INSTITUTE for REPLICATION

No. 3
I4R DISCUSSION PAPER SERIES

Can the Replication Rate Tell Us About Selective Publication?

Patrick Vu

October 2022

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 3

Can the Replication Rate Tell Us About Selective Publication?

Patick Vu[1]

[1] *Brown University, Providence, RI/USA*

OCTOBER 2022

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Can the Replication Rate Tell Us About Selective Publication?

By Patrick Vu*

Selective publication is among the most-cited reasons for widespread replication failures. I show in a simple model of the publication process that the replication rate is completely unresponsive to the suppression of insignificant results. I then show that the expected replication rate falls below its intended target owing to issues with common power calculations in replication studies, even in the absence of other factors such as p-hacking or heterogeneous treatment effects. I estimate an empirical model to evaluate if issues with power calculations alone are sufficient to explain the low replication rates observed in large-scale replication studies. The model produces replication rate predictions (using only data from original studies) that are almost identical to observed replication rates in experimental economics and social science. In psychology, the model explains two-thirds of the gap between the replication rate and its intended target. I conclude by discussing alternative measures of replication that are more responsive to selective publication.

In a 2016 survey conducted by *Nature*, 90% of researchers across various fields agreed that the scientific community faces a 'reproducibility crisis' (Baker, 2016). Growing consensus has been supported by evidence of widespread replication failures. In a replication of 18 experimental economics studies, 61% of significant results were replicated with the same sign and significance (Camerer et al., 2016). In psychology, only 36% of significant results were successfully replicated (Open Science Collaboration, 2015).

Explaining the source of low replication rates has become a topic of intense interest. The most frequently cited reason is selective reporting, with over 90% of researchers identifying it as contributing factor to irreproducible research (Baker, 2016). Its most salient form is censoring statistically insignificant results, either by journals in the editorial process or by researchers who do not write up null findings in anticipation of low chances of publication (Open Science Collaboration, 2015; Maxwell et al., 2015; Camerer et al., 2016; Anderson and Maxwell, 2017;

* *This version:* October 14, 2022. Brown University. patrick_vu@brown.edu. I am especially grateful for the feedback, advice, and encouragement of Jonathan Roth. For helpful comments, suggestions and conversations, I thank Johannes Abeler, Daniel Björkegren, Pedro Dal Bó, Anna Dreber, Peter Hull, Toru Kitagawa, Soonwoo Kwon, and Jesse Shapiro, as well as seminar participants at Brown University.

Camerer et al., 2018; Stanley et al., 2018; Andrews and Kasy, 2019). Contrary to common perceptions, my first main theoretical result shows, using a simple model of selective publication, that the replication rate does not depend on the extent to which null results are or are not published. While this result is somewhat counter-intuitive ex-ante, its explanation is simple. The replication rate is typically defined as the share of *significant* results that are replicated with significance and the same sign. Since the replication rate definition does not depend on insignificant results, it is unaffected by their prevalence in the published literature. It is important to note that this result does not depend on whether or not insignificant findings are chosen for replication; even when they are, the replication rate calculation does not include them (e.g. Open Science Collaboration (2015)). Moreover, large-scale replication studies implementing high-powered designs to detect some fraction of the original effect size are subject to the same problem (e.g. Camerer et al. (2018)). The replication rate is thus ill-suited to uncovering the most salient form of selective publication.

If selective publication is unlikely to explain low replication rates, then what does? The literature offers numerous explanations. Prominent theories include researcher p-hacking in response to selective publication, for example by specification searching to obtain statistically significant findings (Ioannidis, 2005, 2008; Simonsohn et al., 2014; Brodeur et al., 2016, 2020, 2022); heterogeneity across original studies and replications in research design and experimental subjects (Higgins and Thompson, 2002; Cesario, 2014; Simons, 2014; Stanley et al., 2018; Bryan et al., 2019); and measurement error in small samples (Gelman and Carlin, 2014; Loken and Gelman, 2017; Gelman, 2018). My second main theoretical result shows that the replication rate would be expected to fall short of its intended target even in 'ideal' conditions where none of these issues are present. This is because of three issues with the common approach of setting replication power to detect original effect sizes with a pre-specified intended power target (typically around 90%). First, original estimates included in the replication rate are not a random sample of published findings, but instead a selected sample of significant findings. It is well known that samples selected on extreme characteristics (e.g. height, test scores, statistical significance) will regress to the mean in repeated samples (Galton, 1886; Hotelling, 1933; Barnett et al., 2004; Kahneman, 2011). In replication settings, this means that significant original estimates used to calculate the replication rate are mechanically inflated in expectation, and that replication estimates will regress to the mean. Power calculations in replications calibrated to detect inflated original estimates may therefore be underpowered for recovering smaller true effects, leading to low replication rates. Again, whether or not insignificant findings are chosen for replication has no bearing on this conclusion. A second issue is that common power calculations lead to very low replication probabilities when original estimates are the opposite sign of the true effect, which occurs with positive probability due to random sampling

variation. This is because a 'successful' replication in this case relies on the highly unlikely outcome that a replication reproduces a statistically significant result with the 'wrong' sign. Finally, and most importantly, common power calculations do not account for the non-linearity of the power function. I show that this non-linearity implies that the expected replication rate falls below its intended target even in the optimistic scenario where original estimates are completely unbiased for true effects, all findings are published irrespective of significance, replications are a random sample of the published literature, and original studies are highly powered. Overall, these three issues suggest that intended replication rate targets in large-scale replication studies do not provide a meaningful benchmark against which to judge replication rates observed in practice; low replication rates are what we should expect.

To what extent can issues with power calculations alone explain the low replication rates actually observed in large-scale replication studies? To answer this question, I estimate a model of selective publication based on Andrews and Kasy (2019) to produce replication rate predictions for large-scale replication studies in experimental economics (Camerer et al., 2016), psychology (Open Science Collaboration, 2015) and experimental social science (Camerer et al., 2018). Estimation only uses data from original studies, and predictions are based on the power calculations that were actually implemented in replications. The model does not incorporate researcher manipulation, heterogeneous treatment effects, or measurement error. The empirical exercise asks, in effect, whether observed replication rates could have been predicted by issues with common power calculations *alone*, before the replication studies themselves were actually undertaken.

The predicted replication rate is almost identical to observed replication rates in experimental economics (60% vs. 61%) and experimental social science (55% vs. 57%).[1] This suggests that low power in original studies in conjunction with replication power issues is sufficient to explain observed replication rates in these fields. This is consistent with evidence of a low propensity of p-hacking in experimental settings, perhaps because of fewer researcher degrees of freedom compared to observational studies (Brodeur et al., 2016, 2020; Imai et al., 2020). Additionally, these results provide further evidence in support of recommendations to focus greater attention on statistical power for improving the credibility of published research (Ioannidis, 2005; Gelman and Carlin, 2014; Anderson and Maxwell, 2017; Camerer et al., 2019; DellaVigna et al., 2019). In psychology, the model predicts a replication rate of 55%. This is well below mean intended power of 92%, but still above the observed replication rate of 35%. Here, the model can account for around two-thirds of the observed replication rate gap. This discrepancy

[1] In social science experiments, concerns over low power in previous replication studies motivated a higher-powered design consisting of two stages (Camerer et al., 2018). I predict replication outcomes in the first stage, where replication power was calibrated to detect three-quarters of the original effect size with 90% power.

suggests that factors not included in the model – for example, heterogeneous treatment effects, p-hacking, differences across subfields – may be important in psychology.

To be clear, the results in this article do not suggest that there is no problem of selective publication, only that the replication rate is not a meaningful measure for detecting it. The prevalence of the so-called 'file-drawer' problem and its distortions are well documented (Ioannidis, 2008; Franco et al., 2014; Gelman and Carlin, 2014; Landis et al., 2014; Mervis, 2014; Gelman, 2018; Andrews and Kasy, 2019; Abadie, 2020). Responses to mitigate these distortions include results-blind peer review (Chambers, 2013; Foster et al., 2019), journals dedicated to publishing insignificant findings[2], and even cash incentives for publishing null findings (Nature 2020). The results in this article suggest that the replication rate is a poor metric to gauge whether such reforms are successful in reducing selective publication.

I conclude by discussing alternative replication measures that are more responsive to the suppression of significant results. I conduct policy simulations using the estimated model to evaluate how various measures change as selective publication varies in intensity. In line with the first main theoretical result, the replication rate is completely unresponsive to changes in the probability of publishing insignificant results. The same conclusion holds for common alternative measures of replication if only significant results are chosen for replication. Instead, I examine three alternative measures calculated over significant *and* insignificant results: whether a replication's 95% confidence interval covers the original result; replication based on meta-analysis; and the prediction interval approach (Patil et al., 2016). For evaluating efforts to reduce selective publication, the simulation results show that the prediction interval approach may provide a useful alternative to the replication rate, the confidence interval measure, and the meta-analysis approach. The prediction interval measure performs well because it explicitly incorporates variation in both original and replication estimates. In particular, it accounts for the fact that low-powered original studies are relatively uninformative about true effects, and hence a large range of replication estimates are statistically consistent with them.

Related Literature.—This article contributes to the large literature on metascience and publication bias (Card and Krueger, 1995; Ioannidis, 2005; Rothstein et al., 2006; Gorroochurn et al., 2007; Ioannidis, 2008; Button et al., 2013; Franco et al., 2014; Gelman and Carlin, 2014; Landis et al., 2014; Mervis, 2014; Maxwell et al., 2015; Anderson and Maxwell, 2017; Ioannidis et al., 2017; Stanley et al., 2018; Gelman, 2018; Klein et al., 2018; Miguel and Christensen, 2018; Shrout and Rodgers, 2018; Amrhein et al., 2019a,b; Tackett et al., 2019; Andrews and Kasy, 2019; Christensen et al., 2019; Frankel and Kasy, 2022; DellaVigna and Linos, 2022; Nosek et al., 2022). It is not the first to question the replication rate. Amrhein et al. (2019b) criticize the

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

replication rate because it emphasizes statistical significance over scientific significance. This can lead to incongruous conclusions. For example, two studies with identical point estimates, but where one is statistically significant and the other is not due to sample size differences, will be counted as 'inconsistent' under the current definition of the replication rate. Separately, Andrews and Kasy (2019) and Kasy (2021) provide stylized examples showing that the replication rate can vary widely depending on the latent distribution of studies (i.e. the joint distribution of true effects and standard errors for published and unpublished studies). This article contributes to these criticisms. First, it shows that the replication rate is completely insensitive to the degree of selective publication on insignificant results for a fixed latent distribution of studies. Second, it establishes formally that the expected replication rate is bounded above by its nominal target owing to issues with common power calculations in replication studies. This result does not rely on any distributional assumptions about latent studies and holds even for highly-powered original studies (although the *size* of the gap is sensitive to power in original studies). It then shows empirically that the interaction of low power in original studies and issues with replication power alone can adequately explain observed replication rates in experimental economics and social science. Empirically, this article builds on Anderson and Maxwell (2017), which calculates replication rates using fully simulated data. This article predicts the replication rate using a model empirically calibrated on data from real-world replication studies. This allows for a comparison between model-based predictions and observed replication rates.

This article also contributes to the growing literature on predicting research outcomes (Dreber et al., 2015; Camerer et al., 2016; Altmejd et al., 2019; DellaVigna et al., 2019; Camerer et al., 2018; DellaVigna et al., 2020; Gordon et al., 2020). In the replication literature, the main focus is on predicting the outcomes of individual replications, as well as the aggregate measures like the overall replication rate, using alternative methods such as surveys, prediction markets, and machine learning. Altmejd et al. (2019) use 'black-box' machine learning methods to predict individual replication outcomes and find that the most important features predicting replication are measures of statistical power. This accords well with the results in this article. The structural model developed here provides a theoretical underpinning for this atheoretical machine learning result, namely, that low-powered original studies and the non-linearity of the power function lead to low replication probabilities.

I. Simple Example

A simple stylized example illustrates the key ideas. Consider research on the impact of a new drug on health outcomes. Assume the (unobserved) true treatment effect is $\theta = 2.5$.

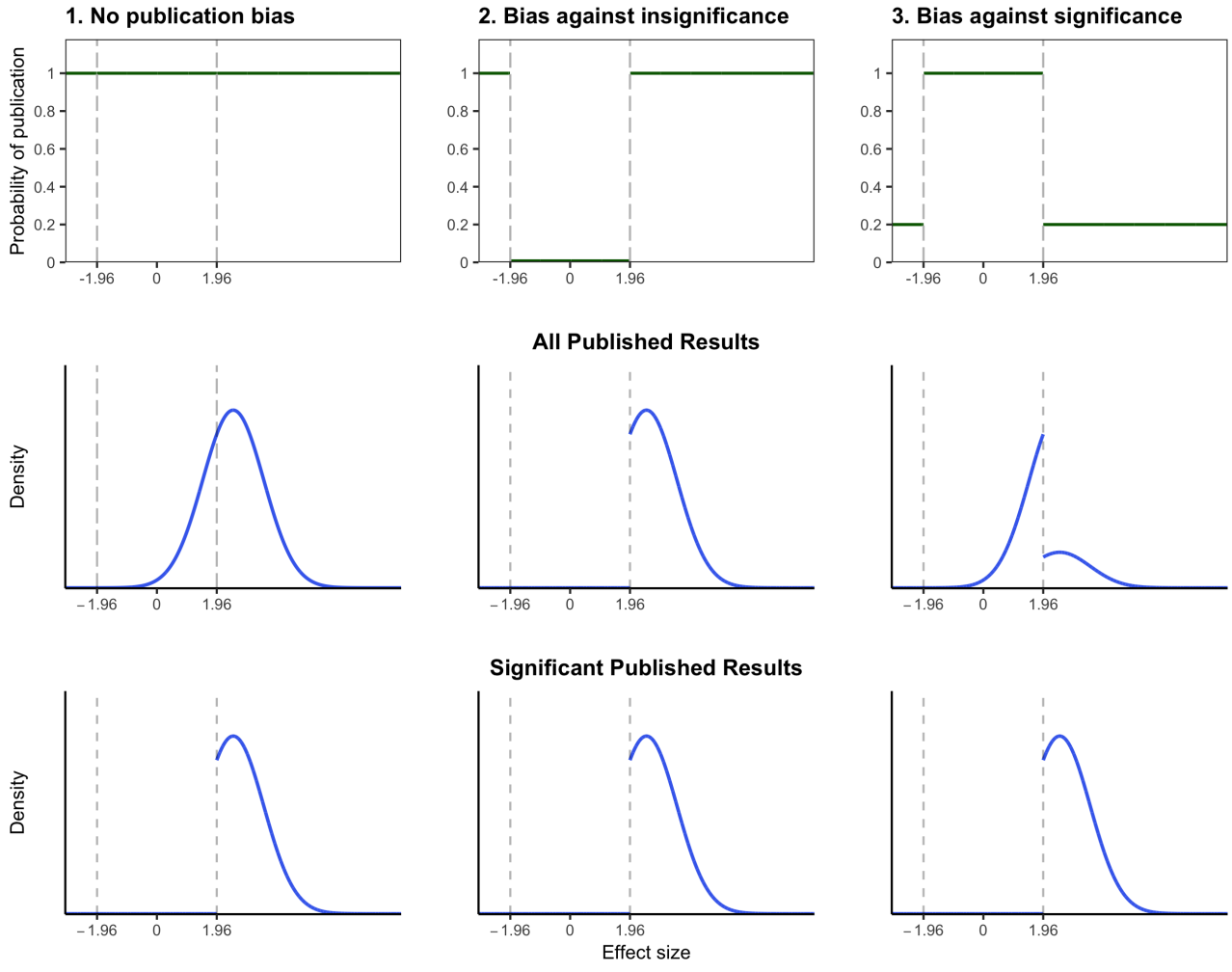
Researchers conduct a large number of independent studies to learn about θ, each producing an estimated effect size X^* drawn from a $N(2.5, 1)$ distribution. However, only a subset may be published because publication depends on a study's finding. Denote published studies as X, which come from the distribution of X^* conditional on publication. Now suppose a large-scale replication is conducted on published studies. Replication estimates X_r are drawn from a $N(2.5, \sigma_r(X)^2)$ distribution, where replication standard errors $\sigma_r(X)$ are calculated to detect the original estimates X with 90% power. This method of calculating power is perhaps the most common approach in replications (e.g. Open Science Collaboration (2015); Camerer et al. (2016)). The question we are interested in answering is: What is the replication rate under different standards for publishing statistically significant and insignificant results?

