PA

Publication Biases in Replication Studies

Adam J. Berinsky[1], James N. Druckman[2] and Teppei Yamamoto[1]

[1] *Department of Political Science, Massachusetts Institute of Technology, Cambridge, MA, USA. Email: berinsky@mit.edu, teppei@mit.edu*
[2] *Department of Political Science, Northwestern University, Evanston, IL, 60208. Email: druckman@northwestern.edu*

Abstract

One of the strongest findings across the sciences is that publication bias occurs. Of particular note is a "file drawer bias" where statistically significant results are privileged over nonsignificant results. Recognition of this bias, along with increased calls for "open science," has led to an emphasis on replication studies. Yet, few have explored publication bias and its consequences in replication studies. We offer a model of the publication process involving an initial study and a replication. We use the model to describe three types of publication biases: (1) file drawer bias, (2) a "repeat study" bias against the publication of replication studies, and (3) a "gotcha bias" where replication results that run contrary to a prior study are more likely to be published. We estimate the model's parameters with a vignette experiment conducted with political science professors teaching at Ph.D. granting institutions in the United States. We find evidence of all three types of bias, although those explicitly involving replication studies are notably smaller. This bodes well for the replication movement. That said, the aggregation of all of the biases increases the number of false positives in a literature. We conclude by discussing a path for future work on publication biases.

Keywords: publication bias, replication studies, file drawer bias, open science, conjoint experiment

Publication bias of academic research occurs when publication decisions are based on criteria unrelated to research quality. A vast array of evidence suggests a prevalent publication bias that privileges statistically significant results, across the sciences (Franco, Malhotra and Simonovits 2014; Brown, Mehta and Allison 2017) including political science (e.g., Gerber, Green, and Nickerson 2000; Gerber *et al*. 2010; Malhotra 2021). One response has been to call for more replications—by which we mean emulating the extant study's procedures but with new data drawn from the same population (Freese and Peterson's 2017 "repeatability"). Replications can correct initial study publication bias and facilitate research transparency by forcing authors to make study materials public (e.g., Nosek *et al*. 2015). There have been a number of large-scale replication studies (Mullinix *et al*. 2015; Open Science Collaboration 2015; Camerer *et al*. 2016, 2018; Coppock *et al*. 2018; Coppock 2019) and ongoing debate about the existence of a "replication crisis" (c.f., Baker 2016; Fanelli 2018). Moreover, some journals, including those in political science, now invite the submission of replications and some even incorporate "replication" sections.[1]

Yet, in political science, we have little sense of whether publication biases also exist when it comes to replications. Are replications less likely to be published than original studies? Do the same file drawer publication biases (i.e., privileging statistically significant results) documented in nonreplications exist with replication studies? Do other publication biases exist such that replications that contradict prior work are more likely to be published? Addressing these questions will allow us to assess whether replications can be an antidote to initial study publication biases and to isolate, and potentially vitiate, sources of biases in the replication literature.

Political Analysis (2021)
vol. 29: 370–384
DOI: 10.1017/pan.2020.34

Published
16 November 2020

Corresponding author
James N. Druckman

Edited by
Jeff Gill

1 Examples from three different disciplines are the *Journal of Experimental Political Science*, the *Journal of Applied Econometrics*, and the *Journal of Experimental Psychology: General*.

We start, in the next section, with a brief motivation for our approach, as it differs from many extant studies of publication bias. We then present a basic model of publication decisions. The model incorporates a single original study and then a single replication attempt—both of which could be published (or not) as individual papers. We use the model to study how publication bias affects false-positive rates (FPR). We next estimate parameters of the model with data from a survey experiment conducted with political scientists. We find that: (1) counter to what one might expect, political scientists seem nearly as likely to recommend publishing a replication study as they would an original study (there is a statistically significant but substantively small difference); (2) there is a file drawer publication bias—a preference for statistically significant results—but that bias is smaller in replications than in original studies; and (3) there is a significant but not large bias toward preferring replication results that contradict original studies (what we call a "gotcha" bias). All of these findings bode well for the "replication movement," but there is a catch. Small biases exist, and the aggregation of the biases inflates the number of false positives entering the literature. There is thus room for improvement and healthy debate about further reforms such as more replication sections in journals. There is also an incentive for scholars, as more replications enter the literature, to track the distinct types of publication biases we identify.

1 Our Approach to Studying Publication Bias

Detecting publication bias is difficult for at least two reasons. First, it involves making an inference about published studies relative to an unobserved set of unpublished studies. Do publication decisions, all else constant, depend on statistical significance? Approaches for answering that question include examining the relationship between published studies' effect sizes and their amount of precision (e.g., funnel plot) and using forensic analyses to look for anomalous statistical patterns in published studies (e.g., caliper test, *p*-curve analysis) (see Malhotra 2021). These methods face a second challenge, which is that they do not control for research quality.

Franco *et al*. (2014) introduce an approach that overcomes both challenges by tracking similar studies, some of which were published and some of which were not. They do this by using a set of high-quality experiments carried out by the Time-sharing Experiments for the Social Sciences (TESS) program (https://tessexperiments.org/). This program accepts proposed survey experiments, on a highly competitive basis, and then fields them on national probability samples in the United States. The program publicly posts all accepted projects and data collections, regardless of whether the results ever led to a publication. Franco *et al*. (2014) leverage the practice by comparing the publication rate of accepted TESS studies that vary in whether they achieved statistical significance. They find that overall studies with significant results are 40 percentage points more likely to be published than null results. Much of that bias stems from authors discontinuing the paper writing process upon discovering null results; when it comes to the set of papers that have been written, they find a roughly 5 percentage point bias toward significant results in the publication process (see Franco *et al*. 2014, appendix, Table S2).

For us, the aforementioned problems are compounded by the reality that in many disciplines, including political science, a sizeable corpus of replication studies does not exist (see Christensen *et al*. 2019, 164–165). For instance, in the Franco *et al*. data, none of the studies were explicit replications. A search of the TESS archives for projects in the years after those analyzed by Franco *et al*. (i.e., after 2012) suggests that only three projects are explicit replication studies. This prevents direct investigation of replication study publication bias à la the Franco *et al*. approach (i.e., there are not enough replication studies from TESS to explore publication bias among them). More generally, replication studies are a relatively recent phenomenon in political science. For instance, a content analysis of all experiments published in the *American Political Science Review* since its

founding to May 2019 finds that only about 5% could be construed as general replications.[2] Also, the *Journal of Experimental Political Science* (*JEPS*) only began a special section for replications (see note 1) in 2019, and as of May 2020, had not yet accepted a paper for the section.[3] This suggests that there do not yet exist a sufficient number of published replication studies that would enable us to use one of the aforementioned methods (e.g., funnel plot, caliper test).

