# Supporting Information for:
# Publication Biases in Replication Studies

May 28, 2020

# Contents

## S.1 Details of the Model

To understand the consequences of publication bias, we consider a simple model of the publication process which consists of both the original publication stage and the replication stage. Specifically, we consider a scenario where a null hypothesis – whether true (i.e., treatment has no effect) or false (i.e., treatment has a nonzero effect) – is first tested with a sample from the population of interest ("original study") and, if published, tested again with a new sample from the same population ("replication study"), with pre-specified error rates associated to each of the two tests. At each stage, we allow for the possibility that the test result may not get published. We assume that an original study is subjected to a replication study if and only if it is published. In other words, there is no study in our model that gets published but goes unreplicated, and there is no replication study that tests an originally unpublished result.

The model consists of two sets of parameters: error rates and publication probabilities. The error rate parameters can further be categorized into two types – type-I error rates and type-II error rates – and their relevance depends on the alternative scenarios with respect to the true state of the world. Table S1 summarizes our notation for these parameters as well as other elements of the model. First, in the case where the true state is such that there is no treatment effect (i.e., the null hypothesis $H_0$ is true; see Figure 1 in the main text), the original study incorrectly classifies the null to be false (i.e., a false positive result) with probability $\alpha_1$, resulting in a type-I error. Conversely, the test correctly fails to reject the null at probability $1 - \alpha_1$. The test result is published with probability $p_1$ if it is a positive result and with $p_0$ if it is a negative result. Similarly, the replication study incorrectly rejects the null hypothesis with probability $\alpha_2$ and correctly fails to reject it at probability $1 - \alpha_2$. The replication result is again published with some probability, but the probability now varies both depending on whether the original study that it attempts to replicate is positive or negative, and on whether the replication study result itself is positive or negative ($q_{11}$, $q_{10}$, $q_{01}$, and $q_{00}$).

We also consider parallel scenarios for the state of the world where the null hypothesis is false, i.e., when there is a nonzero treatment effect (Figure S1). In this case, the test results depend on the type-II error rates ($\beta_1$ and $\beta_2$) or one minus the statistical power of the tests. That is, the original (replication) study correctly classifies the null to be false (i.e., a true positive result) with probability $1 - \beta_1$ ($1 - \beta_2$) and incorrectly fails to reject the null with probability $\beta_1$ ($\beta_2$). We assume that the publication probabilities for the original and replication studies are given by the same parameters ($p_0$, $p_1$, $q_{11}$, $q_{10}$, $q_{01}$ and $q_{00}$) regardless of whether the null hypothesis is true or false; this assumption appears plausible because the truthfulness of the null hypothesis is not directly observable by the agents who govern the publication process. Also note that we assume these parameters to be set by the mechanism exogenous to the model: for example, we do not consider the possibility that publication probabilities are determined as a result of strategic considerations by individual agents who make actual publication decisions. Note that the case of a false null hypothesis is irrelevant when we only consider the AFPR;

| Notation | Definition |
|---|---|
| $H_0$ | null hypothesis |
| $\alpha_k$ | nominal type-I error rate of the test in the original ($k = 1$) or replication ($k = 2$) study |
| $\beta_k$ | nominal type-II error rate of the test in the original ($k = 1$) or replication ($k = 2$) study |
| $p_i$ | probability that the original study is published when the test result is negative ($i = 0$) or positive ($i = 1$) |
| $q_{ij}$ | probability that the replication study is published when the original test result is negative ($i = 0$) or positive ($i = 1$) and the replication test result is negative ($j = 0$) or positive($j = 1$) |

Table S1: Notation for the Model.

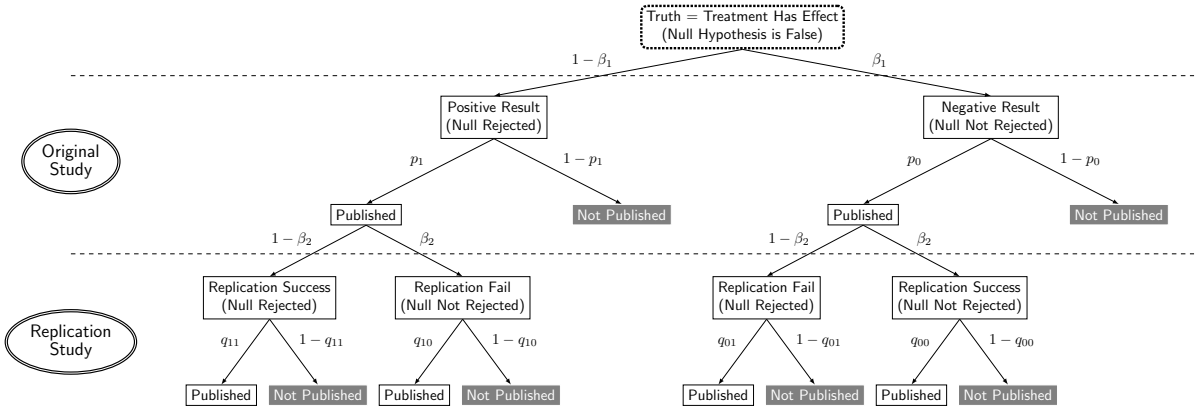


Figure S1: Model of Publication Process with Two Stages, the False Null Case.

however, it plays a crucial role when we analyze other metrics of evidence quality, such as the reproducibility rate (see Section S.5).

Our framework accommodates both two-sided and one-sided tests. This is important because, for a study that attempts to replicate a previous finding that shows an effect in one direction, it is arguably more appropriate to use a one-sided test where the null hypothesis is such that the effect is in the other direction. A replication of an originally insignificant result, on the other hand, would normally be conducted by a two-sided test. Thus, in our model, it is reasonable to interpret the null hypothesis in the replication study as one-sided for a significant original result, and as two-sided for a non-significant original result, both with the nominal type-I error probability of $\alpha_2$.

