Northwestern | INSTITUTE FOR POLICY RESEARCH

Working Paper Series

WP-19-04

Why Replications Do Not Fix the Reproducibility Crisis: A Model and Evidence from a Large-Scale Vignette Experiment

Adam J. Berinksy
Professor of Political Science
Massachusetts Institute of Technology

James N. Druckman
Payson S. Wild Professor of Political Science and IPR Fellow
Northwestern University

Teppei Yamamoto
Assistant Professor of Political Science
Massachusetts Institute of Technology

Version: January 29, 2019

DRAFT

Please do not quote or distribute without permission.

ABSTRACT

Scientists have become increasingly concerned that "most published research findings are false" Ioannidis (2005), and have emphasized the need for replication studies. Replication entails a researcher repeating a prior research study with newly collected data. The mixed results of large-scale replication efforts have led some to conclude there is a "reproducibility crisis": false positives are pervasive. One solution is to encourage more replications. Yet, replication studies can alter the published literature only if they actually are published. And it may well be that replication studies themselves are subject to "publication bias." The researchers offer a micro-level model of the publication process involving an initial study and a replication. The model incorporates possible publication bias both at the initial and replication stages. This enables them to investigate the implications of publication biases on various statistical metrics of evidence quality. They then estimate the key parameters of the model with a large-scale vignette experiment conducted with political science professors teaching at Ph.D.-granting institutions in the United States. Their results show substantial evidence of publication bias: on average, respondents judged statistically significant results about 20 percentage points more likely to be published than statistically insignificant results. They further find evidence of what they call a "gotcha bias." Replication studies that run contrary to the existing literature are more likely to be published than those consistent with past research. Publication biases at the replication stage also can lead to the appearance of increased reproducibility even when there are actually more false positive results entering the published literature.

The authors thank James Dunham, Shiyao Liu, Chris Peng, Robert Pressel, Blair Read, and Jacob Rothschild for research assistance. They are grateful to Donald P. Green, Melissa Sands, 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.

1 Introduction

Replication is a hallmark of science. In the ideal, all empirical research findings would be subject to replication with knowledge accumulating as replications proceed. There has been increasing concern that such an "ideal" would paint an unflattering portrait of science – a recent survey of scientists found that 90% of respondents agreed there is a reproducibility crisis Baker (2016). Evidence of such a crisis comes, in part, from the Open Science Collaboration (OSC) project that replicated just 36% of initially statistically significant results from 100 previously published psychology experiments Open Science Collaboration (2015).

One known possible driver of the "replication crisis" is publication bias at the initial stage: that is, the published literature overstates statistically significant results because those are the only kind that survive the publication process Brown, Mehta and Allison (2017). Non-significant results are instead relegated to the discipline's collective "file drawer" Rosenthal (1979). When publication decisions depend on factors beyond research quality – such as statistical significance – the emergent scientific consensus may be skewed. Encouraging replication seems to be one way to correct a biased record of published research resulting from this file drawer problem Klein (2014); Bohannon (2015); Nosek et al. (2015). Here we consider a broader set of problems. In the current landscape, one must also consider potential publication biases at the replication stage. The reality is that individual replication studies also face a publication hurdle.

We present a model and a survey experiment that captures the publication process for initial and replication studies. In so doing, we introduce a distinct type of publication bias, what we call "gotcha bias." This bias occurs only for replication studies such that the likelihood of publication increases if the replication contradicts the findings of the original study. After all, not all replications see print; it may be that a study that runs contrary to published work is more likely to see the light of day.

Our model also shows that the common metric used to assess the replication success – the "reproducibility" rate (i.e., proportion of published replication results that successfully reproduce the original positive finding) – is not affected by initial study publication bias (i.e., the file drawer). In other words, low rates of reproducibility such as that found by the OCS study do *not* stem from a file drawer problem in initial studies. Further, our empirical results suggest that publication biases occur equally frequently at the replication phase than the initial study phase. Moreover, the

gotcha bias may, in practice, exacerbate the false positive rate. In short, encouraging replication by itself will not fix problems due to publication bias or the reproducibility crisis more broadly.