I consider three vastly different publication regimes and show that all produce exactly the same replication rate. First, the no selective publication regime, where all results are published irrespective of their statistical significance. Second, a regime that publishes all significant results and censors all insignificant results. Third, a regime where insignificant results are five times *more* likely to be published than significant results. The first row of Figure 1 shows the relative publication probabilities for each of these regimes at different t-ratios. The second row shows the implied distribution of published estimates X. As expected, the distribution of published estimates is very different across the three regimes. However, the key observation is that conditional on a published result being significant, the distributions are identical (row 3 in Figure 1). This implies that the replication rate – defined as the share of *significant* findings with the same sign and significance in replications – must be the same under all publication regimes.

There are three main takeaways. Each is highlighted by the descriptive statistics at the bottom of Figure 1. First, the replication rate does not depend on the probability of publishing insignificant results. Selective publication against insignificant findings is therefore unlikely to explain the low replication rates observed in practice. A caveat is that selective publication favoring significant results may incentivize researchers to manipulate results to obtain significance. This analysis shows that the replication rate will fall short of intended power even in the absence of such manipulation. The empirical results show that a model without manipulation produces accurate predictions of the replication rate in experimental economics and experimental social science.

The second takeaway is that conditioning published studies on significance induces upward bias in original estimates. The replication rate definition imposes this conditioning, such that the statistic itself induces inflationary bias. Unbiased replication estimates regress to the mean. This is a consequence of the replication rate definition and is the same across all regimes.

Third, the replication rate is below its 90% intended target in all three regimes, even in this



		Regime 1	Regime 2	Regime 3
All published results				
Expectation of original estimates	$\mathbb{E}(X)$	2.50	2.99	1.87
Bias of original estimates	$\mathbb{E}(X) - \theta$	0.00	0.49	−0.63
Significant published results				
Expectation of original estimates	$\mathbb{E}(X \mid X \geqslant 1.96)$	2.99	2.99	2.99
Expectation of replication estimates	$\mathbb{E}(X_r \mid X \geqslant 1.96) = \theta$	2.50	2.50	2.50
Replication probability (90% target)	$\mathbb{P}(\lvert X_r \rvert \geqslant 1.96\sigma_r, \text{sgn}(X_r) = \text{sgn}(X) \mid \theta = 2.5, \sigma_r)$	0.77	0.77	0.77

FIGURE 1. PUBLICATION REGIMES, DISTRIBUTIONS OF PUBLISHED RESULTS, AND THE REPLICATION RATE

Notes: Published estimates X are assumed to be drawn from a normal distribution centered at $\theta = 2.5$ with standard error $\sigma = 1$, which may be reweighted based on the conditional publication function shown in the first row. In Regime 1, all results are published. In Regime 2, only statistically significant results are published. In Regime 3, insignificant results are five times more likely to be published than significant results. Replication estimates X_r are drawn from a $N(2.5, \sigma_r(X)^2)$ distribution, where $\sigma_r(X) = 3.242/|X|$ is set to detect the original effect size X with 90% power using the common power rule. Statistics are based on 10^6 simulation draws.

simple example with no researcher manipulation or heterogeneity in true effects. I will show in Section II below that this is the result of issues with the common approach to calculating power in replication studies.

An additional observation concerns the size of the gap between the replication rate and its intended target. In this example, the true effect is $\theta = 2.5$ for all studies, which implies that power in original studies to detect a positive significant effect is $[1-\Phi(1.96-2.5)]\times 100 = 71\%$. This corresponds to a replication rate of 77%, which falls short of the 90% target. An important question is whether this gap persists for different distributions of true effects, or, equivalently, different distributions of power in original studies.[3] Figure 2 shows the relationship between power in original studies and the expected replication rate. For any level of original power, the expected replication rate is below its intended target of 90%. Importantly, the size of the gap is very sensitive to power in original studies. If the null hypothesis is true ($\theta = 0$), then original power is equal to 2.5% and the probability of 'successful' replication is also equal to 2.5%. On the other hand, the probability of replication approaches its intended target of 90% as original power approaches 100% (or equivalently, as $\theta/\sigma \to \infty$).

It is noteworthy that estimates of power in the literature are substantially lower than the 71% assumed in this simple example. For instance, median power is estimated to be 18% or less in empirical economics (Ioannidis et al., 2017); 18% in neuroscience (Button et al., 2013); 10% or less in political science (Arel-Bundock et al., 2022); and 36% in psychology (Stanley et al., 2018). This suggests that very low replication rates should be expected in practice when using the common power rule to set replication power.

II. General Case

Conclusions in the simple example hold more generally. This section formalizes these ideas in a general setting, building on the model of selective publication in Andrews and Kasy (2019).

A. Model of Large-Scale Replication Studies

Suppose a large-scale replication study is conducted and we observe the estimated effect sizes and standard errors for original studies and their replications. The data-generating process of these studies is modelled as a truncated sampling process. The model is presented here in general form, while the empirical applications make distributional and functional form assumptions. Upper case letters denote random variables, lower case letters realizations. Latent studies

[3] Statistical power to detect an effect with the 'correct' sign, $1-\Phi(1.96-\theta/\sigma) \in (0.025, 1)$, is a strictly increasing function of the ratio of the true effect and the standard error over the positive real line. Moreover, the expected replication rate depends only on this ratio and the rule for setting replication power.

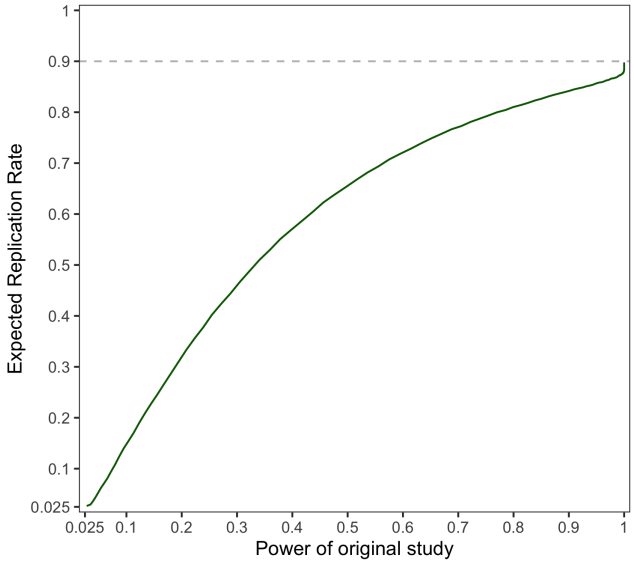


FIGURE 2. ORIGINAL POWER AND THE EXPECTED REPLICATION RATE UNDER THE COMMON POWER RULE

Notes: Original power and the expected replication rate under the common power rule are both functions of $\omega = \theta/\sigma$ (normalized to be positive). Original power to obtain a significant effect with the same sign as the true effect is equal to $1 - \Phi(1.96 - \omega)$. The expected replication rate is calculated by taking 10^6 draws of Z from $N(\omega, 1)$ and then calculating $10^{-6} \sum_{i=1}^{10^6} \left[1 - \Phi\big(1.96 - \text{sign}(z_i)\frac{\omega}{\sigma_r(z_i,\beta^n)}\big)\right]$, with intended power equal to $1 - \beta^n = 0.9$ and depicted by the horizontal dashed line. The replication standard error is calculated using the common power rule to detect original effect sizes with 90% power, which is given by $\sigma_r(z_i, \beta^n) = |z_i|/3.242$. Further details on these formulas are provided in Section II.

(published or unpublished) have a superscript * and published studies have no superscript. The model has five stages:

1. **Draw a population parameter and standard error:** Draw a research question with population parameter (Θ^*) and standard error (Σ^*):

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

   where $\mu_{\Theta,\Sigma}$ is the joint distribution of these random variables.

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

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

3. **Publication selection:** Selective publication is modelled by the function $p()$, which returns the probability of publication for any given t-ratio. Let D be a Bernoulli random

variable equal to 1 if the study is published and 0 otherwise, where

$$\mathbb{P}(D = 1|X^*/\Sigma^*) = \begin{cases} p_{sig}(X^*/\Sigma^*) & \text{if } S_X^* = 1 \\ p_{insig}(X^*/\Sigma^*) & \text{if } S_X^* = 0 \end{cases}$$

where S_X^* is an indicator variable that equals one if $\left|X^*/\Sigma^*\right| \geqslant 1.96$ and zero otherwise.[4]

4. **Replication selection:** Replications are sampled from published studies (X, Σ, Θ); that is, the distribution of latent studies $(X^*, \Sigma^*, \Theta^*)$ conditional on publication $(D = 1)$. Replication selection is modelled by the function $r()$, which returns the probability of being chosen for replication for any given t-ratio. Let R be a Bernoulli random variable equal to 1 if the study is chosen for replication and 0 otherwise, where

$$\mathbb{P}(R = 1|X/\Sigma) = \begin{cases} r_{sig}(X/\Sigma) & \text{if } S_X = 1 \\ r_{insig}(X/\Sigma) & \text{if } S_X = 0 \end{cases}$$

where S_X is an indicator variable that equals one if $\left|X/\Sigma\right| \geqslant 1.96$ and zero otherwise.

5. **Replication:** For results chosen for replication, a replication draw is made with

$$X_r|X, \Sigma, \Theta, \sigma_r^2(X, \Sigma, \beta^n), D = 1, R = 1 \sim N\Big(\Theta, \sigma_r(X, \Sigma, \beta^n)^2\Big)$$

where replication standard errors $\sigma_r(X, \Sigma, \beta^n)$ are chosen by replicators as a function of the original estimate, standard error, and intended statistical power $1 - \beta^n$.

We observe i.i.d draws of $\big(X, \Sigma, X_r, \sigma_r(X, \Sigma, \beta^n)\big)$ from the conditional distribution of $\big(X^*, \Sigma^*, X_r, \sigma_r(X^*, \Sigma^*, \beta^n)\big)$ given $D = 1$ and $R = 1$. The Andrews and Kasy (2019) model consists of the first three steps, which are used to identify and estimate $p()$. Subsequent replication steps are introduced to analyze the replication rate.

Step 4 models the replication selection mechanism. This differs across replication studies. For the theory, we assume that the set of significant results chosen for replication is a random sample from published, significant results; selection of insignificant findings for replication, $r_{insig}()$, has no impact on the conclusions and can take any form.

Step 5 models how replicators set statistical power. This is a critical factor in determining replication probabilities. Note that replication estimates are assumed to be unbiased estimates

[4]Notation distinguishing the publication probability function $p()$ over significant and insignificant regions is convenient for presenting and proving the formal results; the 1.96 cutoff corresponds to the threshold over which results are included in the replication rate.

of the true effect and generated from exact replications (i.e. there is no treatment effect heterogeneity across original and replication studies).

In what follows, we normalize θ to be positive and assume that the distribution of true effects, μ_Θ, has support on an open set on the positive real line.[5] The joint probability of publication and being chosen for replication is identified up to scale. Proofs make use of properties of replication probability function in Appendix A. Proposition proofs are in Appendix B.

B. The Replication Rate and Selective Publication

We begin by defining the replication probability of a single study and then use this to define the expected replication probability over multiple studies.

Definition 1 (Replication probability of a single study). *The replication probability of a published study* (X, Σ, Θ) *chosen for replication* $(R = 1)$ *is*

$$RP\Big(X, \Theta, \sigma_r(X, \Sigma, \beta^n)\Big) = \mathbb{P}\left(\frac{|X_r|}{\sigma_r(X, \Sigma, \beta^n)} \geqslant 1.96, \text{sign}(X_r) = \text{sign}(X)\Big| X, \Theta, \sigma_r(X, \Sigma, \beta^n), R = 1\right) \tag{1}$$

This definition captures the dual requirement that the replication estimate is statistically significant and has the same sign as the original study. The replication rate is an aggregate statistic based the fraction of 'successful' replications across multiple original studies. The population analogue of the replication rate is defined next:

Definition 2 (Expected replication probability). *The expected replication probability is defined over published studies* (X, Σ, Θ) *which are chosen for replication* $(R = 1)$ *and statistically significant* $(S_X = 1)$*. It is equal to*

$$\mathbb{E}\Big[RP\big(X, \Theta, \sigma_r(X, \Sigma, \beta^n)\big)\Big| R = 1, S_X = 1\Big]$$

$$= \int RP\Big(x, \theta, \sigma_r(x, \sigma, \beta^n)\Big) f_{X*, \Sigma*, \Theta*|D, R, S_X^*}\big(x, \sigma, \theta | D = 1, R = 1, S_X^* = 1\big) dx d\sigma d\theta \tag{2}$$

This definition highlights an important distinction between being chosen for replication and being included in the replication rate calculation: while insignificant results may be replicated, they are not, by definition, included in the replication rate. This is the main definition of the replication rate reported in most large-scale replication studies (Klein et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Klein et al., 2018).[6] With this, we can state our first main result.