With these constraints in mind, we take a two-pronged approach. The first prong is a decision-theoretic model of the publication process that involves two studies: an original study and a replication study. The model facilitates the transparent and precise presentation of three distinct types of publication bias, two of which have gone unstudied. Further, it allows us to derive a metric (i.e., the actual false-positive rate [AFPR]) to look at the impact of each type of publication bias.[4] The model is a simplification of reality: it sets aside strategic considerations that may lead one to conduct a replication study in the first place as well as the decision calculus involved in deciding whether to subsequently proceed in writing a paper upon observing the results (that is, it assumes a paper is written).[5] While these are meaningful choices, our approach has the advantage of focusing on biases that emerge strictly from the publication process rather than scholars' decisions, which could stem from misguided perceptions (e.g., a self-enforcing process where authors think nonsignificant results will not be published and thus they do not write such papers).[6] The model also collapses the stages of the publication process that, in theory, involve the author's submission, the reviewers' recommendations, and the editorial decision. The model therefore serves as a starting point or heuristic for studying replication study publication bias, which thus far has been virtually ignored.[7]

The second prong in our approach involves a vignette survey experiment to empirically estimate the types of publication biases we define in the model. This is perhaps the only practicable approach given the aforementioned low volume of replication studies in many disciplines including political science. After all, one cannot empirically study publication bias in replications without a large number of replication studies. We believe it is best to start studying publication bias in replication studies now, rather than waiting years for a corpus of replications to emerge.[8] In sum, traditional approaches to studying publication bias do not straightforwardly apply to studying replication bias (at this time). Ours offers a viable method for establishing an understanding on which others can build.

2 The content analysis comes from Druckman and Green (2021). We focus on experiments since, given many of the large-scale replication studies utilize experiments, it maximizes the likelihood of finding replication studies. Most of the examples we find are experiments that replicated prior work and then extended it in a particular direction. One such example is Mummulo and Peterson (2019) who replicate a series of survey experiments and extend the designs to study the impact of demand effects.

3 The journal has also published only two replications otherwise in its history (since 2014).

4 In Supporting Information, we also identify analytic relationships between types of publication bias and FPR, as well as factors that affect the reproducibility rate (a metric discussed in the Supporting Information).

5 The decision of whether to write up results is a distinct process that involves time, availability, career incentives, and risk attitudes. Regarding the latter point, Franco *et al*. (2015, 2016) report authors make substantial decisions to present limited experimental conditions and outcome variables—these may be frowned upon and thus there is a risk calculus involved in deciding what to present. Overall, this would entail a model and study different from ours. As mentioned, an advantage of our approach is that we speak directly to the publication process itself and offer direct evidence of its effect. This provides insight into whether authors who do not write up results are generating a self-enforcing file drawer bias.

6 Of course, we recognize that these perceptions themselves are likely the product of the publication process.

7 The exception is meta-analytic approaches that assess original and follow-up studies, mostly in medicine (e.g., Ioannidis and Trikalinos 2005). We do not explore massive replication efforts or meta-analyses; these are understood to play a critical role in scientific progress—our focus though is what happens to individual replication studies. We see this as a more pressing question given the recent promotion of conducting replications and the reality that many researchers do not have the resources to conduct massive replications of many studies at once.

8 Using a hypothetical vignette experiment begs the question of the extent to which the results map onto actual publication decisions. However, we can compare a subset of our results to other publication bias studies to assess their generalizability.

2 Model of Publication Decisions

In our model, we consider three distinct types of publication bias. First, we examine the well-known file drawer problem (Rosenthal 1979). *File drawer bias* occurs if a positive test result (i.e., a statistical test that rejects the null hypothesis of no effect) is more likely to be published than a negative test result (i.e., a test that does not reject the null hypothesis), *ceteris paribus*.[9] In other words, the published record of research is skewed away from the true distribution of the research record, overstating the collective strength of positive findings. For example, if one of 10 studies shows that sending varying types of health-related text messages leads people to eat less fatty food, and only that one study is published, the result is a misportrayal of the effect of text messages. The file drawer bias reflects an entrenched culture that prioritizes statistical significance and novelty, as well as a funding system that rewards positive findings (Brown *et al*. 2017). As mentioned, there is a large theoretical and empirical literature that documents this type of publication bias and its consequences in many disciplines including political science (Gerber *et al*. 2000; Gerber and Malhotra 2008; Gerber *et al*. 2010; Franco *et al*. 2014; Fanelli, Costas and Ioannidis 2017; Malhotra 2021).

Second, we introduce and consider what we call a *repeat study bias*. This bias takes the following form: if two studies are identical, except one is original and the other is a replication, the first study is more likely to be published in the peer-reviewed literature, *ceteris paribus*. Suppose, for example, that two sequential studies that are identical in every way other than timing (e.g., they have the same design and sampling procedure) show that contact with people of different ethnicities reduces prejudice by 10%. These studies provide equivalent evidence, but only the first study is published. This type of repeat study bias has not received attention in the literature on publication bias; indeed, common definitions of publication bias focus exclusively on a study's result, ignoring whether it is a replication or not (e.g., Brown *et al*. 2017, 94). This repeat study bias would reflect a system where "incentives to publish positive replications are low and journals can be reluctant to publish negative findings" (Baker 2016, 454).[10]

On its face, repeat study bias may seem innocuous since it does not directly skew the scientific record from the true distribution of research results. Yet, it matters for two reasons: (1) it abates the weight of the aggregate evidence for or against a given phenomenon (and could affect meta-effect sizes in light of other publication biases discussed below) and (2) if repeat studies are published at a lower rate, this process could discourage researchers from engaging in replications. This, in turn, could facilitate the production of false negatives and false positives because results that happen by chance are not retested. As the social sciences continue to encourage replication, understanding repeat study bias is crucial, yet "little empirical research exists to demonstrate that journals explicitly refuse to publish replications" (Martin and Clarke 2017, 1).

The third type of bias is also one we introduce for the first time; it stems from the possibility that the *process* of replication itself could induce bias. We define a *gotcha bias* as occurring when a result in a replication study is more likely to be published when the result contradicts that of an existing prior study that tested the same hypothesis, *ceteris paribus*. That is, replications are more likely to be published if they overturn extant findings. The published record of research therefore overly emphasizes replications that run counter to existing findings, as compared to the true distribution of the research record. This differs from the file drawer bias since it is contingent on the outcomes of both the original study and the replication; it differs from the repeat study bias which is agnostic to the outcome of the replication study.