Several additional remarks are in order regarding the issue of two-sided versus one-sided tests. First, in some fields, a two-sided test may be more commonly used even in a replication study than a one-sided test. However, such a test is effectively one-sided in this context, because

only a result that is statistically significant in the same direction as the original finding would be considered a "positive" result (i.e., a successful replication). Second, although we use a single parameter $\alpha_2$ to represent the type-I error probability for the replication studies of both originally significant and non-significant studies, it is straightforward to allow this probability to vary between the two scenarios in our model. For example, one might find it more reasonable to set the test for an originally non-significant result to have twice as large the nominal type-I error rate, as it would be if the replication study had been designed in exactly the same way. Introducing additional notation to allow for such generalization, however, does not affect the substantive conclusions of our analysis. We therefore opt to adopt the simpler representation for the sake of clarity.

We now derive several metrics of evidence quality under the assumptions implied by the model. We assume all of our probability and error rate parameters to be strictly bounded between 0 and 1. First, we introduce two additional parameters – $\tilde{\alpha}_{21}$ and $\tilde{\alpha}_{20}$ – that denote the actual false positive rates (AFPRs) in replications of originally significant and non-significant studies, respectively. That is,

$$\begin{aligned}
\tilde{\alpha}_{21} &= \Pr(\text{replication test significant} \mid \text{the null is true, original test significant,} \\
&\qquad\qquad \text{replication published}), \\
\tilde{\alpha}_{20} &= \Pr(\text{replication test significant} \mid \text{the null is true, original test insignificant,} \\
&\qquad\qquad \text{replication published}).
\end{aligned}$$

We can show that these conditional AFPRs are related to the nominal FPR as well as the publication probability parameters in the following way.

**Lemma 1.**

$$\tilde{\alpha}_{2i} = \frac{\alpha_2 q_{i1}}{(1-\alpha_2)q_{i0} + \alpha_2 q_{i1}} \quad \textit{for} \quad i \in \{0, 1\}. \tag{1}$$

*Proof.* Note that $\tilde{\alpha}_{20} = \Pr(\text{original test fails to reject } H_0, \text{ replication test rejects } H_0 \text{ and gets published}|H_0) / \Pr(\text{original test fails to reject } H_0 \text{ and replication test gets published}|H_0)$. Expressing both the numerator and the denominator with respect to the model parameters yields equation (1) for $i = 0$. The expression for $\tilde{\alpha}_{21}$ can be obtained analogously. □

Lemma 1 immediately leads to the following important corollary.

**Corollary 1.** $\tilde{\alpha}_{2i} = \alpha_2$ *for any* $\alpha_2$ *if and only if* $q_{i1} = q_{i0}$.

*Proof.* The result is immediate by substituting $\alpha_2 = \tilde{\alpha}_{2i}$ in equation (1) and vice versa. □

The corollary implies that the AFPRs for replication studies will always deviate from their nominal FPR unless there is no publication bias in these studies.

Lemma 1 also implies the following proposition for the overall AFPR.

**Proposition 1.** *The overall AFPR $\tilde{\alpha}_2$, defined in the main text, is given by the following expression:*

$$\tilde{\alpha}_2 \;=\; \frac{\alpha_1 p_1 \alpha_2 q_{11} + (1-\alpha_1) p_0 \alpha_2 q_{01}}{\alpha_1 p_1 \{(1-\alpha_2) q_{10} + \alpha_2 q_{11}\} + (1-\alpha_1) p_0 \{(1-\alpha_2) q_{00} + \alpha_2 q_{01}\}}. \tag{2}$$

*Proof.* Note that

$$\begin{aligned}
\tilde{\alpha}_2 \;&=\; \tilde{\alpha}_{21} \Pr(\text{original test rejects } H_0 \mid \text{replication test published}, H_0) \\
&\quad + \tilde{\alpha}_{20} \Pr(\text{original test fails to reject } H_0 \mid \text{replication test published}, H_0) \\
&=\; \frac{\tilde{\alpha}_{21} \alpha_1 p_1 \{\alpha_2 q_{11}(1-\alpha_2) q_{10}\} + \tilde{\alpha}_{20}(1-\alpha_1) p_0 \{\alpha_2 q_{01} + (1-\alpha_2) q_{00}\}}{\alpha_1 p_1 \{\alpha_2 q_{11} + (1-\alpha_2) q_{10}\} + (1-\alpha_1) p_0 \{\alpha_2 q_{01} + (1-\alpha_2) q_{00}\}}.
\end{aligned}$$

Substituting equation (1) to $\tilde{\alpha}_{21}$ and $\tilde{\alpha}_{20}$ and simplifying yields the desired expression. □

Proposition 1 allows us to simulate the AFPR for assumed values of the publication probabilities. In the main text, we use the data from our vignette experiment to estimate the publication probabilities empirically and employ the formula to calculate the AFPRs (Figure 4).

Moreover, Proposition 1 can also be used to characterize the relationships between the AFPR and the three types of publication biases: the file drawer bias, the repeat study bias, and the gotcha bias. To analyze those relationships, we reparameterize our publication probability parameters in terms of these biases, such that

$$\begin{aligned}
p_1 &= p, \\
p_0 &= p - f, \\
q_{11} &= p - r, \\
q_{10} &= p - r - f + g, \\
q_{01} &= p - r + g, \\
q_{00} &= p - r - f,
\end{aligned}$$

where $p$, $f$, $r$, $g \in (0, 1)$. In our new parameterization, $f$ represents the file drawer bias, $r$ the repeat study bias, and $g$ the gotcha bias. We also rewrite the publication probability for an original positive result as $p$ for simplicity. The intuition behind the new parameterization, and the implied definitions for the three types of publication biases, are straightforward. For example, the file drawer bias ($f$) enters as a penalty on the publication probabilities for the negative results (i.e., $p_0$, $q_{10}$ and $q_{00}$), whereas the repeat study bias ($r$) is a common penalty for the replication tests (i.e., $q_{11}$, $q_{10}$, $q_{01}$ and $q_{00}$). The gotcha bias ($g$), on the other hand, is a bonus for replication results that overturn existing findings (i.e., $q_{01}$ and $q_{10}$).