[5] Large-scale replications include studies that examine different questions and outcomes. Normalizing true effects to be positive is justified because relative signs across studies are arbitrary.

[6] Replication power calculations themselves are typically designed for using this definition. Alternative def-

Proposition 1 (The replication rate does not depend on selective publication of null results). *The expected replication probability depends on the probability of publishing significant results, $p_{sig}()$, and does not depend on the probability of publishing insignificant results, $p_{insig}()$.*

Proposition 1 is somewhat counter-intuitive ex-ante, with over 90% of researchers citing 'selective reporting' as a contributing factor to irreproducibility (Baker, 2016). It follows because the replication rate definition does not include statistically insignificant results, which makes it uninformative about the degree to which such results are or are not published. Thus, even if insignificant results are published, they are not included in the replication rate, which focuses only on the replication probability for significant published results. The replication rate instead depends on replication power, the distribution of latent original studies, and the relative probability of publication when the absolute value of the t-ratio is greater than 1.96, $p_{sig}()$. Appendix C provides an example showing how the replication rate varies as we change $p_{sig}()$.

A noteworthy feature of Proposition 1 is that it applies to any rule for setting replication power that depends on the original study's effect size, its standard error, and statistical power target $1 - \beta^n$. This covers the two most common methods of setting power in large-scale replication studies. The first is the common power rule to detect original effect sizes with power equal to $1 - \beta^n$ (e.g. Open Science Collaboration (2015); Camerer et al. (2016)). The second is the fractional power rule, a high-powered variant that sets replication power to detect some fraction of the original effect size (e.g. Camerer et al. (2018, 2022)).

A caveat is that the model assumes a fixed distribution of latent studies, whereas in practice it may be endogenous. For example, changes in the publication of insignificant results could alter the behaviour of researchers, by changing the likelihood that they engage in specification searching or manipulation (Simonsohn et al., 2014; Brodeur et al., 2016, 2020, 2022). Incorporating such behavior in the model would allow the publication probability function $p_{insig}()$ to affect either the joint distribution $\mu_{\Theta,\Sigma}$ (researchers changing the questions they ask); the interdependence of draws from $\mu_{\Theta,\Sigma}$ within a study (multiple hypothesis testing and specification searching); or the distribution of the estimated effect (manipulation of findings), e.g. if $X^* \in (1.96\sigma - \epsilon, 1.96\sigma)$ with $\epsilon > 0$, then with some probability the researcher misreports $X^* \in (1.96\sigma, 1.96\sigma + \epsilon)$. I examine the possible impact of manipulation on the replication rate in Section III.

initions for defining successful replications frequently reported alongside the main definition are: the relative effect size; whether the 95% confidence interval of the replication effect size includes the original effect size; replication based on meta-analytic estimates; the 95% prediction interval approach (Patil et al., 2016); the 'small telescopes' approach (Simonsohn, 2015); and the one-sided default Bayes factor (Wagenmakers et al., 2016).

Finally, note that the insights of Proposition 1 apply more generally to any measure of replication that can be written in the form $\mathbb{E}\big[g(X,\Sigma,X_r,\beta^n)|R=1,S_X=1\big]$.[7] Intuitively, any measure that excludes null results will contain limited information about their prevalence in the published literature. Statistics exclude null results for different reasons. For example, the replication rate excludes them in its definition. On the other hand, large-scale replication studies may only select significant results for replication (i.e. $R=1 \Rightarrow S_X^*=1$). If this is the case, then common alternative measures of replication will also be unresponsive to the degree of selective publication on null results e.g. the relative effect size $\big($i.e. $g(X,X_r)=\frac{X_r}{X}\big)$ and whether or not the replication confidence interval contains the original estimate $\big($i.e. $g(X,\Sigma,X_r,\beta^n)=\mathbb{1}\big[X\in\big(X_r-1.96\sigma_r(X,\Sigma,\beta^n),X_r+1.96\sigma_r(X,\Sigma,\beta^n)\big)\big]\big)$.

C. Common Power Calculations and Low Replication Rates

This section defines the common power rule, and then shows how it leads to replication rates that fall consistently below intended power.

Definition 3 (Common power rule). *The common power rule to detect original effect size x with intended power $1-\beta^n$ sets the replication standard error to*

$$\sigma_r(x,\beta^n)=\frac{|x|}{1.96-\Phi^{-1}(\beta^n)} \tag{3}$$

This is equivalent to setting the replication sample size to $N\times\frac{\sigma}{|x|}\big[1.96-\Phi^{-1}(\beta^n)\big]$, where N and σ are the original study's sample size and standard deviation, respectively.

Lemma 1 (Justification of the common power rule). *Consider a published study (x,σ,θ). If $x=\theta$ and a replication uses the common power rule to detect the original effect with intended power $1-\beta^n$, then*

$$RP\Big(\theta,\theta,\sigma_r(\theta,\beta^n)\Big)=1-\beta^n \tag{4}$$

Proof. Substitute the common power rule in the replication probability function derived in Lemma A1.1 in Appendix A. If $x=\theta$, then

$$RP\big(\theta,\theta,\sigma_r(\theta,\beta^n)\big)=1-\Phi\left(1.96-\text{sign}(\theta)\frac{\theta}{\sigma_r(\theta,\beta^n)}\right)=1-\Phi\left(1.96-\frac{\theta}{\theta}\big(1.96-\Phi^{-1}(\beta^n)\big)\right)=1-\beta^n \tag{5}$$

□

[7] This is Proposition B1 in Appendix B.

Lemma 1 provides the justification for the common power rule. Its reasoning is as follows. Replication probabilities depend crucially on the unobserved true effect θ. If there are no issues with the original study, then its effect size x should be a reasonable proxy for the true effect θ for setting statistical power. Extending this idea to multiple studies suggests that the replication rate should be close to intended power $1-\beta^n$. In practice, it consistently falls below this benchmark (Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Klein et al., 2018). This is commonly interpreted as an indicator of problems with original studies, replication studies, or both.

My next main result shows that even in the absence of any such problems, the expected replication rate will fall short of its intended target:

Proposition 2 (The common power rule implies the expected replication rate is below its intended target) *Suppose $p_{sig}()$ is symmetric about zero, non-zero, and weakly increasing in absolute value. Allow $p_{insig}()$ to take any form. If replication standard errors are set by the common power rule to detect original estimates with intended power $1-\beta^n \geqslant 0.8314$, then*

$$\mathbb{E}\Big[RP\big(X,\Theta,\sigma_r(X,\Sigma,\beta^n)\big)\Big|R=1,S_X=1\Big] < 1-\beta^n \tag{6}$$

Proposition 2 holds under fairly general conditions. It does not rely on any distributional assumptions for latent studies and remains true even under 'ideal' conditions of no selective publication, no researcher manipulation, replications with identical designs and comparable samples (i.e. no heterogeneity in true effects), no measurement error, random sampling in replication selection, and high-powered original studies (although the size of the gap will depend on this). The expected replication probability still falls below intended power in this case, and thus points to fundamental difficulties in interpreting replication rate gaps observed in large-scale replication studies.

There are three main factors underlying this result. I discuss each in reference to the decomposition of the replication rate gap in equation 7, which I implement in the empirical section to quantify the relative contribution of each factor (for clarity, the notation in the decomposition omits the conditioning on $R=1$).

$$\underbrace{(1-\beta^n)-\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1\big]}_{\text{replication rate gap}}$$

$$=\underbrace{(1-\beta^n)-\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|r(t)=1\ \forall t,p(t)=1\ \forall t,X\geqslant 0\big]}_{\text{(i) non-linearity gap}}$$

$$+\underbrace{\mathbb{P}\Big(X<0|S_X=1\Big)\Big(\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1,X\geqslant 0\big]-\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1,X<0\big]\Big)}_{\text{(ii) 'wrong' sign gap}}$$

$$+\underbrace{\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|r(t)=1\ \forall t,p(t)=1\ \forall t,X\geqslant 0\big]-\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1,X\geqslant 0\big]}_{\text{(iii) regression-to-the-mean gap}} \quad (7)$$

Proof. Write the expected replication probability as

$$\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1\big]=\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1,X\geqslant 0\big]$$
$$+\mathbb{P}\Big(X<0|S_X=1\Big)\Big(\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1,X<0\big]\Big)-\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1,X\geqslant 0\big]\Big) \quad (8)$$

To arrive at equation (7), substitute equation (8) into the replication rate gap; add and subtract $\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|r(t)=1\ \forall t,p(t)=1\ \forall t,X\geqslant 0\big]$; and rearrange the terms. □

The first issue with the common power rule is that it does not account for the fact that the replication probability is a non-linear function of the original estimate X. This implies that even for unbiased original estimates, the expected replication probability does not in general equal the replication probability evaluated at true effect (i.e. for $X\sim N(\theta,\sigma^2)$, in general: $\mathbb{E}[RP(X,\Theta,\sigma_r(X,\beta^n))|\Theta=\theta]\neq RP(\theta,\theta,\sigma_r(\theta,\beta^n))=1-\beta^n$). In fact, one can show that the replication probability function is concave in X around the true effect Θ, so Jensen's inequality implies that the expected probability of the unbiased original estimates will tend to fall below intended power. This undermines the justification of the common power rule in Lemma 1, which does not account for the fact that the replication probability function is non-linear and X is a random variable. The first term in the decomposition quantifies the importance of this insight. It is equal the difference between intended power $1-\beta^n$ and the expected replication rate assuming that all results are published and chosen for replication. Note that original estimates X in this term are unbiased for true effects Θ because there is no publication or replication selection based on results (i.e. $r(t)=1\ \forall t,p(t)=1\ \forall t$). Note also that conditioning this expectation on original results with the 'correct' sign allows us to identify the impact of non-linearity on replication probabilities, separate from the impact of attempting to replicate original estimates with the 'wrong' sign, which we turn to next.

The second issue with common power calculations is that random sampling variation means that original estimates will occasionally have the 'wrong' sign. In this case, the replication probability is bounded above by 0.025 since $X<0$ implies $RP\big(X,\Theta,\sigma_r(X,\beta^n)|X<0\big)=\Phi\big(-1.96-\frac{\Theta}{\sigma_r(X,\beta^n)}\big)<0.025$. The likelihood that original estimates have the opposite sign is higher in settings with low power and small true effects (Gelman and Carlin, 2014; Ioannidis et al., 2017). The contribution of this explanation to the replication rate gap is equal to the difference between the expected replication probability of estimates with the 'correct' and those with the 'wrong' sign, weighted by the probability that original estimates have the 'wrong'

sign. Both expectations condition on statistical significance, in line with the definition of the replication rate.

Third, the replication rate conditions on statistical significance, which mechanically induces upward bias in the set of results included in the replication rate.[8] This is an inevitable consequence of the replication rate definition itself. Regression to the mean in (unbiased) replication attempts is to be expected, and common power calculations calibrated to detect inflated original effects may be underpowered for recovering smaller true effects. This statistical fact invalidates the assumption underpinning the justification of the common power rule in Lemma 1. In general, a significant original estimate X is not an unbiased proxy for the unobserved true effect Θ. To measure the importance of this explanation, the decomposition compares the difference in expected replication probabilities between: (i) a regime that publishes and replicates all results, and includes all of them in its replication rate calculation; and (ii) a regime which only includes significant results in its replication rate calculation, as is typically the case. The sign of this term is ambiguous. For any fixed value of Θ, effect sizes will be exaggerated under regime (ii), which lowers the probability of replication relative to regime (i). However, by conditioning on statistical significance, regime (ii) tends to select higher values of Θ, which has the impact of increasing replication probabilities relative to regime (i). The sign and magnitude of the term are determined by the net effect.

III. Empirical Applications

To what extent can Proposition 2 explain low replication rates observed in practice? To evaluate the extent to which issues with power calculations can explain observed replication rates, I conduct the following empirical exercise:

1. I estimate the latent distribution of studies using an augmented version of the Andrews and Kasy (2019) model applied to three large-scale replications. Estimation does not use any data from replications.

2. I use the estimated model to predict what fraction of significant results would replicate, absent any other issues such as p-hacking or heterogeneity.

3. I compare these predictions (which do not use any data from the replications) to the actual replications.

Accurate predictions provide evidence that the issues with power calculations underlying Proposition 2 can adequately explain low observed replication rates. Discrepancies suggest factors not included in the model may also be important.

[8] For a formal statement and proof, see Proposition B2 in Appendix B.

A. Replication Studies

I examine three replication studies. Camerer et al. (2016) replicate results from all 18 between subjects laboratory experiments published in *American Economic Review* and *Quarterly Journal of Economics* between 2011 and 2014. Open Science Collaboration (2015) replicate results from 100 psychology studies in 2008 from *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Following Andrews and Kasy (2019), I consider a subsample of 73 studies with test statistics that are well-approximated by z-statistics. Camerer et al. (2018) replicate 21 experimental studies in the social sciences published between 2010 and 2015 in *Science* and *Nature*.

In Camerer et al. (2016), replicators set power to detect original effects with at least 90% power. In Open Science Collaboration (2015), replication teams were instructed to achieve at least 80% power, and encouraged to obtain higher power if feasible. Reported mean intended power was 92% in both cases. Camerer et al. (2018) implemented a higher-powered design to counter concerns over low statistical power in earlier replication studies. This design consists of two stages. In the first stage, replicators implemented power to detect 75% of the original effect with 90% power. In the second stage, further data collection was undertaken for insignificant results from the first stage, such that the pooled sample from both stages was calibrated to detect half of the original effect size with 90% power. I predict replication outcomes in the first stage.[9]

B. Estimation

To calculate the expected replication rate, it is necessary to estimate the latent distribution of studies $\mu_{\Theta,\Sigma}$. To do this, I estimate an augmented version of the empirical model in Andrews and Kasy (2019). Specifically, Andrews and Kasy (2019) develop an empirical model to estimate the marginal distribution of true effects Θ^*, but not of standard errors Σ^*. Since predictions of the replication rate also require knowledge of the distribution of Σ^*, I augment the model to estimate the joint distribution of (Θ^*, Σ^*). Estimation is based on the 'metastudy approach', which only uses data from original studies. Identification requires that true effects are statistically independent of standard errors, a common assumption in meta-analyses. I

[9]Predicting second-stage outcomes is complicated by the fact that one study that was 'successfully' replicated in the first stage was erroneously included in the second stage. There are two additional reasons for predicting first-stage outcomes. First, Andrews and Kasy (2019) estimate their empirical model for social science experiments on first-stage outcomes, and this article uses identical model specifications to minimize concerns about specification searching to obtain accurate forecasts. Second, ongoing replication studies use the fractional power rule in a single-stage design analogous to the first stage in Camerer et al. (2018). For example, in the ongoing MTurk Replication Project (Camerer et al., 2022), the replication procedure is to 'carry out the data collection based on having 90% power to detect 67% of the effect size reported in the original study.'

assume that Σ^* follows a gamma distribution: $\Gamma(\kappa_\sigma, \lambda_\sigma)$.