9 We focus on the peer-reviewed scientific literature and thus do not consider the so-called gray literature that includes conference papers, dissertations, etc.

10 Reflective of the incentives to not publish replication is journalists who "preferentially cover initial findings although they are often contradicted by meta-analyses and (then the journalists) rarely inform the public when they are disconfirmed" (Dumas-Mallet *et al*. 2017, 1).

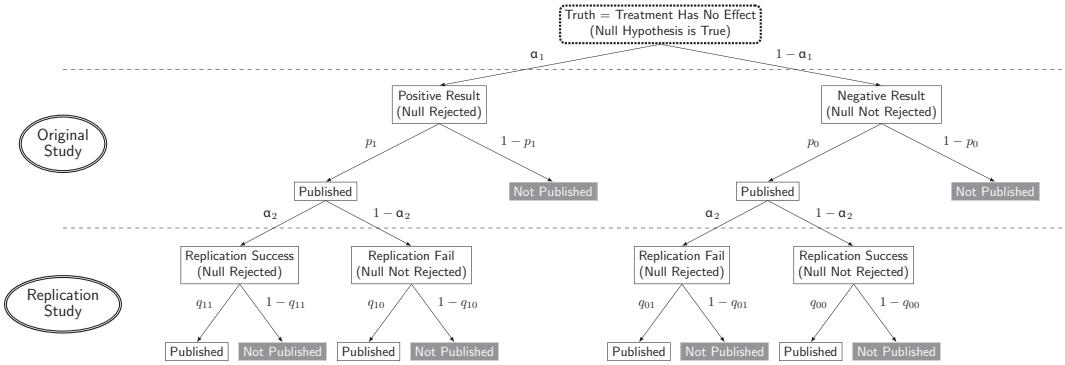

Figure 1. A model of publication process with two stages.

An example of the gotcha bias is as follows—there are 10 similar studies of text messaging effects that each purport to replicate corresponding previous studies, and they all find significant effects. Yet, nine of them are successful replications of the original studies that found equally significant results while one follows on a study that found a nonsignificant result. If only the latter study—that finds a significant effect contrary to a previous nonsignificant study—is published, the larger research community will observe a distorted portrayal of the research record.

The gotcha bias is a concept applicable to replication studies that collect new data in an attempt to replicate results of previous, original studies using similar study designs. The hypothesized mechanism behind this bias is again a proclivity for novelty and sensationalism. Some authors have alluded to a similar phenomenon, most notably the *Proteus phenomenon* that occurs when extreme opposite results are more likely to be published (Ioannidis and Trikalinos 2005). Yet, we are the first (as far as we are aware) to formulate our discussion of this form of bias in the same positive–negative test result terms used to describe file drawer bias.

We employ a simplified model of the publication process in order to study the consequences of these three types of publication biases. We display the model in Figure 1. As explained, we recognize the model does not capture the full complexity of the scientific process. For example, it does not model the decision to write up results, and it does not explicitly distinguish authors', reviewers', and editors' decisions about publication. Yet, the model serves as a useful heuristic for understanding replication and publication biases, and we will use it to identify a clear metric to assess the impact of publication bias.[11]

The model starts with a null hypothesis (e.g., no treatment effect) as the true state of the world and proceeds as follows. (We consider a parallel model for the case where the null is false, as discussed in the Supporting Information.) Then, the "original study" tests the null hypothesis with the nominal type-I error probability of α_1, based on a simple random sample of size N drawn from the target population. The result, whether (false) positive or (true) negative, is written up, goes through a peer-review process and is published, with probability p_1 for a positive result and p_0 for a negative result. The anticipated discrepancy between p_1 and p_0, such that $p_1 > p_0$, represents the file drawer bias (for the original study).

Next, only the published results from the first stage are subjected to replication studies, which we assume to be designed identically to the original study but conducted on a newly collected sample from the same population. With the type-I error rate of α_2, the result is a (false) positive. The results are written up and then go through a peer-review process similar to the first stage, except that the publication probability now depends on test results from the current *and* previous stages (q_{11}, q_{10}, q_{01}, q_{00}). A repeat study bias would be present, for the case of a positive initial

11 Moreover, as mentioned, in the Supporting Information, we provide various analytic results about the relationships between types of publication biases and performance metrics such as the FPR and reproducibility rate.

result, if $p_1 > q_{11}$, and for a negative initial result, if $p_0 > q_{00}$. A gotcha bias occurs if $q_{10} > q_{00}$ (such that a negative replication result is more likely to be published when it contradicts a previous positive result than when it confirms an existing negative result) for insignificant replication results. Similarly, $q_{11} < q_{01}$ represents a gotcha bias for significant replication results.

We realize not all published results will ever be replicated with fresh samples, even with the current push toward credible research. And, as suggested, researchers might strategically choose which existing studies to replicate and which replication results to write up and submit for review, given their perception of publication biases. But our goal here instead is to examine how the idealized model of replication science, as exemplified by the Open Science Collaboration's widely discussed replication study of 100 psychology experiments (Open Science Collaboration 2015), would differ if we "perturbed" it by adding possible publication biases for each individual replication. What would have happened if the Open Science Collaboration had tried to submit each of the 100 experiments separately? Moreover, the model addresses the challenges discussed earlier insofar as it considers a set of published and unpublished studies, holding quality constant (Malhotra 2021).

To study the consequences of the three types of publication biases, we consider the *AFPR in replication results* as a metric of evidence quality in published studies:

$$\widetilde{\alpha}_2 = \Pr(\text{replication test significant} | \text{replication published, null is true}).$$

The AFPR represents the proportion of the positive results in published replication studies that are actually false, that is, where the null hypotheses are in fact true. In the ideal world, this rate would be equal to the nominal FPR of α_2 that the tests in replication studies are theoretically designed to achieve. However, $\widetilde{\alpha}_2$ will diverge from their designed type-I error rate due to the three kinds of publication biases we consider. These biases could shape the scientific record in important ways. It is well known that when studied in isolation, the file drawer bias tends to inflate the FPR by disproportionately "shelving" negative results that correctly identify true null hypotheses (Rosenthal 1979). However, the effects of the other two types of biases—the repeat study bias and the gotcha bias—have not been documented, let alone the effect of file drawer bias when these other biases are also present.