The resulting relationship between the AFPR and the biases turn out to be rather complex. Figure S2 illustrates how the AFPR can be either increasing or decreasing as a function of the biases, using a hypothetical scenario where $p = 0.6$, $r = 0.05$, and $\alpha_1 = \alpha_2 = 0.05$. It is clearly seen from the figure that the AFPR is monotonically increasing in the file drawer bias when the gotcha bias is close to zero, but it can actually *decrease* as the file drawer bias increases if the

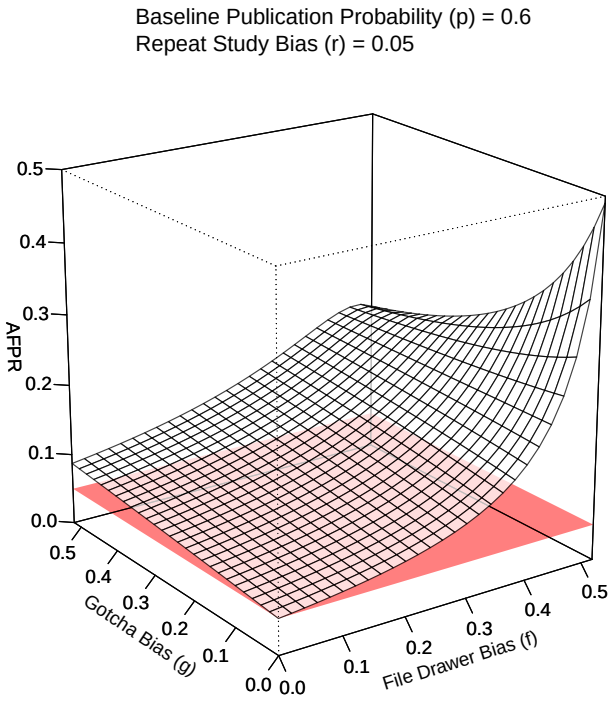


Figure S2: Illustration of the Relationship between the AFPR and the Publication Biases. The figure shows how the AFPR changes as a function of the file drawer and gotcha biases. The red flat plane represents the nominal FPR of the original and replication tests assumed in the simulation ($\alpha_1 = \alpha_2 = 0.05$).

gotcha bias is highly prevalent (e.g. $g = 0.5$). Likewise, the AFPR is monotonically increasing in the gotcha bias for small values of $f$ (i.e. low file drawer bias), but the relationship actually reverses if there is large file drawer bias (e.g. $f = 0.5$).

Still, it is possible to derive an important result: the AFPR is always greater than the nominal FPR if the file drawer bias is larger than the gotcha bias. This result can be formally stated as follows.

**Proposition 2.** *$\tilde{\alpha}_2 > \alpha_2$ if $f > g$.*

*Proof.*

$$\tilde{\alpha}_2 - \alpha_2 = \frac{\alpha_2 A}{\alpha_1 p_1\{(1-\alpha_2)q_{10} + \alpha_2 q_{11}\} + (1-\alpha_1)p_0\{(1-\alpha_2)q_{00} + \alpha_2 q_{01}\}},$$

where

$$\begin{aligned} A &= \alpha_1 p_1 q_{11} + (1-\alpha_1)p_0 q_{01} - \alpha_1 p_1\{(1-\alpha_2)q_{10} + \alpha_2 q_{11}\} - (1-\alpha_1)p_0\{(1-\alpha_2)q_{00} + \alpha_2 q_{01}\} \\ &= \alpha_1(1-\alpha_2)p_1(q_{11} - q_{10}) + (1-\alpha_1)(1-\alpha_2)p_0(q_{01} - q_{00}) \\ &= (1-\alpha_2)\{\alpha_1 p(f-g) + (1-\alpha_1)(p-f)(f+g)\}. \end{aligned}$$

Since $p - f = p_0 > 0$, $A > 0$ if $f > g$. Therefore, $\tilde{\alpha}_2 - \alpha_0 > 0$ if $f > g$. This implies Proposition 2. □

Proposition 2 implies that the classic publication bias problem – the inflation of false positive rates in the published body of evidence – still holds true even in the presence of publication biases other than the file drawer bias, as long as the latter is a dominant mode of publication bias. This also implies, however, that the AFPR could be *smaller* than the nominal FPR should the gotcha bias be more important. The bottom line is that it is crucial to empirically ascertain whether the file drawer bias or the gotcha bias is larger.

## S.2 Details of the Survey Experiment

We began with a population of 5,394 political science faculty at Ph.D. granting institutions. We obtained a list of such departments from the American Political Science Association, and then accessed each department's webpage and obtained the names and contact information for all full and part time faculty. Our data collection occurred between January 25 and February 16, 2017, by e-mail. Specifically, we e-mailed the population, of which 289 were undeliverable, asking them to participate in a study focused on factors that influence publication decisions. We sent one reminder. We focused on Ph.D. granting schools to ensure respondents were likely to be engaged in research; while this leads to the exclusion of non-Ph.D. granting departments with active researchers, it does prevent us from including a number of schools where active research is less common. Among the 5,105 individuals we successfully delivered an email, 1,236 individuals opened the survey, and 993 of these respondents completed at least one vignette. Of

| Variable | Percent |
|---|---|
| Female | 32.89 |
| Served as Editor | 26.69 |
| Current Position | |
| Full Professor | 44.78 |
| Associate Professor | 24.17 |
| Assistant Professor | 26.42 |
| Continuing Non-TT Lecturer | 2.77 |
| Adjunct Professor | 1.32 |
| Post-Doctoral Researcher | 0.40 |
| Graduate Student | 0.13 |
| Primary Field of Study | |
| Comparative Politics | 25.76 |
| American Politics | 36.79 |
| International Relations | 19.32 |
| Political Theory | 3.94 |
| Methodology | 2.37 |
| Other | 11.83 |
| **Variable** | **Mean** |
| Age | 49.20 |
| Ph.D. Students Advised | 9.44 |

Table S2: Sample Characteristics.

those who answered at least one vignette, respondents who had not previously served as editors answered an average of 8.75 vignettes ($\sigma^2 = 6.81$; median $= 10$; mode $= 10$), of the 10 they could have answered. Those who answered at least one vignette and had experience as editors answered an average of 11.26 vignettes ($\sigma^2 = 26.57$; median $= 15$; mode $= 15$) of the 15 they could have answered. Descriptive statistics on key demographic covariates of the sample are provided in Table S2.