2 Model of Publication Decisions

We consider two 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 hypothesis test that rejects the null hypothesis of no effect) is more likely to be published than a negative test result (i.e., a hypothesis test that does not reject the null hypothesis), *ceteris paribus*. In other words, the published record of research is skewed away from the true distribution of the research record, overstating the collective strength of the findings. For example, if 1 out of 10 studies showed that sending varying types of health-related text messages leads people to eat less fatty food, and only that 1 study is published, the result is a mis-portrayal 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, Mehta and Allison (2017). There is a large theoretical and empirical literature that documents this type of publication bias and its consequences Gerber and Malhotra (2008); Franco, Malhotra and Simonovits (2014); Fanelli, Costas and Ioannidis (2017).

Second, we consider the possibility that the process of replication itself could induce bias. We define *gotcha bias* to be the phenomenon that, *ceteris paribus*, a negative test result is more likely to be published when there exists a prior study that tested the same hypothesis and had a positive result than when a prior study also showed a negative test result (and vice versa). 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.

Consider the following example that roughly captures what the gotcha bias entails. A published study shows that exposure to imagery of natural disasters causes people to worry more about climate change (i.e., there is a statistically significant effect). There is a quality replication study that fails to find a statistically significant result and it is published – i.e., the editors and/or reviewers find the failed replication "exciting enough" to publish it. Then consider a distinct

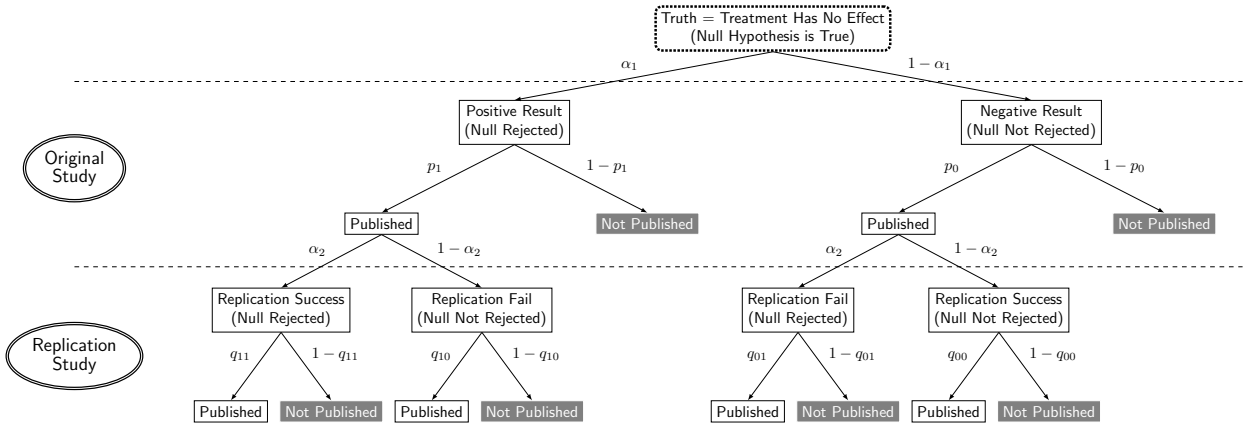

Figure 1: A Model of Publication Process with Two Stages

published study that finds no evidence that reminding people about hot days causes them to worry more about climate change (i.e., there is a no statistically significant effect). A replication takes place that is of equal quality to the natural disaster imagery replication. Like the other replication, it also fails to find a significant result; however, this one is not published since its consistency with the original study is not sufficiently "exciting." This is an example of the gotcha bias – two replications of equal quality fail to find significant results but only the one that contradicted prior work is published. The result is an inaccurate portrait of accumulated knowledge.

The gotcha bias is a concept applicable to replication studies that collect new data in an attempt to reproduce results of previous, original studies using similar study designs. The hypothesized mechanism behind this bias is again a proclivity for novelty and sensationalism. Although 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) – 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 rather simplified model of the publication process in order to study the consequences of these two types of publication biases (Figure 1). The model starts with a null hypothesis (e.g., no treatment effect) as the true state of the world and proceeds as follows. (We also consider a parallel model for the case where the null is false, as discussed in the Supporting Information.) First, 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, goes through a peer-review process and gets 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 what we call the file drawer bias.

Second, 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 then go through a peer-review process similar to the first stage, except that the publication probability now depends on both test results from the current and previous stages (q_{11}, q_{10}, q_{01}, q_{00}). 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), then we call it the gotcha bias for insignificant replication results. Similarly, $q_{11} < q_{01}$ would represent a gotcha bias for significant replication results.