Our model allows us to study the concurrent effects of these three types of publication biases on various indicators of evidence quality, such as the AFPR.[12] In the Supporting Information, we provide an exact mathematical expression for $\widetilde{\alpha}_2$ as a function of our model parameters and investigate the net effects of the three kinds of biases. Our analysis (in the Supporting Information) reveals that the effect of one type of bias is contingent on the other two biases as well as other model parameters. For example, we find that an increase in file drawer bias can affect the overall deviation of the AFPR from the desired FPR either positively or negatively, depending on the size of the other two biases. Gotcha bias can also either increase or decrease the AFPR depending on the other biases.[13] Importantly, our model of the publication process enables the calculation of implied AFPR values given a set of assumed publication probabilities as well as significance levels for the hypothesis tests. In the rest of the paper, we draw on an original large-scale survey experiment to illustrate this exercise with empirical data on publication probabilities for different types of hypothetical studies.

12 In the Supporting Information, we also consider an alternative metric, reproducibility rate, which some previous authors use (e.g., Nosek *et al*. 2015). We derive analytical results and present empirical findings for that metric as well.

13 Although the effects of publication biases are complicated, our model implies several important relationships that are easily interpretable, as we show in the Supporting Information. In particular, when the file drawer bias is larger than the gotcha bias, the AFPR is always greater than the nominal FPR. Thus, the well-known inflation of falsely positive evidence due to publication bias still holds under our model *as long as* the gotcha bias is not too severe.

3 Survey Experiment

To illustrate how our model can shed light on publication biases, we conducted a (conjoint) vignette survey experiment. Our goal is to estimate the amount of file drawer bias, repeat study bias, and gotcha bias. We also seek to display how these biases influence the AFPR.

Our population consists of all political science department faculty at Ph.D. granting institutions based in the United States. The Supporting Information contains a description of our data collection procedure and a demographic portrait of our respondents. Political scientists make for an intriguing sample since discussions of open science and replication are certainly ongoing (Lupia and Elman 2014; Monogan 2015) but, in some sense, are more emergent than in other disciplines, particularly psychology and to some extent economics (e.g., Christensen *et al*. 2019). This may increase the experimental realism of the study, with respondents being cognizant of possible biases but not so ingrained in debates that they would be guarded in how they responded to our survey.

We sent participants a link where they were provided with a set of vignettes that described a paper (on the validity of using vignettes, see Hainmueller, Hangartner and Yamamoto 2015). Each respondent was provided with five different vignettes, each concerning a single distinct paper that has been written. We asked respondents to act as if they were an author and asked whether they would submit the paper to a journal. Respondent then received other five vignettes where we asked them to play the role of a reviewer. Here, we asked whether they would recommend the paper be published. Finally, we asked whether the respondent had ever edited a journal. If the respondent had, we provided five additional vignettes. These vignettes asked the respondent whether he or she would publish the paper (as an editor).

Each vignette randomly varied a host of features, most importantly, the statistical significance of the results and whether the study is an original study or a replication study (and if replication, whether the results of the original study were statistically significant). It also provided (randomly assigned) information about whether the study involves experimental or observational data, the sample size, whether the hypothesis is exciting/important, and whether the results are surprising. Prior work suggests these factors are meaningful for replication studies (Open Science Collaboration 2015, aac4716–2); they also provide respondents with more details about the study and prevent us from overinterpreting the impact of statistical significance and replication status. Finally, we told respondents that the study is "seemingly sound in terms of methods and analysis." This keeps constant research quality, discouraging respondents from interpreting statistical significance as an indicator of the quality (Malhotra 2021). Figure 2 presents the vignette with all possible variations indicated in square brackets.

After each vignette, we posed a question for our main outcome variable. For example, for the reviewer conditions, we asked "If you were a journal reviewer on this paper, what is the percent chance you would recommend publication of this paper (assuming no new data can be collected)?" The author and editor versions asked analogous questions. In sum, our experimental set-up has the crucial advantage of holding study quality constant and can isolate the impact of statistical significance in original and replication studies on publication decisions and the AFPR.

We focus our analyses on the treatment variations that have direct bearing on our model parameters (i.e., statistical significance and whether the study was an original or a replication study). Setting aside variations in other factors (e.g., if the hypothesis is "exciting") does not cause bias in our estimates because they are randomized independently of our main factors. We present main effect and moderating results regarding the other factors in Supporting Information. There we show that the other variables have predictable effects. In particular, experimental data, larger sample sizes, exciting hypotheses, and surprising results all increase the likelihood of publication. These findings validate the experimental realism of our study. Furthermore, these factors show only minor moderating effects for our main findings reported below.

We are interested in [how you, *as an author*, decide to submit your research to a journal/ how you evaluate papers as *a journal reviewer*/ how you evaluate papers as *a journal editor*]. To do this, we will present you with five descriptions of papers. After each description, we will ask you some questions about it.

{For the first vignette in each role:} [Suppose that you were an author of a paper reporting the results of an empirical study./ Suppose that you are evaluating a paper reporting the results of an empirical study in your own subfield./ Suppose that you are evaluating a paper reporting the results of an empirical study.] The study aims at testing a hypothesis with quantitative data and has the following characteristics.
{For the second vignette and on:} Now consider a different paper reporting the results of a study with the following characteristics. Please continue to evaluate this paper as if you are [an author/ a reviewer/ an editor].

- [Analysis of new experimental data (i.e., the study involved an intervention)./ Analysis of new observational data (i.e., the study did not involve an experimental intervention).]
- [There is no existing empirical study that tests the same hypothesis./ It is a replication of an earlier study that had reported a result that is highly significant by conventional standards (e.g., p-value of less than .01) on the test of the same hypothesis./ It is a replication of an earlier study that had reported a result that is significant by conventional standards (e.g., p-value of less than .05) on the test of the same hypothesis./ It is a replication of an earlier study that had reported a result that is not significant (e.g., p-value of greater than .75) on the test of the same hypothesis.]
- A sample size of [50/ 150/ 1000/ 5000].
- The test result is [highly significant by conventional standards (e.g., p-value of .01)/ significant by conventional standards (e.g., p-value of .05)/ not significant by conventional standards (e.g. p-value of .75)].
- The hypothesis is about [an extremely exciting and important/ a moderately exciting and important/a not at all exciting or important] effect.
- The result is [extremely surprising and counterintuitive/ somewhat surprising and counterintuitive/ not at all surprising or counterintuitive] given past work on the topic.
- Seemingly sound in terms of methods and analysis.

Figure 2. Experiment vignette. The phrases in square brackets separated by slashes represent alternative texts that are randomly and independently assigned for each vignette.

Also, notably, we recognize one must be cautious in interpreting the effect magnitude of a given factor. Respondents may overstate the extent to which they would submit or accept nonsignificant results due to recent prominent discussions about publication bias (e.g., Brown *et al*. 2017; Malhotra 2021)—that may make it desirable to support nonsignificant findings. That said, our interest lies in comparisons *between* scenarios (e.g., significant versus nonsignificant) rather than the baseline magnitudes; generally, conjoint experiments minimize social desirability bias by allowing for multiple comparisons across versions (Horiuchi *et al*. 2020).

