Do Standard Error Corrections Exacerbate Publication Bias?

Patrick Vu[†] *(Job market paper)*

Please click here for latest version

Abstract

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

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

[†] *This version:* October 26, 2023. Brown University. Email: patrick_vu@brown.edu. I am especially grateful for invaluable advice and encouragement from Jonathan Roth, Peter Hull, and Toru Kitagawa. I also thank Daniel Björkegren, Kenneth Chay, Soonwoo Kwon, Susanne Schennach, and Jesse Shapiro for helpful comments, as well as seminar participants at Brown University, the University of Canterbury, the Econometrics Society North American Summer Meeting 2023, and the 2023 MAER-Net Colloquium. I gratefully acknowledge financial support from the Orlando Bravo Center for Economic Research.

1. Introduction

Over the past several decades, econometrics research has devoted substantial efforts to improving the accuracy of estimated standard errors in a wide variety of settings (White, 1980; Moulton, 1986; Newey and West, 1987; Staiger and Stock, 1997). In practice, these improvements often (although not always) 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.

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

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

Using this framework, I show that average bias in published studies can either increase or decrease following standard error corrections, but that increases are inevitable when correc-

tions are sufficiently large. Moreover, I show that analagous results hold for changes in true and estimated treatment effects. The case of large corrections is empirically relevant because uncorrected standard errors have been shown in many instances to be severely downward biased.[1] Intuitively, in a regime where standard errors are severely downward biased, a relatively high share of estimates will be reported as statistically significant (often erroneously). This means that relatively few studies are censored by selective publication, leading to little bias in published studies. By contrast, in a regime where standard errors are correctly measured, and hence larger, a greater share of estimates will be insignificant and censored through the publication process, resulting in higher bias (Ioannidis, 2008; Andrews and Kasy, 2019; Frankel and Kasy, 2022).

Despite the possibility of increased bias, I show that standard error corrections nevertheless unambiguously improve average coverage for published confidence intervals. This holds under very general conditions. In particular, it holds for any degree of selective publication against null results, any sized correction, and for arbitrary distributions of true treatment effects. In practical terms, this means that we can extend the common intuition that standard error corrections improve 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 unambiguously enhance the credibility of published confidence intervals, but can also lead to a deterioration in the credibility of published point estimates.

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

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

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

Descriptive statistics reveal two striking patterns. First, the adoption of clustered standard errors in the empirical DiD literature over the 2000's was associated with a near doubling in the magnitude of estimated treatment effects. This large gap remains even after controlling for differences in research topics, sample size, and including year and journal fixed effects. Second, the data exhibit strong evidence of publication bias favoring statistical significance. Following the metaregression approach in Card and Krueger (1995), I find, for both unclustered and clustered studies, a strong positive association between standard errors and effect sizes. This positive relationship means that the overwhelming majority of published studies report statistically significant results. Following Brodeur et al. (2016), I also plot the distributions of test statistics for unclustered and clustered studies. Both distributions show substantial bunching around the 5% significance threshold, which is suggestive of publication bias and p-hacking.

Taken together, the descriptive statistics are consistent with clustered standard errors interacting with the selective publication process to alter the distribution of estimated treatment effects in the empirical DiD literature. However, the theory emphasizes that we cannot make inferences about the sign of the change in bias or the magnitude of the improvement in coverage from these reduced-formed facts alone.

To learn about the impact of clustering on bias and coverage, I estimate an augmented version of the Andrews and Kasy (2019) model using data from clustered studies.[3] Consistent with estimates in alternative settings, I find a high degree of publication bias in the empirical DiD literature, with significant findings at the 5% level over 60 times more likely to be published than insignificant findings.

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

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

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

improvement but still below the nominal coverage of 0.95 due to publication bias.

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

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

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

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

quality of the decision rule is especially costly when the policymaker exhibits a high degree of loss aversion with respect to implementing an ineffective or harmful policy (i.e. Type I error).

Related Literature. This paper contributes to, and connects, two large literatures: the metascience literature on publication bias (Card and Krueger, 1995; Ioannidis, 2005, 2008; Franco et al., 2014; Gelman and Carlin, 2014; Ioannidis et al., 2017; Miguel and Christensen, 2018; Amrhein et al., 2019; Andrews and Kasy, 2019; Frankel and Kasy, 2022; DellaVigna and Linos, 2022) and the econometrics literature on robust measures of uncertainty (Anderson and Rubin, 1949; White, 1980; Moulton, 1986, 1990; Bertrand et al., 2004; Lee et al., 2022; Abadie et al., 2023). While both literatures are guided by the overarching goal of improving the credibility of empirical analysis, little attention has been paid to how they interact with one another. This article builds on existing publication selection models to provide general theoretical results on how standard error corrections 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 due to selective publication.

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

This paper proceeds as follows. Section 2 introduces the theoretical framework and states the main results. Section 3 introduces the empirical setting and presents descriptive statistics. Section 4 presents the results from the empirical model. Section 5 develops the treatment choice model with clustering and presents the main welfare results. Section 6 concludes.

2. Theory

2.1. Model of Publication Bias and Standard Error Corrections

I begin by introducing a model of how studies are generated and published in an empirical literature of interest. This could be a literature addressing many different research questions (e.g. the DiD literature). Alternatively, it could be a meta-analysis focused on a single question (e.g. the impact of job training programs on employment outcomes). The model builds on the

selective publication model in Andrews and Kasy (2019) to incorporate the possibility that reported standard errors are downward biased. While much of the discussion is framed around clustering to match the empirical application, the same model applies more generally to any method correcting for downward bias in standard errors. For proofs of the propositions in this section, see Appendix A.

Notationally, let upper case letters denote random variables and let lower case letters denote their realizations. Latent studies (published or unpublished) have a superscript * and published studies have no superscript.

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

1. **Draw latent true treatment effect and standard error:** Draw a research question with true treatment effect (Θ^*) and standard error (Σ^*):

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

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

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

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

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

$$\widetilde{\Sigma}^* = r \cdot \Sigma^*$$

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

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

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

We observe i.i.d. draws $(X, \widetilde{\Sigma}, C)$ from the conditional distribution of $(X^*, \widetilde{\Sigma}^*, C^*)$ given

$D = 1$. In the corrected regime, standard errors are accurately measured with $r = 1$ and the model is identical to the Andrews and Kasy (2019) model. However, the model differs in the uncorrected regime, where standard errors are downward biased with $r \in (0, 1)$. This implies that reported t-ratios are upward biased since $|X^*|/\widetilde{\Sigma}^* > |X^*|/\Sigma^*$. Imposing a constant downward bias factor of r permits a simple exposition of the model.[4]

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

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

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

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

Assumption 2 (Independence of True Effects and Standard Errors)**.** *Let* $\Theta^* \perp\!\!\!\perp \Sigma^*$.

This is commonly assumed in meta-analyses and is also assumed in the 'meta-study' estimation approach proposed in Andrews and Kasy (2019), which I implement in the empirical section. This assumption is unlikely to hold when researchers choose sample sizes based on predicted effect sizes in power analyses e.g. in experimental settings. However, it may be more likely to hold in observational settings where available datasets are a primary determinant of the sample size.

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

Assumption 3 (Publication Selection Function)**.** *Let* $p(X^*/\widetilde{\Sigma}^*) = 1 - (1 - \beta_p) \cdot \mathbb{1}[|X^*|/\widetilde{\Sigma}^* < 1.96]$ *with* $\beta_p \in [0, 1)$.

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

[4]Note however that all theoretical results can be generalized to the case where r is a random variable with support on $(0, 1)$, provided that $r \perp\!\!\!\perp (X^*, \Theta^*, \Sigma^*)$.

2.1.1. Motivating Example

Consider a simple example to illustrate the model and motivate the general theoretical results to follow. Suppose researchers are interested in studying the impact of a health reform on average life expectancy.

For the first stage of the model, suppose the true average treatment effect (ATE) of the reform is a one year improvement in life expectancy, $\theta = 1$, and that the standard error is $\sigma = 1$ across all studies (i.e. the joint distribution of latent true effects and standard errors, $\mu_{\Theta,\Sigma}$, is degenerate). In the second stage, researchers conduct a large number of independent DiD studies to learn about the (unobserved) ATE, each producing an unbiased DiD estimate X^* drawn from a $N(1, 1)$ distribution. For the third stage, we consider clustered and unclustered regimes for calculating standard errors. Suppose the reform is only implemented in some states and the correct approach is to cluster standard errors by state. Thus, in the clustered regime, reported standard error equals the true standard errors ($\widetilde{\Sigma}^* = \Sigma^*$). 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 $\widetilde{\Sigma}^* < \Sigma^*$). In the fourth and final stage, only a subset are published because journals prefer publishing statistically significant results. Suppose the publication process censors all insignificant findings at the 5% level (i.e. $\beta_p = 0$ in Assumption 3).

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

First, consider bias and recall that the true ATE is a one-year improvement in life expectancy. In the unclustered regime, all DiD estimates X^* whose absolute value is less than $1.96 \times \frac{1}{2} = 0.98$ are censored by selective publication; this clearly leads to upward bias in effect sizes such that the average DiD estimate conditional on publication is $\mathbb{E}[X^*|D = 1, r = \frac{1}{2}] = 1.6$ years. Clustering makes matters worse. Increases in reported standard errors raise the effective threshold for statistical significance. Now, DiD estimates whose absolute value is less than 1.96 are censored such that the average estimate conditional on publication increases to $\mathbb{E}[X^*|D = 1, r = 1] = 2.5$ years. Thus, clustering exacerbates publication bias by 150%.

This raises additional concerns. Higher bias implies that estimates are, on average, further away from the true ATE. This raises the question of whether clustering could possibly fail to meet its primary goal of improving coverage in published studies (in this example, and also more generally). It turns out that coverage does in fact increase in this example, by 19 percentage points (0.65 to 0.84). Lemma A.6 in Appendix A proves this for the case where true effects

are degenerate at any value and $\beta_p = 0$. The proof shows that higher coverage is equivalent to showing that the hazard function of the normal distribution is increasing.

This example illustrates a key tension emphasized throughout this paper: for the studies selected for publication, improvements in the credibility of confidence intervals through better coverage (↑ 41 ppts) can come at the cost of a deterioration in the credibility of point estimates due to increased bias (↑ 150%). The example illustrates that these effects can be large. This tension, of course, has only been shown here for a special case with $(\mu_{\Theta,\Sigma}, \beta_p, r) = \left(\mathbb{P}[\Theta^* = 1, \Sigma^* = 1] = 1, 0, \frac{1}{2}\right)$. In the remainder of this section, I move beyond this special case to answer, in general, what happens to bias, coverage, and other statistical properties when standard error corrections for downward bias are applied. In particular, I derive exact conditions under which the tension between increased bias and coverage generalizes to other settings.

2.2. Estimated Treatment Effects, True Treatment Effects and Bias

The example showed that it is possible for standard error corrections to increase bias in published studies. Under what conditions does this conclusion hold more generally? Moreover, what is the impact of standard error corrections on other measures that may be of interest, such as true and estimated treatment effects? This subsection answers both questions. For the first, I find that a sufficient condition for increased bias is that corrections are 'sufficiently' large, and present an example where small corrections can in fact lead to a decrease in bias. For the second, I establish analogous results for true and estimated treatment effects.

Before presenting these results, I first define the key measures of interest. Throughout, I normalize the true standard error to $\Sigma^* = 1$ and omit it from the notation for clarity. First, define the expected estimated treatment effect in standard error regime r as $\mathbb{E}[X^*|D = 1, \widetilde{\Sigma}^* = r]$ and the expected true treatment effect in regime r as $\mathbb{E}[\Theta^*|D = 1, \widetilde{\Sigma}^* = r]$. Consider two definitions of bias. Define *individual-study bias* in standard error regime r as $\mathbb{E}[X^* - \Theta^*|D = 1, \widetilde{\Sigma}^* = r]$. This measure asks how far, on average, published estimates are from the questions they answer. It is appropriate when examining an empirical literature addressing different research questions. The second measure is *meta-study bias*, which is defined as $\mathbb{E}[X^*|D = 1, \widetilde{\Sigma}^* = r] - \mathbb{E}[\Theta^*]$. This measure may be preferable when examining an empirical literature examining a single question (e.g. the impact of a change in the minimum wage on employment).[5] It asks how far published estimates are from the average true effect across all latent studies. Theoretical results apply to both definitions. The empirical results in the

[5]Latent true effects Θ^* may still follow a distribution in the case where only a single question is examined. This is because latent true treatment effects may differ across studies due to heterogeneity in populations, research design, policies etc. Note also that meta-study bias can be decomposed into the sum of individual-study bias and question-selection bias: $\mathbb{E}[X^*|D = 1, \widetilde{\Sigma}^* = r] - \mathbb{E}[\Theta^*] = \mathbb{E}[X^* - \Theta^*|D = 1, \widetilde{\Sigma}^* = r] + \left(\mathbb{E}[\Theta^*|D = 1, \widetilde{\Sigma}^* = r] - \mathbb{E}[\Theta^*]\right)$.

following section focus on individual-study bias because the empirical DiD literature covers many different research questions.

Finally, note that in general Θ^* follows a distribution. This means that standard error corrections can lead to different research questions being addressed in the published literature due to selective publication. While changes in bias have clear implications for statistical credibility, the impact of changes in the magnitude of true treatment effects are less clear.

With this we can state the main result of this subsection:

Proposition 1 (Large Corrections Increase Bias)**.** *Under Assumptions 1, 2, and 3, there exists an* $r^* \in (0, 1]$ *such that for any* $r \in (0, r^*)$*, individual-study bias, meta-study bias, estimated treatment effects, and true treatment effects all increase with standard error corrections.*[6]

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

For intuition underlying Proposition 1, consider individual-study 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 are relatively unbiased. However, corrections increase standard errors, which leads to more studies with small effect sizes being censored by the publication process and hence higher bias. It follows that moving from the uncorrected regime with little bias to the corrected regime must necessarily increase bias.

To see the necessity of the sufficient condition, consider an example where small standard error corrections lead to a *decrease* in individual-study bias.[7] Let the latent distribution of true effects Θ^* take on two possible values $(\theta_1, \theta_2) = (1, 4)$ with probabilities $\frac{4}{5}$ and $\frac{1}{5}$, respectively. This can be thought of as a literature addressing two questions, one with a small true effect and one with a large true effect. Assume that only one in twenty insignificant studies are published ($\beta_p = \frac{1}{20}$) and unclustered standard are 80% of their true value ($r = \frac{4}{5}$).

[6] All inequalities are strict except for true treatment effects which weakly increase. If the latent distribution of true treatment is non-degenerate, then the inequality for true treatment effects is also strict.

[7] Similar examples can also be constructed for meta-analytic bias, estimated treatment effects and true treatment effects. See Appendix B for more details.