We note that our model is not intended to be an accurate, descriptive portrayal of the actual scientific practice. For example, not all published results will ever be replicated with fresh samples, even with the current push towards credible research. It is indeed possible that researchers might strategically choose what existing studies to replicate and which replication results to submit for review, given their perception about publication biases. Instead, our goal here is to examine how the idealized model of replication science, as exemplified by the OSC study Open Science Collaboration (2015), would differ if we "perturbed" it by adding possible publication bias for replication studies themselves.

To study the consequences of the two types of publication biases, we specifically consider the following metrics of evidence quality in published studies.

Definition 1 *[Actual False Positive Rate (AFPR) in published replication studies]*

$$\tilde{\alpha}_2 = \Pr(\textit{replication test significant} \mid \textit{replication published}, \textit{the null is true})$$

Definition 2 *[Reproducibility Rate]*

$$R = \Pr(\textit{replication test significant} \mid \textit{original test significant and published, replication published})$$

The AFPR represents the proportions of the positive results in published replication studies that are actually false, i.e., where the null hypotheses are in fact true. In the ideal world, this rate would be equal to the nominal FPR (α_2) that the tests in replication studies are theoretically designed to achieve. However, $\tilde{\alpha}_2$ will diverge from their designed type-I error rate due to the two kinds of publication biases we consider. It is well known that file drawer bias tends to inflate the false positive rate by disproportionately "shelving" negative results that correctly identifies true null hypotheses Rosenthal (1979). The effect of gotcha bias, however, has not been documented.

Our analysis reveals that the gotcha bias affects the AFPR in replication results in several interconnected ways. Specifically, *ceteris paribus*, gotcha bias for significant replication results exacerbates the inflation of the AFPR in replication results, while gotcha bias for insignificant replication results has the opposite effect of deflating the AFPR closer to the nominal FPR in replication results. Intuitively, this tends to occur because gotcha bias makes publication of false positive results in replication studies more likely when the original studies (correctly) accepted the same null hypotheses, but less likely when the original test (incorrectly) also rejected the hypothesis. Moreover, in the presence of gotcha bias, we find that the file drawer bias in *original* studies has the effect of decreasing AFPR in replication results. The net effect of gotcha bias on the replication-study AFPR is thus ambiguous and depends on which of these mutually countervailing effects is dominant. Note, however, that our model also implies that the AFPR can never be less than the nominal FPR of replication tests as long as the replication results themselves are also subject to non-negative file drawer bias (i.e., $q_{01} > q_{00}$ and $q_{11} > q_{10}$). The Supporting Information contains a more precise discussion, with reference to the exact mathematical expression for $\tilde{\alpha}_2$ in terms of our model parameters.

The reproducibility rate refers to the proportion of the published replication test results that successfully reproduce the positive original results. In other words, R asks "How often do replication studies that are published confirm the positive results of the original published studies?" This is the central metric used in the aforementioned OSC study that reported statistically significant results for 36% of initially statistically significant effects. The authors concluded "there is room to improve reproducibility in psychology," attributing the low rate to publication bias among other factors.

In the Supporting Information, however, we provide an exact formula for R in terms of the

model parameters that casts serious doubt that low reproducibility stems, at all, from the file drawer problem. Our model implies that the reproducibility rate should have no direct relationship with the file drawer bias in the *original* studies. This may be surprising given that a file drawer bias in original studies makes for the over-representation of false positives in the published literature. Intuitively, one might thus expect fewer successful replications. However, the twist is that file drawer bias in original studies also makes true positives more likely to enter the published literature. In fact, both false positives and true positives are equally overrepresented compared to true negatives and false negatives. The implication is that the ratio of false positives to true positives among original published studies is the same as it was prior to the initial publication process (i.e., before a file drawer bias). That is, file drawer bias does not differentially overrepresent false positives compared to true positives.

Instead of original study file drawer bias, our analysis show that what determines the reproducibility rate more is the power of the original and replication studies, publication bias in the replication studies themselves, and what Ioannidis calls the "pre-study odds" of a true relationship (i.e., proportion of false nulls in the field) Ioannidis (2005). In particular, publication bias in replication studies can either increase or decrease R, depending on the relative importance of file drawer bias and gotcha bias. Moreover, regardless of the presence of publication bias, the reproducibility rate can be easily close to 20% or even lower in low-power studies or when researchers are testing mostly true nulls.