## S.3 Effects of Study Attributes on Publication Probability

Figure S3 presents the estimated average marginal component effects (AMCEs) of the hypothetical study attributes on the respondents' chance of taking an action in favor of publication for the paper (i.e. submitting the paper as an author, recommending publication as a reviewer, or supporting publication as an editor) from our vignette experiment. (For the sake of brevity, we will refer to this outcome variable as the "chance of publication" hereafter.) Overall, our study attributes have effects on the chance of publication in expected directions. Specifically, on average, an observational study has 3 percentage points lower chance of publication than

an experimental study (s.e.= .58); a moderately and extremely exciting/important hypothesis has a 16 (s.e.= .69) and 20.7 (s.e.= .74) percentage points higher chance of publication than a not at all exciting/important hypothesis, respectively; a somewhat and extremely surprising/counterintuitive result has a 10 (s.e.= .69) and 13.4 (s.e.= .75) percentage points higher chance of publication, respectively; and studies with sample sizes of 500, 1,000, and 5,000 have higher chance of publication than a study with 50 observations by 8.7 (s.e.= .79), 14.6 (s.e.= .85), and 17.8 (s.e.= .88) percentage points, respectively.

## S.4 Additional Results on Publication Biases

Figures S4 to S8 show the additional empirical results on publication biases mentioned in the main text. Overall, there is little evidence that the magnitudes of the publication biases are moderated by either respondents' hypothetical role (Figure S4) or other study attributes included in our vignette (Figures S5 to S8).

We note, however, two possible moderational effects of other study attributes with respect to the gotcha bias that are of potential interest. First, the gotcha bias for an insignificant replication result (Figure S6, middle plot) appears to be larger for studies with larger sample sizes. While the gotcha bias is estimated to be 4.87 percentage points when the study has only 50 observations, the bias increases to 13.25 for studies with $N = 1000$ (difference significant with $p < .002$) and to as large as 15.34 when $N = 5000$ ($p < .0002$). A possible interpretation is that respondents may perceive insignificant "gotcha" replication results particularly publishable when the study is high powered and the null result is thus highly convincing as evidence of lack of an effect.

Second, the gotcha bias for a significant replication result (Figure S8, right plot) disappears for a result that is extremely surprising and counterintuitive. That is, while the estimated gotcha bias is 5.17 ($p < .002$) for a not at all surprising or counterintuitive replication result that is statistically significant, it becomes statistically indistinguishable from zero ($p = 0.706$) for an extremely surprising and counterintuitive result. This is consistent with our hypothesized mechanism behind the gotcha bias, since the "gotcha" factor may not make an already highly surprising result more publishable.

In sum, our analysis of possible moderation effects suggests only minor levels of moderation by other study attributes included in the vignette, and the small number of significant relationships give support to our hypothesized mechanism causing the publication biases.

## S.5 Analysis of the Reproducibility Rate

In this appendix, we turn to an alternative metric of evidence quality, *reproducibility rate*, defined as follows.