For all other aspects of the model, I implement identical parametric assumptions and model specifications as Andrews and Kasy (2019), who examine the same three applications with a focus on estimating publication bias. Matching their model specifications, I assume that $|\Theta^*|$ follows a $\Gamma(\kappa_\theta, \lambda_\theta)$ distribution; and that the selection function $p(X/\Sigma) \times r(X/\Sigma)$, which measures the joint probability of being published and chosen for replication, is a step-function parameterized by $\beta_{\mathbf{p}}$. The inclusion of steps at common significance levels $(1.64, 1.96, 2.58)$ varies slightly across applications owing to different approaches for choosing which studies to replicate.[10]

Table 1 – Maximum Likelihood Estimates

	Latent true effects Θ^*		Latent standard errors Σ^*		Selection parameters		
	κ_θ	λ_θ	κ_σ	λ_σ	β_{p1}	β_{p2}	β_{p3}
Economics experiments							
Augmented model	1.426	0.148	2.735	0.103	0.000	0.039	–
	(1.282)	(0.072)	(0.536)	(0.031)	(0.000)	(0.05)	–
Andrews and Kasy (2019)	1.343	0.157	–	–	0.000	0.038	–
	(1.285)	(0.075)	–	–	(0.000)	(0.05)	–
Psychology experiments							
Augmented model	0.782	0.179	4.698	0.044	0.012	0.303	–
	(0.782)	(0.179)	(4.698)	(0.044)	(0.012)	(0.303)	–
Andrews and Kasy (2019)	0.734	0.185	–	–	0.012	0.300	–
	(0.408)	(0.056)	–	–	(0.007)	(0.134)	–
Social science experiments							
Augmented model	0.070	0.663	5.792	0.028	0.000	0.000	0.584
	(0.091)	(0.326)	(1.754)	(0.009)	(0.000)	(0.000)	(0.419)
Andrews and Kasy (2019)	0.070	0.663	–	–	0.000	0.000	0.583
	(0.091)	(0.327)	–	–	(0.000)	(0.000)	(0.418)

Notes: Maximum likelihood estimates for economics (Camerer et al., 2016), psychology (Open Science Collaboration, 2015) and social sciences (Camerer et al., 2018). Robust standard errors are in parentheses. Latent true effects and standard errors are assumed to follow a gamma distribution; parameters (κ, λ) are the shape and scale parameters, respectively. In economics and psychology, joint publication and replication probability coefficients are measured relative to the omitted category of studies significant at 5 percent level. For example, in experimental economics, an estimate of $\beta_{p2} = 0.038$ implies that results which are significant at the 5% level are 26.3 times more likely to be published and chosen for replication than results that are significant at the 10% level but insignificant at the 5% level. In social sciences, the omitted category is studies significant at the 1% level. Andrews and Kasy (2019) estimates are reproduced from accessible data and code from their analysis.

[10]Details on mechanisms for replication selection are outlined in Appendix D. With $Z = X/\Sigma$, the selection functions in each application are: $r(X/\Sigma) \times p(X/\Sigma) \propto \mathbb{1}\big(1.64 \leqslant |Z| < 1.96\big)\beta_{p2} + \mathbb{1}\big(|Z| \geqslant 1.96\big)$ in economics; $r(X/\Sigma) \times p(X/\Sigma) \propto \mathbb{1}\big(|Z| < 1.64\big)\beta_{p1} + \mathbb{1}\big(1.64 \leqslant |Z| < 1.96\big)\beta_{p2} + \mathbb{1}\big(|Z| \geqslant 1.96\big)$ in psychology; and $r(X/\Sigma) \times p(X/\Sigma) \propto \mathbb{1}\big(1.96 \leqslant |Z| < 2.58\big)\beta_{p3} + \mathbb{1}\big(|Z| \geqslant 2.58\big)$ for social science experiments. Separate identification of the publication probability function, $p()$, requires that we specify the replication selection function $r()$.

Table 1 presents the maximum likelihood estimates. For each large-scale replication study, the first set of rows shows parameter estimates for the augmented model used in this article, and the second set of rows reports estimates from reproducing the results for the standard model in Andrews and Kasy (2019).[11] For common parameters, estimates are very close.

C. The Predicted Replication Rate

This section describes how the estimated models in Table 1 can be used to generate replication rate predictions by simulating replications. Replication probabilities depend on power calculations. I assume that simulated replications implement the power calculations actually used in each application. The procedure is as follows:

1. Draw 10^6 latent (published or unpublished) research questions and standard errors $(\theta^{*sim}, \sigma^{*sim})$ from the estimated joint distribution $\hat{\mu}_{\Theta,\Sigma}(\hat{\kappa}_\theta, \hat{\lambda}_\theta, \hat{\kappa}_\sigma, \hat{\lambda}_\sigma)$.

2. Draw estimated effects $x^{*sim}|\theta^{*sim}, \sigma^{*sim} \sim N(\theta^{*sim}, \sigma^{*sim2})$ for each latent study.

3. Use the estimated selection parameters $(\hat{\beta}_{p1}, \hat{\beta}_{p2}, \hat{\beta}_{p3})$ to determine the subset of studies that are published and chosen for replication.

4. For the subset of replication studies, calculate the replication standard error σ_r^{sim} according to the following rule

$$\sigma_r^{sim}(x^{sim}, \beta^n, \psi) = \frac{\psi \cdot |x^{sim}|}{1.96 - \Phi^{-1}(\beta^n)} \tag{9}$$

   where $\psi = 1$ and $1 - \beta^n = 0.92$ in economics and psychology, which corresponds to the common power rule; and $\psi = \frac{3}{4}$ and $1 - \beta^n = 0.9$ in social science experiments, which corresponds to a fractional power rule.[12]

5. Simulate replications by drawing replication effect sizes $x_r^{sim}|\theta^{sim}, \sigma_r^{sim} \sim N(\theta^{sim}, \sigma_r^{sim2})$

[11] Estimates for psychology in this article are slightly different to the meta-study estimates reported in Andrews and Kasy (2019) (their Table 2). The difference is due to a misreported p-value in the raw psychology data for one study, which leads to an erroneous outlier in the distribution of original study standard errors. Table 1 in this article reproduces estimates of their model after correcting data for this study. Excluding this study in the augmented model leads to very similar replication rate predictions.

[12] This assumes that all simulated replications set intended power equal to the mean of reported intended power. In practice, there was some variation in the application of the power rule around the mean. Appendix E reports predicted replication rates allowing for variation in intended power across studies that matches the empirical variation in each application. Results are very similar and in fact slightly more accurate in all three applications (61.5% in economics; 52.3% in psychology; and 56.5% in social science).

Let $\{x_i, \sigma_i, x_{r,i}, \sigma_{r,i}\}_{i=1}^{M_{sig}}$ be the (simulated) set of published, replicated original studies that are significant at the 5% level, and their corresponding replication results.[13] M_{sig} is the size of the set. The predicted replication rate is equal to

$$\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \mathbb{1}\Big(|x_{r,i}| \geqslant 1.96\sigma_{r,i}, \text{sign}(x_{r,i}) = \text{sign}(x_i)\Big) \tag{10}$$

D. Results

Common Power Rule (Economics and Psychology).—Panel A in Table 2 presents the replication rate predictions. First, consider the replication studies implementing the common power rule. In experimental economics, the predicted replication rate is 60%, which is very close to the observed rate of 61.1%. The accuracy of this prediction provides evidence that low power in original studies in conjunction with replication power issues is sufficient to explain the observed replication rate. This is also consistent with little evidence of researcher manipulation in experimental settings (Brodeur et al., 2016, 2020; Imai et al., 2020)

The second row shows the results for psychology, where the model predicts a replication rate of 54.5%. This is well below mean intended power of 92%, but higher than the observed replication rate of 34.8%. In this case, the model can account for around two-thirds of the replication rate gap. The unexplained portion of the gap in psychology suggests that problems with calculating replication power can account for some but not all of the replication rate gap. Other factors discussed in the literature and not incorporated in the model may be important, including heterogeneity in true effects, p-hacking, and measurement error. Another possibility is that the model should account for differences in replicating main effects and interaction effects, in addition to differences across subfields. For example, Open Science Collaboration (2015) and Altmejd et al. (2019) point out that that interaction effects have a substantially lower probability of replication than main effects, and that the same is true for findings in social psychology compared to those in cognitive psychology.

Calculating the decomposition of the replication rate gap from equation (7) shows that failing to account for the non-linearity of the power function explains the overwhelming majority of the explained replication rate gap in both economics and psychology (Panel B in Table 2). For intuition, note that the replication probability of a null effect is 0.025. Continuity implies that true effects close to zero also have very low replication probabilities. Thus, the non-linearity gap

[13] In both experimental economics and psychology, a small number of original results whose p-values were slightly above 0.05 were treated as 'positive' results and included in the replication rate calculation. To match this, I set the cutoff for significant findings for the purposes of replication equal to the smallest z-statistic that was treated as a 'positive' result for replication. This is 1.81 in economics and 1.86 in psychology. Predictions are almost identical with a strict 0.05 significance threshold.

is large when there is a lot of mass of true effects near zero, which is what is found in practice. Attempts to replicate original estimates with the 'wrong' sign account for between 5.7–8% of the gap, which also reflects relatively low power in original studies. Regression-to-the-mean in replication attempts explains 3.1% of the gap in economics, while it actually *increases* the replication rate in psychology. The latter outcome arises because conditioning on statistical significance tends to select larger true effects, which have higher replication probabilities than small effects; in psychology, this outweighs the fact that for any *fixed* true effect, conditioning on significance induces inflationary bias. For more details on the intuition underlying the decomposition results, see Appendix F.

Fractional Power Rule (Social Sciences).—Concerns over underpowered replication studies have led to modifications of the common power rule to obtain higher statistical power. A popular approach is the fractional power rule, where replication power is set to detect some fraction of the original effect size with a given level of statistical power (e.g. Camerer et al. (2018) and Camerer et al. (2022)). Consider two theoretical observations before examining the empirical results for Camerer et al. (2018). First, Proposition 1 applies to the fractional power rule, namely, the replication rate remains invariant to selective publication of null results. Second, under the specific rule applied in Camerer et al. (2018), the expected replication rate

	Economics experiments	Psychology	Social sciences
A. Replication rate predictions			
Nominal target (intended power)	0.92	0.92	–
Observed replication rate	0.611	0.348	0.571
Predicted replication rate	0.600	0.545	0.553
B. Decomposition of explained gap			
Predicted replication rate gap	0.320 (100%)	0.375 (100%)	–
Non-linearity gap	0.292 (91.16%)	0.364 (97.16%)	–
Wrong-sign gap	0.018 (5.72%)	0.030 (8.03%)	–
Regression-to-the-mean gap	0.010 (3.12%)	-0.019 (-5.18%)	–

Table 2 – Replication Rate Predictions

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social sciences to Camerer et al. (2018). The replication rate is defined as the share of original estimate whose replications have statistically significant findings of the same sign. Figures in the first row report the mean intended power reported in both applications. The second row shows observed replication rates. The third row reports the predicted replication rate in equation (10) calculated using parameter estimates Table 1. Panel B calculates the decomposition of the explained replication rate gap in equation (7) using Monte Carlo methods. In social sciences, power is set to detect three-quarters of the original effect size with 90% power. This approach does not have a fixed nominal target for the replication rate and thus the decomposition is not well-defined. Panel C shows average relative effect size predictions. The relative effect size is defined as the ratio of the replication effect size to the original effect size in Pearson correlation coefficient units.

can range anywhere between 0.025 and 0.99 depending on the power of original studies.[14]

Turning to the empirical results for social science experiments, the predicted replication rate is 55.3%, which is very close to the observed rate of 57.1%. Similarly to economics, this suggests that low power in original studies and issues with power calculations in replications alone can adequately explain the observed replication rate.[15]

Extensions.—I examine four extensions. First, I augment the model to include p-hacking. The augmented model incorporates an egregious form of manipulation, where researchers who obtain a marginally insignificant results misreport their findings as significant with some prespecified probability β_h.[16] The main takeaway is that p-hacking lowers the replication rate, but that its total impact is relatively small when compared to problems with calculating power. A useful benchmark for what might constitute a realistic value for β_h comes from Brodeur et al. (2016) and Brodeur et al. (2022), who estimate that the proportion of 'wrongfully claimed significant results' is around 10%. In the augmented model, this implies that β_h is between 0.22–0.26 across applications (Table G1 in Appendix G). Under these specifications, the replication rate falls by between 2-4 percentage points compared to the case with no manipulation. In economics and psychology, this implies that p-hacking accounts for between 5-6% of the total gap between the predicted replication rate and mean intended power of 92%. Issues related to calculating power account for the remainder of the gap. For further details, see Appendix G.

A second extension uses the estimated models in Table 1 to predict the average relative effect size, a complementary continuous measure of replication defined as the average of the ratio of the replication effect size and the original effect size. Recall that non-random replication selection of statistically significant results implies that this measure is below one in expectation, even in the absence of p-hacking (Proposition B2). I use the estimated models to generate predicted average relative effect sizes using a similar procedure to the replication rate predictions. The predicted average relative effect size is below one and relatively close to observed average relative effect size in economics and social science, although somewhat higher in both cases. In psychology, the predicted average relative effect size is much higher compared to the observed value. See Appendix G for more details.