In the clustered regime, a higher share of studies addressing the question with the larger effect ($\theta_2 = 4$) are published relative to the unclustered regime. This is because studies addressing the question with the smaller true effect ($\theta_2 = 1$) are more likely to be insignificant with clustering and hence censored by selective publication. This decreases average individual-study bias because studies addressing questions with very large effect sizes have bias close to zero.[8] The intuition behind this is that when true effects are large, the probability of obtaining an insignificant result, and thus being subject to publication bias, is low. Overall, then, clustering shifts the distribution of published studies toward those with larger true effects and hence smaller bias.

This example highlights a second important point: it is possible for estimated treatment effects to increase with clustering, despite the fact that individual-study bias decreases. This possibility underscores the limitations of what we can learn about bias from reduced-form statistics calculated on observed effect sizes. To see why, note that the change in estimated treatment effects can be decomposed into the sum of the change in individual-study bias and the change in true treatment effects:

$$\begin{aligned} &\mathbb{E}[X^*|D=1,\widetilde{\Sigma}^*=1]-\mathbb{E}[X^*|D=1,\widetilde{\Sigma}^*=r] \\ &=\Big(\mathbb{E}[X^*-\Theta^*|D=1,\widetilde{\Sigma}^*=1]-\mathbb{E}[X^*-\Theta^*|D=1,\widetilde{\Sigma}^*=r]\Big)+\Big(\mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=1]-\mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=r]\Big) \end{aligned} \tag{2}$$

where the expectations in all three expressions are taken with respect to $(\Theta^*, \Sigma^*) \sim \mu_{\Theta,\Sigma}$ and $X^*|\Theta^*, \Sigma^* \sim N(\Theta^*, \Sigma^{*2})$.

Clustering in this example leads to an overall increase in estimated treatment effects (0.30) that reflects an increase in true treatment effects (0.31) which outweighs a decrease in individual-study bias (-0.01). Thus, by observing higher effect sizes in clustered studies, it is not possible, in general, to infer the sign of the change in bias. Proposition 1, of course, guarantees that bias must increase if corrections are sufficiently large. Figure 1 illustrates this by tracing out the change in individual-study bias from adopting different sized standard error corrections (r). In this example, we have that $r^* = 0.77$, meaning that corrections that enlarge standard errors by more than about 30% will lead to an increase in bias.

In summary, bias, estimated treatment effects, and true treatment effects can in general increase or decrease with corrections, but must always increases when corrections are sufficiently large.

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

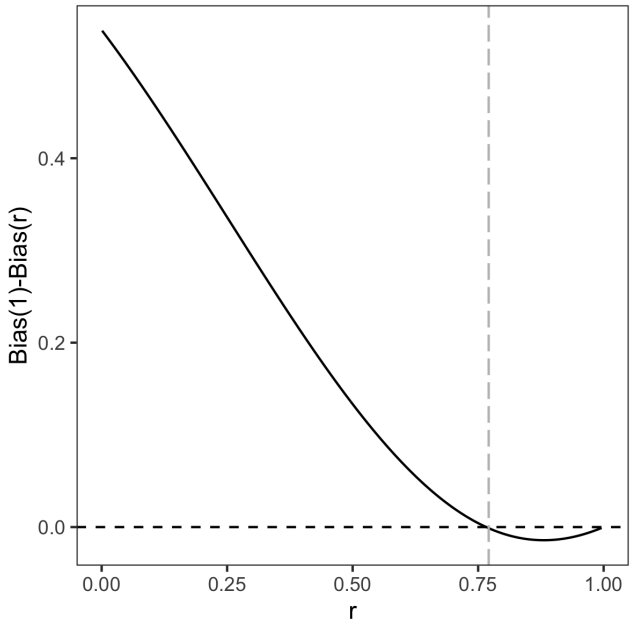

FIGURE 1. Plot of the change in bias from adopting standard error corrections for different values of r. The number $r^* = 0.77$ denotes the value of below which bias always increases with standard error corrections.

2.3. Coverage

We turn next to how standard error corrections impact coverage probabilities in the presence of publication bias. First, define expected coverage conditional on publication for standard error regime $r \in (0, 1]$ as $\text{Coverage}(r) = \mathbb{P}[\Theta^* \in (X^* - 1.96r, X^* + 1.96r) | D = 1, \widetilde{\Sigma}^* = r]$ i.e. the probability that *reported* 95% confidence intervals contain the true effect Θ^* conditional on publication. Compare this to expected coverage in a standard econometric analysis without publication bias: $\mathbb{P}[\Theta^* \in (X^* - 1.96r, X^* + 1.96r)]$. In this latter case, it is clear that standard error corrections improve coverage.

The presence of publication bias, however, introduces several complications. In the definition of $\text{Coverage}(r)$, see that the degree of downward bias affects not only the width of reported confidence intervals, but also the studies (X^*, Θ^*) that end up making it into the published literature, since uncorrected standard errors are more likely to lead to statistically significant findings. This can complicate comparisons between uncorrected and corrected regimes. Consider Figure 2, which graphically illustrates three possible data realizations for estimates which would be treated differently under corrected and uncorrected regimes. Confidence intervals with unclustered standard errors are in yellow and those with corrections applied are twice the width and in purple. Consider the three cases:

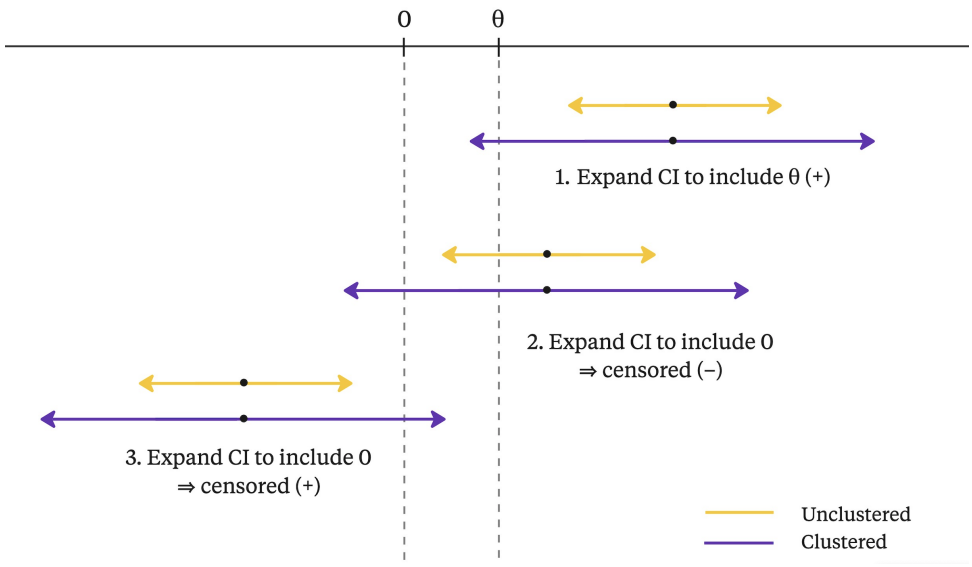

FIGURE 2. Three Effects of Clustering on Coverage

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

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

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

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

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

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

Proposition 2 (Standard Error Corrections Improve Coverage)**.** *Under Assumptions 2, and 3, Coverage*$(1) - $*Coverage*$(r) > 0$.

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

The proof of Proposition 2 builds on the the special case where the distribution of true effects Θ^* is degenerate and $\beta_p = 0$.[9] The proof shows that this conclusion hold more generally, in particular, for (i) arbitrary levels of selective publication against null results, $\beta_p \in (0, 1)$; and for (ii) arbitrary distributions of latent studies μ_Θ. Both generalizations are non-trivial extensions of the special case. This is because the distribution of published studies, $X^*, \Theta^* | D = 1, \widetilde{\Sigma} = r$, on which expected coverage is calculated, depends jointly on the degree of selective publication β_p, the extent to which standard errors are downward biased by r, and the latent distribution of true effects μ_Θ.

The generalization to any level of selective publication makes use of a result which shows that any publication regime $\beta_p \in [0, 1]$ can be expressed as a mixture of a publication regime which publishes all insignificant results ($\beta_p = 1$) and one that censor all insignificant results ($\beta_p = 0$). Loosely speaking, since coverage trivially improves in the former regime, we only need to focus on the latter case where $\beta_p = 0$. Generalizing the result to non-degenerate distributions of Θ^* uses the shape of the coverage probability curve as a function of Θ^* and the fact that when $\beta_p = 0$ the distribution of published true treatment effects in the corrected regime, $\Theta^* | D = 1, \widetilde{\Sigma} = 1$, first-order stochastically dominates the corresponding distribution in the uncorrected regime, $\Theta^* | D = 1, \widetilde{\Sigma} = r$. For more details on the proof, see Appendix A.

3. Setting and Data

I turn now to analyzing the implications of the theoretical results in a particular setting: the adoption of clustered standard errors in the empirical DiD literature. There are several

[9] Coverage is shown to improve in this case in the proof of Lemma A.6 in Appendix A. The proof shows there are two cases to consider, one where θ is relatively 'large' and another where it is relatively 'small'. For large true effect, only effects one and three in Figure 2 occur and thus coverage must improve with corrections. For 'small' true effects, it can be shown that better coverage is equivalent to showing that the hazard function for normal distribution is increasing.

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

3.1. Data

I collected data from 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 identify articles using DiD using a string-search algorithm. I collect data on the 'main' DiD estimate in each study, and exclude placebo tests and tests of alternative hypotheses. The 'main' estimate is chosen from the first DiD table in the paper. When there are multiple estimates, I choose the one emphasized in the discussion of the results or the abstract. When there are several specifications, I select the one with full controls. For DiD articles that fit the inclusion criteria described below, I manually collected data on the estimated DiD treatment effect; the reported standard error; an indicator for whether a correction for serial correlation is implemented; an indicator for policy evaluations[10]; and the number of observations. I also obtained JEL classification codes for each study from *EconLit.*

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

[10]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.

stronger assumption that reported clustered standard errors reflect the true standard error.

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.[11] For example, consider a study estimating the impact of an educational program on the drop-out rate. I convert the estimated treatment effect into percent units by dividing it by the mean drop-out rate of the treated group before the intervention. When this is not available, I instead normalize with the mean of the dependent variable for the whole sample. Two studies did not report an average for the dependent variable and were therefore 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).

Table 1 presents the summary statistics for the DiD studies meeting this inclusion criteria. The sample consists of 96 studies, with 66 implementing clustering. 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 $4.250/6.497 = 0.654$ i.e. published clustered standard errors are on average 53% larger than published unclustered standard errors. It is important to note that 0.654 is not an estimate of the degree of downward bias in unclustered standard errors (r), which would be equal to the ratio of unclustered to clustered standard errors in latent studies, not published studies.[12]

Clustered studies are also associated with much larger effect sizes (19.5% vs. 12.2% in unclustered studies). That larger standard errors are accompanied by higher effect sizes is consistent with the main mechanism emphasized in the theory in Section 2, namely, that clustering raises the bar for statistical significance and results in the selection of larger effect sizes due to publication bias. More detailed descriptive statistics consistent with this interpretation are presented further below.

The remaining rows of Table 1 show summary statistics on study characteristics for unclustered and clustered studies. Studies may list multiple JEL categories are the average for

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

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

TABLE 1 – Summary Statistics: Unclustered and Clustered Studies using Difference-in-Differences

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

Notes: The sample is DiD literature over 2000-2009 based on inclusion criteria described in the main text. All reported figures are means and standard errors are 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.

both types of studies is around three. The most common primary JEL categories are H (Public Economics), I (Health, Education, and Welfare), and J (Labor and Demographic Economics). While a high share of unclustered and clustered studies belong to these categories, clustered studies are somewhat less likely to report categories I and J.[13] Similarly, while the majority of all studies are policy evaluations, the fraction for unclustered studies (0.87) is somewhat higher than in clustered studies (0.80). These statistics are consistent with DiD research designs being used in a wider variety of settings over time.

3.2. Three Stylized Facts

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

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

1. The adoption of clustering in the empirical DiD literature rose from almost no use in the 1990's to near universal adoption by the end of the 2000's;

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

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

3.2.1. Adoption of Clustering in the 2000's

Despite earlier emphasis in the econometrics literature on the importance of accounting for serial correlation when calculating standard errors (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 1990–2000. Specifically, Bertrand et al. (2004) identified 65 papers with a potential serial correlation problem and found only five (7.7%) that implement some form of correction for serial correlation.[14]

In Figure 3, I show the fraction of DiD articles implementing a correction for serial correlation over the next decade. The 2000's 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 Bertrand et al. (2004) study, which was highly influential and released as a working paper in the early 2000's.

[14]Moreover, four of these five studies use GLS for corrections, which they argue is relative ineffective.

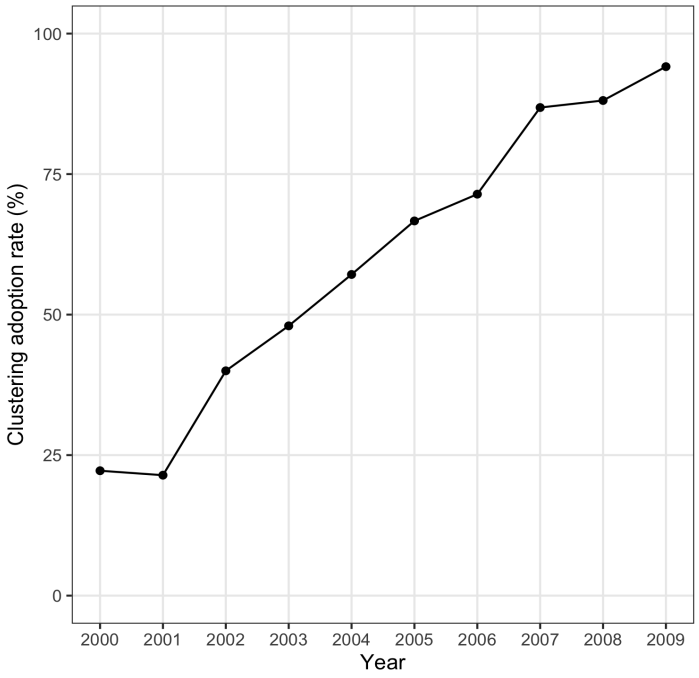

FIGURE 3. Three-Year Centered Moving Average of the Clustering Adoption Rate

3.2.2. Effect Size Gap for Clustered and Unclustered Studies

As noted in Table 1, there is a large difference in the magnitude of estimated treatment effects between unclustered and clustered studies. Differences in observable study characteristics cannot explain this gap. Table 2 reports results from a regression of the effect size on an indicator for clustering, adding additional controls with each successive column. The final specification includes year and journal fixed effects and controls for sample size, research topic (JEL categories), and an indicator for policy evaluations. The estimated coefficient in the specification with full controls implies that effect sizes in clustered studies are larger than those in unclustered studies by a factor of 1.84 (22.36% vs. 12.18%).