3 Survey Experiment

To illustrate how our simple model can shed light on the "reproducibility crisis" in a scientific discipline, we conducted a large-scale vignette survey experiment among members of one discipline. Our goal is to estimate, to the extent possible using hypothetical scenarios, the amount of file drawer bias and gotcha bias in a scientific discipline. We also seek to display how these biases influence the AFPR and the reproducibility rate.

Our population constituted 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. While caution should be

taken in generalizing to other disciplines, it is noteworthy that, as with other scientific disciplines, questions of publication bias have become central to ongoing discussions and initiatives in political science Lupia and Elman (2014); Monogan (2015). We thus expect many respondents were cognizant of biases introduced by basing publication decisions on statistical significance.

Participants were sent to a link where they were provided 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 5 different vignettes, each concerning a single paper. We asked them to act as if they were an author and asked whether they would submit the paper to a journal. Each respondent then received another 5 vignettes where they were asked to play the role of a reviewer. Here, we asked whether they would recommend the paper be published. Finally, we asked whether the responded had ever edited a journal. If they had, we gave them 5 additional vignettes. These vignettes asked the respondent whether he or she would publish the paper.

Each vignette randomly varied a host of features; Figure 2 presents a vignette for the "author" condition with all possible variations for those features indicated in square brackets. After each vignette, we asked the following question as our main outcome variable: "If you were the author of this paper, what is the percent chance you would send this paper to a peer-reviewed journal?" Versions for the "reviewer" and "editor" conditions are provided in the Supporting Information. In reporting the results, we focus on treatment variations that have direct bearing on our model parameters (i.e., statistical significance, whether the study was an original or an replication study, and sample size). Ignoring 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. The data and analysis code will be made available at a public repository upon publication of the paper.

4 Results

4.1 Estimating Two Types of Publication Bias

We begin by asking how much evidence our data show of the two types of publication bias – file drawer bias and gotcha bias. Figure 3 presents the average percent chance of taking an action toward publication (e.g., sending out a paper as an author, recommending publication as

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

Suppose that you were an author of a paper reporting the results of an empirical study. The study aims at testing a hypothesis with quantitative data and has the following characteristics.

- [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 less than .01)/ significant by conventional standards (e.g., p-value of less than .05)/ not significant by conventional standards (e.g., p-value of greater than .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: Sample Vignette from the Experiment. The phrases in square brackets separated by slashes represent alternative texts that are randomly and independently assigned for each vignette.

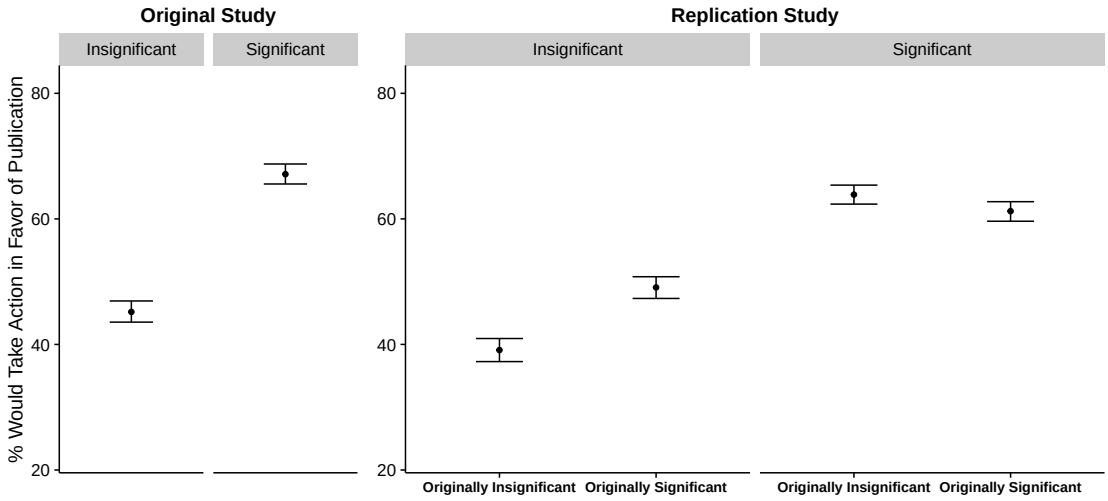

Figure 3: Evidence of Two Types of Publication Bias. In each plot, the solid dot represents the estimated probability of a respondent taking an action in favor of publishing the hypothetical study for a given combination of study characteristics indicated at the top and bottom. The vertical bars represent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

a reviewer, and supporting publication as an editor) that the respondents gave to different types of hypothetical papers described in our randomly generated vignettes, along with 95% confidence intervals. 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. We also combine the two conditions in which test results are described as statistically significant (i.e., .01 or .05 level) into a single category in our analysis.