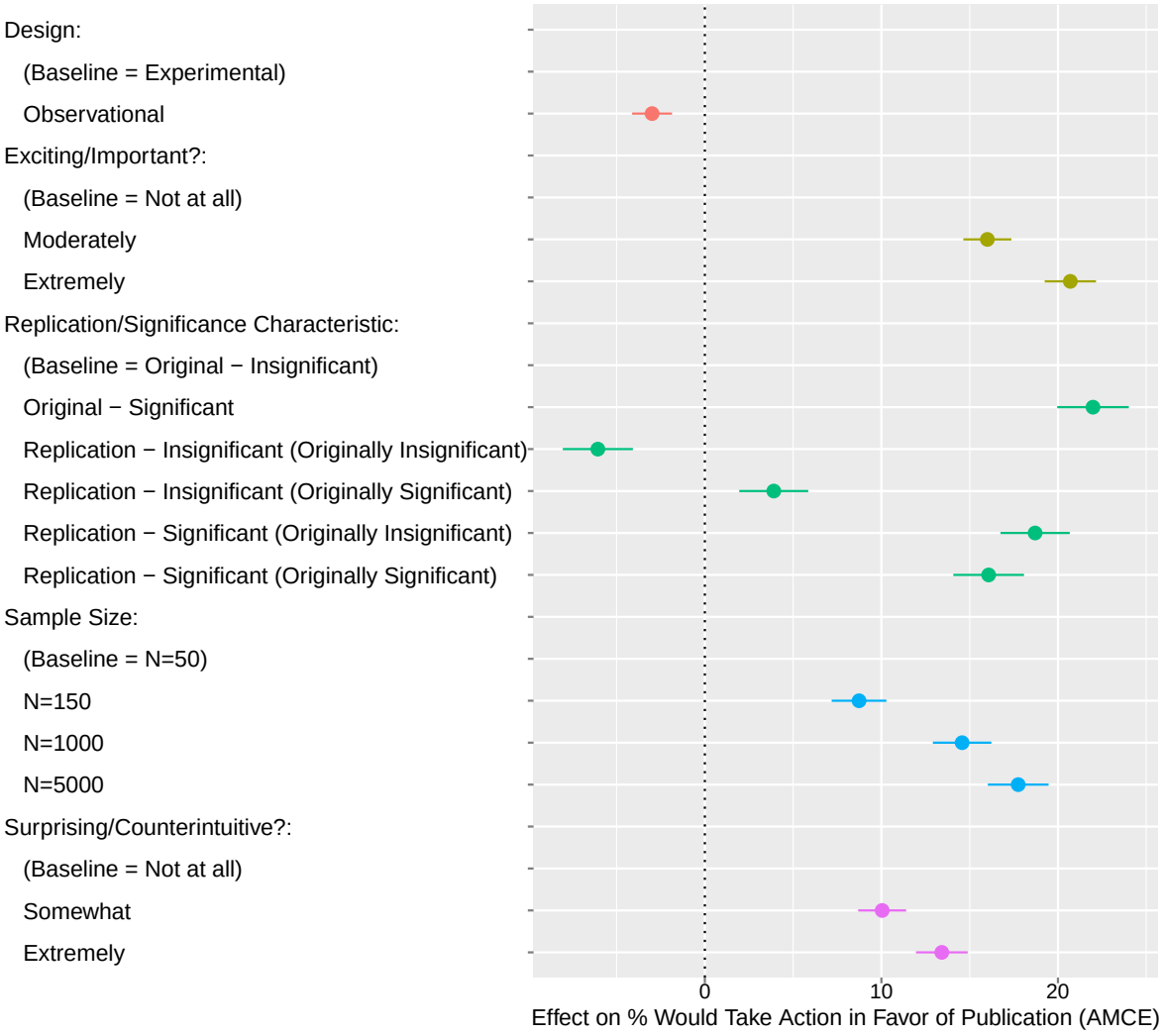


Figure S3: Estimated Average Marginal Component Effects (AMCEs) of the Study Attributes on the Chance of Taking an Action in Favor of Publication for the Paper. Horizontal bars represent 95% confidence intervals using cluster-robust standard errors at the respondent level.

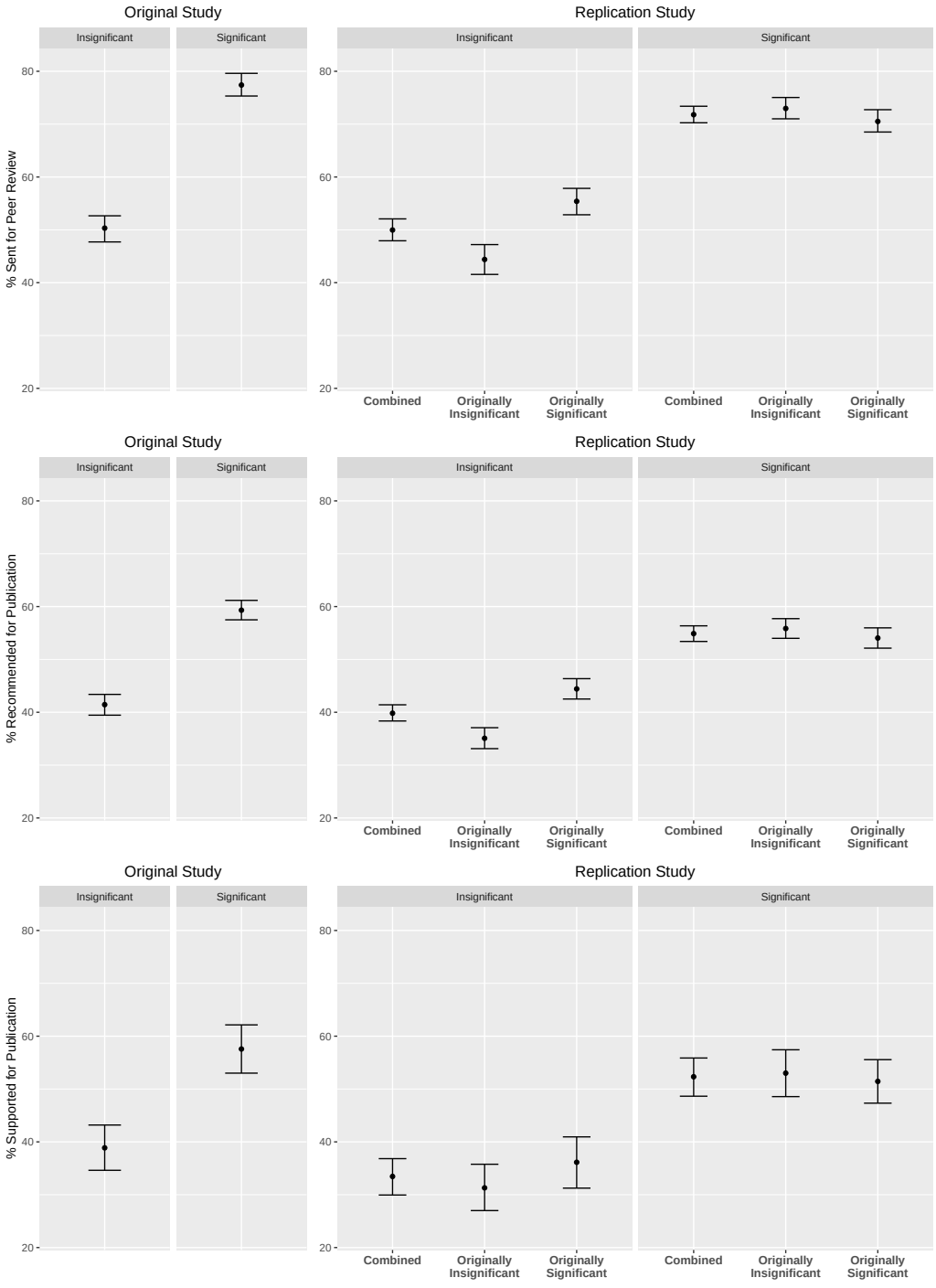


Figure S4: Estimates of Three Types of Publication Biases by Role. The estimates are for the author (top), reviewer (middle) and editor (bottom) conditions.