This is a striking gap and consistent with a substantial shift in the distribution of published studies. However, it is important to emphasize that the theoretical results in Subsection 2.2 show that observing larger estimated treatment effects in clustered studies does not, in and of itself, tell us whether bias has actually increased. The example presented there shows that higher effect sizes can also be consistent with a decrease in bias.[15] To make inferences about

[15]Strictly speaking, the example shows that the *unnormalizaed* difference in effect sizes, $\mathbb{E}[X^*|D=1,C^*=1]-\mathbb{E}[X^*|D=1,C^*=0]$, is positive. However, it is also true in this example that the difference in the magnitude of estimated treatment effects, $\mathbb{E}[|X^*||D=1,C^*=1]-\mathbb{E}[|X^*||D=1,C^*=0]$ is positive. This section focuses on absolute effect sizes because we do not in fact observe unnormalized effect sizes X^* conditional on our normalization that Θ^* is positive (Assumption 1). For a concrete example, consider a study with an observed estimate X^*, and an unobserved true effect Θ^*, which could be positive or negative. Now normalize the

TABLE 2 – Impact of Clustering on Estimated Treatment Effects

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

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

changes in bias, it is therefore necessary to estimate the latent distribution of studies, which we do in the following section.

One potential concern with the results in Table 2 is the possibility of strategic clustering. This is a particular form of endogeneity where researchers p-hack their standard errors to increase the chances of publication. In particular, suppose that researchers strategically choose not to cluster if doing so would overturn a statistically significant result. This behavior would also generate a positive correlation between clustering and estimated treatment effects. Thus, the observed increase in estimated treatment effects in Table 2 might reflect the impact of clustering on estimated treatment effect via selective publication process *and* strategic clustering by researchers.

To test whether strategic clustering is driving this result, I examine effect sizes of unclustered studies in the 1990–1999 period from the same set of journals. During this period, the overwhelming majority of studies reported unclustered standard errors (Bertrand et al., 2004) and hence strategic clustering is unlikely to be affecting the distribution of effect sizes. If strategic clustering was absent in the 1990–1999 period, but present during the 2000–2009 period, then, all else equal, we might expect effect sizes to be smaller in the 2000–2009 period. This is because strategic clustering would increase the fraction of published studies in the unclustered regime with relatively small effect sizes that would be 'just significant' without clustering, but insignificant with it. Instead, I find that the mean effect size in the 1990–1999 period is close to, and in fact slightly lower than, the mean effect size in the 2000–2009 period (10.6% and

true effect to be positive $|\Theta^*|$. Whether or not we switch the sign of X^* to be consistent with this normalization requires knowledge of the sign of unnormalized Θ^*, which we do not observe.

12.2%). The difference is statistically indistinguishable from zero, although statistical power is somewhat limited.[16] Controlling for differences in observable study characteristics, including JEL topics and sample sizes, does not change this conclusion. This supports the idea that strategic clustering of the simple form discussed here is not driving observed differences in effect sizes across clustered and unclustered regimes. This, of course, covers only one form of endogeneity and other forms could in principle be present. For more details, see Appendix E.

Ultimately, the primary goal of the empirical analysis is to estimate the changes in bias and coverage that occur due to clustering, not simply changes in effect sizes. To this end, I propose in the following section an estimation approach for the empirical model that yields unbiased estimates of the model parameters irrespective of whether or not there is strategic clustering of the simple form described here. See Subsection 4.1 for further discussion.

3.2.3. Selective Publication on Statistical Significance

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

The first is the metaregression approach proposed in Card and Krueger (1995). Figure 4 visualizes a regression of estimated treatment effects on reported standard errors. Panels (a) and (b) separate articles using clustered and unclustered standard errors, respectively. These plots are consistent with selective publication on the basis of statistical significance, for at least three reasons. First, there are simply 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 one percentage point increase in standard errors is associated with a little over a two percentage point increase in estimated effect sizes – this is, approximately the increment required for maintaining statistical significance. In the absence of selective publication, there is little reason to expect a systematic relationship between estimated treatment effects and standard errors, because the sample size in observational studies is not typically chosen but instead predetermined by available datasets.[17] Finally, the estimated slope coefficient on reported standard errors is very similar across clustered and unclustered regimes. Given that unclustered standard errors are systemat-

[16] This is because the number of studies in the 1990's which are (i) detected by the (Currie et al., 2020) string-search algorithm and (ii) meet the inclusion criteria in Subsection 3.1 is relatively small.

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

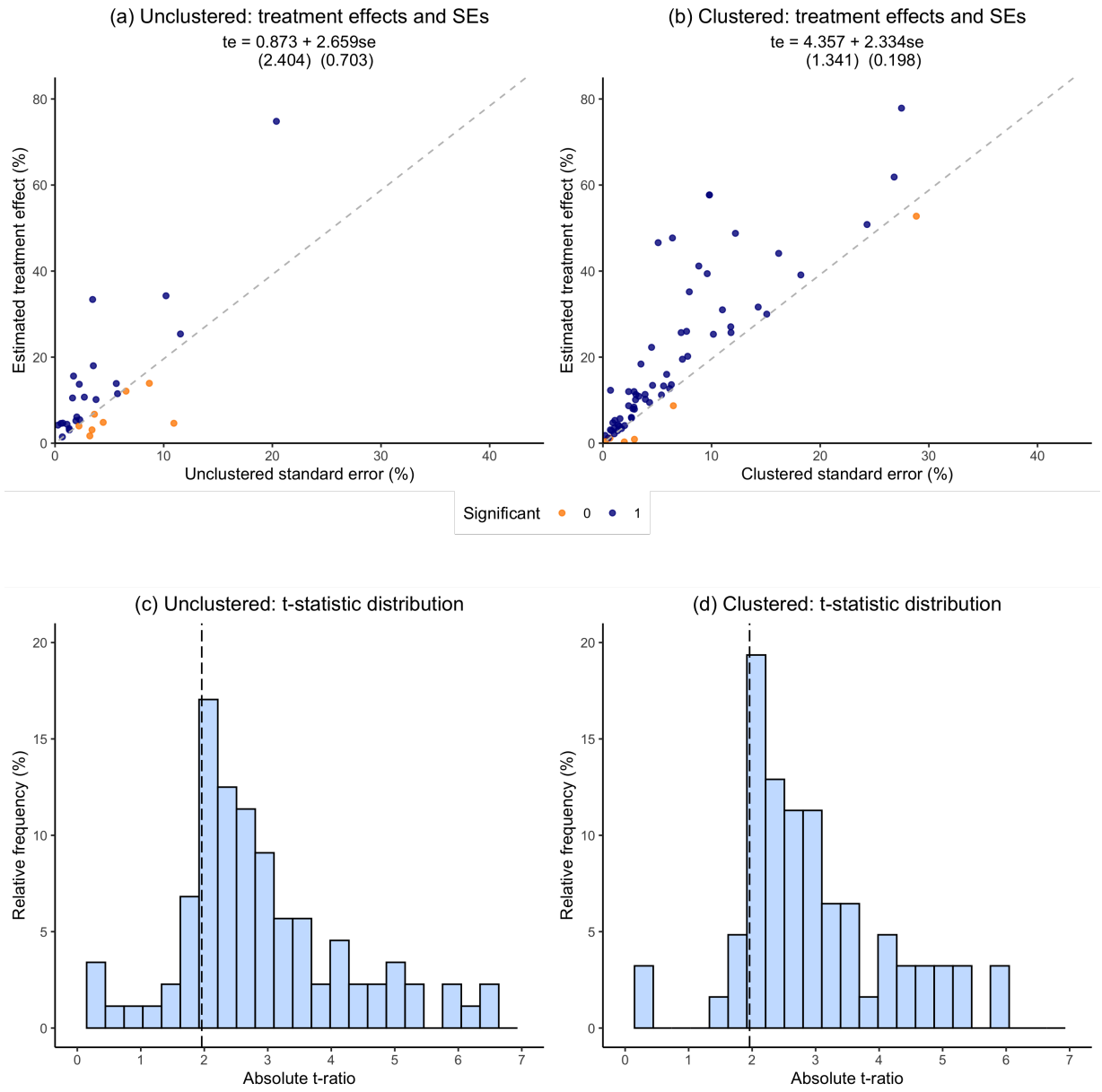

FIGURE 4. SELECTIVE PUBLICATION AND *p*-HACKING

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

ically downward biased, one would expect, under the null hypothesis of no selective publication, that clustering would lead to a decrease in the slope coefficient on standard errors. Instead, the estimated linear relationship between treatment effects and reported standard errors is similar

across regimes.

Following Brodeur et al. (2016), a second test examines the distribution of t-statistics to determine if there is a bunching around critical significance thresholds. Panel (c) shows the distribution of test statistics for unclustered studies, while Panel (d) shows the same for clustered studies. The vertical dashed line marks the 5% threshold significance level. In both figures, there is a large mass of t ratio values just above this threshold, and a 'missing' mass just below it. Despite the fact that standard errors are systematically higher in clustered studies, the distributions appear very similar in both regimes, providing additional evidence of selective publication (or p-hacking).

4. Empirical Model

Descriptive statistics provide evidence that clustering led to a change in the distribution of estimated treatment effects via selective publication. However, from these descriptives alone, we cannot make inferences about some of the main quantities of interest, namely, bias and coverage. To do this, I follow an empirical strategy consisting of two steps. In the first, I estimate the model in Section 2 using data from clustered DiD studies. This gives parameters governing the latent distribution ($\mu_{\Theta,\Sigma}$) and selective publication (β_p) for clustered studies. With these model estimates, we can analyze counterfactual scenarios of what would have happened had clustered studies instead reported unclustered standard errors which were downward biased by any specified factor r. In the second step, I describe two approaches for calibrating reasonable values for r. I then present the main results.

4.1. Estimation

First, I estimate the model of selective publication in Section 2 using data from clustered studies. Following Andrews and Kasy (2019), I estimate the latent distribution of true effects assuming that $\Theta^* \perp\!\!\!\perp \Sigma^*$ (Assumption 2) and $\Theta^*|\lambda_\theta, \kappa_\theta \sim \text{Gamma}(\lambda_\Theta, \kappa_\Theta)$. Following Vu (2023), I augment the baseline model to jointly estimate the distribution of standard errors and assume that this also follows a gamma distribution: $\Sigma^*|\lambda_\sigma, \kappa_\sigma \sim \text{Gamma}(\lambda_\sigma, \kappa_\sigma)$. This is necessary in order to simulate studies with the estimated model to calculate, for example, coverage. In line with the theory, I assume publication probabilities follow a step function where the relative probability of publishing a statistically insignificant result at the 5% level is given by β_p.[18] Finally, note that clustered standard errors are assumed in estimation to reflect

[18]This is similar to Assumption 3 in that selective publication follows a step function at the 5% level. It differs, however, in that it does not impose that $\beta_p \in [0, 1)$. In particular, estimation allows the possibility that $\beta_p \geqslant 1$ such that the *relative* probability of publishing insignificant results is the same as, or higher than, for significant results. Note also that publication probabilities are only identified up to scale.

TABLE 3 – Maximum Likelihood Estimates

Latent true effects Θ^*		Latent standard errors Σ^*		Selection
κ_θ	λ_θ	κ_σ	λ_σ	β_p
0.154	17.802	1.426	6.475	0.016
(0.035)	(2.692)	(0.167)	(1.282)	(0.007)

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

the true variation of estimated treatment effects.

Consistency of the model parameters requires that $C^* \perp\!\!\!\perp X^* | \Theta^*$. This assumption is violated if there is strategic clustering, which I address below in an alternative estimation approach that delivers similar estimates to the baseline model. The assumption is not violated, however, by non-random clustering with respect to study characteristics. For example, there is suggestive evidence in Table 1 that DiD studies outside of Health, Education & Welfare (JEL:I) and Labor & Demographics (JEL:J) are more likely to use clustered standard errors. If this were indeed the case, then estimation would still yield consistent estimates of the latent distribution of studies in the clustered regime, although the latent distribution in the unclustered regime would differ. Finally, note that I restrict attention to clustered studies to avoid imposing strong assumptions about the mapping between unclustered standard errors and (unobserved) clustered standard errors for unclustered studies in the likelihood function.[19]

Table 3 presents the maximum likelihood estimates. The estimate $\hat{\beta}_p = 0.016$ implies a high degree of selective publication. In particular, it means that statistically significant results are around 60 times more likely to be published than insignificant results. This is broadly similar to estimates of publication bias in Andrews and Kasy (2019) for replication studies in economics ($\hat{\beta}_p = 0.038$) and psychology ($\hat{\beta}_p = 0.017$).

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

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

The key idea is to estimate the parameters governing the latent distribution of studies on the selected subset of *statistically significant* clustered studies.[20] The rationale is that the distribution of significant, clustered studies, $X^*, \Sigma^* | D = 1, C^* = 1, |X^*|/\Sigma^* \geqslant 1.96$, is completely invariant to this form of strategic clustering. This is because strategic clustering only affects studies whose results are insignificant when clustered but significant when unclustered. However, none of these studies are included in the subsample of statistically significant clustered studies. Thus, the distribution of studies, and hence the likelihood, is unaffected by whether or not strategic clustering is present. For a formal statement and proof of this claim, see Lemma F.1 in Appendix F. Robust estimates for the latent distribution of studies are presented in Table F1 and statistically indistinguishable from the baseline estimates in Table 3. This suggests that strategic clustering of the form discussed here does not bias baseline parameter estimates.[21]

4.2. Unclustered Counterfactuals

With the model estimates in Table 3, we can calculate expected bias, coverage, true treatment effects and estimated treatment effects under the counterfactual scenario where clustered studies report unclustered standard errors that are downward biased for some specified factor $r \in (0, 1)$. We can then compare these statistics – and thus the statistical credibility of published studies – across unclustered and clustered regimes. To interpret these results, note that this counterfactual comparison is, in a certain sense, analogous to an ATT measure of the impact of clustering on clustered studies, rather than an ATE measure which would be the impact of clustering on both unclustered and clustered studies.

This ATT measure can be computed for any specified value of $r \in (0, 1)$ using only the model estimates in Table 3. Figure G1 in Appendix G shows the results as a function of r over the unit interval. This can be connected usefully to Proposition 1, which states that bias must increase for sufficiently large standard error corrections i.e. for any r less than some model-dependent value r^*. Based on the estimates in the DiD setting, I find that $r^* = 0.95$. This implies that any corrections enlarging standard errors by 5.3% or more would lead to an increase in bias in published DiD studies. Since Proposition 2 guarantees improved coverage, it follows that the qualitative conclusion of better coverage but higher bias will exist for all but very small standard error corrections. The quantitative results, however, will depend on r, with larger corrections leading to greater increases in both bias and coverage.

[20] This involves setting $\beta_p = 0$ and not estimating it.