Consistent with extant work Franco, Malhotra and Simonovits (2014), our estimates for original studies show clear evidence of file drawer bias. 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 (95% CI = [65.6%, 68.7%]), they gave only a 45.2% chance of doing the same for a paper with a non-significant finding ([43.6%, 46.9%]).

More interestingly, our result clearly suggests 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. On average, respondents reported a 62.5% chance of moving a significant test result in a replication study toward publication ([61.3%, 63.8%]), whereas they only gave a 44.1% chance for a non-significant replication test result ([42.6%, 45.6%]). Thus, regardless of whether a replication study "succeeds" or "fails" to reproduce the original finding, that replication is more

likely to be published when its result is statistically significant than when it is a null finding. It is also noteworthy that replication studies are less likely to be published than original studies: averaging across the significance conditions, respondents indicated less chance of taking an action toward publishing a replication result than an original test result by 2.8% points (t=-4.65, p<0.00). Our findings therefore imply that extra efforts may need be made in encouraging publication of replication studies in general.

Turning to gotcha bias, our results show clear evidence that this more subtle form of publication bias occurs. Respondents assigned a 49.1% chance of submitting/recommending/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 non-significant finding ([37.3%, 40.9%]). 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 ([62.5%, 65.4%]). This percentage drops to 61.2% ([59.6%, 62.7%]) for a significant replication result that successfully reproduces an earlier significant finding (t=-2.92, p<0.01). Thus, for replication studies, there is an increased probability of supporting publication of surprising results in either direction – findings that overturn previously published studies are privileged in the publication process. Importantly, however, the gotcha effect only goes so far; even at the replication stage, the standard file drawer problem exerts influence. Our overall evidence indicates that the file drawer bias is of larger magnitude than the gotcha effect for replication studies. Thus, replication results that find statistically significant effects are more likely to move towards publication than insignificant results, no matter what the original results may be.

In sum, in a world in which people are unlikely to seek to publish null results, there is a danger of biased collective findings because only significant results find an audience – not just in the initial stage, but in replications as well. This fact is further compounded by gotcha bias, of which we find smaller but still substantial evidence. These results suggest that, for example, had the Open Science Collaboration's replications been independently submitted, *more than half might not have been published*.

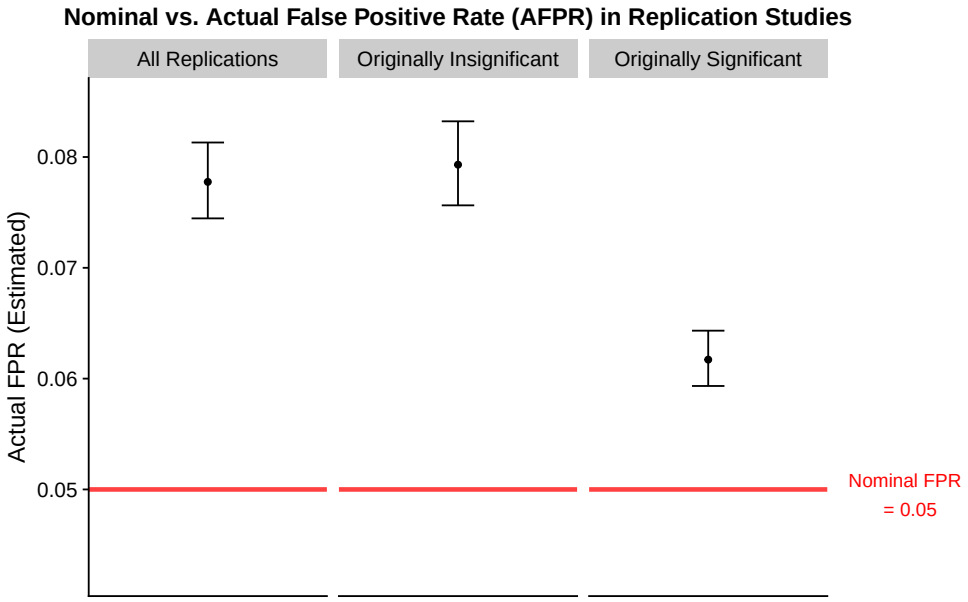

Figure 4: Estimates of AFPR for Published Replication Study Results. The solid dots represent the estimated AFPR for all published replication results (left), published replications of originally insignificant results (middle) and published replications of originally significant findings (right) based 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.