4 Results

4.1 Estimating Three Types of Publication Bias

We begin by asking how much evidence our data shows of each of the three types of publication biases: file drawer bias, repeat study bias, and gotcha bias.[14] Figure 3 presents the average percent

14 For the data, see Berinsky *et al*. (2020). The study was assessed as exempt by the Northwestern University Institutional Review Board.

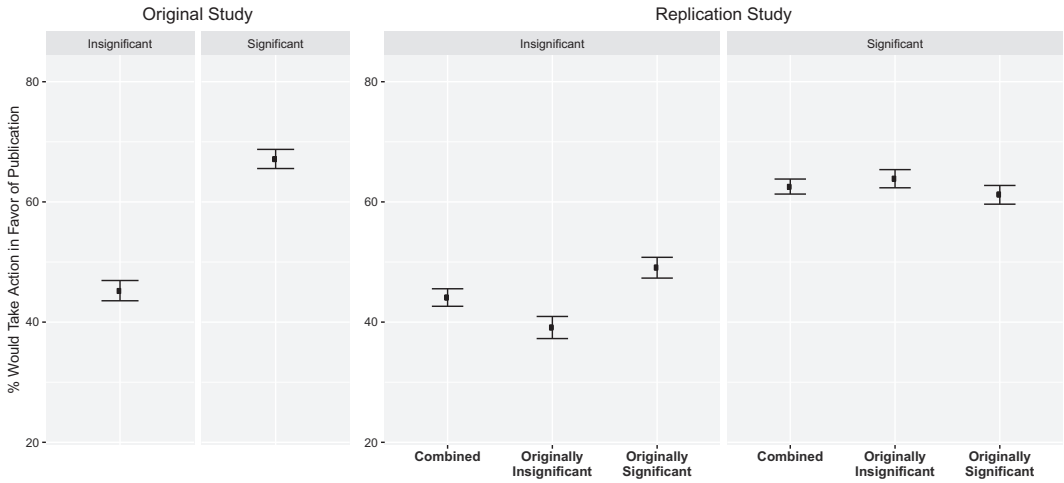

Figure 3. Evidence of three types of publication bias. In each plot, the solid dot represents the estimated probability of a respondent taking an action in favor of publishing the hypothetical study for a given combination of study characteristics indicated at the top and bottom. The vertical bars represent 95% confidence intervals.

chance of taking an action toward publication (e.g., sending out a paper as an author, recommending publication as a reviewer, and supporting publication as an editor) that the respondents allotted different types of hypothetical papers described in our randomly generated vignettes, along with 95% confidence intervals (CI). Here, we pool the author, reviewer, and editor conditions in our analysis; the results broken down for these roles are provided in the Supporting Information, and in each case, the results echo those we present with the data merged (i.e., our findings are substantively unchanged with respect to publication bias). We also combine the two conditions in which test results are described as statistically significant (i.e., 0.01 or 0.05 level) into a single category in our analysis (the results are robust to not merging these conditions).

We begin with the file drawer bias in original studies; these results are presented in the left panel of Figure 3. While respondents, on average, indicated a 67.1% chance of submitting, recommending, or supporting a paper with a significant test result ([65.6%, 68.7%]), they gave only a 45.2% chance of doing the same for an identical paper with a nonsignificant finding ([43.6%, 46.9%]). Thus, as others have shown (e.g., Franco *et al*. 2014), a large file drawer bias exists in original studies (i.e., approximately 21.9 percentage points). On the one hand, our effect is smaller than the overall 40 percentage points reported by Franco *et al*. (2014) but larger than the effect they find contingent on a resulting paper (i.e., a paper being written), which, as noted above, is just 5 percentage points.[15]

Our results further show that replication studies are subject to the same kind of file drawer bias as the original research studies. These results are presented in the right panel of Figure 3. Combining both originally significant and insignificant test results, respondents reported a 62.5% chance of moving a significant result in a replication study toward publication ([61.3%, 63.8%]), whereas they only gave a 44.1% chance for a nonsignificant replication result ([42.6%, 45.5%]). Thus, regardless of whether a replication study "succeeds" or "fails" to reproduce the original finding, replication is more likely to be published when its result is statistically significant than when it is a null finding. That said, this difference of roughly 18.4 percentage points is significantly smaller than the file drawer bias in original studies ($t = -2.95$, $p < 0.00$). Thus, there exists a file drawer bias in replication studies, but it differs in size from the file drawer bias in original studies.

15 We suspect the Franco *et al*.'s (2014) small 5% file drawer bias reflects: 1) a very small sample size of null results that were written up in their data (just 18 studies), and 2) a self-selection of those 18 such that they likely involved particularly intriguing findings from what is a very high quality data collection program (i.e., TESS).

We next turn to the repeat study bias. Respondents indicated a 61.2% chance of moving a significant replication result that successfully reproduces an earlier significant finding toward publication ([59.6%, 62.7%]). This represents a 5.9 percentage point drop compared to the probability for an original significant result, which is statistically significant ($t = 6.32$, $p < 0.00$). Similarly, respondents indicated only a 39.1% chance of submitting, recommending, or supporting publication of an insignificant replication test result that confirms a previously nonsignificant finding ([37.3%, 40.9%]), a 6.1 percentage point decrease ($t = 5.99$, $p < 0.00$) from the probability of doing the same for an original insignificant result. Thus, we find clear evidence of a repeat study bias, a distinctive and previously unstudied type of publication bias. That said, the file drawer bias dramatically dwarfs the repeat study bias: there is some but not a substantial bias against publishing replication studies as a general rule. Political scientists seem to be quite open to the publication of replication studies.

Finally, turning to gotcha bias, our results show evidence that this more subtle form of publication bias occurs in replication studies. Respondents assigned a 49.1% chance of submitting, recommending, or supporting publication of an insignificant test result ([47.3%, 50.8%]) when the study fails to replicate an earlier significant test result, compared to only 39.1% when it successfully replicates a previously nonsignificant finding. This 10 percentage point difference reveals a potentially pernicious form of replication study bias where contradicting extant knowledge is privileged, holding research quality constant. Likewise, respondents indicated a 63.9% chance of making a decision in favor of publishing a replication test result when that replication finds a significant effect which runs contrary to a previous insignificant test result of the same hypothesis ([62.4%, 65.4%]). This percentage drops to 61.2% for a significant replication result that successfully reproduces an earlier significant finding, a small but statistically significant decrease of 2.6 percentage points ($t = -2.92$, $p < 0.01$). Thus, for replication studies, there is an increased probability of supporting publication of contradictory results in either direction; scholars privilege the publication of replication studies that overturn extant results. Importantly, however, the gotcha effect only goes so far: even at the replication stage, the standard file drawer bias exerts much more influence. Replication results that find statistically significant effects are more likely to move toward publication than insignificant results, no matter what the original results may be.