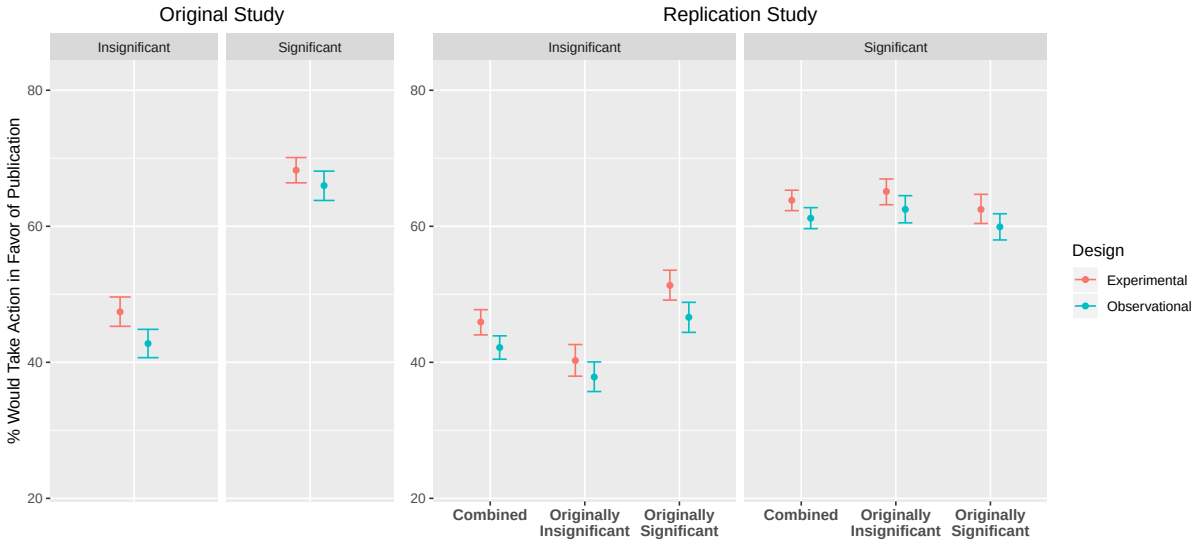


Figure S5: Estimates of Three Types of Publication Biases by Study Design.

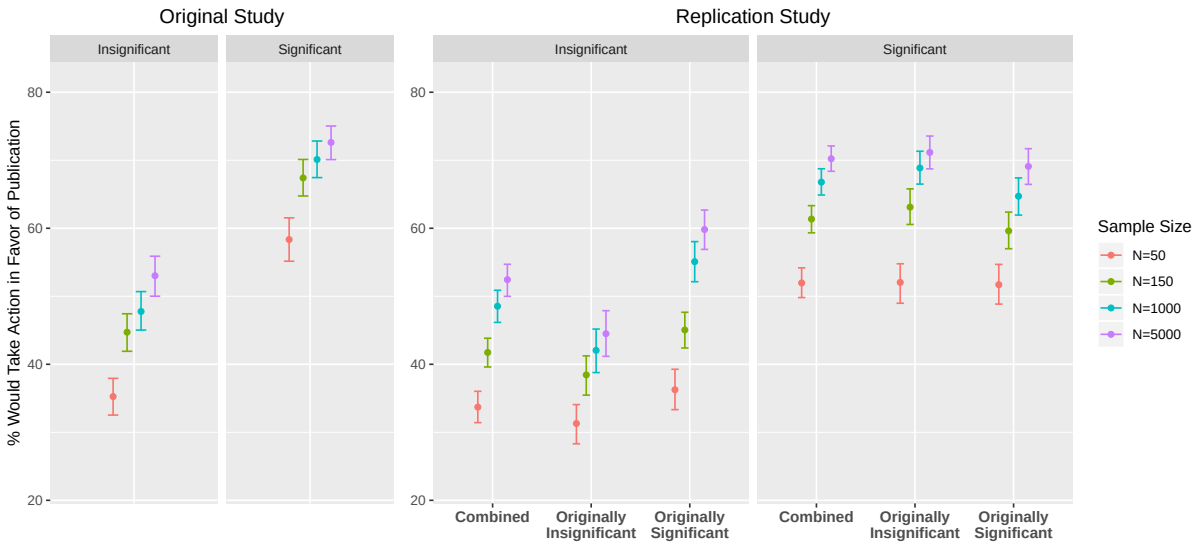


Figure S6: Estimates of Three Types of Publication Biases by Sample Size.

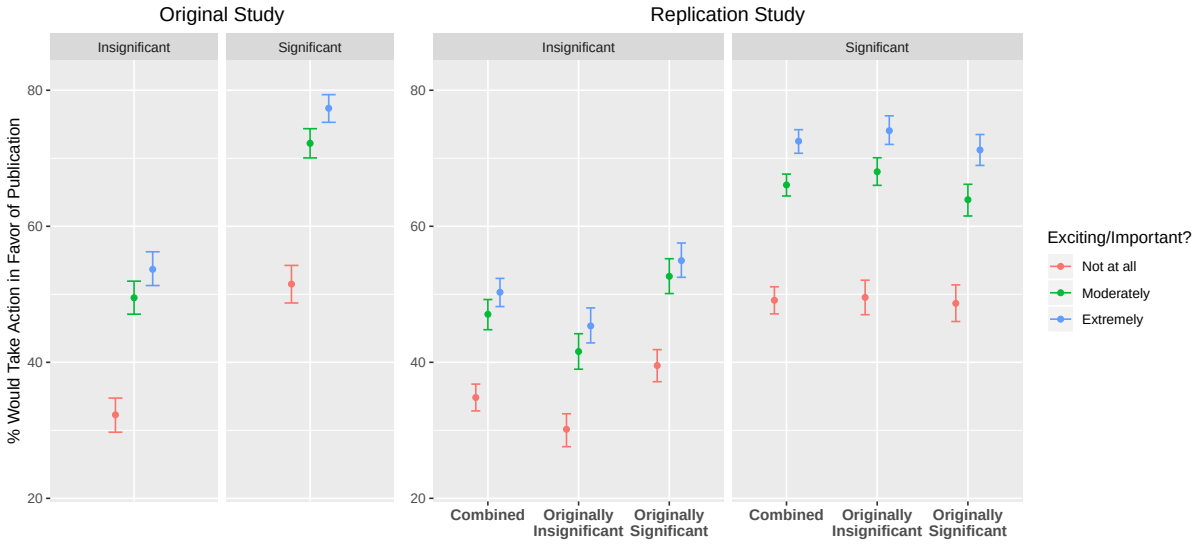


Figure S7: Estimates of Three Types of Publication Biases by Degree of Excitement.