[14]Under the fractional power rule, the expected replication probability approaches $1 - \Phi[1.96 - \frac{1}{\psi}(1.96 - \Phi^{-1}(\beta^n))]$ as θ/σ approaches infinity (or equivalently, as original power approaches one). With the fraction of original effect size to detect equal to $\psi = 3/4$, and intended power set to $1 - \beta^n = 0.9$, this limit equals 0.99.

[15]Applying the same fractional power rule in social science experiments to economics yields a predicted replication rate of 71.2%. This is higher than in social science experiments because mean power of significant studies in larger in economics.

[16]The augmented model for p-hacking assumes that the estimated models in Table 1 accurately reflect the DGP, which notably do not incorporate p-hacking. This assumption is more plausible for experimental economics and social science, where replication rate predictions are very accurate, compared to psychology, where they are not.

A third extension considers the proposed rule of setting replication power equal to original power in Appendix E. In a review of 108 psychology replications by Anderson and Maxwell (2017), 19 (17.6%) implemented this approach. In all three applications, this approach leads to lower predicted replication rates than under the common power rule.

Given the problems that stem from conditioning on statistical significance, I consider a final extension where the replication rate is extended to include null results that are 'replicated' if their replications are also insignificant. For empirical models in economics and psychology, this 'extended' replication rate remains below intended power under the common power rule.[17] For more details, see Appendix H.

E. Alternative Measures of Selective Publication

Proposition 1 shows that the replication rate is unresponsive to the most salient form of selective publication. For journals and policymakers seeking to change current norms, this highlights the need for more informative measures. In this section, I conduct policy simulations using the estimated model to show how three alternative measures respond to changes in the selective publication of null results:

1. **Replication CI:** This measure counts a replication as 'successful' if its 95% confidence interval covers the original estimate: $\mathbb{1}\big[X \in \big(X_r - 1.96\Sigma_r, X_r + 1.96\Sigma_r\big)\big]\big)$.

2. **Meta-analysis:** The standard criterion of replication with the same sign and significance is applied to a fixed-effect meta-analytic estimate combining the original and replication estimate (uncorrected for selective publication): $\mathbb{1}\big[|X_m| \geqslant 1.96\Sigma_m, \text{sign}(X_m) = \text{sign}(X)\big]$ where X_m and Σ_m are the meta-analytic estimate and standard error, respectively.[18]

3. **Prediction interval:** Original and replication estimates are counted as 'consistent' under this approach if their difference is not statistically different from zero at the 5% level (Patil et al., 2016). This is equivalent to estimating a 95% 'prediction interval' for the original estimate and then determining if it covers the replication estimate: $\mathbb{1}\big[X_r \in$

[17] The extended replication rate is actually *lower* when there is no selective publication, because a higher share of small insignificant original effects are selected for replication. The probability that small, insignificant original effects are also insignificant in replications is relatively low. This is because small original effect sizes will have large sample sizes in replications under the common power rule. Appendix H also considers this extended replication rate when setting replication power equal to original power.

[18] The fixed-effects meta-analytic estimate is a weighted average of original and replication estimates: $X_m = \big(\omega_o X + \omega_r X_r\big)/(\omega + \omega_r)$, where the weights are equal to the precision of each estimate i.e. $(\omega, \omega_r) = (\Sigma^{-2}, \Sigma^{-2})$. These weights minimize the mean-squared error of X_m (Laird and Mosteller, 1990). The variance of this estimator is given by $\Sigma_m^2 = 1/(\omega + \omega_r)$.

$\left(X - 1.96\sqrt{\Sigma^2 + \Sigma_r^2}, X + 1.96\sqrt{\Sigma^2 + \Sigma_r^2}\right)\Big]\Big)$.[19]

These alternative replication measures are frequently reported in large-scale replication studies (Open Science Collaboration, 2015; Camerer et al., 2016, 2018). In simulations, I calculate these measures over significant and insignificant published results, since conditioning on statistical significance makes them unresponsive to selective publication on null results (Proposition B1).

Simulations assume that all results significant at the 5% level are published, and that results insignificant at the 5% level are published with probability β_p. I then calculate how the various measures change with β_p to see how well they capture changes in selective publication (e.g. because of policy changes that reduce selective publication). Policymakers' successful efforts to increase the probability of publishing null results leads to an increase in the policy variable, β_p. Note that while model estimation assumes multiple cutoffs, policy simulations are performed assuming policymakers influence publication probabilities at a single cutoff (1.96) for simplicity (i.e. in the policy simulations I set $\beta_p = \beta_{p1} = \beta_{p2}$ and $\beta_{p3} = 1$ in social science).

Figure 3 shows the results. In line with Proposition 1, the replication rate is completely unresponsive to changes in the probability of publishing null results, making it a poor measure to evaluate efforts to reduce selective publication. Turning to alternative measures, note that the replication CI and meta-analysis measures actually *worsen* when more null results are published ($\beta_p \to 1$). This is because less selective publication leads to more small effects being selected for replication, which have relatively low replication probabilities under these approaches. By contrast, the prediction interval measure is low when selective publication is high, and approaches close to 95% as the probability of publishing null results approach one.[20] The prediction interval measure performs well because it explicitly accounts for the decline in original power as more small effects are selected for replication. Noisy low-powered original studies contain limited information about true effects, which implies that a large range of replication estimates are statistically consistent with them.

Overall, for the purpose of evaluating efforts to reduce selective publication, these results suggest that calculating the prediction interval measure over a random sample of all published results could provide a useful alternative to the replication rate.

[19] This approach assumes that original and replication estimates share the same true effect and are statistically independent. For more details, see the Supplementary Materials for Patil et al. (2016).

[20] When $\beta_p = 1$, the prediction interval measure is slightly higher than 95% in all applications. This is because it assumes that the original estimate X and the replication estimate X_r are uncorrelated. In practice, the replication standard error is a function of the original estimate via the common power rule, which generates some correlation between X and X_r.

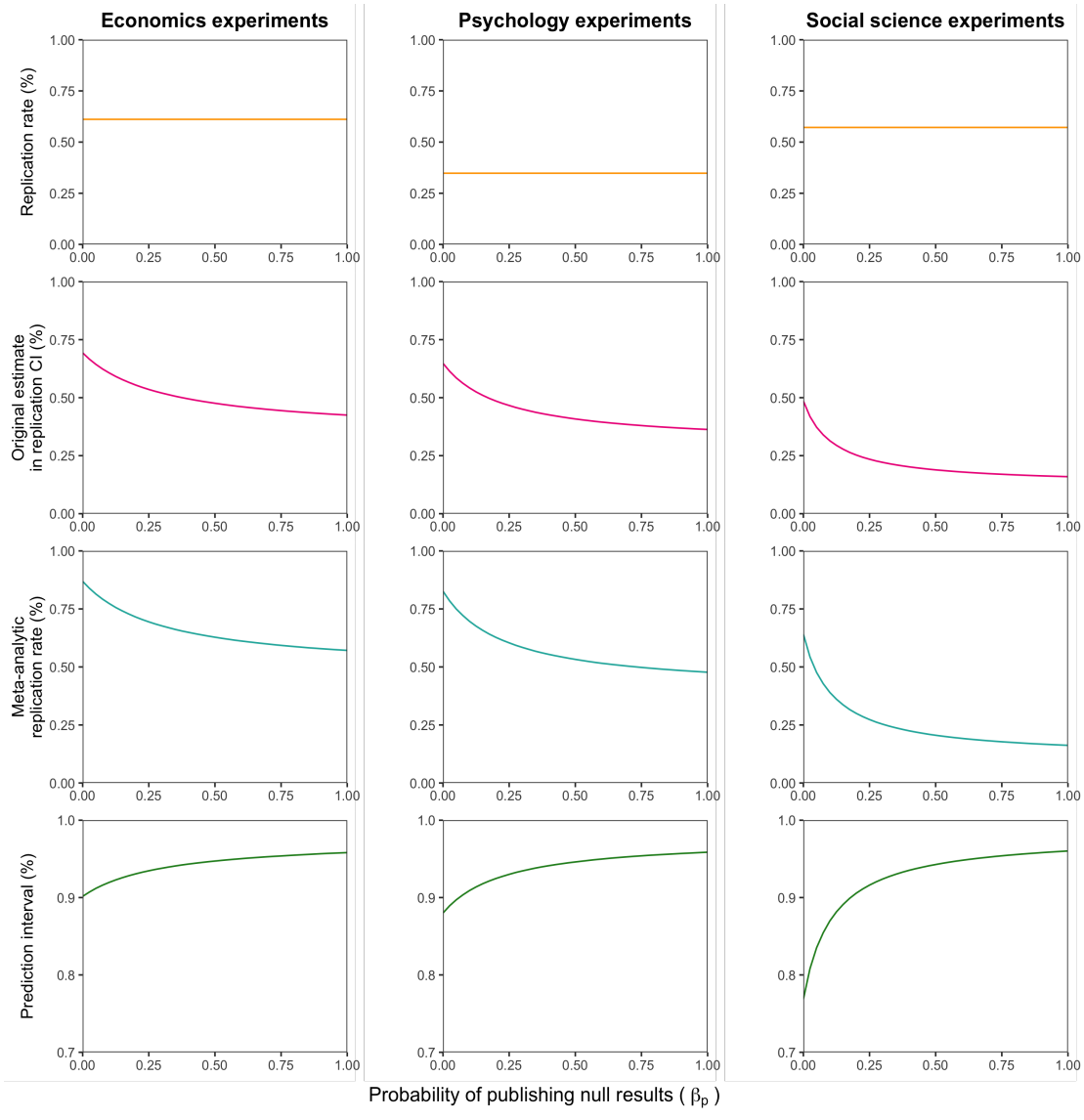


FIGURE 3. POLICY SIMULATIONS – ALTERNATIVE MEASURES OF REPLICATION AND SELECTIVE PUBLICATION

Notes: Details of each measure are provided in the main text. All measures except for the replication rate are calculated over significant and insignificant published results. Simulations use model estimates of the latent distribution of studies from Table 1 and set different levels of selective publication β_p. The first column reproduces replication rate predictions in Table 2.

IV. Conclusion

The prominence of the replication rate stems in part from its apparent transparency and ease of interpretation. However, in reality, it poses substantial difficulties for inference and interpretation. In this article, the first main result shows that the replication rate provides very limited information about selective publication, despite it being the most frequently cited factor contributing to irreproducible research (Baker, 2016).

The second main result shows that problems with power calculations imply that low replication rates should be expected, even in the case where there is no p-hacking or heterogeneity in true effects. In particular, low replication rates arise from the interaction between low power in original studies and the failure to account for non-linearity in the power function when setting replication power. Accurate model-based replication rate predictions suggest that problems with power calculations alone are sufficient to explain observed replication rates in experimental economics and social science.

In light of these results, a reassessment of the empirical content of the replication rate is necessary. In particular, caution should be applied when interpreting the replication rate from large-scale replication studies that use the common power rule to detect original effect sizes with some prespecified power target. In general, these targets are not attainable in expectation, and observed replication rates say little about the most salient form of selective publication. Simulations show that the prediction interval approach proposed by Patil et al. (2016) may provide a useful alternative for measuring selective publication. Finally, these results provide additional evidence in support of recommendations to place greater focus on statistical power for judging the credibility of, and uncertainty surrounding, published research findings (Ioannidis, 2005; Gelman and Carlin, 2014; Anderson and Maxwell, 2017; Camerer et al., 2019).

References

Abadie, Alberto, "Statistical Nonsignificance in Empirical Economics," *American Economic Review: Insights*, 2020, *2* (2), 193–208.

Altmejd, Adam, Anna Dreber, Eskil Forsell et al., "Predicting the replicability of social science lab experiments," *PLoS ONE*, 2019, *14* (12).

Amrhein, Valentin, David Trafimow, and Sander Greenland, "Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication," *The American Statistician*, 2019, *73* (1), 262–270.

_ , Sander Greenland, and Blake McShane, "Retire Statistical Significance," *Nature*, 2019, *567*, 305–307.

Anderson, Smantha F. and Scott E. Maxwell, "Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power ," *Multivariate Behavioral Research*, 2017, *52* (3), 305–324.

Andrews, Isaiah and Maximilian Kasy, "Identification of and Correction for Publication Bias," *American Economic Review*, 2019, *109* (8), 2766–2794.

Arel-Bundock, Vincent, Ryan C. Briggs, Hristos Doucouliagos, and T. D. Avi na Marco Mendoza Stanley, "Quantitative Political Science Research is Greatly Underpowered," *OSF Preprint*, 2022.

Baker, Monya, "1,500 scientists lift the lid on reproducibility," *Nature*, 2016, *533*, 452–454.

Barnett, Adrian G., Jolieke C. Van Der Pols, and Annette J. Dobson, "Regression to the mean: what it is and how to deal with it," *Journal of Business and Psychology*, 2004, *34* (1), 215–220.

Benjamin, Daniel J., James O. Berger, and Valen E. Johnson, "Redefine statistical significance," *Nature Human Behaviour*, 2018, *2*, 6–10.

Brodeur, Abel, Mathias Lé, Marc Sangnier, and Yanos Zylberberg, "Star Wars: The Empirics Strike Back," *American Economic Journal: Applied Economics*, 2016, *8* (1), 1–32.

_ , Nikolai Cook, and Anthony Heyes, "Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics," *American Economic Review*, 2020, *110* (11), 3634–3660.

_ , _ , and _ , "We Need to Talk about Mechanical Turk: What 22,989 Hypothesis Tests Tell Us about Publication Bias and p-Hacking in Online Experiments," *IZA Discussion Paper 15478*, 2022.

Bryan, Christopher J., David S. Yeager, and Joseph M. O'Brien, "Replicator degrees of freedom allow publication of misleading failures to replicate," *Proceedings of the National Academy of Sciences of the United States of America*, 2019, *116* (51), 25535–25545.

Button, Katherine S., John Ioannidis, Claire Mokrysz et al., "Power failure: why small sample size undermines the reliability of neuroscience," *Nature reviews neuroscience*, 2013, *14* (5), 365–376.

Camerer, Colin F., Anna Dreber, and Magnus Johannesson, "Replication and other practices for improving scientific quality in experimental economics," in Arthur Schram and Aljaž Ule, eds., *Handbook of Research Methods and Applications in Experimental Economics*, Edward Elgar Publishing, 2019, chapter 5, p. 83–102.

_ , _ et al., "Evaluating replicability of laboratory experiments in economics," *Science*, 2016, *351* (6280), 1433–1437.

_ , _ , and _ , "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015," *Nature*, 2018, *2*, 637–644.