In sum, our results echo a large literature that reveals file drawer publication bias in original studies. But, we also identify two new types of publication biases that occur with regard to replication studies. One the one hand, that these two biases do not rival the size of the file drawer bias bodes well for the potential corrective nature of replication studies. On the other hand, that these biases still do exist along with the file drawer problem at the replication stage accentuates the need for institutional efforts to vitiate all three types of biases.

Estimating Actual False-Positive Rates

In addition to estimating the three types of publication biases, our survey experiment data allow us to make inferences about the key metric of evidence quality: the AFPR. Here, we provide our estimates of the AFPR in replication studies for a nominal 0.05-level significance test. We calculate these estimates using our model-based formula, as well as the publication bias estimates based on the vignette data. We display the estimates in Figure 4.

Consider a published study that tries to replicate an earlier published study by testing the same hypothesis with a new sample at the 0.05 significance level. The estimated AFPR for such a replication test ($\widetilde{\alpha}_2$) is 0.078 (95% CI = [0.074, 0.081], left plot) based on our vignette data. That is, the net effect of the file drawer bias, gotcha bias, and repeat study bias turns out to be a 2.8 percentage point inflation of the AFPR compared to the nominal type-I error rate for which the

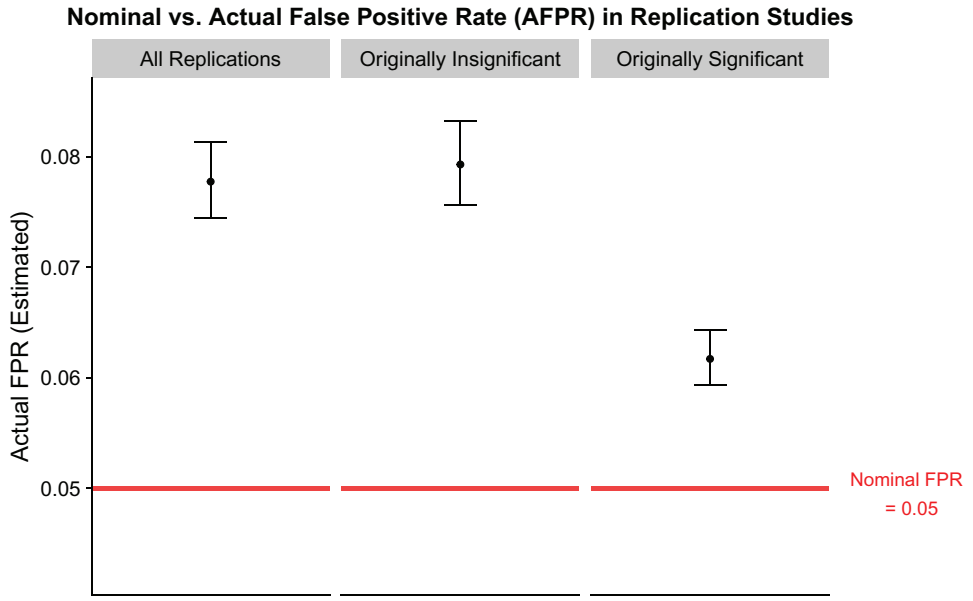

Figure 4. Estimates of actual false-positive rate for published replication study results. The solid dots represent the estimated AFPR for all published replication results (left), published replications of originally insignificant results (middle), and published replications of originally significant findings (right) based on the vignette survey data, assuming the nominal FPR (i.e., the alpha level) of 0.05 for the significance tests. The vertical bars represent 95% confidence intervals.

replication test is designed. Publication biases thus lead to an overrepresentation of false positives in the published body of replication evidence.

To better understand how publication bias affects the AFPR of replication tests, we also estimate the AFPR conditional on whether the original result was statistically significant. The result reveals a clear dependence of the AFPR on different types of publication biases. If the original study failed to reject the null hypothesis, the AFPR for its replication test is estimated to be 0.079 (95% CI = [0.076, 0.083], middle plot), an estimate similar to but slightly larger than the overall AFPR estimate. This is because both the file drawer bias and the gotcha bias operate in the same direction under this scenario: both types of biases make the positive replication result more likely to be published than in the absence of publication bias. In contrast, if the original study rejected the null hypothesis, the estimated AFPR for the replication result drops to 0.062 ([0.059, 0.064], right plot), a value much closer to the nominal FPR of the test. This 0.016-point decrease occurs because the gotcha bias partially offsets the upward pressure caused by the file drawer bias for the replication result. That is, the file drawer bias increases the probability that a false-positive replication result is published, but the gotcha bias counteracts this effect (i.e., it makes the result less likely to be published compared to the case where the false-positive result was a surprise). The bottom line is that our data suggest that replications are not a panacea to correct the scientific record—publication bias in replication studies leads to an AFPR that exceeds what would occur by chance, even in the best possible scenario. That said, the inflation of AFPR is not as large as one may have suspected.

5 Conclusion

For many, replication is a hallmark of scientific progress. The recent so-called "replication crisis" has accentuated the role of replications and generated much discussion about how replications should be conducted. Yet, virtually no attention has been paid to how publication biases manifest in the case of replications. This is a crucial question given the extent to which such biases can skew accumulated knowledge. A challenge to studying publication bias in replication studies is that extant methods for detecting publication bias have not considered the additional types of

biases we introduced, and, regardless, in many disciplines including political science, there does not yet exist a sufficiently large corpus of replication studies to explore bias. We thus introduced an alternative approach using a model and survey experiment. While we recognize limitations in the approach, it allowed us to offer novel insights regarding how to think about publication bias in replication studies—differentiating three types of biases including the file drawer bias, the repeat study bias, and the gotcha bias. The latter two types of bias have not previously been considered. We also were able to offer initial estimations of the extent of these biases.

Overall, the results show that each of these biases exists. From one perspective, our results may seem sobering. Publication biases are notable and, in the aggregate, skew the number of false positives entering the literature. Moreover, the two new types of publication biases we identify—the repeat study bias and the gotcha bias—are present. From another perspective, however, our results could be seen as reassuring. The bias against replication studies, all else constant, is present but not large, the file drawer problem in replication studies is smaller than that in original studies, and the gotcha bias also is not as large as one may have expected. In short, scholars seem open to replications, and this is auspicious for the replication movement.