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

4.3. Calibrating r

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

In the first approach, I make the simplifying assumption that all unclustered standard errors are downward biased by a constant factor $r \in (0, 1)$. I then calibrate r using the method of simulated moments (McFadden, 1989). Specifically, I select the value of r which minimizes the distance between moments predicted by the model and the actual moments observed in the data. Given that r measures the degree of downward bias in unclustered standard errors, the moment I choose for calibration is the difference in average reported standard errors between clustered and unclustered studies in the published literature. Carrying out this procedure gives $\hat{r} = 0.51$. In other words, clustered standard errors are estimated to be around two times larger than unclustered standard errors. This approach assumes that the distribution of latent studies in clustered studies is the same as in unclustered studies.[22] This would be violated, for example, if there are differences in the datasets which tend to be used in *latent* unclustered and clustered studies, since this would imply differences in the latent distribution of standard errors. Nevertheless, if the assumption is violated, then we still obtain a valid counterfactual for what would have occurred if clustered studies had instead been unclustered and were around half the size of true standard errors.

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

The studies are those identified as DiD articles in Brodeur et al. (2020). I collected data on standard errors from six of the 25 journals sampled in that study.[23] While code is available for

[22] Based on the earlier analogy, this would imply that the ATT and ATE coincide.

[23] The journal are *Applied Economic Journal: Applied Economics*, *Applied Economic Journal: Economic Policy*, *American Economic Review*, *Journal of Labor Economics*, *Journal of Political Economy* and the *Quarterly Journal of Economics*. Four overlap with journals from the main analysis. The two excluded journals are:

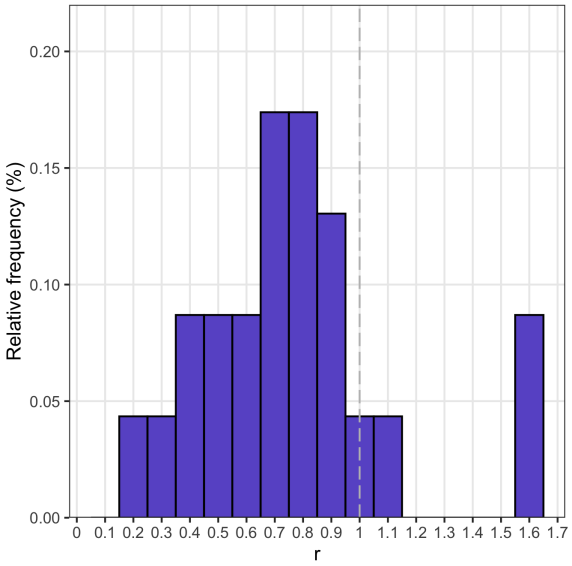

FIGURE 5. Empirical Distribution of r from 2015–2018 DiD Studies

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

almost all studies, only a subset use publicly available data. Overall, I calculate r in 23 out of 72 DiD studies (31.9%) sing non-proprietary data. Figure 5 shows the empirical distribution. The mean is 0.76 and a small fraction of studies have clustered standard errors which are *larger* than unclustered standard errors ($r > 1$). For calculating the counterfactual scenario for unclustered studies, we can draw randomly from this distribution to determine the degree bias for each study individually. This is useful because in reality, r varies across studies and depends on the within-cluster correlation of the regressor, the within-cluster correlation of the error, and the number of observations in each cluster (Cameron and Miller, 2015).

4.4. Results on the Impact of Clustering

Panel A in Table 4 presents the results for how the statistical properties of published studies change with clustering based on the method of simulated moments for calibrating r. First, see that clustering increased coverage dramatically, from only 0.28 in the unclustered regime to 0.70 in the clustered regime. This implies severe mismeasurement of standard errors prior to the adoption of clustering, with fewer than one in three published studies having intervals

(i) the *Industrial and Labor Relations Review*, which is not in the Brodeur et al. (2020) sample; and the *Journal of Public Economics*, which did not require authors to submit data and code over the 2015–2018 period. I included data from *Applied Economic Journal: Applied Economics* and *Applied Economic Journal: Economic Policy* due to a small sample size based on the four overlapping journals alone. These journals: (i) published a high share of DiD studies over this period; and (ii) required replication materials for publication.

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

	Unclustered ($\hat{r} = 0.51$)	Clustered ($r = 1$)	Change
A Selective Regime ($\hat{\beta}_p = 0.016$)			
Coverage	0.28	0.70	0.41
Estimated Treatment Effect	6.25 (100%)	12.74(100%)	6.48(100%)
Expected Bias	1.23 (19.6%)	2.44 (19.2%)	1.21 (18.7%)
Expected Θ	5.03 (80.4%)	10.3 (80.8%)	5.27 (81.3%)
B Non-Selective Regime ($\beta_p = 1$)			
Coverage	0.68	0.95	0.27
Estimated treatment effect	2.74(100%)	2.74(100%)	0
Expected Bias	0.00 (0%)	0.00 (0%)	0
Expected Θ	2.74 (100%)	2.74 (100%)	0

Notes: These figures are based on the parameter estimates of the empirical model in Table 3. Figures are calculated by simulating published studies under unclustered and clustered regimes. Panel A shows results for the estimated level of publication bias against insignificant results at the 5% level. Panel B shows results for a counterfactual scenario where publication is entirely non-selective with respect to statistical significance.

containing the true effect. Note also that while coverage improves substantially, it still remains, at 0.70, below nominal coverage of 0.95 due to selective publication.

The model shows that estimated treatment effects increased from 6.25% to 12.74% with clustering. This change can be decomposed into the sum of the change in bias and the change in the size of true treatment effects. Based on this decomposition, around four-fifths of the change in estimates reflects a shift in the types of questions addressed toward those with larger true effects. Larger true effect sizes are selected for in the clustered regime because they are, all else equal, more likely to lead to statistically significant results. The remainder of the increase in estimated treatment effects reflects an increase in bias, which accounts for about one fifth of the total change. This is a large increase and represents a doubling in the magnitude of bias, from 1.23 ppts to 2.44 ppts. One way to gauge the size of this change is to ask what level of publication bias would be required to induce the same increase in bias. In other words, what fraction of insignificant results (with correctly measured standard errors) would need to be censored to increase bias by 1.21 ppts? I find that 85% of null results would need to be censored (i.e. $\beta_p = 0.15$). This implies that the increase in bias from clustering is comparable to quite severe levels of publication bias against null results.

Robustness results based on the empirical distribution for r are presented in Table H1 in Appendix H. In this alternative approach, the degree of bias of unclustered studies is drawn randomly from the distribution of r in Figure 5, such that r varies across unclustered studies.

Results are quantitative similar to those in Table 4. In particular, clustering improves coverage from 0.36 to 0.70 and bias increases by 1.07 ppts. Alternatively, assuming that unclustered studies are downward biased by a constant factor equal to the mean of the empirical distribution ($\hat{r} = 0.76$) yields qualitatively similar results, but somewhat smaller changes in both coverage and bias. For more details, see Appendix H.

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

4.5. Non-Selective Publication

A common recommendation to combat distortions arising from publication bias is to implement reforms to publish all results, irrespective of their statistical significance. For example, implementing results-blind peer review (Chambers, 2013; Foster et al., 2019), launching journals dedicated to publishing insignificant findings[24], and even offering cash incentives for publishing null findings (Nature 2020).

To analyze the impact of these reforms in the DiD literature, I perform a counterfactual analysis where we assume no selective publication. In other words, I perform the same empirical exercise as for the main results, but set $\beta_p = 1$ such that no insignificant studies are censored. Panel B in Table 4 presents the results. When publication is non-selective, there exists no trade-off between coverage and bias when clustering. Coverage increases from 0.68 to reach nominal coverage of 0.95, and estimated treatment effects are unbiased in either regime. The welfare implications, however, are not clear. In particular, publishing all results is not necessarily without drawbacks. This is because non-selective publication leads to many published studies with small true treatment effects that are very imprecisely measured, and hence relatively uninformative for decision-makers who rely on empirical evidence from published studies to make policy choices. As noted in Frankel and Kasy (2022), if publication comes at a cost (e.g. the opportunity cost of drawing attention away from other studies due to limited journal space), then it is not necessarily the case that the non-selective regime is preferable to the selective regime. To better understand the impact of clustering on welfare, I develop a treatment choice model in the next section to evaluate the impact of clustering on decision-making in a policy context.

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

5. Impact of Clustering on Evidence-Based Policy

The empirical model in Section 4 suggest that clustering led to large improvements in coverage but also substantially higher bias. What are the implications of this for evidence-based policy? In this section, I develop a model of a policymaker who chooses whether to implement a policy based on evidence from published studies, but who may overestimate the precision of estimates when standard errors are unclustered. I consider a policymaker who makes decisions with the aim to minimize maximum regret i.e. the expected welfare loss from making an inferior treatment choice. I derive the minimax decision rule in the clustered and unclustered regimes, and then compare minimax regret across regimes. The main finding is that clustering lowers minimax regret if and only if the policymaker has sufficiently high loss aversion with respect to mistakenly implementing an ineffective or harmful policy i.e. of committing Type I error. Overall, the results suggest that clustering is beneficial if the cost of Type I error is specified in a way that is consistent with hypothesis testing using a 5% significance threshold.

5.1. Setup

The basic setup is the same as in Kitagawa and Vu (2023), who extend the model of minimax regret decision-makers in Manski (2004) and Tetenov (2012) to include publication bias. The model presented here makes a further extension to include the possibility that reported standard errors are downward biased (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 θ. 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 ($\theta \leqslant 0$). The policymaker's utility from an action a with average treatment effect θ is given by

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

where $K \geqslant 1$ measures the policymaker's loss aversion. As K increases, the policymaker weighs

[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 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.

the utility cost of committing Type I error (implementing the policy when $\theta \leqslant 0$) increasingly high relative to Type II error (not implementing the policy when $\theta > 0$). As a benchmark, note that classical hypothesis testing is consistent with a high degree of loss aversion from Type I error. In particular, regret from committing Type I error would need to be weighed around 100 times more than Type II regret for a decision rule that minimizes maximum regret to be consistent hypothesis testing with a 5% statistical significance threshold (Tetenov, 2012).

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

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

We turn next to see how these beliefs affect the policymaker's decision rule. Let $\delta_1 : X \to [0, 1]$ be the statistical decision rule in the event that a study is published. 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 X is above some chosen threshold T i.e. $\delta_1^T(X) = \mathbb{1}\{X > T\}$.[26] Thus the expected welfare of the threshold rule δ_1^T under the misspecified belief that X is normal and the *potentially* misspecified belief about Σ, is equal to

[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.

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

Before deriving a decision rule, it is necessary to adopt a framework for dealing with the uncertainty of θ. 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 θ and chooses a threshold T to maximize (misspecified) expected welfare: $\int \widetilde{W}\big(\delta_1^T, \theta, \widetilde{\sigma}\big)\pi(\theta)d\theta$.

However, in many situations, policymakers may have insufficient information to form a reasonable prior or priors may conflict when decisions are made by members of a group. In this situation, a common alternative, which I use here, 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_t^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:

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

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

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

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

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

Note that this expression is also misspecified, in that the policymaker makes no inferences about the fact that a study might have been censored. Similarly to the event where a study is

published, the no-data decision rule is obtained by the following optimization

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

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

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

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

Overall, the policymaker's minimax decision rule (T^*, δ_0^*) covers both realizations of the publication process and is chosen according to (6) and (8). Next, we compare minimax regret decision rules between clustered and unclustered regimes.

5.2. Minimax Regret Decision Rule

The follow result gives the minimax decision rule under misspecified regret:

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

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

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

Figure 6 illustrates Lemma 1 assuming the level of publication bias ($\hat{\beta}_p = 0.016$) and downward bias in standard errors ($\hat{r} = 0.51$) estimated in the empirical model for 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 $K > 1$. For intuition, note that the threshold rule is closely connected to the reported standard error (equation (9)). When results are perceived by the policymaker to be more precise, the estimate is believed to convey more information about the average treatment effect and hence a less conservative threshold rule can be implemented. Thus, in the unclustered regime, the policymaker overestimates the precision of evidence from published studies and is therefore too lenient with their threshold rule for implementing the policy. The absolute size of the difference increases with Type I error loss aversion. This is

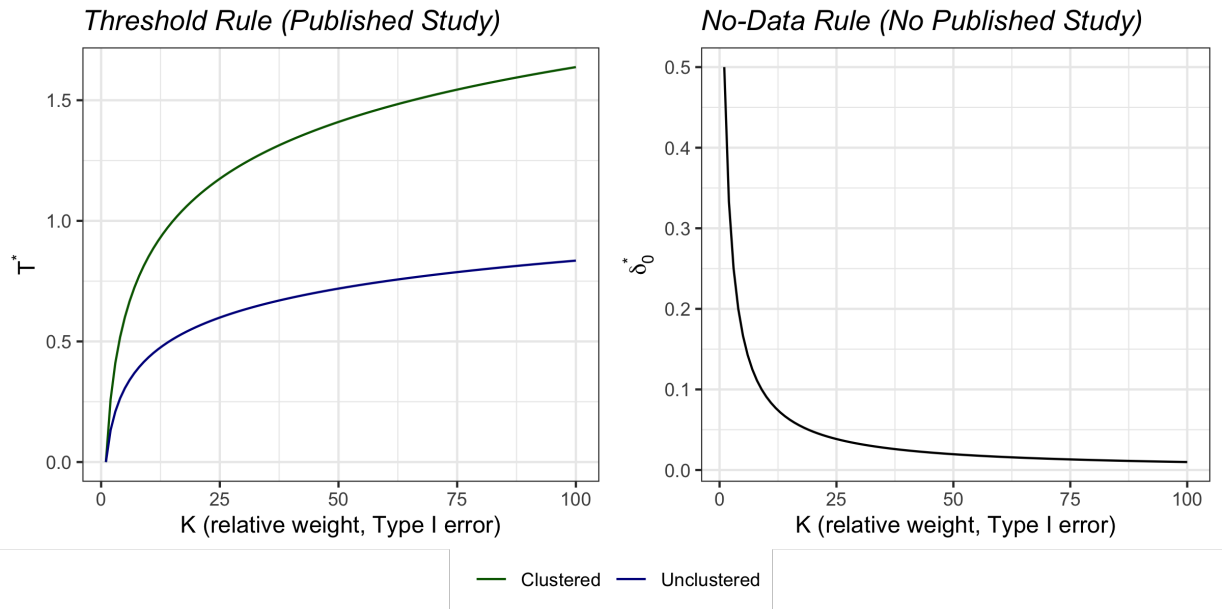

FIGURE 6. Minimax Regret Decision Rule in Clustered and Unclustered Regimes

Notes: The first panel shows the threshold rule in the event that a study is published and given by equation (6). The second panel shows the no-data rule in even that a study is not published. The level of publication bias $\hat{\beta}_p = 0.016$ 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.