Camerer, Colin, Yiling Chen, Anna Dreber et al., "Mechanical Turk Replication Project," 2022.

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

Cesario, Joseph, "Priming, Replication, and the Hardest Science," *Perspectives on Psychological Science*, 2014, *9* (1), 40–48.

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

Christensen, Garret, Jeremy Freese, and Edward Miguel, *Transparent and Reproducible Social Science Research*, University of California Press, 2019.

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

_ , Devin Pope, and Eva Vivalt, "Predict science to improve science," *Science*, 2019, *336* (6464), 428–429.

_ , Nicholas Otis, and Eva Vivalt, "Forecasting the Results of Experiments: Piloting an Elicitation Strategy," *AEA Papers and Proceedings*, 2020, *110*, 75–79.

Dreber, Anna, Thomas Pfeiffer, Johan Almenberg et al., "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences of the United States of America*, 2015, *112* (50), 15343–15347.

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

Fisher, Ronald A., "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," *Biometrika*, 1915, *10* (4), 507–521.

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

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

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

Galton, Francis, "Regression Towards Mediocrity in Hereditary Stature," *The Journal of the Anthropological Institute of Great Britain and Ireland*, 1886, *15*, 246–263.

Gelman, Andrew, "The Failure of Null Hypothesis Significance Testing When Studying Incremental Changes, and What to Do About It," *Personality and Social Psychology Bulletin*, 2018, *44* (1), 16–23.

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

Gordon, Michael, Domenico Viganola, Michael Bishop et al., "Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme," *Royal Society Open Science*, 2020, *7*.

Gorroochurn, Prakash, Susan E. Hidge et al., "Non-replication of association studies: "pseudo-failures" to replicate?," *Genet Med*, 2007, *9* (6), 325–331.

Higgins, Julian P.T. and Simon G. Thompson, "Quantifying heterogeneity in a meta-analysis," *Statistics in Medicine*, 2002, *21* (11), 1539–1558.

Hotelling, Harold, "Review: The Triumph of Mediocrity in Business, By Horace Secrist," *Journal of the American Statistical Association*, 1933, *28* (184), 463–465.

Imai, Taisuke, Klavdia Zemlianova, Nikhil Kotecha, and Colin F. Camerer, "How Common are False Positives in Laboratory Economics Experiments? Evidence from the P-Curve Method," *Working Paper*, 2020.

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

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

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

Kahneman, Daniel, *Thinking, Fast and Slow*, Farrar, Straus and Giroux, 2011.

Kasy, Maximilian, "Of Forking Paths and Tied Hands: Selective Publication of Findings, and What Economists Should Do about It," *Journal of Economic Perspectives*, 2021, *35* (3), 175–192.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello et al., "Investigating Variation in Replicability: A "Many Labs" Replication Project," *Social Psychology*, 2014, *45* (3), 142–152.

_ **, Michelangelo Vianello, Fred Hasselman et al.**, "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings," *Advances in Methods and Practices in Psychological Science*, 2018, *1* (4), 443–490.

Laird, Nan M. and Frederick Mosteller, "Some statistical methods for combining exper- imental results," *International Journal of Technology Assessment in Health Care*, 1990, *6* (1), 5–30.

Landis, Ronald S., Lawrence R. James, Charles E. Lance, Charles A. Pierce, and Steven G. Rogelberg, "When is Nothing Something? Editorial for the Null Results Special Issue of Journal of Business and Psychology," *Journal of Business and Psychology*, 2014, *29*, 163–167.

Loken, Eric and Andrew Gelman, "Measurement error and the replication crisis," *Science*, 2017, *355* (6325), 584–585.

Maxwell, Scott E., Michael Y. Lau, and George S. Howard, "Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? ," *American Psychologist*, 2015, *70* (6), 487–498.

McShane, Blakeley B., David Gal, Andrew Gelman, Christian Robert, and Jennifer L. Tackett, "Abandon Statistical Significance," *The American Statistician*, 2019, *73* (1), 235–245.

Mervis, Jeffrey, "Why null results rarely see the light of day," *Science*, 2014, *345*, 992.

Miguel, Edward and Garret Christensen, "Transparency, Reproducibility, and the Credibility of Economics Research," *Journal of Economic Literature*, 2018, *56* (3), 920–980.

Nosek, Brian A., Tom E. Hardwicke, Hannah Moshontz et al., "Replicability, Robustness, and Reproducibility in Psychological Science," *Annual Review of Psychology*, 2022, *73*, 719–748.

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

Patil, Prasad, Roger D. Peng, and Jeffrey T. Leek, "What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science," *Perspectives on Psychological Science*, 2016, *11* (4), 539–544.

Rothstein, Hannah R., Alexander J. Sutton, and Michael Borenstein, *Publication bias in meta-analysis: Prevention, assessment and adjustments*, John Wiley & Sons, 2006.

Shrout, Patrick E. and Jospeh L. Rodgers, "Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis," *Annual Review of Psychology*, 2018, *69*, 487–510.

Simons, Daniel J., "The Value of Direct Replication," *Perspectives on Psychological Science*, 2014, *9* (1), 76–80.

Simonsohn, Uri, "Psychological Science," *Small telescopes: detectability and the evaluation of replication results*, 2015, *26* (5), 559–69.

_ , Leif D. Nelson, and Joseph P. Simmons, "P-Curve: A Key to the File-Drawer," *Journal of Experimental Psychology: General*, 2014, *143* (2), 534–547.

Stanley, T. D., Evan C. Carter, and Hristos Doucouliagos, "What meta-analyses reveal about the replicability of psychological research," *Psychological Bulletin*, 2018, *144* (12), 1325–1346.

Tackett, Jennifer L., Cassandra M. Brandes, Kevin M. King, and Kristian E. Markon, "Psychology's Replication Crisis and Clinical Psychological Science," *Annual Review of Clinical Psychology*, 2019, *15*, 579–604.

Wagenmakers, Eric-Jan, Josine Verhagen, and Alexander Ly, "Behavior Research Methods," *How to quantify the evidence for the absence of a correlation*, 2016, *48* (2), 413–26.

Appendix

A. Properties of the Replication Probability Function

This Appendix derives properties of the replication probability function (Definition 1). The first 'property' simply provides a convenient, compact notation. The remaining properties consider the replication probability function under the common power rule to detect original effect sizes with $1-\beta^n$ intended power (Definition 3). Recall that the replication probability for original study (x, σ, θ) is equal to

$$RP\big(x, \theta, \sigma_r(x, \sigma, \beta^n)\big) = \mathbb{P}\left(\frac{|X_r|}{\sigma_r(x, \beta^n)} \geqslant 1.96, \text{sign}(X_r) = \text{sign}(x)\right) \tag{11}$$

To provide intuition of the properties, Figure A1 provides an illustration of the replication probability function for different values of x under the common power rule for $1-\beta^n = 0.9$ and a fixed value of θ.

Lemma A1 (Properties of the replication probability function). *The replication probability function satisfies the following properties:*

1. *For any replication standard error $\sigma_r(x, \sigma, \beta^n)$, the replication probability for an original study (x, σ, θ) can be written compactly as*

$$RP\big(x, \theta, \sigma_r(x, \sigma, \beta^n)\big) = 1 - \Phi\left(1.96 - \text{sign}(x)\frac{\theta}{\sigma_r(x, \sigma, \beta^n)}\right) \tag{12}$$

The remaining properties assume the replication standard error $\sigma_r(x, \beta^n)$ is set using the common power rule in Definition 3 with intended power $1-\beta^n$:

2. *If $1-\beta^n > 0.025$, then $RP\big(x, \theta, \sigma_r(x, \beta^n)\big)$ is strictly decreasing in x over $(-\infty, 0)$ and $(0, \infty)$.*

3. *If $(1-\beta^n) > 0.6628$, then $RP\big(x, \theta, \sigma_r(x, \beta^n)\big)$ is strictly concave with respect to x over the open interval $(\max\{0, [1-r^*(\beta^n)]\theta\}, [1+r^*(\beta^n)]\theta)$, where*

$$r^*(\beta^n) = -\big(2 + 1.96.h(\beta^n)\big) + \sqrt{\frac{\big(2 + 1.96.h(\beta^n)\big)^2 - 4 \times \big(1 + 1.96.h(\beta^n) - h(\beta^n)^2\big)}{2}} > 0 \tag{13}$$

with $h(\beta^n) = \big(1.96 - \Phi^{-1}(\beta^n)\big)$.

4. *The limits of the replication probability function with respect to x are*

$$\lim_{x\to\infty} RP\big(x,\theta,\sigma_r(x,\beta^n)\big) = 0.025 \text{ and } \lim_{x\to-\infty} RP\big(x,\theta,\sigma_r(x,\beta^n)\big) = 0.025 \tag{14}$$

$$\lim_{x\uparrow 0} RP\big(x,\theta,\sigma_r(x,\beta^n)\big) = 0 \text{ and } \lim_{x\downarrow 0} RP\big(x,\theta,\sigma_r(x,\beta^n)\big) = 1 \tag{15}$$

5. *Suppose $X^* \sim N(\theta,\sigma^2)$. Then $\mathbb{E}\big[RP\big(X,\theta,\sigma_r(X,\beta^n)\big)\big] \to 1-\beta^n$ as $\theta\to\infty$ for fixed σ.*

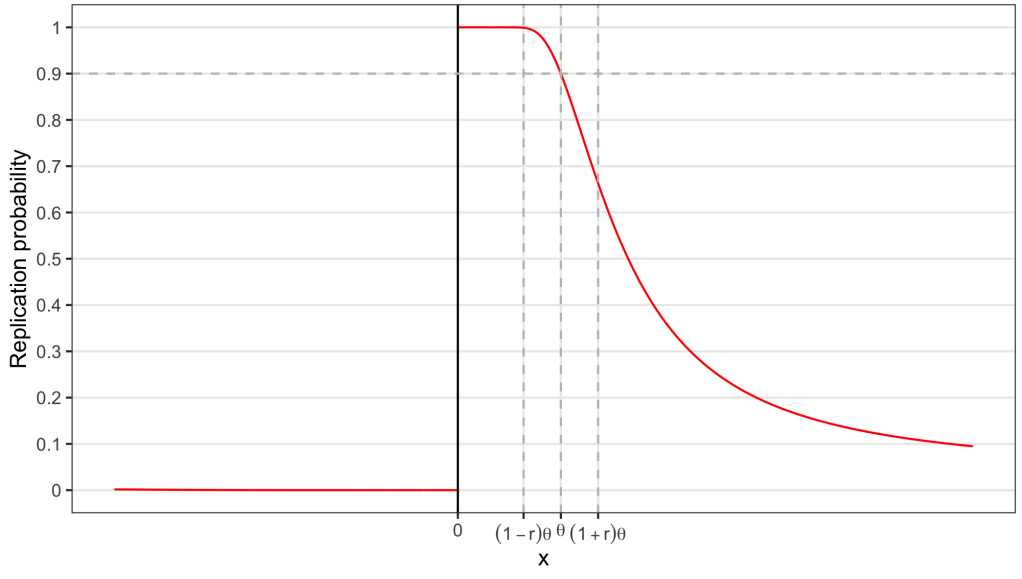


FIGURE A1. Example of the replication probability function under the common power rule with intended power $(1-\beta^n) = 0.9$. The two vertical lines around θ marks the open interval over which the replication probability function is strictly concave, where r^* is given by equation (13).

Proof of 1. The probability in equation (11) equals $\big[\mathbb{1}(x/\sigma \geqslant 1.96) \times \big(1-\Phi\big(1.96-\frac{\theta}{\sigma_r}\big)\big] + \big[\mathbb{1}(x/\sigma \leqslant -1.96) \times \Phi\big(-1.96-\frac{\theta}{\sigma_r}\big)\big]$. This captures the two requirements for 'successful' replication: the replication estimate must attain statistical significance and have the same sign as the original estimate. Equation (12) is obtained using the symmetry of the normal distribution, which implies that $\Phi(t) = 1-\Phi(-t)$ for any t. □

Proof of 2. The first derivative of the replication probability function with the common power rule is

$$\frac{\partial RP\big(x,\theta,\sigma_r(x,\beta^n)\big)}{\partial x} = \begin{cases} -\frac{\theta}{x^2}\big(1.96-\Phi^{-1}(\beta^n)\big) \times \phi\Big(1.96-\frac{\theta}{x}\big(1.96-\Phi^{-1}(\beta^n)\big)\Big), & x>0 \\ -\frac{\theta}{x^2}\big(1.96-\Phi^{-1}(\beta^n)\big) \times \phi\Big(-1.96-\frac{\theta}{|x|}\big(1.96-\Phi^{-1}(\beta^n)\big)\Big), & x<0 \end{cases} \tag{16}$$

These are strictly negative whenever $\big(1.96-\Phi^{-1}(\beta^n)\big)>0 \iff (1-\beta^n)>0.025$. ▫

Proof of 3. First, note that for $x>0$, the second derivative of the replication probability function with the common power rule is

$$\frac{\partial^2 RP\big(x,\theta,\sigma_r(x,\beta^n)\big)}{\partial x^2} = \left(\frac{h(\beta^n)\theta}{x^3}\right)\phi\left(1.96-\frac{h(\beta^n)\theta}{x}\right)\left[1+\left(\frac{h(\beta^n)\theta}{x}\right)\left(1.96-\frac{h(\beta^n)\theta}{x}\right)\right] \tag{17}$$

Let $x=(1+r)\theta$. Substituting this into the previous equation and simplifying shows that equation (17) is strictly negative when the following inequality is satisfied

$$r^2+\big(2+1.96h(\beta^n)\big).r+\big(1+1.96h(\beta^n)-h(\beta^n)^2\big)<0 \tag{18}$$

The solution to the quadratic equation has a unique positive solution $r^*(\beta^n)$ whenever $(1-\beta^n)>0.6628$. To see this, note that there exists a unique positive solution when $\big(1+1.96h(\beta^n)-h(\beta^n)^2\big)<0$. This quadratic equation in $h(\beta^n)$ must have a unique positive and negative solution in turn, since the parabola opens downwards and equals 1 when $h(\beta^n)=0$. The positive root can be obtained from the quadratic formula, which gives 2.38014. Since the quadratic function opens downward, this implies that for any $h(\beta^n)>2.38014$, we have $\big(1+1.96h(\beta^n)-h(\beta^n)^2\big)<0$. Thus, a unique positive solution to equation (18) exists whenever this condition is satisfied. In particular, a unique positive solution exists whenever

$$\begin{aligned} & h(\beta^n)=1.96-\Phi^{-1}(\beta^n)>2.38014 \\ \iff\ & \Phi(1.96-2.38014)>\beta^n \\ \iff\ & (1-\beta^n)>0.6628 \end{aligned} \tag{19}$$