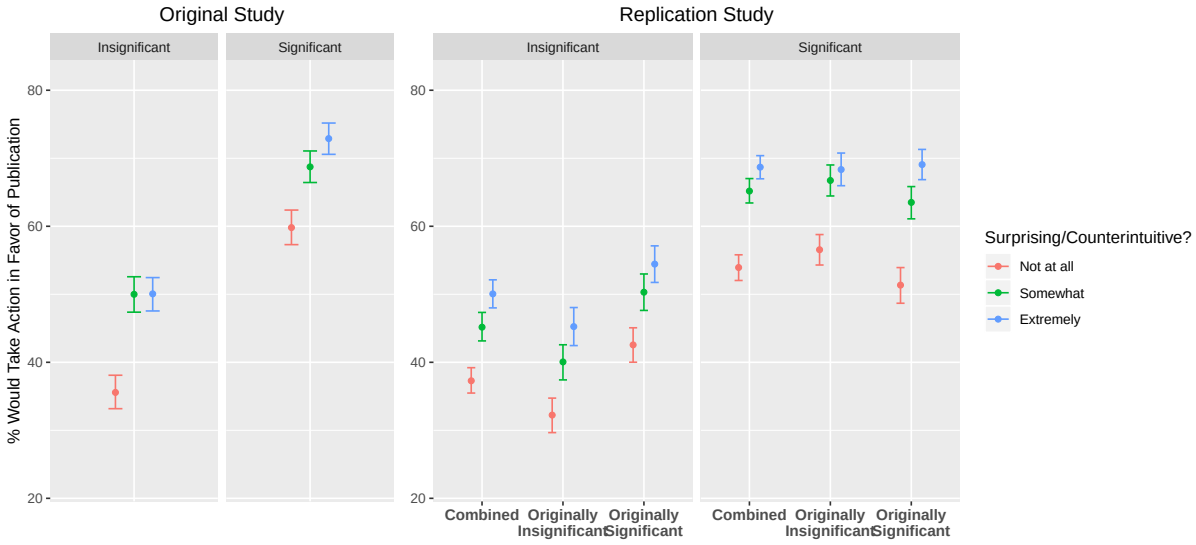


Figure S8: Estimates of Three Types of Publication Biases by Degree of Surprise.

**Definition 1.** *[Reproducibility Rate]*

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

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 Open Science Collaboration study (Open Science Collaboration, 2015). 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.

However, our model implies that the reproducibility rate should have no direct relationship with the file drawer bias in the *original* studies. The following proposition provides an exact formula for $R$ in terms of the model parameters to illuminate this point.

**Proposition 3.**

$$R \;=\; \frac{(1-\pi)(1-\beta_1)(1-\beta_2)q_{11} + \pi\alpha_1\alpha_2 q_{11}}{(1-\pi)(1-\beta_1)\{\beta_2 q_{10} + (1-\beta_2)q_{11}\} + \pi\alpha_1\{(1-\alpha_2)q_{10} + \alpha_2 q_{11}\}}, \tag{3}$$

*where $\pi = \Pr(H_0)$, the prior probability of the null hypothesis being true.*

*Proof.* First, let $R_0$ and $R_1$ represent reproducibility for the true and false null hypotheses, respectively, such that

$$\begin{aligned} R_0 \;&=\; \Pr(\text{test rejects } H_0 \mid H_0, \text{original test rejects } H_0 \text{ and published, replication published}), \\ R_1 \;&=\; \Pr(\text{test rejects } H_0 \mid H_1, \text{original test rejects } H_0 \text{ and published, replication published}), \end{aligned}$$

where $H_1$ denotes the event that $H_0$ is false. Note that $R_0 = \tilde{\alpha}_{21}$, the AFPR for the replication of an originally significant result. Now, consider $R_1$. We have

$$\begin{aligned} R_1 \;&=\; \frac{\Pr(\text{original test rejects } H_0 \text{ and published, replication test rejects } H_0 \text{ and published} \mid H_1)}{\Pr(\text{original test rejects } H_0 \text{ and published, replication test published} \mid H_1)} \\ &=\; \frac{\Pr(\text{original test rejects } H_0 \text{ and published, replication test rejects } H_0 \text{ and published} \mid H_1)}{\left\{ \begin{array}{l} \Pr\left( \begin{array}{l} \text{original test rejects } H_0 \text{ and published,} \\ \text{replication test rejects } H_0 \text{ and published} \end{array} \;\middle|\; H_1 \right) \\ +\Pr\left( \begin{array}{l} \text{original test rejects } H_0 \text{ and published,} \\ \text{replication test fails to reject } H_0 \text{ and published} \end{array} \;\middle|\; H_1 \right) \end{array} \right\}} \\ &=\; \frac{(1-\beta_1)p_1(1-\beta_2)q_{11}}{(1-\beta_1)p_1(1-\beta_2)q_{11} + (1-\beta_1)p_1\beta_2 q_{10}} \;=\; \frac{(1-\beta_2)q_{11}}{(1-\beta_2)q_{11} + \beta_2 q_{10}}. \end{aligned} \tag{4}$$

Therefore,

$$
\begin{aligned}
R &= \sum_{i\in\{0,1\}} R_i \Pr(H_i \mid \text{original test rejects } H_0 \text{ and published, replication test published}) \\
&= \frac{\sum_{i\in\{0,1\}} R_i \Pr(\text{original test rejects } H_0 \text{ and published, replication test published} \mid H_i)\Pr(H_i)}{\sum_{i\in\{0,1\}} \Pr(\text{original test rejects } H_0 \text{ and published, replication test published} \mid H_i)\Pr(H_i)} \\
&= \frac{\alpha_1\alpha_2 p_1 q_{11}\pi + (1-\beta_1)p_1(1-\beta_2)q_{11}(1-\pi)}{\{\alpha_1 p_1 \alpha_2 q_{11} + \alpha_1 p_1(1-\alpha_2)q_{10}\}\pi + \{(1-\beta_1)p_1(1-\beta_2)q_{11} + (1-\beta_1)p_1\beta_2 q_{10}\}(1-\pi)} \\
&= \frac{\alpha_1\alpha_2 p_1 q_{11}\pi + (1-\beta_1)p_1(1-\beta_2)q_{11}(1-\pi)}{\alpha_1\{\alpha_2 q_{11} + (1-\alpha_2)q_{10}\}\pi + (1-\beta_1)\{(1-\beta_2)q_{11} + \beta_2 q_{10}\}(1-\pi)}.
\end{aligned}
$$

□

As is clear from Proposition 3, the reproducibility rate has no direct relation with the file drawer bias in original studies, as the formula for $R$ does not contain neither $p_1$ or $p_0$. This result casts serious doubt that low reproducibility stems, at all, from the file drawer problem. 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), a monotonic transformation of the $\pi$ parameter (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.

