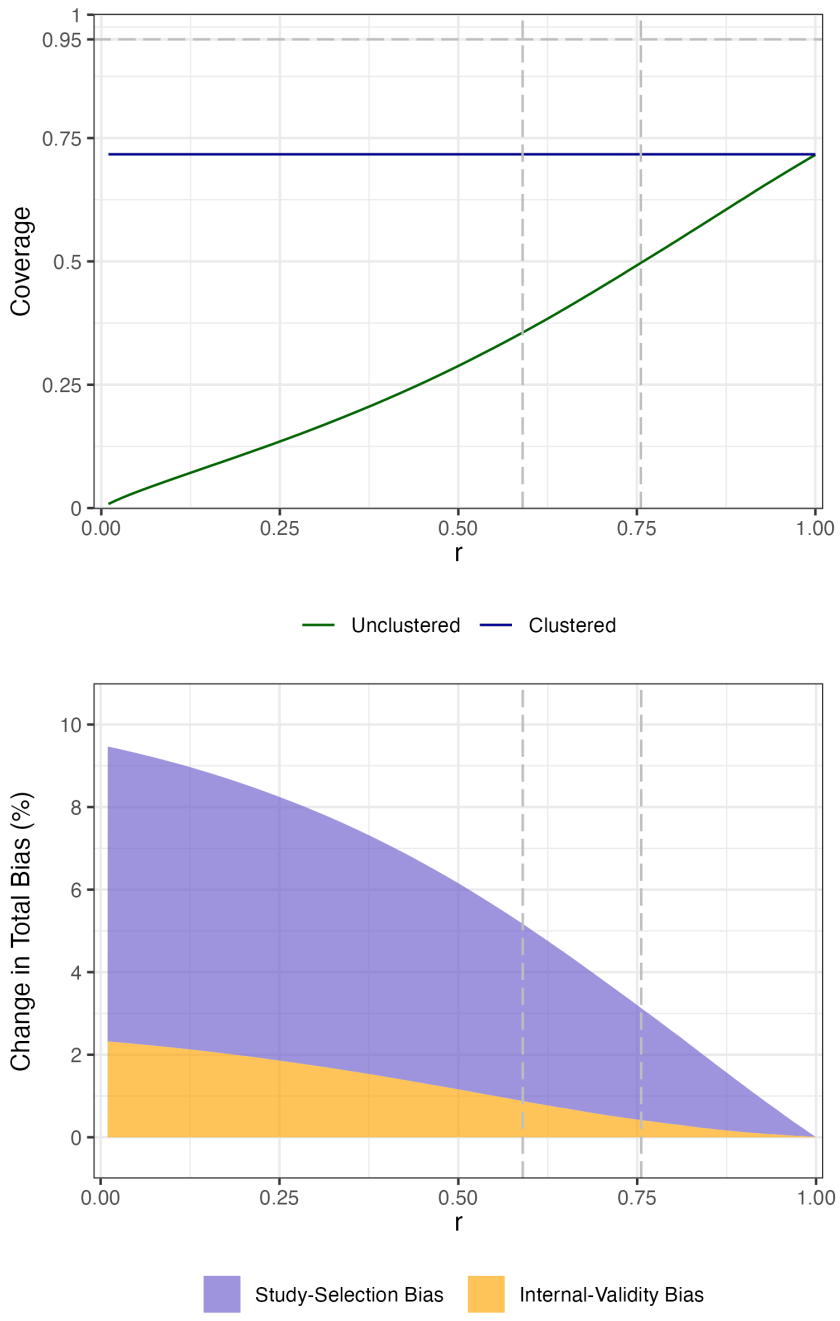

FIGURE F1. Results on the Impact of Clustering for Different Values of r

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 $\hat{r} = 0.59$ 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.