4.2 Estimating Actual False Positive Rates

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

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 ($\tilde{\alpha}_2$) 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 and gotcha 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 replication test is designed. Publication bias thus leads to 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 also 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 an elixir to correct the scientific record – publication bias in replication studies lead to an AFPR that exceeds what would occur by chance.

4.3 Estimating Reproducibility

Finally, we look at another important metric of evidence quality: the reproducibility rate for originally positive results in published replication studies. In addition to publication bias, the key parameters that determine reproducibility are power and the proportion of true null hypotheses that are tested in a given scientific field. We therefore first simulate the reproducibility rate under different scenarios with respect to those two key parameters, in the assumed absence of publication bias. These simulated theoretical values of the reproducibility rate are plotted by dashed lines in Figure 5, assuming 0.05-level significance tests in both stages. We provide four sets of reproducibility simulations, each corresponding to a specific sample size (50, 150, 1,000 and 5,000) in our vignette experiment. The sample sizes are translated to implied power values in these simulations. (Results for different combinations of assumed alpha levels are provided in the Supporting Information.)

As mentioned above, the reproducibility rate varies widely depending on these parameters.

When hypotheses are tested with a small sample size (such as N =50), these tests have low statistical power. The reproducibility rate therefore remains low even when researchers are all testing for effects that are true. This result occurs because a large majority of replication studies with such low power will fail to detect those effects. In contrast, high-powered replication studies can reproduce original positive results with high probability even when the pre-study odds of true effects are rather low, because such studies are unlikely to mis-classify those few true effects as insignificant. Of course, the reproducibility rate eventually converges to the nominal type-I error rate of the replication test as the proportion of true nulls approaches one, at which point the tests are merely "replicating" the wrong results at their designed false positive rate.

What happens to the reproducibility rate when we incorporate our estimated levels of publication bias in its calculation? Here, we again use our vignette survey data to produce such estimates corresponding to each of the simulated scenarios (solid lines in Figure 5) along with their 95% confidence bands (shaded regions). Somewhat counterintuitively, we find that the publication bias exhibited in our experiment would *improve* the reproducibility rate by statistically significant margins across all possible values of statistical power and the pre-study odds of true effects. This result stems from the predominance of file drawer bias that we find even in replication studies. That is, because positive results are published more often than negative results, the "successful" reproduction of original positive results are overrepresented in published replication studies compared to negative reproduction results. The gotcha bias does counterbalance this tendency to some extent, but because this bias is smaller than the file drawer bias, the net effect is to increase the reproducibility rate.

To be clear, as is suggested by our model, the reproducibility rate is unaffected by the original study file drawer bias (recall this is because original study file drawer bias does not differentially overrepresent false positives compared to true positives). What does matter for the reproducibility rate is the nature of the file drawer bias and the gotcha bias *in the replication study*. We find that, empirically, publication biases in replication studies actually *increase* the reproducibility rate. This result should not be taken as a recommendation to encourage publication bias in replication studies, however. Recall that the same replication study biases that increase the reproducibility rate also increase the AFPR. This is a stark reminder that reproducibility is not a direct indicator of whether study results represent true effects or not. It is rather a metric of the regularity of