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

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

5.3. Comparing Regimes Based on True Regret

While the minimax regret decision rule in Lemma 1 is based on misspecified regret, we evaluate any given decision rule (T, δ_0) based on its *true regret*. True regret is derived from accurate beliefs about X, 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 β_p). The utility of action a_1 when a study is published and action a_0 when it is not, is given by

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

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

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

where $\mathbb{P}[D = 1|\theta, \widetilde{\sigma}]$ is the ex-ante publication probability conditional on $(\theta, \widetilde{\sigma})$; and $F(\cdot|\theta, \sigma, \widetilde{\sigma}, D = 1)$ is the cdf of a truncated normal distribution.[27] See that the probability of publication is based on the *reported* standard error and thus effect significance thresholds 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 θ, true (i.e. clustered) standard error σ, and the Type I error loss aversion parameter K, regret is given by the following expression:

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

Thus, true regret is equal to the ex-ante probability of making an the incorrect treatment choice multiplied by the cost of the mistake $|\theta|$, 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 'should' be using to choose their decision rule in order to minimize maximum regret. The minimax regret of any decision rule (T, δ_0) given σ is given by

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

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

[27] Specifically, the cdf is given by

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

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

Figure 7 plots this for different values of the Type I error loss aversion parameter K. Results show that clustering lowers minimax regret if and only if $K > 63$. Recall that classical hypothesis testing at the 5% level entails a much larger level of loss aversion to Type I error i.e. $K = 102.4$. Thus, the model suggests that clustering increased welfare if we use the benchmark cost implicitly implied by 5% hypothesis testing.

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 chances that policymakers are forced to make decisions without evidence, since larger reported standard errors lead to more insignificant results being censored. Suppose that $K = 1$. In this unique case, the threshold rule is identical across regimes ($T^* = 0$) and thus clustering provides no advantage. However, the probability of publication is lower in the clustered regime such that minimax regret is substantially larger than in the unclustered regime. As K increases, however, the trade-off described above gradually favors clustering. This is because the threshold rule in the unclustered regime becomes increasingly miscalibrated as K increases, which leads to larger

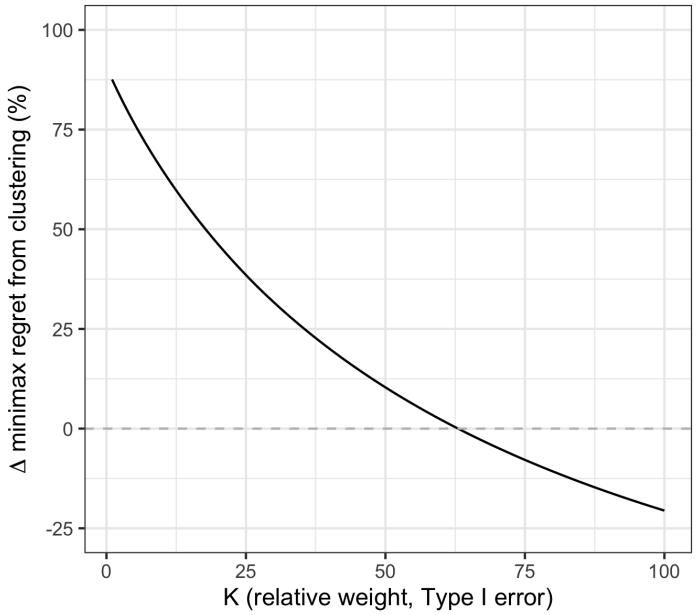

FIGURE 7. Percent Change in Minimax Regret from Clustering

Notes: The percent change in minimax regret moving from the unclustered regime to the clustered regime is calculated according to equation (14). The level of publication bias $\hat{\beta}_p = 0.016$ 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.

costs in terms of minimax regret. When K is above than 63, minimax regret in the clustered regime is lower than in the unclustered regime.

6. Conclusion

The econometrics literature on standard error corrections and the meta-science literature on publication bias share the same goal of improving the credibility of published research. However, an underappreciated issue is that the interaction between corrections and publication bias can have important implications for the statistical credibility of published studies. This paper systematically studies these issues both theoretically and empirically.

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

References

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey M. Wooldridge, "When Should You Adjust Standard Errors for Clustering?," *The Quarterly Journal of Economics*, 2023, pp. 1–35.

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.

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.

Cameron, Miller A. and Douglas L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources*, 2015, *50* (2), 317–372.

Card, David and Alan B. Krueger, "Time-Series Minimum-Wage Studies: A Meta-analysis," *American Economic Review: Papers and Proceedings*, 1995, *85* (2), 238–243.

Chambers, Christopher D., "Registered Reports: A new publishing initiative at Cortex," *Cortex*, 2013, *49* (3), 609–610.

Currie, Janet, Henrik Kleven, and Esmee Zwiers, "Technology and Big Data Are Changing Economics: Mining Text to Track Methods," *AEA Papers and Proceedings*, 2020, *110*, 42–48.

DellaVigna, Stefano and Elizabeth Linos, "RCTs to Scale: Comprehensive Evidence From Two Nudge Units," *Econometrica*, 2022, *90* (1), 81–116.

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

Foster, Andrew, Dean Karlan, Edward Miguel, and Aleksandar Bogdanoski, "Pre-results Review at the Journal of Development Economics: Lessons Learned So Far," *World Bank Development Impact Blog*, 2019.

Franco, Annie, Neil Malhotra, and Gabor Simonovits, "Publication bias in the social sciences: Unlocking the file drawer," *Science*, 2014, *345* (6203), 1502–1505.

Frankel, Alexander and Maximilian Kasy, "Which Findings Should Be Published?," *American Economic Journal: Microeconomics*, 2022, *14* (1), 1–38.

Gelman, Andrew and John Carlin, "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors," *Perspectives on Psychological Science*, 2014, *9* (6), 641–651.

Ioannidis, John P. A., T. D. Stanley, and Hristos Doucouliagos, "The Power of Bias in Economics Research," *The Economic Journal*, 2017, *127* (605), 236–265.

Ioannidis, John P.A., "Why Most Published Research Findings Are False," *PLoS Med*, 2005, *2* (8).

_ , "Why Most Discovered True Associations Are Inflated," *Epidemiology*, 2008, *19* (5), 640–648.

Kahneman, Daniel and Amos Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 1979, *47* (2), 263–292.

Karlin, Samuel and Herman Rubin, "The Theory of Decision Procedures for Distributions with Monotone Likelihood Ratio," *The Annals of Mathematical Statistics*, 1956, *27* (2), 272–299.

Kitagawa, Toru and Alex Tetenov, "Who Should Be Treated? Empirical Welfare Maximization Methods for Treatment Choice," *Econometrica*, 2018, *86* (2), 591–616.

_ **and Patrick Vu**, "At What Level Should One Cluster Standard Errors in Paired Experiments, and in Stratified Experiments with Small Strata?," *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?," *Working Paper*, 2023.

Wald, Abraham, *Statistical Decision Functions*, New York: John Wiley & Sons, 1950.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 1980, *48* (4), 817–838.

Appendix

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

A. Proofs

Proof of Proposition 1: This 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 $r \in (0, r_1)$ individual-study bias increases:

$$\mathbb{E}[X^* - \Theta^* | D = 1, \widetilde{\Sigma}^* = 1] - \mathbb{E}[X^* - \Theta^* | D = 1, \widetilde{\Sigma}^* = r] > 0$$

Next, Lemma A.3 claims that there exists an $r_2 \in (0, 1]$ such that for any $r \in (0, r_2)$ true treatment effects weakly increase:

$$\mathbb{E}[\Theta^* | D = 1, \widetilde{\Sigma}^* = 1] - \mathbb{E}[\Theta^* | D = 1, \widetilde{\Sigma}^* = r] \geqslant 0$$

Define $r^* = \min\{r_1, r_2\}$. It follows that for any $r \in (0, r^*)$, individual-study bias will increase and true treatment effects will weakly increase. This immediately implies that the change in estimated treatment effects, $\mathbb{E}[X^* | D = 1, \widetilde{\Sigma}^* = 1] - \mathbb{E}[X^* | D = 1, \widetilde{\Sigma}^* = r]$, is positive since it is simply equal to the sum of the change in individual study bias and true treatment effects. Finally, since the change in meta-study bias is identical to the change in estimated treatment effects, it must also increase.

Below, I present Lemmas A.2 and A.3 on which this argument is based. Lemma A.1 is presented beforehand and used in Lemma A.2.

Lemma A.1 (Expression for Bias)**.** *For a given $\theta \in [0, \infty)$, $\beta_p \in [0, 1)$ and $r \in (0, 1]$,*

$$Bias(\theta, \beta_p, r) = \frac{(1-\beta_p)\big[\phi(1.96r-\theta) - \phi(\theta+1.96r)\big]}{\Phi(-1.96r-\theta) + \beta_p\big[\Phi(1.96r-\theta) - \Phi(-1.96r-\theta)\big] + 1 - \Phi(1.96r-\theta)} \quad (15)$$

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

Proof. Define $Z^* = X^* - \theta$ such that $Z^* \sim N(0,1)$. See that bias is equal to $\mathbb{E}[Z^*|D=1, \widetilde{\Sigma}^* = r] = \mathbb{E}[X^*|D=1, \widetilde{\Sigma}^* = r] - \theta$. We can write bias as the weighted sum of three truncated standard normals:

$$\mathbb{E}[Z^*|D=1, \widetilde{\Sigma}^* = r] = \mathbb{P}[Z^* \leqslant -1.96r - \theta | D=1, \widetilde{\Sigma}^* = r] \cdot \mathbb{E}[Z^*|Z^* \leqslant -1.96r - \theta]$$

$$+\mathbb{P}[-1.96r - \theta < Z^* \leqslant 1.96r + \theta | D=1, \widetilde{\Sigma}^* = r] \cdot \mathbb{E}[Z^*| -1.96r - \theta < Z^* \leqslant 1.96 * r + \theta]$$

$$+\mathbb{P}[Z^* > -1.96r - \theta | D=1, \widetilde{\Sigma}^* = r] \cdot \mathbb{E}[Z^*|Z^* > -1.96r - \theta]$$

$$= -\frac{\phi(-1.96r-\theta)}{\mathbb{P}(D=1|\widetilde{\Sigma}^* = r)} + \frac{\beta_p\big[\phi(-1.96r-\theta) - \phi(1.96r-\theta)\big]}{\mathbb{P}(D=1|\widetilde{\Sigma}^* = r)} + \frac{\phi(1.96r-\theta)}{\mathbb{P}(D=1|\widetilde{\Sigma}^* = r)}$$

where the second equality uses Bayes' Rule on the probability terms and the formula for the expectation of a truncated standard normal on the expectation terms. Simplifying this expression gives the result. □

Lemma A.2 (Sufficient Condition for Increase in Individual-Study Bias)**.** *Under Assumptions 1, 2, and 3, there exists an $r_1 \in (0,1]$ such that for any $r \in (0, r_1)$ individual-study bias increases.*

Proof. First, I show that $\mathbb{E}[X^*|D=1, \widetilde{\Sigma}^* = r] \to \mathbb{E}[\Theta^*]$ as $r \to 0$. Using Bayes Rule, we have that

$$\mathbb{E}[X^*|D=1, \widetilde{\Sigma}^* = r] = \int \left(\frac{x \cdot p\left(\frac{x}{r}\right) \int_\theta \phi(x-\theta) f_\Theta(\theta) d\theta}{\int_\theta \mathbb{P}[D=1|\widetilde{\Sigma}^* = r, \Theta^* = \theta] f_\Theta(\theta) d\theta} \right) dx \quad (16)$$

Consider the integrand. First, see that the numerator approaches $x \int_\theta \phi(x-\theta) f_\Theta(\theta)$ as $r \to 0$. Next, see that the denominator satisfies

$$\lim_{r \to 0} \int_\theta \mathbb{P}[D=1|\widetilde{\Sigma}^* = r, \Theta^* = \theta] f_\Theta(\theta) d\theta = 1$$

This equality uses the dominated convergence theorem to move the limit inside the integral and the fact that the probability of publication conditional on $\Theta^* = \theta$ approaches one as $r \to 0$ (since no results are censored in the limit). To see that the conditions for the dominated

convergence theorem are met, first see that the integrand converges pointwise to $f_\Theta(\theta)$ as $r \to 0$. Second, see that for any $r \in (0,1]$ and $\theta \geqslant 0$, the integrand is bounded above by $f_\Theta(\theta)$ because $\mathbb{P}[D=1|\widetilde{\Sigma}=r,\Theta^*=\theta] \leqslant 1$.

From this, it follows that the integrand in equation (16) converges pointwise to $x\int_\theta \phi(x-\theta)f_\Theta(\theta)$ as $r \to 0$.

Next, see that for any $r \in (0,1]$ and $x \in \mathbb{R}$, the absolute value of the integrand in equation (16) satisfies

$$\frac{|x| \cdot p\left(\frac{x}{r}\right)\int_\theta \phi(x-\theta)f_\Theta(\theta)d\theta}{\int_\theta \mathbb{P}[D=1|\widetilde{\Sigma}^*=r,\Theta^*=\theta]f_\Theta(\theta)d\theta} \leqslant \frac{|x|\cdot\phi(0)}{\int_\theta \mathbb{P}[D=1|\widetilde{\Sigma}^*=1,\Theta^*=\theta]f_\Theta(\theta)d\theta}$$

where the bound follows from the fact that $p\left(\frac{x}{r}\right) \leqslant 1$ and $\int_\theta \phi(x-\theta)f_\Theta(\theta)d\theta \leqslant \phi(0)$ in the numerator, and $\mathbb{P}[D=1|\widetilde{\Sigma}=r,\Theta^*=\theta]$ is strictly decreasing in r.

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

$$\lim_{r\to 0}\mathbb{E}[X^*|D=1,\widetilde{\Sigma}^*=r] = \int_x x\int_\theta \phi(x-\theta)f_\Theta(\theta)d\theta dx$$

$$= \int_\theta\left(\int_x x\phi(x-\theta)dx\right)f_\Theta(\theta)d\theta = \int_\theta \mathbb{E}[X^*|\Theta^*=\theta]f_\Theta(\theta)d\theta = \mathbb{E}[\Theta^*] \tag{17}$$

which is what we wanted to show.

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

$$\mathbb{E}[\Theta^*|D=1\widetilde{\Sigma}^*=r] = \int_\theta\left(\frac{\theta\cdot\mathbb{P}[D=1|\widetilde{\Sigma}^*=r,\Theta^*=\theta]f_\Theta(\theta)}{\int_\theta \mathbb{P}[D=1|\widetilde{\Sigma}^*=r,\Theta^*=\theta]f_\Theta(\theta)d\theta}\right)d\theta$$

Now see that the integrand converges pointwise to $\theta f_\Theta(\theta)$ as $r \to 0$. This follows because $\lim_{r\to 0}\mathbb{P}[D=1|\widetilde{\Sigma}=r,\Theta^*=\theta]=1$ in the numerator and because the denominator converges to one, as shown earlier.

