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

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

[a]Department of Political Science, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139; [b]Department of Political Science, Northwestern University, 601 University Place, Evanston, IL 60208

This manuscript was compiled on January 9, 2018

There has recently been a dramatic increase in concern about whether "most published research findings are false" (Ioannidis 2005). While the extent to which this is true in different disciplines remains debated, less contested is the presence of "publication bias," which occurs when publication decisions depend on factors beyond research quality, most notably the statistical significance of an effect. Evidence of this "file drawer problem" and related phenomena across fields abounds, suggesting that an emergent scientific consensus may represent false positives. One of the most noted responses to publication bias has involved increased emphasis on replication where a researcher repeats a prior research study with different subjects and/or situations. But do replication studies suffer themselves from publication bias, and if so, what are the implications for knowledge accumulation? In this study, we contribute to the emerging literature on publication and replication practices in several important ways. First, we 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, enabling us to investigate the implications of such bias on various statistical metrics of quality of evidence. Second, we estimate the key parameters of the model with a large-scale vignette experiment conducted with the universe of political science professors teaching at Ph.D.-granting institutions in the United States. Our results show substantial evidence of both types of publication bias in the discipline: On average, respondents judged statistically significant results about 20 percentage points more likely to be published than statistically insignificant results, and contradictory replication results about six percentage points more publishable than replications that are consistent with original results. Based on these findings, we offer practical recommendations for the improvement of publication practices in political science.

Keyword 1 | Keyword 2 | Keyword 3 | Keyword 4 | Keyword 5

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 (1). 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 (2).

One 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 (3). Non-significant results are instead relegated to the discipline's collective "file drawer" (4). When publication decisions depend on factors beyond research quality, 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 (5–7). Yet, in the current landscape, one must also consider potential publication biases at the replication stage. While this was not an issue for OSC since they relied on over 250 scholars to replicate the 100 studies, the reality is that individual replication studies also face a publication hurdle.

In what follows, 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 original study. We also show 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) according to our model. In other words, a low reproducibility rate provides no insight into whether there is a file drawer problem. Finally, our empirical results from a survey of political scientists suggest that publication biases occur with greater frequency at the replication phase and the gotcha bias may, in practice, exacerbate the false positive rate.

Model of Publication Decisions

We consider two distinct types of publication bias. First, we examine the well-known file drawer problem (4). *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 is published, the result is a mis-portrayal of the effect of text messages. There is a large theoretical and empirical literature that documents the evidence and possible consequences of this

Please provide details of author contributions here.

Please declare any conflict of interest here.

[1]A.J.B., J.N.D. and T.Y. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: teppei@mit.edu

type of publication bias (8–10). The file drawer bias reflects an entrenched culture that prioritizes statistical significance and novelty, as well as a funding system that rewards positive findings (3).

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 positive test result is more likely to be published when there exists a prior study that tested the same hypothesis and had a negative result than when a prior study showed a positive 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 replication that run counter to existing findings, as compared to the true distribution of the research record. For example, there might be 10 similar studies of text messaging effects that each purport to replicate corresponding previous studies, and they all find significant effects but 9 of them are successful replications of the original studies that found equally significant results. If only the one study that finds a significant effect contrary to a previous non-significant study is published, the larger research community will see a distorted portrayal of the research record. The gotcha bias is a concept applicable to replication studies that 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 (11) – 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 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 Supplementary Materials.) 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 their publication probability now depends on both test results from the current and previous stages (q_{11}, q_{10}, q_{01}, q_{00}). If $q_{11} < q_{01}$ (such that a positive replication result is more likely to be published when it contradicts a previous negative result than when it confirms an existing positive result), then we call it the gotcha bias for significant replication results. Similarly, $q_{10} > q_{00}$ would represent a gotcha bias for insignificant 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 in reality, 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 (2), 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]*

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

Definition 2 *[Reproducibility]*

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

The AFPRs represent 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, these rates would be equal to the nominal type-I error rate 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 (4). 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 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 $q_{11} > q_{10}$). Supplementary Materials contain a more precise discussion, with reference to the exact mathematical expression for $\tilde{\alpha}_2$ in terms of our model parameters.