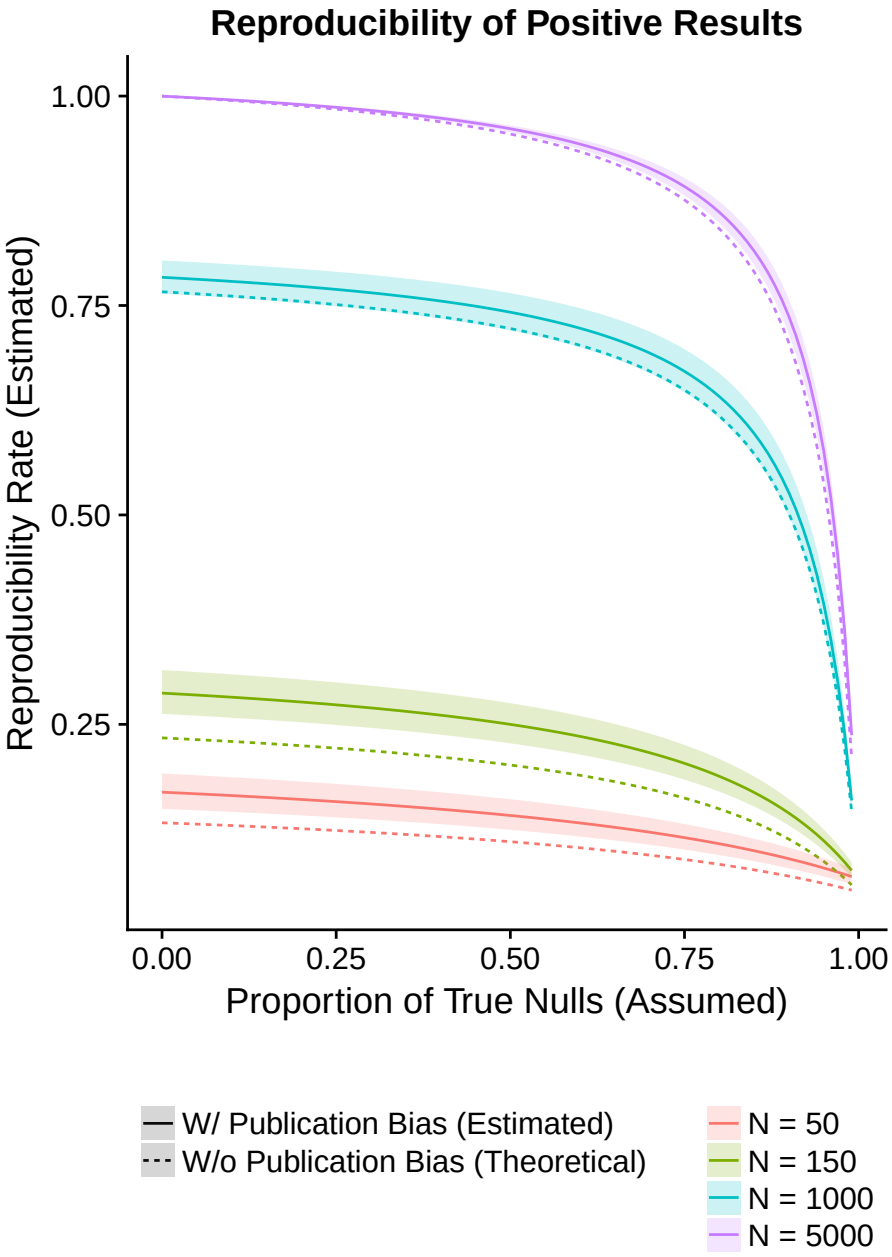

Figure 5: Estimates of Reproducibility as Function of Power and "Pre-Study Odds." The dashed lines represent the simulated theoretical reproducibility rate for a given combination of the assumed proportion of the true null hypotheses (horizontal axis) and the power value implied by a sample size (four different colors, as indicated in the plot legend) in the absence of publication bias. The solid lines show the estimated reproducibility rates with the publication bias estimated from the vignette data, with 95% confidence bands indicated by the shaded areas. The results are for the assumed nominal significance level of 0.05 for both the original and replication tests.

finding positive test results whether or not they are indicative of the true state of the world. When publication biases persist, the reproducibility rate metric should be used with great care.

5 Conclusion

The "reproducibility crisis" in science has generated substantial discussion and a number of efforts to encourage wide-scale replications. Indeed, one can only assess and ultimately address such a crisis if replications occur and become part of the scientific literature. This process of learning is not as straightforward as often thought. Our model isolates how an idyllic replication process works, showing two distinct types of biases that can skew the published literature. Moreover, the model shows that reproducibility is not contingent on publication bias in initial studies but rather is affected by power and bias in replication publication, *inter alia*.

Our survey experiment results show that even in the midst of widespread discussions about the importance of replication and publication bias, scholars still exhibit these biases. Moreover, changing incentives and behaviors is not easy: researchers, like everyone else, exhibit confirmation biases that lead them to privilege the preconception that statistical significance is critical Bollen et al. (2015). We find these preconceptions carry over to their evaluation of replication studies. Moreover, incentivizing journals to be "more encouraging" of replications Nosek et al. (2015) could perhaps backfire since the gotcha bias might lead to a mis-portrayal of accumulated knowledge and possibly mis-incentivize researchers conducting replications. Put simply, there are no easy solutions.