G. Minimax Regret

In this appendix, I develop an alternative decision-theoretic model based on minimax regret, as a complement the Bayesian welfare analysis in the main text. The model considers a policymaker faces a binary choice about whether or not to implement a policy based on evidence from published studies, but who overestimates the precision of estimates when standard errors are unclustered. The policymaker aims to minimize maximum regret i.e. the expected welfare loss from making an inferior treatment choice. The main finding is that clustering lowers 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. Taking high levels of loss inversion implied by standard hypothesis testing as a benchmark (Tetenov, 2012) would suggest that clustering is beneficial for policymaking.

G.1. Setup

The model extends the model of minimax regret decision-makers in Manski (2004) and Tetenov (2012) in two ways. First, to include publication bias. Second, to allow for 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 ($a = 0$).[25] 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 ($\beta \leq 0$). Specifically, the policymaker's utility from an action a with average treatment effect β is given by

$$U(a, \beta|K) = \begin{cases} Ka\beta & \text{if } \beta \leq 0 \\ a\beta & \text{if } \beta > 0 \end{cases} \tag{19}$$

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

[25] 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 the 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.

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 standard case where a study is published and the policymaker uses the evidence contained in it to inform their policy choice. Second, the case where no study is published and the policymaker must rely on a default action.

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

Importantly, the policymaker's statistical decision rule is chosen based on their beliefs about how a study's results, $(\hat{\beta}, \widetilde{\sigma})$, were generated. In the main analysis, I consider a naive policymaker who believes $\hat{\beta}$ is normally distributed on $\mathcal{B} = \mathbb{R}$ according to $N(\beta, \widetilde{\sigma}^2)$, since approximate normality is widely assumed in practice for inference, including in all the DiD papers I examine. This belief may 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, naivety means that policymakers simply 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. $\delta_1^T(\hat{\beta}) = \mathbb{1}\{\hat{\beta} > T\}$.[26] Thus the expected welfare of the threshold rule δ_1^T under the

[26]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.

misspecified belief that $\hat{\beta}$ is normal and the observed, but potentially mismeasured, standard error $\widetilde{\sigma}$, is equal to

$$\widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) = \begin{cases} K\beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta \leq 0 \\ \beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta > 0 \end{cases} \tag{20}$$

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

$$\begin{aligned} \widetilde{R}_1\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) &= W\big(\mathbb{1}\{\beta > 0\}\big) - \widetilde{W}\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) \\ &= \begin{cases} -K\beta\big[1 - \Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big)\big] & \text{if } \beta \leq 0 \\ \beta\Phi\big(\frac{T-\beta}{\widetilde{\sigma}}\big) & \text{if } \beta > 0 \end{cases} \end{aligned} \tag{21}$$

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:

$$T^* = \arg\min_{T \in \mathbb{R}} \max_{\beta \in \beta} \widetilde{R}_1\big(\delta_1^T, \beta, \widetilde{\sigma}|K\big) \tag{22}$$

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

$$\widetilde{R}_0\big(\delta_0, \beta | K\big) = \begin{cases} -K\beta\delta_0 & \text{if } \beta \leq 0 \\ \beta(1-\delta_0) & \text{if } \beta > 0 \end{cases} \tag{23}$$

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

$$\delta_0^* = \arg\min_{\delta_0 \in [0,1]} \max_{\beta \in \beta} \widetilde{R}_0\big(\delta_0, \beta | K\big) \tag{24}$$

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

Assumption G.1 (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 (22) and (24).