Reproducibility refers to the proportion of the published replication test results that successfully reproduce the positive

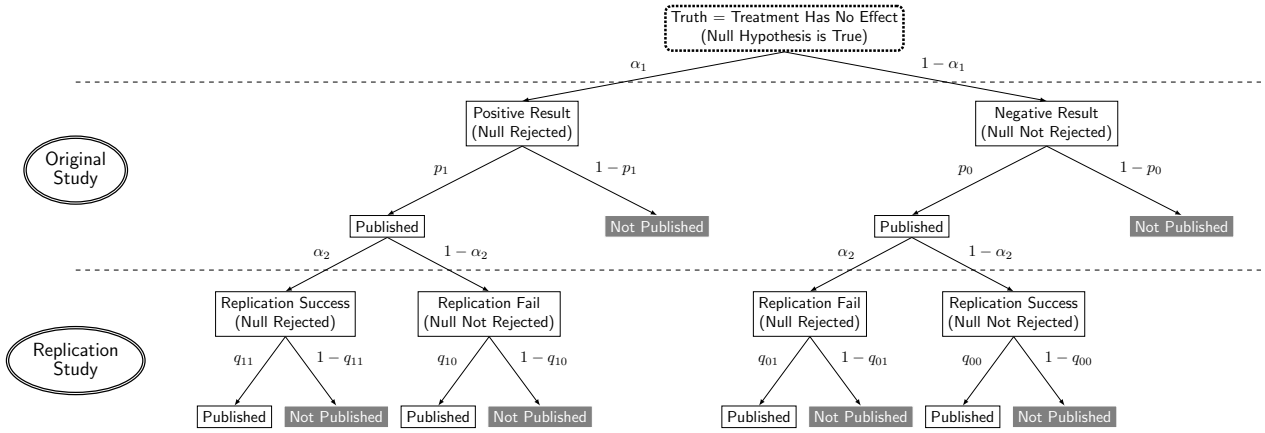

Fig. 1. A Model of Publication Process with Two Stages

original results. In other words, R asks "How often do replication studies that are published confirm the positive results of the original published test results?" This is one of the metrics employed by the Open Science Collaboration (2); that project involved more than 250 scholars who attempted to replicate 100 psychology experiments from three highly ranked journals. They reproduced statistically significant results for 36% of initially statistically significant effects and concluded "there is room to improve reproducibility in psychology," attributing the low rate to publication bias among other factors.

In Supplementary Materials, however, we provide an exact formula for R in terms of the model parameters that casts serious doubt on their interpretation of the low reproducibility. Indeed, our model implies that reproducibility should have *no direct relationship with the file drawer bias in the original studies*, because *ceteris paribus*, increase in FPR due to file drawer bias should be offset by a corresponding decrease in the type-II error rate. Our analysis instead points at more important determinants of reproducibility, including 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) (12). 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, reproducibility can be easily close to 20% or even lower in low-power studies or when researchers are testing mostly true nulls.

Survey Experiment

To illustrate how our simple model can shed light on the "reproducibility crisis" in a scientific discipline, we conduct a large-scale vignette survey experiment. Our primary goal in the calibration analysis here is not to provide an exact estimate of evidence quality in a scientific discipline, but to illustrate the utility of our framework with data that have reasonable empirical relevance.

Our population constituted all political science department faculty at Ph.D. granting institutions based in the United States. The Supplementary Materials contain 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 (13, 14).

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 (15)). Each respondent was provided with 5 different vignettes, each concerning a single paper. We requested 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, many of which are not relevant to our focus here (see Supplementary Materials for the full text of the vignette with all possible variations). For our purposes, a sample vignette for the "author" condition is presented in Figure 2. After each vignette, we asked the following question as our main our variable: "If you were the author of this paper, what is the percent chance you would send this paper to a peer-reviewed journal?" The entire wording included some normalization and is also provided in Supplementary Materials along with versions for the "reviewer" and "editor" conditions.