Proposition 3 also allows us to interpret findings from large-scale controlled replication studies such as Open Science Collaboration's well-known study in more precise terms. Note that such a replication study is not subject to either file drawer bias or gotcha bias, because it by definition reports all the replication test results. Further, suppose that the replications are conducted in an idealized setting, such that $\beta_2 \simeq 0$ and $\alpha_2 \simeq 0$ (e.g., think of an infinitely large sample). Under this scenario, $R \simeq (1-\pi)(1-\beta_1)/\{(1-\pi)(1-\beta_1) + \pi\alpha_1\}$. This implies that reproducibility in this type of replication study will be high if (1) there are many true relationships to be discovered (i.e., $\pi$ is small), (2) original studies are conducted with large

power (i.e., $\beta_1$ is small) or (3) original positive results passed stringent statistical tests (i.e., $\alpha_1$ is small). These agree with many of the factors that Open Science Collaboration pointed out as possible causes of the low reproducibility in their study, but it is again noteworthy that these do not include publication bias in the original studies, contrary to what Open Science Collaboration suspected.

We now turn to our vignette survey experiment to investigate the reproducibility rate empirically. 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 each of the four panels in Figure S9, assuming different levels of significance tests at each stage. For each combination of the significance levels, 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.

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 S9) 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

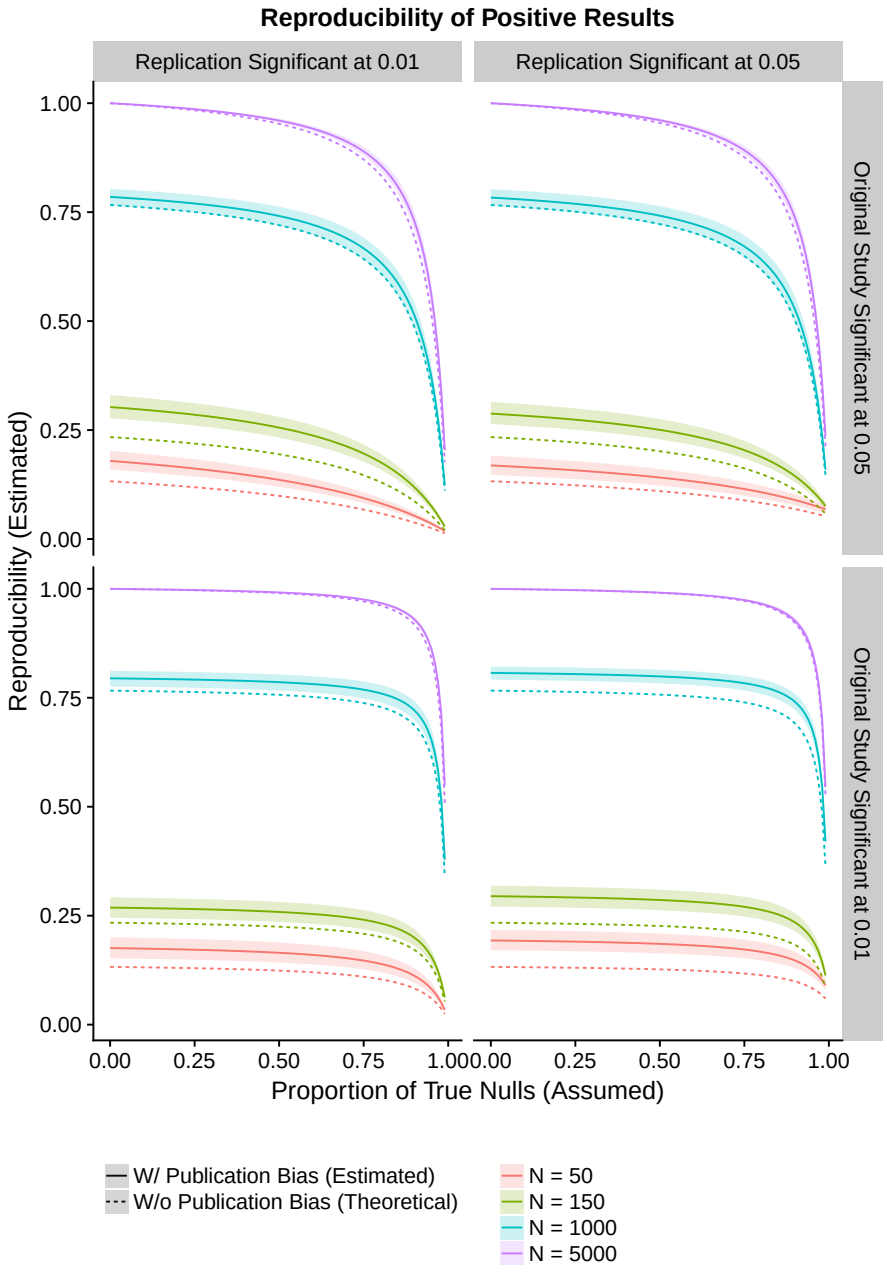


Figure S9: Estimates of Reproducibility as Function of Power and "Pre-Study Odds." In each panel, dashed lines represent the simulated theoretical reproducibility rate for a given combination of the assumed proportion of the true null hypotheses (horizontal axis) and the power value implied by a sample size (four different colors, as indicated in the plot legend) in the absence of publication bias. 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 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 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.

# References

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

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