G.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 G.1 (Minimax Regret Decision Rule)**.** *Under Assumptions* **??** *and G.1, 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*

$$(T^*, \delta_0^*) = \left(g(K) \cdot \widetilde{\sigma},\ \frac{1}{1+K} \right) \tag{25}$$

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

Proof. 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). □

Figure G1 illustrates Lemma G.1 calibrating to the level of publication bias ($\hat{\gamma} = 0.023$) and downward bias in standard errors ($\hat{r} = 0.59$) 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 $K > 1$.[27] For intuition, see that the threshold rule in equation (25) 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 chosen. 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. Note also that the absolute size of the difference increases with Type I error loss aversion.[28]

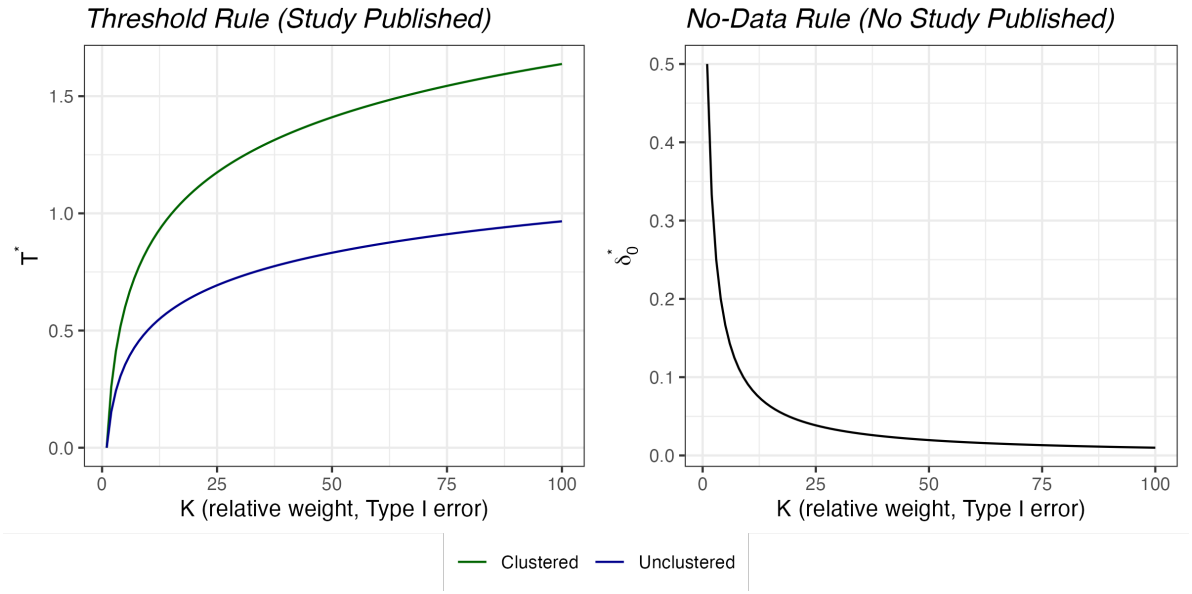

FIGURE G1. Minimax Regret Decision Rule in Clustered and Unclustered Regimes

Notes: The first panel shows the threshold rule in the event that a study is published and given by equation (22). 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.

[27] 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.

[28] This is because Lemma G.1 implies that the threshold rule in the unclustered regime is downward biased by a constant factor r, since $T^*_{C=0}/T^*_{C=1} = g(K) \cdot \widetilde{\sigma}/g(K) \cdot \sigma = r$. Hence, if $T^*_{C=1}$ increases with K, then

The second panel shows the minimax regret decision rule when no study is published. 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 implementing the policy. 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.

G.3. Comparing Regimes Based on True Regret

We would like to compare decision-making outcomes in the unclustered and clustered regimes on the basis of regret. However, recall that the minimax regret decision rule in Lemma G.1 is based on *misspecified* regret. Hence, to evaluate any given decision rule (T, δ_0), we instead use *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

$$U\big(a_1, a_0, \beta | K\big) = \begin{cases} K\beta D a_1 + \beta(1-D)a_0 & \text{if } \beta \leq 0 \\ \beta D a_1 + \beta(1-D)a_0 & \text{if } \beta > 0 \end{cases} \tag{26}$$