We believe our results highlight several ways forward. There is of course the question of why we do not see more replications in the published literature given the relatively small estimate of our repeat study bias. The answers likely stem from a variety of sources. First is the existence of a file drawer bias in replication studies, which while smaller than that in original studies, is still nontrivial (18.4% in our data). The second is the recency of the replication movement within the discipline. Third is a possible self-enforcing dynamic such that scholars may anticipate a larger repeat study bias than might actually exist and thus do not replicate studies.

We recognize scholars may be reticent to take our results from a survey experiment as definitive. Indeed, we strongly encourage scholars to investigate the magnitude of *each type of bias* in the actual production and submission of papers. This will become increasingly possible as more replications enter the literature and we have good reason to think that will occur. For example, a search on "replication" in the Evidence in Governance and Politics (EGAP) registry yields nearly 40 studies, with about half coming from the last 2 years. These projects likely will enter the published literature in the near future. The new *JEPS* replications section and other journals that may incorporate similar initiatives also will facilitate replications. There also are some large-scale replication projects such as the Center for Open Science's Systematizing Confidence in Open Research and Evidence (SCORE) program. This involves replicating hundreds of prior claims across the social sciences. Also, there is EGAP's Metaketa Initiative that involves replicating studies across contexts (e.g., Dunning *et al*. 2019).[16] It will be crucial to monitor the actual fate of replication study submissions to assess whether our small repeat study bias finding reflects actual behavior (a la the open science push) or is an overstatement (reflecting a desire to self-report but not enact open science initiatives).

Our framework offers guidance on how to explore publication bias going forward, making clear that one need not only attend to file drawer bias but also repeat study bias and gotcha bias. Isolating these biases not only helps to assess the production of knowledge, but also allows researchers to gain a handle on the incentives and hurdles that evolve as researchers move toward replicating more work. We emphasize too that while some subfields are more amenable to the type of replication on which we focus (involving the collection of new data) than others, the basic logic of the model may carry over to cases of reproducibility (i.e., reanalyzing extant data).[17]

16 It is worth noting that some of the previously discussed techniques for studying publication bias do not require a large number of studies. For example, Gerber *et al*. (2010) use just 16 and 19 articles to study publication bias using the "caliper test."

17 Further, Graham *et al*. (2020) offer a framework for non-experimental replications that may stimulate replications in more fields.

Another avenue for future work concerns other aspects of the research process. We explained the advantage of not dealing with the processes by which scholars decide to pursue replications and/or write-up results. These involve complicated decision about career incentives and time investments. We have suggested that the lack of replications may stem partially from a self-enforcing process where authors do not purse them—the same can be said for the lack of null results (Franco *et al*. 2014) and other biases in the publication process such as gender bias (Teele and Thelen 2017, 443). Clearly, more work needs to unravel how scholars make these decisions.

The aforementioned choices clearly reflect anticipation of professional incentives and policies. Journals and institutions might encourage meta-replications projects such as the ones that have received some attention by aiming to replicate a large set of studies (e.g., Mullinix *et al*. 2015; Open Science Collaboration 2015; Camerer *et al*. 2016, 2018; Coppock 2019). This is a way to dampen some of the publication biases. To see why, consider that the aforementioned Open Science Collaboration (2015) replicated just 36% of initially statistically significant results from 100 previously published psychology experiments. Few papers have generated as much discussion and debate (e.g., Anderson *et al*. 2016; Gilbert *et al*. 2016); yet, according to our results had each of the 100 attempted replications been submitted individually, more than half might not have been published. These undertakings are large, however, and require institutional and individual investments that are nontrivial. Moreover, even these large-scale attempts could be subject to bias. One could wonder whether the aforementioned psychology replication would have received as much attention had the replication rate been closer to 95%. Regardless, these large-scale efforts also require that authors make the study materials and details easily accessible (Nosek *et al*. 2018). One related approach is to encourage more self-replications where authors make an effort to replicate their own work (Janz and Freese n.d.).

Of course, the reality is that altering incentives and behaviors are not easy: researchers, like everyone else, exhibit confirmation biases (Bollen *et al*. 2015) that lead them to privilege the preconception that statistical significance is critical and that original work is more important than replication. In terms of the former, one approach is to institutionalize results-blind review where papers are assessed sans the results.[18] For replications, the most straightforward solution would be to have more journals offer brief sections devoted strictly to replications, as many perceive journals as unwelcoming for replications (Martin and Clarke 2017, 3). As mentioned, some journals have already done this, and one additional innovation in light of budgetary constraints is that this section could be published online but still listed in the printed table of contents (as is done, e.g., in the *Journal of Experimental Psychology: General*) (also see Coffman *et al*. 2017). In the end, publication biases will continue to be a barrier to scientific progress: there is not an easily implemented fix. As scientific communities explore different approaches, however, it would be beneficial to account not only for the classical file drawer bias in original studies but also publication biases that occur at the replication stage.

Data Availability Statement

The replication materials for this paper can be found at Berinsky *et al*. (2020).

Supplementary Materials

To view supplementary material for this article, please visit http://dx.doi.org/10.1017/pan.2020.34.

Acknowledgments

We thank Michelle Bueno Vásquez, James Dunham, Shiyao Liu, Chris Peng, Robert Pressel, Blair Read, Jacob Rothschild, and Anna Weissman for research assistance. We are grateful to John

18 This becomes tricky, however, as statistical significance likely correlates with research quality; indeed, one attempt to institute this process generated a mixed outcome (Findley *et al*. 2016).

Bullock, Donald Green, Melissa Sands, the editor, and anonymous reviewers, and the participants at the 2017 Conference of the Society for Political Methodology and the 2018 Midwest Political Science Association Annual Meeting for useful comments and suggestions. Yamamoto developed and analyzed the model, and designed and performed the statistical analysis. All three authors designed and implemented the empirical study, interpreted the results, and wrote the paper.

References

Anderson, C. J. et al. 2016. "Response to Comment on 'Estimating the Reproducibility of Psychological Science'." *Science* 351:1037–1039

Baker, M. 2016. "Is There a Reproducibility Crisis?" *Nature* 533:452–455.

Berinsky, A. J., J. N. Druckman, T. Yamamoto. 2020. "Replication Data for: Publication Biases in Replication Studies." https://doi.org/10.7910/DVN/BJMZNR, Harvard Dataverse, V1.

Bollen, K., J. T. Cacioppo, R. M. Kaplan, J. A. Krosnick, J. L. Olds, and H. Dean. 2015. "Social, Behavioral, and Economic Sciences Perspectives on Robust and Reliable Science." Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences.