Results

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 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 Supplementary Materials. We also combine the two conditions in

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 are evaluating a paper reporting the results of an empirical study in your own subfield. The study aims at testing a hypothesis with quantitative data and has the following characteristics.

- [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)].
- Seemingly sound in terms of methods and analysis.

Fig. 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.

which test results are described as statistically significant into a single category in our analysis.

As expected, given extant work (9), 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 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 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: on average, respondents indicated 2.8% less chance of taking an action toward publishing a replication result than an original test result (t=-4.65, p<0.00). Our findings imply that extra efforts may need be made in encouraging publication of replication studies in general.

Turning to gotcha bias, our results again show clear evidence that this more subtle form of publication bias is fostered among political scientists, at least when asked about hypothetical publication decisions. Respondents assigned 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 standard 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 would not have been published.*

Estimating Actual False Positive Rates. In addition to estimating the two types of publication bias, our survey experimental data allow us to make inferences about the aforementioned key metrics of evidence quality: the AFPR and reproducibility. Here, we provide our estimates of the AFPR based on our formula and the vignette data (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. If the original study reported a significant test result for the hypothesis at the 0.01 level, the AFPR in its replication test is estimated to be 0.063 (95% CI = [0.060, 0.067]). The estimate drops slightly to 0.060 ([0.057, 0.064]) if the original study rejected the null at the 0.05 level. In contrast, a positive replication result that is significant at the 0.05 level is estimated to have an AFPR of 0.079 ([0.075, 0.083]) if the result is contrary to an original null result.

This nearly 0.02 point increase is a direct consequence of publication bias. A 0.05-level positive replication test result is a surprise and therefore a more publishable finding when the original result was non-significant. Moreover, we find that a large majority of true nulls would in fact be classified as non-significant by the original study (more precisely, 92.6 percent of the time if the nominal FPR of the test was 0.05, with 95% CI = [92.2, 92.9]). Thus, according to our model of publication process, those published significant test results that contradict existing non-significant results will contain more false positives than those studies that confirm earlier significant findings. In Supplementary Materials, we show that the same pattern emerges for the positive replication tests that use the nominal FPR of 0.01. 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.

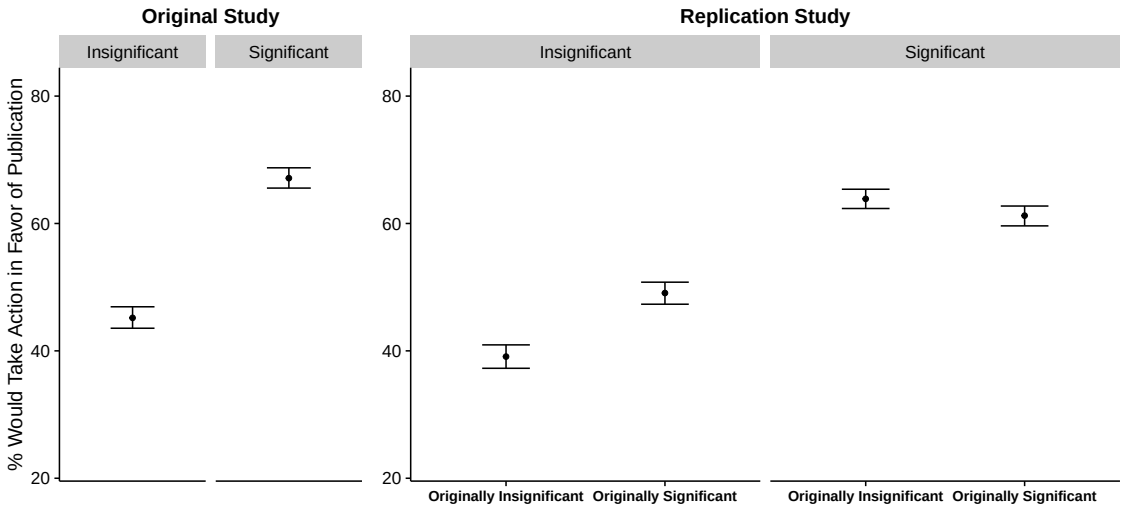

Fig. 3. Evidence of Two Types of Publication Bias.