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

$$\frac{\theta\cdot\mathbb{P}[D=1|\widetilde{\Sigma}^*=r,\Theta^*=\theta]f_\Theta(\theta)}{\int_\theta \mathbb{P}[D=1|\widetilde{\Sigma}^*=r,\Theta^*=\theta]f_\Theta(\theta)d\theta} \leqslant \frac{\theta f_\Theta(\theta)}{\int_{\theta'}\mathbb{P}[D=1|\widetilde{\Sigma}^*=1,\Theta^*=\theta']f_\Theta(\theta')d\theta'}$$

where the inequality follows from the fact that $\mathbb{P}[D=1|\widetilde{\Sigma}=r,\Theta^*=\theta] \leqslant 1$ and $\mathbb{P}[D=1|\widetilde{\Sigma}=r,\Theta^*=\theta]$ is decreasing in r. Note that the upper bound is integrable since Assumption 1 requires Θ^* to have a finite first moment. Thus, appealing again to the dominated convergence theorem, we have

$$\lim_{r\to 0} \mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=r] = \int_\theta \theta f_\Theta(\theta)d\theta = \mathbb{E}[\Theta^*] \tag{18}$$

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

$$\Delta\text{Bias}(r) \equiv \mathbb{E}[X^* - \Theta^*|D=1,\widetilde{\Sigma}^*=1] - \mathbb{E}[X^* - \Theta^*|D=1,\widetilde{\Sigma}^*=r]$$

$$\to \mathbb{E}[X^* - \Theta^*|D=1,\widetilde{\Sigma}^*=1] = \int_\theta \text{Bias}(\theta,\beta_p,1) f_\Theta(\theta)d\theta > 0 \tag{19}$$

as $r \to 0$. The final inequality follows because it is clear from Lemma A.1 that $\text{Bias}(\theta,\beta_p,1) \geqslant 0$ when $\beta_p \in [0,1)$ (Assumption 3) and $\theta \geqslant 0$, and with strict inequality when $\theta > 0$. Assumption 1 requires that there exists some $\theta > 0$ on the support of Θ^*, which gives the strict inequality.

Now we can prove the main claim. Define the 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 show this by contradiction. Suppose instead that there exists an $\bar{r} \in (0, r_1)$ where

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

where the second inequality follows from equation (19). Note that $\Delta\text{Bias}(r)$ is continuous in r over $(0,1)$ and well-defined for all $r \in (0,1]$. Thus, there must exist some $\epsilon \in (0,\bar{r})$ such that $\Delta\text{Bias}(\bar{r}) < 0 < \Delta\text{Bias}(\epsilon)$. 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 True Treatment Effects)**.** *Under Assumptions 1, 2, and 3, there exists an $r_2 \in (0,1]$ such that for any $r \in (0, r_2)$ the expected true treatment effect weakly increases.*

Proof. Consider two cases. First, suppose that the distribution of Θ^* is degenerate at some $\theta > 0$. Then for any $r \in (0,1]$

$$\Delta\text{TTE}(r) \equiv \mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=1] - \mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=r] = 0$$

Let $r_2 = 1$ such that for any $r \in (0,1)$ there is no change in true treatment effects with standard error corrections: $\Delta\text{TTE}(r) = 0$.

Next, consider the case where the distribution of Θ^* is non-degenerate. See that

$$\begin{aligned}\lim_{r\to 0}\Delta\text{TTE}(r) &= \mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=1] - \lim_{r\to 0}\mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=r] \\ &= \mathbb{E}[\Theta^*|D=1,\widetilde{\Sigma}^*=1] - \mathbb{E}[\Theta^*] \\ &= \int_0^\infty [F_\Theta(t) - F_{\Theta|D,\widetilde{\Sigma}}(t|D=1,\widetilde{\Sigma}^*=1)]dt \end{aligned} \tag{20}$$

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

$$\begin{aligned}&\int_0^t f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)d\theta - \int_0^t f_\Theta(\theta)d\theta \\ &= \frac{1}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=1)}\left(\int_0^t \mathbb{P}(D=1|\Theta^*=\theta,\widetilde{\Sigma}^*=1)f_\Theta(\theta)d\theta - \mathbb{P}(D=1|\widetilde{\Sigma}^*=1)\int_0^t f_\Theta(\theta)d\theta\right) \\ &= \frac{F_\Theta(t)}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=1)}\left(\mathbb{E}_\Theta\Big[\mathbb{P}(D=1|\Theta^*,\widetilde{\Sigma}^*=1)\Big|\Theta^*\leqslant t)\Big] - \mathbb{E}_\Theta\Big[\mathbb{P}(D=1|\Theta^*,\widetilde{\Sigma}^*=1)\Big]\right) \leqslant 0\end{aligned}$$

where the first equality uses Bayes' Rule and the last weak inequality follows from the fact that $\mathbb{P}(D=1|\theta,\widetilde{\Sigma}^*=1)$ is an increasing function of θ.[28] Since Θ^* is non-degenerate, there exists some $t \in [0,\infty)$ for which this inequality is strict. This implies that equation (20) is strictly positive, which is what we wanted to show.

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

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

[28] The derivative is given by:

$$\frac{\partial}{\partial\theta}\Big[\mathbb{P}(D=1|\theta,\widetilde{\Sigma}^*=1)\Big] = (1-\beta_p)\Big(\phi(1.96-\theta)-\phi(1.96+\theta)\Big) \geqslant 0$$

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

where the second inequality follows from the arguments above. Note that $\Delta\text{TTE}(r)$ 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 $\Delta\text{TTE}(\bar{r}) < 0 < \Delta\text{TTE}(\epsilon)$. It follows from the intermediate value theorem that there exists an $r' \in (\epsilon, \bar{r})$ such that $\Delta\text{TTE}(r') = 0$ 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: As a starting point, the following Lemma provides an expression for average coverage in published studies for a fixed true effect, which will be used throughout the proof.

Lemma A.4 (Expression for Coverage with Degenerate Θ^*)**.** *For any $\theta \in [0, \infty)$, $r \in (0,1]$ and $\beta_p \in [0,1]$, the expected coverage is published studies is given by*

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

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

$$\begin{aligned} \text{Coverage}(\theta, r) &= \mathbb{P}[X^* - 1.96r \leqslant \theta \leqslant X^* + 1.96r | D = 1, \widetilde{\Sigma}^* = r] \\ &= \int_{\theta-1.96r}^{\theta+1.96r} f_{X|D,\Theta,\widetilde{\Sigma}}(x|D=1, \Theta^* = \theta, \widetilde{\Sigma}^* = r)dx \\ &= \frac{\int_{\theta-1.96r}^{\theta+1.96r} \mathbb{P}(D=1|X^*=x, \Theta^*=\theta, \widetilde{\Sigma}^*=r)\phi(x-\theta)dx}{\mathbb{P}(D=1|\Theta^*=\theta, \widetilde{\Sigma}^*=r)} \end{aligned}$$

using Bayes Rule in the last equality. Recall that statistically significant results are published with probability one and insignificant results with probability $\beta_p \in [0,1)$ (Assumption 3). Taking the integral in both the denominator and numerator of the equation above gives the expression for coverage in equation (21). □

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

Lemma A.5 (Publication Regime as Mixed Distribution)**.** *Let the density of published studies in publication regime $\beta_p \in [0,1]$ be denoted by $f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;\beta_p)$. This density is equivalent to the following mixture of densities:*

$$\begin{aligned} &f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;\beta_p) \\ &= \omega(r) f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;1) + \big(1-\omega(r)\big) f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;0) \end{aligned}$$

with

$$\omega(r) = \frac{\beta_p}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p)} \in [0,1] \tag{22}$$

Proof. This is trivially true in the case where $\beta_p = 0$ or $\beta_p = 1$. Let $\beta_p \in (0,1)$. With Bayes Rule and Assumption 3 which assumes a step-wise publication selection function, we have that

$$\begin{aligned} f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;\beta_p) &= \frac{\mathbb{P}(D=1|X^*=x,\widetilde{\Sigma}^*=r;\beta_p)\phi(x-\theta)f_\Theta(\theta)}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p)} \\ &= \frac{\mathbb{1}\{|x| \geqslant 1.96r\}\phi(x-\theta)f_\Theta(\theta) + \mathbb{1}\{|x|<1.96r\}\beta_p\phi(x-\theta)f_\Theta(\theta)}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p)} \end{aligned} \tag{23}$$

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

$$\begin{aligned} &\omega(r)\cdot f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;1) + (1-\omega(r))\cdot f_{X,\Theta|D,\widetilde{\Sigma}}(x,\theta|D=1,\widetilde{\Sigma}^*=r;0) \\ &= \underbrace{\left(\frac{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p) - \beta_p\big(1-\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;0)\big)}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;0)}\right)}_{\equiv\kappa}\left(\frac{\mathbb{1}\{|x|\geqslant 1.96r\}\phi(x-\theta)f_\Theta(\theta)}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p)}\right) \\ &\quad + \left(\frac{\mathbb{1}\{|x|<1.96r\}\beta_p\phi(x-\theta)f_\Theta(\theta)}{\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p)}\right) \end{aligned}$$

It is clear that this expression equals the density in the publication regime $\beta_p \in (0,1)$ in equation (23) provided that $\kappa = 1$. This is easily verified by using the fact that $\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;\beta_p) = \mathbb{P}(D=1|\widetilde{\Sigma}^*=r;0) + \beta_p(1-\mathbb{P}(D=1|\widetilde{\Sigma}^*=r;0))$. □

In the next step, I show that Lemma A.5 implies we only need to show that coverage

improves with standard error corrections in the publication regime where $\beta_p = 0$. For clarity, let expected coverage in publication regime $\beta_p \in [0,1]$ and standard error regime $r \in (0,1]$ be denoted by

$$c_{\beta_p}(r) \equiv \int \text{Coverage}(\theta, r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1, \widetilde{\Sigma}^* = r; \beta_p) d\theta$$

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

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

where the inequality uses the fact that $c_1(1) - c_1(r) = [\Phi(1.96) - \Phi(-1.96)] - [\Phi(1.96r) - \Phi(-1.96r)] > 0$, and $\omega(1) > \omega(r)$ because the probability of publication in the denominator for the weight is clearly decreasing in r. Thus, we only need to show that coverage improves in the case where $\beta_p = 0$ to show that coverage improves overall in publication regime $\beta_p \in [0,1)$.

Fix $\beta_p = 0$ for the reminder of the proof. We want to show that expected coverage improves with standard error corrections:

$$\begin{aligned} &c_0(1) - c_0(r) \\ &= \int_0^\infty \text{Coverage}(\theta,1) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)d\theta - \int_0^\infty \text{Coverage}(\theta,r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)d\theta \\ &= \left(\int_0^{2\times 1.96r} \text{Coverage}(\theta,1) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)d\theta - \int_0^{2\times 1.96r} \text{Coverage}(\theta,r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)d\theta \right) \\ &+ \left(\int_{2\times 1.96r}^\infty \text{Coverage}(\theta,1) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)d\theta - \int_{2\times 1.96r}^\infty \text{Coverage}(\theta,r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)d\theta \right) \end{aligned} \tag{24}$$

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

Consider the second difference, where the integrals are over $\theta \geqslant 2 \times 1.96r$. Consider the *integrand* in the second term of the difference (and keep the integral limits fixed). Using the

expression for coverage when $\theta \geqslant 2 \times 1.96r$ from Lemma A.4 and Bayes' Rule we have that the integrand is equal to

$$\text{Coverage}(\theta, r) f_{\Theta|D,\widetilde{\Sigma}}(\theta | D = 1, \widetilde{\Sigma}^* = r) = \left(\frac{\Phi(1.96r) - \Phi(-1.96r)}{\mathbb{P}(D = 1 | \widetilde{\Sigma}^* = r; \beta_p)} \right) \cdot f_\Theta(\theta)$$

Consider the term in parentheses. The numerator is clearly increasing in r and the denominator clearly decreasing in r. Since both terms are strictly positive, this implies that the derivative is weakly increasing in r (and strictly positive when $f_\Theta(\theta) > 0$). In equation (24), this immediately implies that the difference in the second parentheses is weakly positive.

Next, I show that the first difference in (24) is positive. To do so, I make use of three Lemmas.

Lemma A.6 (Coverage Improves for Degenerate Θ^*)**.** *Let $\beta_p = 0$. For any $\theta \in (0, \infty)$ and $r \in (0, 1]$, we have*

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

Proof. Let $c = 1.96r$ denote the critical significance threshold. The case where $\theta \geqslant 2c$ has already been shown in the main text of the proof for the more general case where Θ^* follows a distribution. Consider then the case where $\theta \in (0, 2c)$. The expression for coverage (Lemma A.4) when $\beta_p = 0$ is given by

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

Taking the derivative with respect to c gives

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

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

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

$$\frac{\phi(c + \theta) + \phi(c - \theta)}{1 - \Phi(c + \theta) + 1 - \Phi(c - \theta)} \geqslant \frac{\phi(c - \theta) - \phi(c)}{\Phi(c) - \Phi(c - \theta)} \tag{25}$$

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

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

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

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

For $\theta \geqslant 0$, we have that $\mu_1 \geqslant \mu_2 \geqslant \mu_3$. Now let

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

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

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

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

Proof. We will show that this holds for any $\beta_p \in [0,1]$. Let $c = 1.96r$ denote the critical significance threshold for greater clarity in the expressions below. Consider the expressions for coverage in Lemma A.4. Consider first the case where $\theta \leqslant 2c$. Using the quotient rule gives

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

where $d(\theta,c) \equiv \Phi(-c-\theta) + 1 - \Phi(c-\theta) + \beta_p[\Phi(c-\theta) - \Phi(-c-\theta)] > 0$ and $n_1(\theta,c) \equiv \beta_p[\Phi(c-\theta)-\Phi(-c)] + \Phi(c) - \Phi(c-\theta) > 0$. The inequality follows because $d(\theta,c) > n_1(\theta,c)$ and $\phi(c-\theta) > \phi(c-\theta) - \phi(c+\theta) > 0$).