Brown, A. W., T. S. Mehta, and D. B. Allison. 2017. "Publication Bias in Science: What is it, Why is it Problematic, and How can it be Addressed?" In *The Oxford Handbook of the Science of Communication*, edited by K. H. Jamieson, D. Kahan, and D. A. Scheufele, 93–101. Oxford: Oxford University Press.

Camerer, C. F. et al. 2016. "Evaluating Replicability of Laboratory Experiments in Economics." *Science* 351:1433–1436.

Camerer, C. F. et al. 2018. "Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* Between 2010 and 2015." *Nature Human Behavior* 2:637–644.

Christensen, G., J. Freese, and E. Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. Oakland, CA: University of California Press.

Coffman, L. C., M. Niederle, and A. J. Wilson. 2017. "A Proposal to Organize and Promote Replications." *American Economic Review* 107:41–45.

Coppock, A. 2019. "Generalizing from Survey Experiments Conducted on Mechanical Turk: A Replication Approach." *Political Science Research and Methods* 7:613–628.

Coppock, A., T. J. Leeper, and K. J. Mullinix. 2018. "Generalizability of Heterogeneous Treatment Effect Estimates across Samples." *Proceedings of the National Academy of Sciences* 115:12441–12446.

Druckman, J. N., and D. P. Green. 2021. "A New Era of Experimental Political Science." In *Advances in Experimental Political Science*, edited by J. N. Druckman, and D. P. Green. New York: Cambridge University Press.

Dumas-Mallet, E., A. Smith, T. Boraud, and F. Gonon. 2017. "Poor Replication Validity of Biomedical Association Studies Reported by Newspapers." *PLoS One* 12:e0172650.

Dunning, T., G. Grossman, M. Humphreys, S. Hyde, C. McIntosh, and G. Nellis. 2019. *Information, Accountability, and Cumulative Learning: Lessons from Metaketa I*. Cambridge: Cambridge University Press.

Fanelli, D. 2018. "Is Science Really Facing a Reproducibility Crisis, and do We Need it to?" *Proceedings of the National Academy of Sciences* 115:2628–2631.

Fanelli, D., R. Costas, and J. P. A. Ioannidis. 2017. "Meta-Assessment of Bias in Science." *Proceedings of the National Academy of Sciences* 114: 3714–3719.

Findley, M. G., N. M. Jensen, E. J. Malesky, and T. B. Pepinsky. 2016. "Can Results-Free Review Reduce Publication Bias?: The Results and Implications of a Pilot Study." *Comparative Political Studies* 49:1667–1703.

Franco, A., N. Malhotra, and G. Simonovits. 2014. "Publication Bias in the Social Sciences: Unlocking the File Drawer." *Science* 345:1502–1505.

Franco, A., N. Malhotra, and G. Simonovits. 2015. "Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results." *Political Analysis* 23:306–312.

Franco, A., N. Malhotra, and G. Simonovits. 2016. "Underreporting in Psychology Experiments: Evidence from a Study Registry." *Social Psychological and Personality Science* 7:8–12.

Freese, J., and D. Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43:147–165.

Gerber, A. S., D. P. Green, and D. Nickerson. 2000. "Testing for Publication Bias in Political Science." *Political Analysis* 9:385–392.

Gerber, A. S., and N. Malhotra. 2008. "Publication Bias in Empirical Sociological Research: Do Arbitrary Significance Levels Distort Published Results?" *Sociological Methods & Research* 37:3–30.

Gerber, A. S., N. Malhotra, C. M. Dowling, and D. Doherty. 2010. "Publication Bias in Two Political Behavior Literatures." *American Politics Research* 38:591–613.

Gilbert, D. T., G. King, S. Pettigrew, and T. D. Wilson. 2016. "Comment on Estimating the Reproducibility of Psychological Science." *Science* 351:1037.

Graham, M. H., G. A. Huber, N. Malhotra, and C. H. Mo. 2020. "Irrelevant Events and Voting Behavior: Replications Using Principles from Open Science." Working Paper, Yale University.

Hainmueller, J., D. Hangartner, and T. Yamamoto. 2015. "Validating Vignette and Conjoint Survey Experiments Against Real-World Behavior." *Proceedings of the National Academy of Sciences* 112:2395–2400.
Horiuchi, Y., Z. Markovich, and T. Yamamoto. 2020. "Does Conjoint Analysis Mitigate Social Desirability Bias?" Working Paper, Massachusetts Institute of Technology.
Ioannidis, J. P. A., and T. A. Trikalinos. 2005. "Early Extreme Contradictory Estimates May Appear in Published Research: The Proteus Phenomenon in Molecular Genetics Research and Randomized Trials." *Journal of Clinical Epidemiology* 58:543–549.
Janz, N., and J. Freese. n.d. "Replicate Others as You Would Like to Be Replicated Yourself." *Political Science and Politics*, forthcoming.
Lupia, A., and C. Elman. 2014. "Openness in Political Science: Data Access and Research Transparency." *PS: Political Science & Politics* 47:19–42.
Malhotra, N. 2021. "The Scientific Credibility of Experiments." In *Advances in Experimental Political Science*, edited by J. N. Druckman and D. P. Green. New York: Cambridge University Press.
Martin, G. N., and R. M. Clarke. 2017. "Are Psychology Journals Anti-replication? A Snapshot of Editorial Practices." *Frontiers in Psychology* 8:1–6.
Monogan, J. E. 2015. "Research Preregistration in Political Science: The Case, Counterarguments, and a Response to Critiques." *PS: Political Science & Politics* 48:425–429.
Mullinix, K. J., T. J. Leeper, J. N. Druckman, and J. Freese. 2015. "The Generalizability of Survey Experiments." *Journal of Experimental Political Science* 2:109–138.
Mummolo, J., and E. Peterson. 2019. "Demand Effects in Survey Experiments: An Empirical Assessment." *American Political Science Review* 113:517–529.
Nosek, B. A. et al. 2015. "Promoting an Open Research Culture." *Science* 348:1422–1425.
Nosek, B. A., C. R. Ebersole, A. C. DeHaven, and D. T. Mellor. 2018. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences* 115:2600–2606.
Open Science Collaboration . 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349:aac4716-1–aac4716-8.
Rosenthal, R. 1979. "The File Drawer Problem and Tolerance for Null Results." *Psychological Bulletin* 86:638–641.
Teele, D. L., and K. Thelen. 2017. "Gender in the Journals: Publication Patterns in Political Science." *PS: Political Science & Politics* 50:433–447.