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

$$W\big(\delta_1^T, \delta_0, \beta, \sigma, \widetilde{\sigma} | K\big) = \begin{cases} K\Big(\beta \cdot \mathbb{P}\mathrm{r}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + \beta \cdot \big(1 - \mathbb{P}\mathrm{r}[D=1|\beta,\widetilde{\sigma}]\big)\delta_0\Big) & \text{if } \beta \leq 0 \\ \beta \cdot \mathbb{P}\mathrm{r}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + \beta \cdot \big(1 - \mathbb{P}\mathrm{r}[D=1|\beta,\widetilde{\sigma}]\big)\delta_0 & \text{if } \beta > 0 \end{cases} \tag{27}$$

where $\mathbb{P}\mathrm{r}[D = 1|\beta, \widetilde{\sigma}]$ 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.[29] 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:

$T^*_{C=1} - T^*_{C=0}$ must also grow with K.

[29] Specifically, the cdf is given by

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

$$R\big(\delta_1^T, \delta_0, \beta, \sigma, \widetilde{\sigma}|K\big) = \begin{cases} -K \cdot \beta\Big(\mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}] \cdot [1 - F(T|\beta,\sigma,\widetilde{\sigma},D=1)] + (1-\mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}])\delta_0\Big) & \text{if } \beta \leq 0 \\ \beta\Big(\mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}] \cdot F(T|\beta,\sigma,\widetilde{\sigma},D=1) + (1-\mathbb{P}\mathbb{r}[D=1|\beta,\widetilde{\sigma}]) \cdot (1-\delta_0)\Big) & \text{if } \beta > 0 \end{cases} \tag{28}$$

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 regret of any decision rule (T, δ_0) given σ is therefore given by

$$\text{Regret}(T, \delta_0|K) = \max_{\beta \in [-B,B]} R\big(\delta_1^T, \delta_0, \beta, \sigma|K\big) \tag{29}$$

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

$$100 \cdot \left(\frac{\text{Regret}^*_{C=1}(K)}{\text{Regret}^*_{C=0}(K)} - 1\right) \tag{30}$$

Figure G2 plots this quantity for different values of the Type I error loss aversion parameter K. Results show that clustering lowers regret if and only if $K > 73$. 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 implied by 5% hypothesis testing, although this could be overly conservative in certain settings.

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

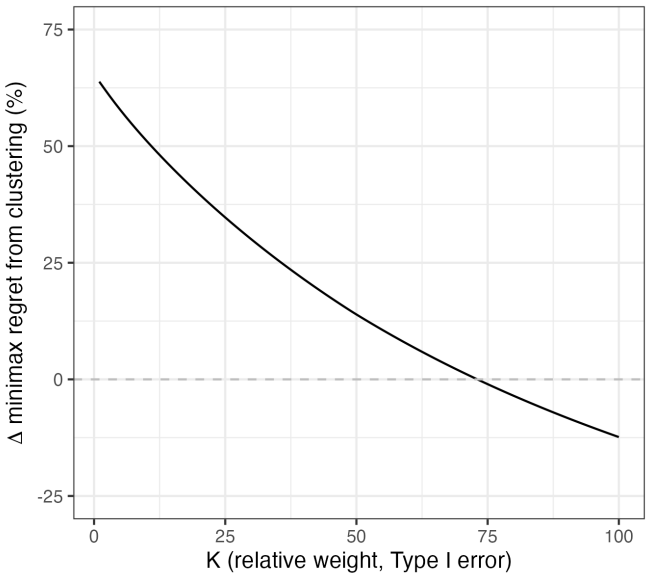

FIGURE G2. Percent Change in Minimax Regret from Clustering

Notes: The percent change in minimax regret moving from the unclustered regime to the clustered regime is calculated according to equation (30). The level of publication bias $\hat{\gamma} = 0.023$ and the extent of downward bias $\hat{r} = 0.59$ are based on the empirical model estimated on studies in the DiD literature in Section 4.

becomes increasingly miscalibrated as K increases, which leads to larger costs in terms of regret. When K is above 73, regret in the clustered regime is lower than in the unclustered regime.