The unique positive solution for equation (18) can again be obtained by the quadratic formula, which gives equation (13). Note that for any $r>0$ where the inequality for concavity in equation (18) is satisfied, the same must also be true of $-r$, since it makes the left-hand-side strictly smaller. This implies that the replication probability function is strictly concave (since its second derivative is strict negative) over $(\max\{0,[1-r^*(\beta^n)]\theta\},[1+r^*(\beta^n)]\theta)$, where the

maximum is taken because the replication probability function is discontinuous at 0. This follows because of the properties of the quadratic function. Specifically, suppose $f(x)$ is a parabola that opens upward and intersects the y-axis at a negative value. Then for any two points (a, b) with $a < b$ and $f(a), f(b) < 0$, it must be that $f(c) < 0$ for any $c \in (a, b)$. ▫

Proof of 4. Substituting the common power rule into the replication probability function gives

$$RP\big(x, \theta, \sigma_r(x, \beta^n)\big) = 1 - \Phi\left(1.96 - \frac{\theta}{x}\big(1.96 - \Phi^{-1}(\beta^n)\big)\right) \tag{20}$$

The values of the limits can be seen immediately from this expression. ▫

Proof of 5. This proof consists of two steps. In the first step, I show that the replication probability function approaches linearity in x in an even interval around θ, as $\theta \to \infty$ for fixed σ. To see this, fix $r \in (0, 1)$. Then the second derivative evaluated at any point $c\theta \in \big(r\theta, (1+r)\theta\big)$ equals

$$\left.\frac{\partial^2 RP\big(x, \theta, \sigma_r(x, \beta^n)\big)}{\partial x^2}\right|_{x=c\theta} = \left(\frac{h(\beta^n)}{c^3\theta^2}\right)\phi\left(1.96 - \frac{h(\beta^n)}{c}\right)\left[1 + \left(\frac{h(\beta^n)}{c}\right)\left(1.96 - \frac{h(\beta^n)}{c}\right)\right] \tag{21}$$

This approaches zero as $\theta \to \infty$, which implies that $RP\big(x, \theta, \sigma_r(x, \beta^n)\big)$ approaches linearity in x over the interval $\big(r\theta, (1 + r)\theta\big)$ in the limit.

For the second step, see that as $\theta \to \infty$ with fixed σ, we have that

$$\mathbb{P}\big[X^* \in \big(r\theta, (1 + r)\theta\big)|\theta, \sigma\big] = \Phi\left(\frac{(1 + r)\theta - \theta}{\sigma}\right) - \Phi\left(\frac{r\theta - \theta}{\sigma}\right) \to 1 \tag{22}$$

That is, the probability of drawing X^* inside of the range $\big(r\theta, (1 + r)\theta\big)$ approaches one in the limit. But from the first step we know that the replication probability function is linear over this range as $\theta \to \infty$ with fixed σ. This implies in the limit that $\mathbb{E}\big[RP\big(X, \theta, \sigma_r(X, \beta^n)\big)\big] = RP\big(\mathbb{E}[X], \theta, \sigma_r(X, \beta^n)\big) = RP\big(\theta, \theta, \sigma_r(X, \beta^n)\big) = 1 - \beta^n$, as shown in Lemma 1 in the main text.

B. Proofs of Propositions

Proposition B1. *Let $g()$ be a function of X, Σ, X_r, β^n. Then $\mathbb{E}\big[g(X, \Sigma, X_r, \beta^n)|R = 1, S_X = 1\big]$ depends on the probability probability of publishing significant results, $p_{sig}()$, and does not depend on the probability of publishing insignificant results, $p_{insig}()$.*

Proof. We can write $\mathbb{E}\big[g(X, \Sigma, X_r, \beta^n)|R = 1, S_X = 1\big]$ as

$$\int g(x, \sigma, x_r, \beta^n) f_{X*,\Sigma*,\Theta*,X_r|D,R,S_X^*}\Big(x, \sigma, \theta, x_r|D = 1, R = 1, S_{X*} = 1\Big) dx d\sigma d\theta dx_r$$

$$= \int_{x,\sigma,\theta} \left(\int_{x_r} g(x, \sigma, x_r, \beta^n) f_{X_r|X*,\Sigma*,\Theta*}\Big(x_r|\theta, \sigma_r(x, \sigma, \beta^n)\Big) dx_r \right) f_{X*,\Sigma*,\Theta*|D,R,S_X^*}(x, \sigma, \theta|D = 1, R = 1, S_X^* = 1) dx d\sigma d\theta \tag{23}$$

The equality uses $f_{X_r|X*,\Sigma*,\Theta*,D,R,S_X^*}\big(x_r|\theta, \sigma_r(x, \sigma, \beta^n)\big) = f_{X_r|X*,\Sigma*,\Theta*}\big(x_r|\theta, \sigma_r(x, \sigma, \beta^n)\big)$. Replication estimates are not subject to selective publication, which implies this is a normal density that does not depend on $p()$. Hence, the term in parentheses can only be affected by $p()$ indirectly through $f_{X*,\Sigma*,\Theta*|D,R,S_X^*}$, which is the joint distribution of original studies conditional on being published, chosen for replication, and statistically significant at the 5% level. However, this distribution does not depend on the probability of publishing insignificant findings. To see this, apply Bayes rule twice to get

$$\begin{aligned} & f_{X*,\Sigma*,\Theta*|D,R,S_X^*}(x, \sigma, \theta|D = 1, R = 1, S_X^* = 1) \\ &= \frac{\mathbb{P}\Big(D = 1\big|X^* = x, \Sigma^* = \sigma, \Theta^* = \theta, R = 1, S_X^* = 1\Big)}{\mathbb{P}\Big(D = 1\big|R = 1, S_X^* = 1\Big)} \times \frac{\mathbb{P}\Big(R = 1\big|X^* = x, \Sigma^* = \sigma, \Theta^* = \theta, S_X^* = 1\Big)}{\mathbb{P}\Big(R = 1\big|S_X^* = 1\Big)} \\ & \quad \times f_{X*,\Theta,\Sigma*|S_X^*}\Big(x, \theta, \sigma\big|S_X^* = 1\Big) \\ &= \frac{p_{sig}(x/\sigma)}{\mathbb{E}\big(p_{sig}(X*/\Sigma*)\big|S_X^* = 1\big)} \cdot \frac{r_{sig}(x/\sigma)}{\mathbb{E}\big(r_{sig}(X*/\Sigma*)\big|S_X^* = 1\big)} \cdot f_{X*,\Sigma*,\Theta*|S_X^*}\Big(\theta, x, \sigma\big|S_X^* = 1\Big) \end{aligned} \tag{24}$$

In the final line, the first factor in the product includes only $p_{sig}()$; the denominator does not condition on R because replication selection is assumed to be random for significant findings. The second factor equals one because replication selection for significant results is assumed to be random. The final factor in the product is the density of latent studies conditional on significance, which is not affected by selective publication. □

Proposition B2 (Regression to the mean in replications). *Suppose $p_{sig}()$ is symmetric about zero, non-zero, differentiable and weakly increasing in absolute value. Allow $p_{insig}()$ to take any form. Published original estimates X and corresponding replication estimates X_r satisfy*

$$\mathbb{E}\big[X\big|\Theta=\theta, S_X=1\big] > \theta = \mathbb{E}\big[X_r|\Theta=\theta\big] \tag{25}$$

Proof. We have $\mathbb{E}\big(X_r\big|\Theta=\theta\big)=\theta$ by assumption. Next, note that

$$\mathbb{E}_{X^*|\Theta^*,S_X^*,D}\Big(X^*|\Theta^*=\theta,|X^*/\Sigma^*|\geqslant 1.96, D=1\Big)=\mathbb{E}_{X|\Theta,S_X}\Big(X|\Theta=\theta,|X/\Sigma|\geqslant 1.96\Big)$$

$$=\mathbb{E}_{\Sigma|\Theta,S_X}\left(\mathbb{E}_{X|\Theta,\Sigma,S_X}\Big(X|\Theta=\theta,\Sigma=\sigma,|X/\sigma|\geqslant 1.96\Big)\right) \tag{26}$$

where the last line uses the Law of Iterated Expectations. We will prove $\mathbb{E}_{X|\Theta,\Sigma,S_X^*}\big(X|\Theta=\theta,\Sigma=\sigma,|X/\sigma|\geqslant 1.96\big)>\theta$, which implies that equation (26) is also greater than θ. Recall that $X|\theta,\sigma$ is the effect size of published studies and follows a truncated normal distribution:

$$\frac{p\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\big(\frac{x-\theta}{\sigma}\big)\mathbb{1}\big(|\frac{x}{\sigma}|\geqslant 1.96\big)}{\int p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\big(\frac{x'-\theta}{\sigma}\big)\mathbb{1}\big(|\frac{x}{\sigma}|\geqslant 1.96\big)dx'} \tag{27}$$

Define $X=\theta+\sigma Z$. Then the density for the transformed random variable Z is

$$\frac{p\big(z+\frac{\theta}{\sigma}\big)\phi\big(z\big)\mathbb{1}\big(|z+\frac{\theta}{\sigma}|\geqslant 1.96\big)}{\int p\big(z'+\frac{\theta}{\sigma}\big)\phi\big(z'\big)\mathbb{1}\big(|z+\frac{\theta}{\sigma}|\geqslant 1.96\big)dz'} \tag{28}$$

For notational convenience, define the following normalization constants:

$$\bar{\eta}=\mathbb{P}(X\leqslant -1.96\sigma)+\mathbb{P}(X\geqslant 1.96\sigma)=\mathbb{P}\left(Z\leqslant -1.96-\frac{\theta}{\sigma}\right)+\mathbb{P}\left(Z\geqslant 1.96-\frac{\theta}{\sigma}\right) \tag{29}$$

$$\eta_1=\mathbb{P}(X\leqslant -1.96\sigma)=\mathbb{P}\left(Z\leqslant -1.96-\frac{\theta}{\sigma}\right) \tag{30}$$

$$\eta_2=\mathbb{P}(X\geqslant 2\theta+1.96\sigma)=\mathbb{P}\left(Z\geqslant \frac{\theta}{\sigma}+1.96\right) \tag{31}$$

$$\eta_3=\mathbb{P}(1.96\sigma\leqslant X\leqslant 2\theta-1.96\sigma)=\mathbb{P}\left(1.96-\frac{\theta}{\sigma}\leqslant Z\leqslant \frac{\theta}{\sigma}-1.96\right) \tag{32}$$

Case 1. Consider two cases. First, suppose $\theta\in(0,1.96\sigma)$. Conditional on (θ,σ) (where we suppress the conditional notation on (θ,σ) for clarity), the expected value of a published estimate conditional of statistical significance is

$$\mathbb{E}(X|1.96\sigma \leqslant |X|) = \frac{1}{\bar{\eta}}\Bigg(\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma)$$

$$+ \big(\bar{\eta} - \eta_1 - \eta_2\big)\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta + 1.96\sigma)\Bigg) \tag{33}$$

First note that $\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta + 1.96\sigma) > \theta$ since we assume that $\theta \in (0, 1.96\sigma)$ and $p_{sig}() > 0$. If $\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) \geqslant \big(\eta_1 + \eta_2\big)\theta$, it follows that $\mathbb{E}(X|1.96\sigma \leqslant |X|) > \theta$, which is what we want to show. Consider the first expectation in this expression:

$$\mathbb{E}(X|X \leqslant -1.96\sigma) = \mathbb{E}\Big(\theta + \sigma Z|Z \leqslant -1.96 - \frac{\theta}{\sigma}\Big) = \theta + \sigma\mathbb{E}\Big(Z|Z \leqslant -1.96 - \frac{\theta}{\sigma}\Big) \tag{34}$$

Evaluating the expectation in the right-hand-side of equation (34) gives

$$\mathbb{E}\Big(Z|Z \leqslant -1.96 - \frac{\theta}{\sigma}\Big) = \frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} zp_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz = -\frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi'(z)dz$$

$$= -\frac{1}{\eta_1}\Bigg[p_{sig}(-1.96)\phi\Big(-1.96 - \frac{\theta}{\sigma}\Big) - p_{sig}(-\infty)\phi(-\infty) - \int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz\Bigg]$$

$$= -\frac{1}{\eta_1}p_{sig}(-1.96)\phi\Big(-1.96 - \frac{\theta}{\sigma}\Big) + \frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz \tag{35}$$

where the second equality uses $\phi'(z) = -z\phi(z)$; the third equality uses integration by parts; and the final equality follows because $p_{sig}(-\infty)\phi(-\infty) = 0$ since $p_{sig}()$ is bounded between zero and one. Substituting this into equation (34) gives

$$\mathbb{E}(X|X \leqslant -1.96\sigma) = \theta - \frac{\sigma}{\eta_1}p_{sig}(-1.96)\phi\Big(-1.96 - \frac{\theta}{\sigma}\Big) + \frac{\sigma}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz \tag{36}$$

Next, note that

$$\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) = \theta + \sigma\mathbb{E}\Big(Z|Z \leqslant \frac{\theta}{\sigma} + 1.96\Big) \tag{37}$$

where

$$\mathbb{E}\Big(Z|Z \leqslant \frac{\theta}{\sigma} + 1.96\Big) = \frac{1}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} z p_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz \geqslant \frac{1}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} z p_{sig}\bigg(z - \frac{\theta}{\sigma}\bigg)\phi(z)dz \quad (38)$$

since $p_{sig}(z+\theta/\sigma) \geqslant p_{sig}(z-\theta/\sigma)$ for all $z \in (1.96+\theta/\sigma, \infty)$ because $p_{sig}(t)$ is weakly increasing over $t > 1.96$. For the right-hand-side of this equation, we can apply similar arguments used to derive equation (35). Substituting the result into equation (37) gives

$$\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) \geqslant \theta + \frac{\sigma}{\eta_2} p_{sig}(1.96)\phi\bigg(1.96 + \frac{\theta}{\sigma}\bigg) + \frac{\sigma}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(z - \frac{\theta}{\sigma}\bigg)\phi(z)dz \quad (39)$$

Equations (36) and (39) imply

$$\begin{aligned}
&\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) \\
&\geqslant (\eta_1+\eta_2)\theta + \sigma\bigg[p_{sig}(1.96)\phi\bigg(1.96 + \frac{\theta}{\sigma}\bigg) - p_{sig}(-1.96)\phi\bigg(-1.96 - \frac{\theta}{\sigma}\bigg)\bigg] \\
&+ \sigma\bigg[\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz + \int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(z - \frac{\theta}{\sigma}\bigg)\phi(z)dz\bigg] = (\eta_1+\eta_2)\theta \qquad (40)
\end{aligned}$$

In the second line, the second term in the sum equals zero because symmetry of $p_{sig}()$ and $\phi()$ about zero implies that both terms in the brackets are equal. To see why the third term in the sum equals zero, note that

$$\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\bigg(z + \frac{\theta}{\sigma}\bigg)\phi(z)dz = \int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(-u + \frac{\theta}{\sigma}\bigg)\phi(u)du = -\int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\bigg(u - \frac{\theta}{\sigma}\bigg)\phi(u)du \quad (41)$$

The first equality follows from both changing the order of the integral limits and applying the substitution $u = -x$; it also uses the symmetry of $\phi()$. The final equality holds because symmetry of $p_{sig}()$ about zero implies that for any $t > 1.96$, $p'_{sig}(t) = -p'_{sig}(-t)$.