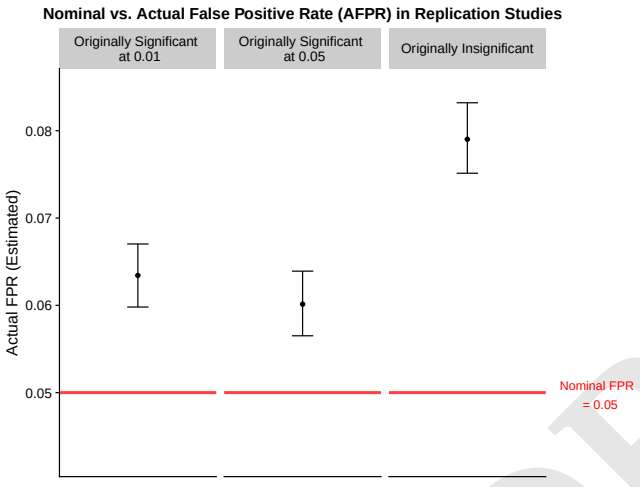

Fig. 4. Estimates of AFPR for Published Replication Study Results.

Estimating Reproducibility. Finally, we look at another important metric of evidence quality: the reproducibility of original 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 reproducibility under different scenarios with respect to those two key parameters, in the assumed absence of publication bias. These simulated theoretical values of reproducibility are plotted by dashed lines in Figure 5.

In Figure 4, we present a given combination of nominal type-I error rates in the original and replication studies. We provide four sets of reproducibility simulations, each corresponding to a specific sample size (50, 150, 1,000 and 5,000), and an implied power value, that was also used in our vignette experiment. As mentioned above, reproducibility 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. Reproducibility 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, reproducibility eventually converges to the nominal type-I error 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 reproducibility 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* reproducibility 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 reproducibility.

Thus, our finding suggests that publication bias in replication studies might actually have a positive impact on the standard metric of reproducibility assuming that the real world mirrors our vignette experiment result and the file drawer bias dominates gotcha bias in replication studies as we find through our vignette experiment. This result, however, should *not* be taken as a recommendation to encourage publication bias in replication studies. For one thing, we also find that reproducibility is more strongly determined by other factors such as power of the studies and the pre-study odds of true effects in a given discipline. More fundamentally, reproducibility is

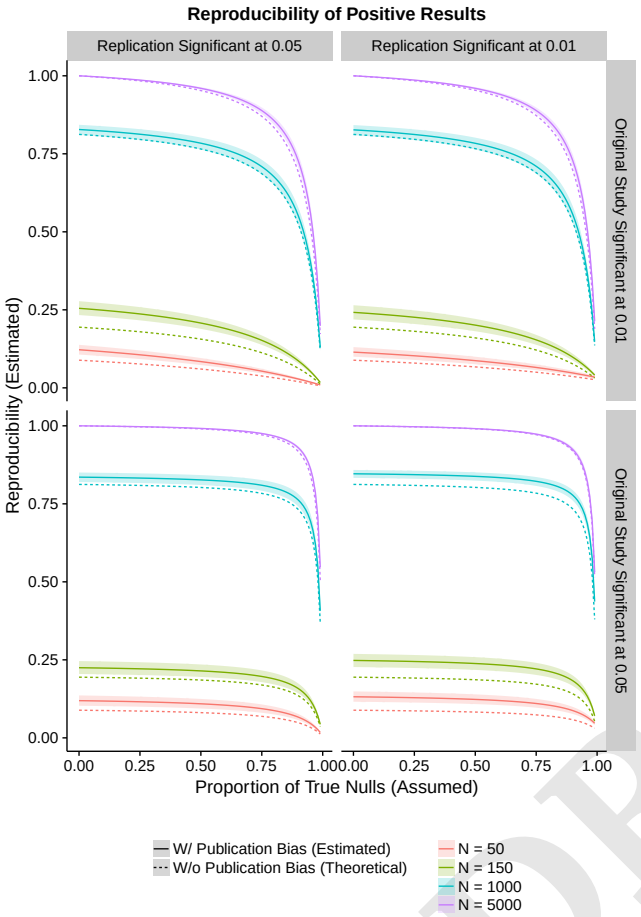

Fig. 5. Estimates of Reproducibility as Function of Power and "Pre-Study Odds."

not a direct indicator of whether study results represent true effects or not, but a metric of regularity in finding positive test results whether or not they are actually indicative of the true state of the world.