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

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

□

Lemma A.8 (First Order Stochastic Dominance in Corrected Standard Error Regime)**.** *Let $\beta_p = 0$ and $F_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)$ denote the distribution of published true treatment effects in standard error regime $r \in (0,1)$. Then for any $r \in (0,1)$, $F_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)$ first-order stochastically dominates $F_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)$.*

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

$$\frac{f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)}{f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)} = \frac{\left(\frac{\mathbb{P}[D=1|\Theta^*=\theta,\widetilde{\Sigma}^*=1]f_\Theta(\theta)}{\mathbb{P}[D=1|\widetilde{\Sigma}^*=1]}\right)}{\left(\frac{\mathbb{P}[D=1|\Theta^*=\theta,\widetilde{\Sigma}^*=r]f_\Theta(\theta)}{\mathbb{P}[D=1|\widetilde{\Sigma}^*=r]}\right)}$$

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

where $c \equiv 1.96r$ and $K \equiv \mathbb{P}[D=1|\widetilde{\Sigma}^*=r]/\mathbb{P}[D=1|\widetilde{\Sigma}^*=1] > 0$. Thus the derivative with respect to θ is given by

$$\frac{\partial}{\partial\theta}\left(\frac{f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=1)}{f_{\Theta|D,\widetilde{\Sigma}}(\theta|D=1,\widetilde{\Sigma}^*=r)}\right) \propto \frac{\partial}{\partial\theta}\Big(\Phi(-1.96-\theta)+1-\Phi(1.96-\theta)\Big)\Big(\Phi(-c-\theta)+1-\Phi(c-\theta)\Big)$$

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

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

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

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

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

Next, define

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

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

$$\frac{\partial}{\partial c}\Big(\mu_1(\theta,c)\cdot\mu_2(\theta,c)\Big) \propto \frac{\partial}{\partial c}\Big(\mu_1(\theta,c)\Big)\Big(\mu_2(\theta,c)\Big) + \Big(\mu_1(\theta,c)\Big)\frac{\partial}{\partial c}\Big(\mu_2(\theta,c)\Big)$$

Showing that each of the four terms in this expression are positive is sufficient for proving the derivative is positive. First, see that $\mu_2(\theta,c)$ is clearly positive. Next, see that $\mu_1(\theta,c)$ is positive because it is the conditional expectation of a standard normal over an even interval centered

at $c > 0$. Moreover, the derivative $\partial\mu_1(\theta, c)/\partial c$ is positive because the conditional expectation increases when the interval over which the expectation is taken increases (i.e. shifts to the right). Finally, see that

$$\frac{\partial}{\partial c}\Big(\mu_2(\theta, c)\Big) \propto \frac{\partial}{\partial c}\Big(n(\theta, c)\Big)\Big(d(\theta, c)\Big) - \Big(n(\theta, c)\Big)\frac{\partial}{\partial c}\Big(d(\theta, c)\Big)$$
$$= \Big(\phi(c+\theta) - \phi(c-\theta)\Big)d(\theta, c) + n(\theta, c)\Big(\phi(c+\theta) + \phi(c-\theta)\Big)$$

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

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

This inequality holds because it is equivalent to $\mathbb{E}[Z|Z > c+\theta] \geqslant \mathbb{E}[Z|Z > c-\theta]$ with $Z \sim N(0, 1)$ by the truncated normal expectation formula, which clearly holds for $c > 0$ and $\theta \geqslant 0$.

Thus, $f_{\Theta|D,\widetilde{\Sigma}}(\theta|D = 1, \widetilde{\Sigma}^* = 1)/f_{\Theta|D,\widetilde{\Sigma}}(\theta|D = 1, \widetilde{\Sigma}^* = r)$ is increasing in θ and therefore satisfies the monotone likelihood ratio property. This implies first-order stochastic dominance, giving the desired result. □

Using these three Lemmas, we have that

$$\int_0^{2\times 1.96r} \text{Coverage}(\theta, 1) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D = 1, \widetilde{\Sigma}^* = 1)d\theta - \int_0^{2\times 1.96r} \text{Coverage}(\theta, r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D = 1, \widetilde{\Sigma}^* = r)d\theta$$
$$\geqslant \int_0^{2\times 1.96r} \text{Coverage}(\theta, r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D = 1, \widetilde{\Sigma}^* = 1)d\theta - \int_0^{2\times 1.96r} \text{Coverage}(\theta, r) f_{\Theta|D,\widetilde{\Sigma}}(\theta|D = 1, \widetilde{\Sigma}^* = r)d\theta \geqslant 0$$

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

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

Proof of Lemma 1: First, consider the threshold rule. Tetenov (2012) considers the case where the estimated treatment effect X 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 the identical beliefs, the results from Tetenov (2012) on page 160 immediately apply, despite the fact that those beliefs happen to be incorrect in this setting. (Note however that regret, which is based on the true distribution of studies, will differ in this setting compared to the setting in Tetenov (2012)).

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

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

Proposition 1 shows that when standard error corrections are sufficiently large, bias, estimated treatment effects and true treatment effects must all increase. This appendix shows that for small standard error corrections it is in fact possible for them to decrease. This is formalized in the following lemma:

Lemma B.1 (Ambiguous Impact on Bias, Estimated Treatment Effects, and True Treatment Effects)**.** *Under Assumptions 1, 2, and 3, the individual signs of the change from standard error corrections to individual-study bias, meta-study bias, estimated treatment effects, and true treatment effects are ambiguous. That is, there exist distinct combinations of* $(\mu_{\Theta,\Sigma}, \beta_p, r)$ *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 individual-study bias, meta-study bias, and estimated treatment effects. In the second, I do the same for true treatment effects.

(1) Individual-Study Bias, Meta-Study Bias, and Estimated Treatment Effects

Suppose that Θ^* follows a degenerate distribution with $\mathbb{P}[\Theta^* = \theta] = 1$ for some $\theta > 0$. This implies that the change in individual-study bias following standard error corrections will be equal to the change in meta-study bias and the change in estimated treatment effects:

$$\underbrace{\mathbb{E}[X^* - \theta | D = 1, \widetilde{\Sigma}^* = 1] - \mathbb{E}[X^* - \theta | D = 1, \widetilde{\Sigma}^* = r]}_{\Delta\text{Individual-Study Bias}=\Delta\text{Meta-Study Bias}} = \underbrace{\mathbb{E}[X^* | D = 1, \widetilde{\Sigma}^* = 1] - \mathbb{E}[X^* | D = 1, \widetilde{\Sigma}^* = r]}_{\Delta\text{Estimated Treatment Effects}} \tag{27}$$

We can use the expression for $\text{Bias}(\theta, \beta_p, r)$ from Lemma A.1 to show that the sign of equation (27) from standard error corrections is ambiguous i.e. the sign of $\text{Bias}(\theta, \beta_p, 1) - \text{Bias}(\theta, \beta_p, r)$ can be positive, negative or zero. Fix $(\beta_p, r) = (0.1, 0.75)$. Then for $\theta = 1.5$ and $\theta = 0.75$, 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 $\theta' \in (0.25, 1.5)$ such that $\text{Bias}(\theta', 0.1, 1) - \text{Bias}(\theta', 0.1, 0.75) = 0$.

(2) True Treatment Effects

Consider a two-point distribution for Θ^* where $\mathbb{P}[\Theta^* = \theta] = p_1^* \cdot \mathbb{1}\{\theta = \theta_1\} + (1 - p_1^*) \cdot \mathbb{1}\{\theta = \theta_2\}$ for $0 \leqslant \theta_1 < \theta_2$ and $p_1^* \in (0, 1)$. Then by Bayes' Rule we have

$$\text{TrueTE}(\theta_1, \theta_2, p_1^*, \beta_p, r) \equiv \mathbb{E}[\Theta^* | D = 1, \widetilde{\Sigma} = r] = \frac{p_1^* \theta_1 C(\theta_1, \beta_p, r) + (1 - p_1^*)\theta_2 C(\theta_2, \beta_p, r)}{p_1^* C(\theta_1, \beta_p, r) + (1 - p_1^*) C(\theta_2, \beta_p, r)}$$

where, as above, $C(\theta, \beta_p, r) \equiv \int_{z'} p\big(\frac{\theta + z'}{r}\big)\phi(z')dz'$ is the probability of publication for a given θ.

Now suppose $\theta_1 = 0$ and $p_1^* = 0.5$. Then the change in true treatment effects is given by

$$\text{TrueTE}(0, \theta_2, 0.5, \beta_p, 1) - \text{TrueTE}(0, \theta_2, 0.5, \beta_p, r)$$

$$= \theta_2 \left(\frac{C(\theta_2, \beta_p, 1)}{C(0, \beta_p, 1) + C(\theta_2, \beta_p, 1)} - \frac{C(\theta_2, \beta_p, r)}{C(0, \beta_p, r) + C(\theta_2, \beta_p, r)} \right)$$

which is strictly positive if and only if

$$\frac{C(\theta_2, \beta_p, 1)}{C(0, \beta_p, 1)} > \frac{C(\theta_2, \beta_p, r)}{C(0, \beta_p, r)}$$

That is, true treatment effects will increase if the probability of publication conditional on $\Theta^* = \theta_2 > 0$ relative to the probability of publication conditional on $\Theta^* = \theta_1 = 0$ is higher in the corrected regime relative to the uncorrected regime.

As in the previous section, fix $(\beta_p, r) = (0.1, 0.75)$. We can use the expression in equation (B) to calculate the change in true treatment effects from standard error corrections for different values of $\theta_2 > 0$. For $\theta_2 = 1.5$ and $\theta_2 = 0.75$, 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 $\theta' \in (0.75, 1.5)$ such that $\text{TrueTE}(0, \theta', 0.5, 0.1, 1) - \text{TrueTE}(0, \theta', 0.5, 0.1, 0.75) = 0$. □

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

Recall that the main text provides an example where individual-study bias decreases with corrections. This example relies on the distribution of published true effects increasing with corrections which outweighs the decrease in bias. Proposition B.1 states that meta-study bias can also decrease, which, by contrast, has a fixed reference point $\mathbb{E}[\Theta^*]$. In other words, decreases in estimated treatment effects can arise solely from selection of estimated effect sizes, rather than in combination with selection of true treatment effects.

For intuition, consider 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, the clustered regime will therefore censor a large share of studies between 1.5 and 2. How this impacts bias depends on whether censoring these studies tends to increase or decrease the expected estimated treatment effect in the uncorrected regime. In the examples given in the proof, we have that $\mathbb{E}[X^*|D = 1, \Theta^* = 1.5, \widetilde{\Sigma}^* = \frac{3}{4}] = 2.13$ and $\mathbb{E}[X^*|D = 1, \Theta^* = \frac{1}{4}, \widetilde{\Sigma}^* = \frac{3}{4}] = 0.62$, where Θ^* 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 Θ^* is degenerate. In the second case, the opposite occurs.

C. Bias and True Treatment Effects

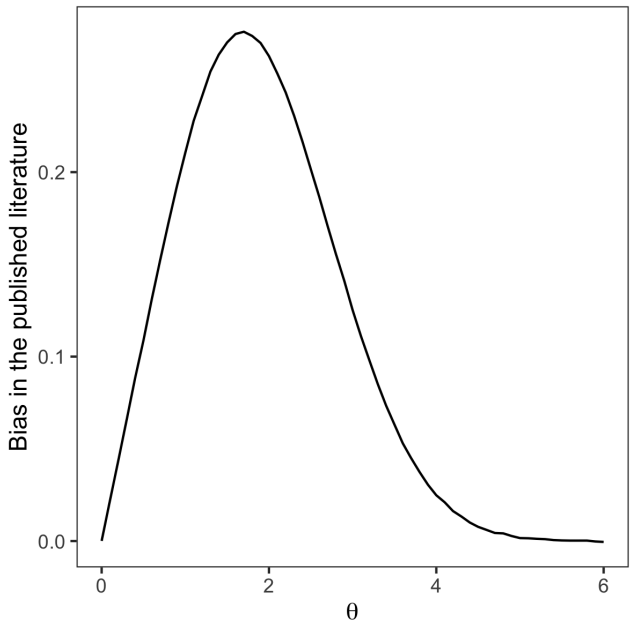

FIGURE C1. Plot of $\mathbb{E}[X^* - \theta^* | \Theta^* = \theta^*, \widetilde{\Sigma} = 1]$ for different values of θ^*, with $\beta_p = 0.1$.

D. Details on Descriptive Statistics

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

Figure C1 shows the distribution of JEL codes. Note that studies typically include multiple JEL codes and Figure C1 plots the distribution at the JEL code level rather than at a study-level e.g. with weighted JEL codes. The results show that clustered articles are less likely to be Health, Education & Welfare (I); and Labor (J), although the difference is not statistically significant. Figure C1 shows that clustered studies are more contain to have JEL codes that are outside the three dominant categories of Public Economics (H); Health, Education & Welfare (I); and Labor (J).

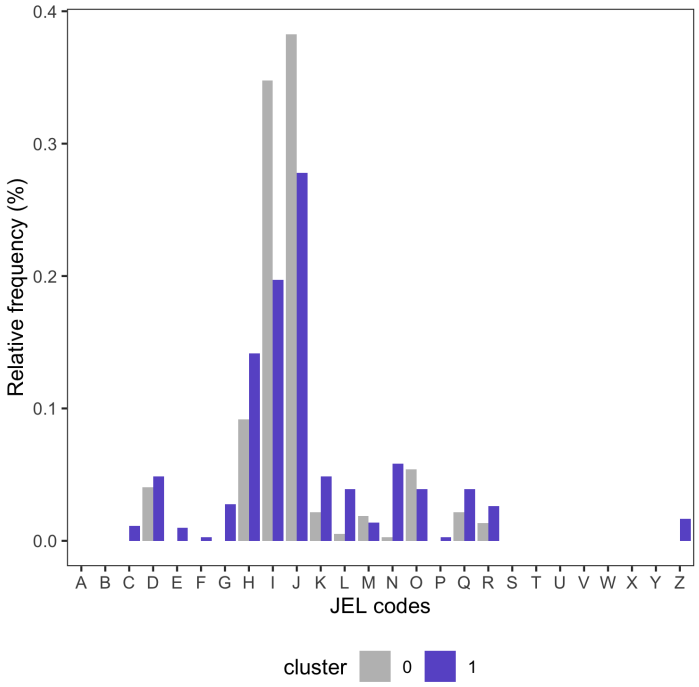

FIGURE C1. Distribution of JEL codes. The most common JEL codes are: Public Economics (H); Health, Education & Welfare (I); and Labor (J)

Figure C2 shows the five-year centered moving average of estimated treatment effects by clustering regime.[29] Effect sizes are considerably larger for studies reporting clustered standard errors. In particular, the magnitude of estimated treatment effects range approximately between 20–25% in the clustered regime and between 12.5–17.5% in the unclustered regime.

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

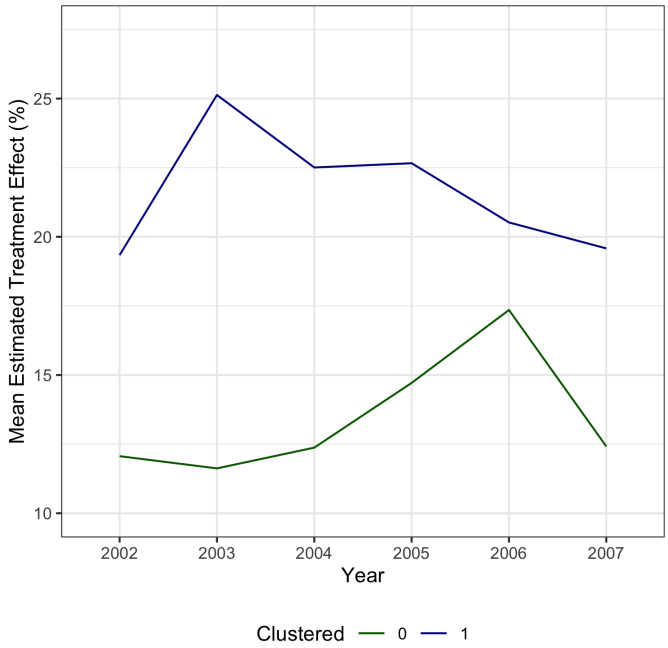

FIGURE C2. Five-Year Centered Moving Average of the Magnitude Estimated Treatment Effects

E. Comparative Descriptive Statistics from 1990–1999

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

Table D1 compares effect sizes between unclustered studies published between 2000–2009 to those published between 1990–1999. The average effect size between 2000-2009 is 12.18%. In the earlier 1990-1999 period, effect sizes were only between 1.5–2 ppts smaller. This difference is statistically indistinguishable from zero, although with relatively few observations there is somewhat limited power to reject the null hypothesis. This provides suggestive evidence that the large increase in effect sizes observed over the the 2000–2009 period is not driven by strategic clustering of the form discussed here.