Addressing publication bias, with respect to replication studies, will likely require broader institutional change such as a collective commitment to pre-registration, a shift to open access journals and/or required publication, or blind review Nosek et al. (2017); Brown, Mehta and Allison (2017). Each of these reforms involve considerable resource investments and have downsides such as potentially prioritizing certain types of research Coffman and Niederle (2015); Freese and Peterson (2017), and/or resulting in an overwhelming amount of information to assess (although see de Winter and Happee (2013)). There are more modest approaches including having journals publish brief "replication" sections and incentivizing citations to replications Coffman et al. (2017). These latter ideas could help attenuate the replication publication bias and we believe they are

worth exploring on a larger-scale.

We also urge scholars to take a step back in assessing the "reproducibility crisis." While we do not question extant evidence from well-known studies and meta-analytic literature which shows replication inconsistencies Ioannidis and Trikalinos (2005); Fanelli, Costas and Ioannidis (2017), we also note that other large-scale replication attempts in political science Mullinix et al. (2015) and economics Camerer et al. (2016) were relatively more successful than the Open Science Collaboration results. The question then is what research areas, theories, methods, and context affect the likelihood of replication, and this understanding, in turn, would facilitate the assessment of replications.

References

Baker, Monya. 2016. "Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the crisis rocking science and what they think will help." Nature 533(7604):452–455.

Bohannon, John. 2015. "Many psychology papers fail replication test." Science 349(6251):910–911.

Bollen, Kenneth, JT Cacioppo, RM Kaplan, JA Krosnick and JL Olds. 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." Technical Report . Retrieved at: www. nsf. gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report. pdf.

Brown, Andrew W, Tapan S Mehta and David B Allison. 2017. "Publication Bias in Science: What Is It, Why Is It Problematic, and How Can It Be Addressed?" The Oxford Handbook of the Science of Science Communication p. 93.

Camerer, Colin F, Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan et al. 2016. "Evaluating replicability of laboratory experiments in economics." Science 351(6280):1433–1436.

Coffman, Lucas C and Muriel Niederle. 2015. "Pre-analysis plans have limited upside, especially where replications are feasible." The Journal of Economic Perspectives 29(3):81–97.

Coffman, Lucas C, Muriel Niederle, Alistair J Wilson et al. 2017. "A Proposal to Organize and Promote Replications." American Economic Review 107(5):41–45.

de Winter, Joost and Riender Happee. 2013. "Why selective publication of statistically significant results can be effective." PloS one 8(6):e66463.

Fanelli, Daniele, Rodrigo Costas and John PA Ioannidis. 2017. "Meta-assessment of bias in science." Proceedings of the National Academy of Sciences p. 201618569.

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

Freese, Jeremy and David Peterson. 2017. "Replication in Social Science." Annual Review of Sociology (0).

Gerber, Alan S and Neil Malhotra. 2008. "Publication bias in empirical sociological research: Do arbitrary significance levels distort published results?" Sociological Methods & Research 37(1):3–30.

Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. "Validating vignette and conjoint survey experiments against real-world behavior." Proceedings of the National Academy of Sciences 112(8):2395–2400.

Ioannidis, John P. A. 2005. "Why Most Published Research Findings are False." PLoS Medicine 2(8):696–701.

Ioannidis, John PA and Thomas 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(6):543–549.

Klein, Stanley B. 2014. "What can recent replication failures tell us about the theoretical commitments of psychology?" Theory & Psychology 24(3):326–338.

Lupia, Arthur and Colin Elman. 2014. "Openness in political science: Data access and research transparency." PS: Political Science & Politics 47(1):19–42.

Monogan, James E. 2015. "Research preregistration in political science: The case, counterarguments, and a response to critiques." PS: Political Science & Politics 48(3):425–429.

Mullinix, Kevin J, Thomas J Leeper, James N Druckman and Jeremy Freese. 2015. "The generalizability of survey experiments." Journal of Experimental Political Science 2(2):109–138.

Nosek, Brian A, Charles R Ebersole, Alexander DeHaven and David Mellor. 2017. "The Preregistration Revolution.".

Nosek, Brian A, George Alter, George C Banks, D Borsboom, SD Bowman, SJ Breckler, S Buck, Christopher D Chambers, G Chin, G Christensen et al. 2015. "Promoting an open research culture." Science 348(6242):1422–1425.

Open Science Collaboration. 2015. "Estimating the reproducibility of psychological science." Science 349(6251):aac4716.

Rosenthal, Robert. 1979. "The file drawer problem and tolerance for null results." Psychological Bulletin 86(3):638.