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 (16). We find these preconceptions carry over to their evaluation of replication studies. Moreover, incentivizing journals to be "more encouraging" of replications (7) could perhaps backfire since the gotcha bias might lead to a mis-portrayal of accumulated knowledge and possibly misincentivize 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 (3, 17). Each of these reforms involve considerable resource investments and have downsides such as potentially prioritizing certain types of research (18, 19), and/or resulting in an overwhelming amount of information to assess (although see (20)). There are more modest approaches including having journals publish brief "replication" sections and incentivizing citations to replications (21). 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 (10, 11), we also note that other large-scale replication attempts in political science (22) and economics (23) 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.

ACKNOWLEDGMENTS. We thank James Dunham, Jacob Rothschild, Robert Pressel, Chris Peng, and Shiyao Liu for research assistance. We are grateful to Melissa Sands and the participants at the 2017 Conference of the Society for Political Methodology for useful comments and suggestions.

1. Baker M (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.
2. Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science* 349(6251):aac4716.
3. Brown AW, Mehta TS, Allison DB (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.
4. Rosenthal R (1979) The file drawer problem and tolerance for null results. *Psychological Bulletin* 86(3):638.
5. Klein SB (2014) What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology* 24(3):326–338.
6. Bohannon J (2015) Many psychology papers fail replication test. *Science* 349(6251):910–911.
7. Nosek BA, et al. (2015) Promoting an open research culture. *Science* 348(6242):1422–1425.
8. Gerber AS, Malhotra N (2008) Publication bias in empirical sociological research: Do arbitrary significance levels distort published results? *Sociological Methods & Research* 37(1):3–30.
9. Franco A, Malhotra N, Simonovits G (2014) Publication bias in the social sciences: Unlocking the file drawer. *Science* 345(6203):1502–1505.
10. Fanelli D, Costas R, Ioannidis JP (2017) Meta-assessment of bias in science. *Proceedings of the National Academy of Sciences* p. 201618569.
11. Ioannidis JP, Trikalinos TA (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.
12. Ioannidis JPA (2005) Why most published research findings are false. *PLoS Medicine* 2(8):696–701.
13. Lupia A, Elman C (2014) Openness in political science: Data access and research transparency. *PS: Political Science & Politics* 47(1):19–42.
14. Monogan JE (2015) Research preregistration in political science: The case, counterarguments, and a response to critiques. *PS: Political Science & Politics* 48(3):425–429.
15. Hainmueller J, Hangartner D, Yamamoto T (2015) Validating vignette and conjoint survey experiments against real-world behavior. *Proceedings of the National Academy of Sciences* 112(8):2395–2400.
16. Bollen K, Cacioppo J, Kaplan R, Krosnick J, Olds J (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.
17. Nosek BA, Ebersole CR, DeHaven A, Mellor D (2017) The preregistration revolution.
18. Coffman LC, Niederle M (2015) Pre-analysis plans have limited upside, especially where replications are feasible. *The Journal of Economic Perspectives* 29(3):81–97.
19. Freese J, Peterson D (2017) Replication in social science. *Annual Review of Sociology* (0).
20. de Winter J, Happee R (2013) Why selective publication of statistically significant results can be effective. *PloS one* 8(6):e66463.
21. Coffman LC, Niederle M, Wilson AJ, , et al. (2017) A proposal to organize and promote replications. *American Economic Review* 107(5):41–45.
22. Mullinix KJ, Leeper TJ, Druckman JN, Freese J (2015) The generalizability of survey experiments. *Journal of Experimental Political Science* 2(2):109–138.
23. Camerer CF, et al. (2016) Evaluating replicability of laboratory experiments in economics. *Science* 351(6280):1433–1436.