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

work. A second reason for the small sample is that studies must meet the inclusion criteria described in Section 3 which ensure comparability of effect sizes (i.e. estimated treatment effects in percent units from a binary treatment) across studies.

TABLE D1 – Effect Sizes of Unclustered Studies: 1990's vs. 2000's

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

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

F. Robust Estimation for Strategic Clustering

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

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

F.1. Model of Strategic Clustering

The model extends the model in Section 2 to incorporate strategic clustering:

1. **Draw a latent study:** $(\Theta^*, \Sigma^*) \sim \mu_{\Theta,\Sigma}$
2. **Estimate the treatment effect:** $X^*|\Theta^*, \Sigma^* \sim N(\Theta^*, \Sigma^{*2})$

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) Endogeneous clustering:

$$\widetilde{\Sigma}^* = \begin{cases} r \cdot \Sigma^* & \text{if } 1.96r \leqslant |X^*|/\Sigma \leqslant 1.96 \\ \Sigma^* & \text{otherwise} \end{cases}$$

   (b) Exogeneous clustering:

$$\widetilde{\Sigma}^* = \begin{cases} r \cdot \Sigma^* & \text{with probability } 1-\beta_{c,2} \\ \Sigma^* & \text{with probability } \beta_{c,2} \end{cases}$$

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

4. **Publication selection:**

$$\mathbb{P}(D=1|X^*/\widetilde{\Sigma}^*) = \begin{cases} \beta_p & \text{if } |X^*|/\widetilde{\Sigma}^* \geqslant 1.96 \\ 1 & \text{otherwise} \end{cases} \tag{28}$$

The extension from the baseline model in Section 2 is in the third step. There exists some probability $\beta_{c,1}$ that researchers will choose whether or not to cluster strategically. Specifically, researchers strategically choose not to cluster with probability when doing so allows them to obtain statistical significance. Otherwise, they always cluster. When $\beta_{c,1} = 0$ clustering is completely exogenous and the model collapses to the baseline model.

F.2. Robust Estimation

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

Lemma F.1. *The distribution of statistically significant, published studies in the clustered regime,* $X^*, \Sigma^*, \Theta^* | D=1, C^*=1, |X^*|/\Sigma^* \geqslant 1.96$*, does not depend on* $(\beta_{c,1}, \beta_{c,2})$*.*

Proof. First, note that conditioning on statistical significant is equivalent to setting $\beta_p = 0$ i.e. of censoring all insignificant results. I will show that the density of published clustered studies

in the endogenous regime is identical to the density in the exogenous regime when $\beta_p = 0$. Since the overall density of published clustered studies is simply a mixture of these two regimes, it follows that the overall density must equal to the density in the exogenous regime, 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

$$f_{X,\Sigma,\Theta|D,\widetilde{\Sigma}}(x, \sigma, \theta | D = 1, \widetilde{\Sigma} = \sigma; E = 1) = \frac{\mathbb{P}[D = 1 | X^* = x, \widetilde{\Sigma}^* = \sigma; E = 1] \frac{1}{\sigma} \phi\big(\frac{x-\theta}{\sigma}\big)}{\mathbb{P}[D = 1 | \widetilde{\Sigma}^* = \sigma; E = 1]}$$

$$\propto \mathbb{1}\{|x|/\sigma \leqslant 1.96r\} \cdot \beta_p \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) + \mathbb{1}\{|x|/\sigma > 1.96\} \cdot \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right)$$

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

Next, consider the density of published clustered studies in the exogenous regime. By similar arguments we have that

$$f_{X,\Sigma,\Theta|D,\widetilde{\Sigma}}(x, \sigma, \theta | D = 1, \widetilde{\Sigma} = \sigma; E = 0)$$

$$\propto \mathbb{1}\{|x|/\sigma \leqslant 1.96\} \cdot \beta_p \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right) + \mathbb{1}\{|x|/\sigma > 1.96\} \cdot \frac{1}{\sigma} \phi\left(\frac{x-\theta}{\sigma}\right)$$

When $\beta_p = 0$, 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 significant with clustering i.e. $|X|/\Sigma \in (1.96r, 1.96)$. But studies with $|X|/\Sigma \in (1.96r, 1.96)$ would never be published in the clustered regime 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 when we condition on statistical significance.

This provides the basis of an approach to get a unbiased estimates of the latent distribution, even in the presence strategic clustering. We do this by estimating the model with the selected sample of statistically significant clustered studies, $X^*, \Sigma^* | D = 1, C^* = 1, |X^*|/\Sigma \geqslant 1.96$, and setting $\beta_p = 0$ such that we only estimate $\mu_{\Theta,\Sigma}$. 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 $\beta_p = 0$ and do not need to estimate it. In other

words, once we condition on the selection of the econometrician, conditioning again by selective publication has no impact since it is also based on statistical significance. Thus, we can recover the latent distribution irrespective of whether or not there is strategic clustering.

F.3. Robust Maximum Likelihood Estimation

Under the null hypothesis of no strategic clustering, the estimated latent distribution using the full sample, $X^*, \Sigma^* | D = 1, C^* = 1$, should be similar to the unbiased estimate with the significant sample, $X^*, \Sigma^* | D = 1, C^* = 1, |X^*|/\Sigma \geqslant 1.96$. 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.[30] Thus if the estimates of the latent distribution are sufficiently different, then we can reject the null of no strategic clustering. Otherwise, we do not reject it.

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

Table F1 – Robust Maximum Likelihood Estimates

	Latent true effects Θ^*		Latent standard errors Σ^*		Selection
	κ_θ	λ_θ	κ_σ	λ_σ	β_p
Restricted (Robust)	0.205	15.126	1.602	6.039	0.000
	(0.102)	(3.220)	(0.260)	(2.006)	–
Standard	0.154	17.802	1.426	6.475	0.016
	(0.0353)	(2.692)	(0.167)	(1.282)	(0.007)

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

F.4. Bias and Coverage Results with Robust Model

Ultimately, we are interested in how differences in parameter estimates from the robust approach could affect our final conclusions about the impact of clustering on bias and coverage. One

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

concern with the statistical test above is that limited power in the above test prevents us from rejecting the null hypothesis despite differences in parameter estimates that have a meaningful impact on the main results examining the impact of clustering on bias and coverage in Section 4. To alleviate these concerns, I perform a robustness exercise where I reproduce the main analysis using parameter estimates from the robust model. This allows us to test the sensitivity of the main results to the (statistically insignificant) differences in parameter estimates in Table F1.

To estimate the parameters of the latent distribution, the robust model sets $\beta_p = 0$ and therefore does not estimate it. Thus, it is necessary to choose the value of β_p to calculate the impact of clustering. For robustness, I choose three different values. The first is setting β_p to the same value estimated in the standard model for DiD studies (A). The second is to set $\beta_{=}0.037$, which is the value estimated by Andrews and Kasy (2019) for replications in experimental economics (B).[31] Finally, to test sensitivity of the results, I set it to $\beta_p = 0.1$, a relatively large value which is 6.25 times larger than the value estimated in DiD studies (C).

Table F2 presents the results. Overall, the conclusion from the 'standard model' that clustering increases coverage by a large amount at the expense of increased bias is maintained across all calibrations of the robust model. This suggests that the main results are unlikely to be driven strategic clustering of the form presented in the model above.

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

TABLE F2 – Results for Model Robust to Strategic Clustering

	Unclustered ($\hat{r} = 0.51$)	Clustered ($r = 1$)	Change
Standard Model ($\hat{\beta}_p = 0.016$)			
Coverage	0.28	0.70	0.41
Estimated Treatment Effect	6.25 (100%)	12.74(100%)	6.48(100%)
Expected Bias	1.23 (19.6%)	2.44 (19.2%)	1.21 (18.7%)
Expected Θ	5.03 (80.4%)	10.3 (80.8%)	5.27 (81.3%)
Robust Model			
A DiD Studies ($\beta_p = 0.016$)			
Coverage	0.31	0.72	0.41
Estimated Treatment Effect	6.9 (100%)	13.29 (100%)	6.39 (100%)
Expected Bias	1.52 (22%)	2.94 (22.2%)	1.42 (22.3%)
Expected Θ	5.38 (78%)	10.34 (77.8%)	4.96 (77.7%)
B Economics Experiments ($\beta_p = 0.037$)			
Coverage	0.33	0.75	0.42
Estimated Treatment Effect	6.7 (100%)	11.96 (100%)	5.26 (100%)
Expected Bias	1.44 (21.5%)	2.56 (21.4%)	1.12 (21.3%)
Expected Θ	5.26 (78.5%)	9.4 (78.6%)	4.14 (78.7%)
C One-in-Ten Censored ($\beta_p = 0.1$)			
Coverage	0.38	0.81	0.43
Estimated Treatment Effect	6.2 (100%)	9.44 (100%)	3.24 (100%)
Expected Bias	1.24 (20%)	1.83 (19.4%)	0.59 (18.2%)
Expected Θ	4.96 (80%)	7.61 (80.6%)	2.65 (81.8%)

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

G. Impact of Clustering for Different Sized Corrections

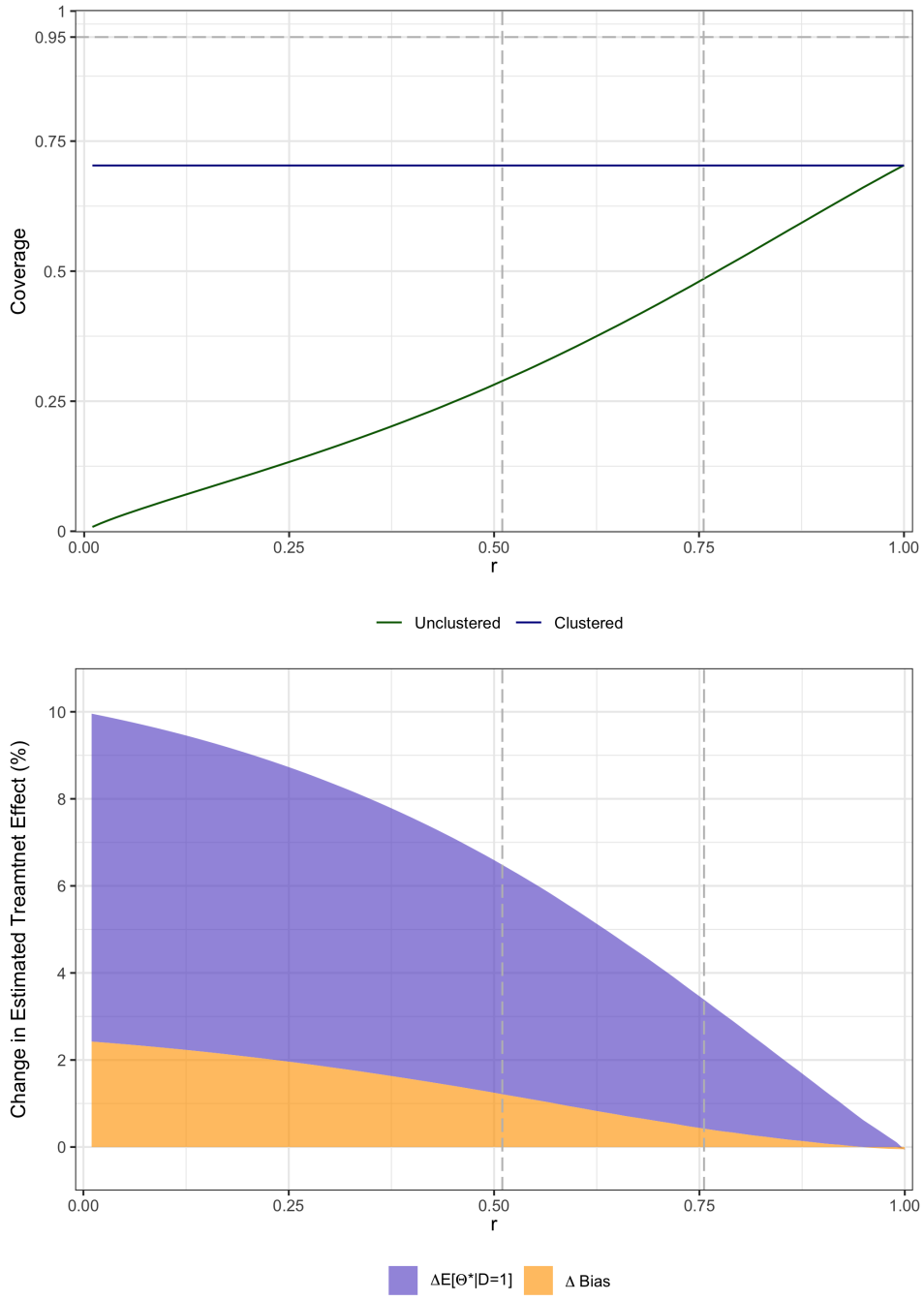

FIGURE G1. Results on the Impact of Clustering for Different Values of r

Notes: Change in coverage, estimated treatment effects, true treatment effects and bias for the estimated model parameters in Table 3 as a function of downward bias in unclustered standard errors r. The vertical dashed line at $\hat{r} = 0.51$ represents the calibrated value using the method of simulated moments. The vertical dashed line at $\hat{r} = 0.76$ represents the mean of the empirical distribution of r from 2015–2018 DiD studies.

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

TABLE H1 – Impact of Clustering Based on 2015–2018 Empirical Distribution of r

	Unclustered	Clustered ($r = 1$)	Change
Random draws of r			
Coverage	0.36	0.7	0.34
Estimated Treatment Effect	7.41(100%)	12.73(100%)	5.32(100%)
Expected Bias	1.38 (18.6%)	2.44 (19.2%)	1.07 (20%)
Expected Θ	6.03 (81.4%)	10.29 (80.8%)	4.26 (80%)
Mean: $\hat{r} = 0.76$			
Coverage	0.49	0.7	0.21
Estimated Treatment Effect	9.41(100%)	12.74(100%)	3.32(100%)
Expected Bias	2.03 (21.6%)	2.44 (19.2%)	0.41 (12.3%)
Expected Θ	7.38 (78.4%)	10.3 (80.8%)	2.91 (87.7%)

Notes: These figures are based on the parameter estimates of the empirical model in Table 3. Figures are calculated by simulating published studies under unclustered and clustered regimes. In the unclustered regime, the degree of bias in unclustered studies is based on the empirical distribution of r from 2015–2018 studies. Panel A shows results based on drawing different values of r from the empirical distribution for unclustered studies. Panel B assumes that all unclustered studies are downward biased by a constant factor equal to the mean of the empirical distribution ($\hat{r} = 0.76$).