Case 2. Consider the second case where $\theta \geqslant 1.96\sigma$. For a given (θ, σ), we have

$$\mathbb{E}(X|1.96\sigma \leqslant |X|) = \frac{1}{\bar{\eta}}\bigg(\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma)$$

$$\eta_3\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma) + \big(\bar{\eta} - \eta_1 - \eta_2 - \eta_3\big)\mathbb{E}(X|2\theta - 1.96\sigma \leqslant X \leqslant 2\theta + 1.96\sigma)\Bigg)$$

$$> \frac{1}{\bar{\eta}}\Bigg(\theta(\eta_1 + \eta_2) + \big(\bar{\eta} - \eta_1 - \eta_2 - \eta_3\big)\theta + \eta_3\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma)\Bigg) \tag{42}$$

The inequality follows from two facts. First, the inequality proved in the first case: $\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) \geqslant (\eta_1 + \eta_2)\theta$. Second, the expectation in the third term of the sum satisfies $\mathbb{E}(X|2\theta - 1.96\sigma \leqslant X \leqslant 2\theta + 1.96\sigma) > \theta$ because $\theta \geqslant 1.96\sigma \iff 2\theta - 1.96\sigma \geqslant \theta$ and we assume that $p_{sig}() > 0$.

It remains to show that $\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma) \geqslant \theta$. Then it follows that $\mathbb{E}(X|1.96\sigma \leqslant |X|) > \theta$, which is what we want to show. First, note that

$$\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma) = \theta + \sigma\mathbb{E}\left(Z\middle|1.96 - \frac{\theta}{\sigma} \leqslant Z \leqslant -1.96 + \frac{\theta}{\sigma}\right) \tag{43}$$

It is therefore sufficient to show that $\mathbb{E}\left(Z\middle|1.96 - \frac{\theta}{\sigma} \leqslant Z \leqslant -1.96 + \frac{\theta}{\sigma}\right) \geqslant 0$. Writing out the expectation in full gives

$$\mathbb{E}\left(Z\middle|1.96-\frac{\theta}{\sigma} \leqslant Z \leqslant -1.96+\frac{\theta}{\sigma}\right) = \frac{1}{\eta_3}\left(\int_{1.96-\frac{\theta}{\sigma}}^{0} zp_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz + \int_{0}^{\frac{\theta}{\sigma}-1.96} zp_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz\right)$$

$$= \frac{1}{\eta_3}\left(\int_{0}^{\frac{\theta}{\sigma}-1.96} z\left[p_{sig}\left(z + \frac{\theta}{\sigma}\right) - p_{sig}\left(-z + \frac{\theta}{\sigma}\right)\right]\phi(z)dz\right) \geqslant 0 \tag{44}$$

The second equality follows because

$$\int_{1.96-\frac{\theta}{\sigma}}^{0} zp_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz = -\int_{0}^{1.96-\frac{\theta}{\sigma}} zp_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz = -\int_{0}^{\frac{\theta}{\sigma}-1.96} up_{sig}\left(-u+\frac{\theta}{\sigma}\right)\phi(u)du \tag{45}$$

which uses the substitution $u = -x$ and the symmetry of $\phi()$. The weak inequality in equation (44) follows because $p_{sig}()$ is assumed to be weakly increasing over positive values. Thus, $z - \theta/\sigma > -z + \theta/\sigma$ for all $z \in (0, \theta/\sigma - 1.96)$ implies $p_{sig}\big(z + \theta/\sigma\big) - p_{sig}\big(-z + \theta\sigma\big) \geqslant 0$.

This covers all cases and proves the proposition. □

Proof of Proposition 1: This follows immediately from the more general result in Proposition B1 with $g(x, \sigma, x_r, \beta^n) = \mathbb{1}\Big[\frac{|x_r|}{\sigma_r(x,\sigma,\beta^n)} \geqslant 1.96, \text{sign}(x_r) = \text{sign}(x)\Big]$.

Proof of Proposition 2: For notational convenience, let $(X_{sig}, \Sigma_{sig}, \Theta_{sig})$ denote the distribution of latent studies $(X^*, \Sigma^*, \Theta^*)$ conditional on being published $(D = 1)$ and statistically

significant ($|X^*/\Sigma^*| \geqslant 1.96$). The expected replication probability (Definition 2) under the common power rule (Definition 3) can be written as

$$\begin{aligned}
&\mathbb{E}_{X^*,\Sigma^*,\Theta^*|D,R,S_X^*}\Big[RP\Big(X^*,\Theta^*,\sigma_r(X^*,\beta^n)\Big)\Big|D=1,R=1,|X^*/\Sigma^*|\geqslant 1.96\Big] \\
&= \mathbb{E}_{X,\Sigma,\Theta|S_X}\Big[RP\big(X,\Theta,\sigma_r(X,\Sigma,\beta^n)\big)\big||X/\Sigma|\geqslant 1.96\Big] \\
&= \mathbb{E}_{X_{sig},\Sigma_{sig},\Theta_{sig}}\Big[RP\Big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta^n)\Big)\Big] \\
&= \mathbb{E}_{\Sigma_{sig},\Theta_{sig}}\Bigg[\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\Big[RP\Big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta^n)\Big)|\Theta_{sig}=\theta,\Sigma_{sig}=\sigma\Big]\Bigg] \qquad (46)
\end{aligned}$$

where the second inequality drops the conditioning on being chosen for replication (R) because it is assumed that replication selection on significant results is random; and the last equality uses the Law of Iterated Expectations. The proof shows that the conditional expected replication probability satisfies $\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\big[RP\big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta^n)\big)|\Theta_{sig}=\theta,\Sigma_{sig}=\sigma\big] < 1-\beta^n$ which implies that the expected replication probability is also less than intended power $1-\beta^n$. For greater clarity in what follows, let $\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\big]$ be shorthand for $\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\big[RP\big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta^n)\big)|\Theta_{sig}=\theta,\Sigma_{sig}=\sigma\big]$.

Note that the conditional expected replication probability can be written explicitly as

$$\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\big]=\int\Bigg(1-\Phi\Big(1.96-\frac{\theta}{x}\big(1.96-\Phi^{-1}(\beta^n)\big)\Big)\Bigg)\frac{p\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}|\geqslant 1.96\big)dx}{\int_{x'}p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}|\geqslant 1.96\big)dx'} \qquad (47)$$

where the integrand in equation (47) is obtained using the compact notation for the replication probability derived in Lemma A1.1 and then substituting the common power rule in Definition 3. The pdf of estimates differs from a normal distribution in two respects: (1) the publication probability function $p\big(\frac{x}{\sigma}\big)$ reweights the distribution; and (2) conditioning on statistical significance truncates original effects falling in the insignificant region $(-1.96\sigma, 1.96\sigma)$. The denominator is the normalization constant.

First, we introduce some notation. Lemma A1.3 shows that if $(1-\beta^n) > 0.6628$, then $RP\big(x,|\theta,\sigma,\beta^n\big)$ is strictly concave over the open interval $(\max\{0,[1-r^*(\beta^n)]\theta\},[1+r^*(\beta^n)]\theta)$, where $r^*(\beta^n)$ is given by equation (13). This Proposition assumes $(1-\beta^n) > 0.8314$, so the condition is satisfied. To simplify the notation, define $(l^*,u^*) = \big((1-r^*)\theta,(1+r^*)\big)$ when $r^* \in (0,1)$ and $(l^*,u^*) = \big(0,2\theta\big)$ when $r^* \geqslant 1$; in both cases, the replication probability function is strictly concave over an interval with mid-point θ.

Consider first the case where $r^* \geqslant 1$ so that $(l^*,u^*) = \big(0,2\theta\big)$. The conditional replication probability can be expressed as a weighted sum

$$\mathbb{E}\Big[\big(RP(X_{sig}|\theta,\sigma,\beta^n)\Big] = \mathbb{P}\Big(X_{sig} < l^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|X_{sig} < l^*\Big]$$
$$+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big]+\mathbb{P}\Big(X_{sig} > u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|X_{sig} > u^*\Big]$$
$$< \mathbb{P}\Big(X_{sig} < l^*\Big)0.025+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big]+\mathbb{P}\Big(X_{sig} > u^*\Big)\big(1-\beta^n\big) \tag{48}$$

In the last line, the first term in the sum uses the fact that the maximum value of the replication probability when $x < l^* = 0$ is 0.025 (Lemma A1.2 and Lemma A1.4 in Appendix A). The third term follows because $RP\big(2\theta|\theta,\sigma,\beta^n\big)$ is the maximum value the function takes over $x > u^* = 2\theta$, since the function is strictly decreasing over $x > 0$ (Lemma A1.2); and therefore that $RP\big(2\theta|\theta,\sigma,\beta^n\big) < RP\big(\theta|\theta,\sigma,\beta^n\big) = 1-\beta^n$, where the equality is shown in Lemma 1. From equation (48), we can see that $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta^n)|l^* \leqslant X_{sig} \leqslant u^*\big] < 1-\beta^n$ is a sufficient condition for $\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\big] < 1-\beta^n$.

Before showing that this sufficient condition is satisfied, we show that the same sufficient condition holds in the second case, where $r^* \in (0,1)$ so that $(l^*,u^*) = \big((1-r^*)\theta,(1+r^*)\theta\big)$. This requires additional steps. First, express the conditional replication probability as a weighted sum

$$\mathbb{E}\Big[\big(RP(X_{sig}|\theta,\sigma,\beta^n)\Big] = \mathbb{P}\Big(X_{sig} \leqslant l^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|X_{sig} \leqslant l^*\Big]$$
$$+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big]+\mathbb{P}\Big(X_{sig} \geqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|X_{sig} \geqslant u^*\Big]$$
$$< \mathbb{P}\Big(X_{sig} \leqslant l^*\Big)+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big]+\mathbb{P}\Big(X_{sig} \geqslant u^*\Big)RP\Big(u^*|\theta,\sigma,\beta^n\Big) \tag{49}$$

The strict inequality follows for two reasons. For the first term in the sum, one is the maximum value the function can take for any x. For the third term, $RP(u^*|\theta,\sigma,\beta^n)$ is the function's maximum value over $x \geqslant u^*$, since the integrand is strictly decreasing over positive values (Lemma A1.2). With an additional step, we can write this inequality as

$$\mathbb{E}\Big[\big(RP(X_{sig}|\theta,\sigma,\beta^n)\Big] < \frac{1}{2}\Big(1-\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\Big)\Big(1+RP\big(u^*|\theta,\sigma,\beta^n\big)\Big)$$
$$+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] \tag{50}$$

This follows because $\mathbb{P}(X_{sig} \leqslant l^*) \leqslant \mathbb{P}(X_{sig} \geqslant u^*)$ and $RP(u^*|\theta,\sigma,\beta^n) < 1$. That is, increasing the relative weight on the maximum value of one, such that both tails are equally weighted, must lead to a (weakly) larger value. The weak inequality $\mathbb{P}(X_{sig} \leqslant l^*) \leqslant \mathbb{P}(X_{sig} \geqslant u^*)$ required for this simplification is shown below:

Lemma B1. *Suppose $X|\theta,\sigma$ follows the truncated normal pdf in equation (47). Then for any*

$r^* \in (0,1)$, *the following inequality holds:* $\mathbb{P}\big(X_{sig} \leqslant (1-r^*)\theta\big) < \mathbb{P}\big(X_{sig} \geqslant (1+r^*)\theta\big)$.

Proof. First, note that $\big((1-r^*)\theta, (1+r^*)\theta\big)$ is an interval over the positive real line centered at θ. Consider two cases:

Case 1: Let $(1-r^*)\theta \leqslant 1.96\sigma$. Define the normalization constant $C = \int_{x'} p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}| \geqslant 1.96\big)dx'$. Then

$$\mathbb{P}\Big(X_{sig} \leqslant (1-r^*)\theta\Big) = \frac{1}{C}\int_{-\infty}^{-1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' \leqslant \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx'$$

$$< \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{\max\{1.96\sigma,(1+r^*)\theta\}}^{2\theta+1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' = \mathbb{P}\Big(X_{sig} \geqslant (1+r^*)\theta\Big) \tag{51}$$

Consider the weak inequality. Note that the mid-point between -1.96σ and $2\theta + 1.96\sigma$ is θ. Thus, with no selective publication (i.e. $p(t) = 1$ for all t), we would have equality owing to the symmetry of the normal distribution. However, recall that $p_{sig}()$ is symmetric about zero and weakly increasing in absolute value. It follows therefore that $|2\theta + 1.96\sigma| > |-1.96\sigma|$ implies $p_{sig}(|2\theta + 1.96\sigma|) \geqslant p_{sig}(|-1.96\sigma|)$; using this fact and symmetry of the normal distribution about θ gives the weak inequality. The strict inequality follows because the additional term is strictly positive, since $p_{sig}()$ is assumed to be non-zero.

Case 2: Let $(1-r^*)\theta > 1.96\sigma$. The argument is similar to the first case:

$$\mathbb{P}\Big(X_{sig} \leqslant (1-r^*)\theta\Big) = \frac{1}{C}\int_{-\infty}^{-1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{1.96\sigma}^{(1-r^*)\theta} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx'$$

$$< \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{(1+r^*)\theta}^{2\theta-1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx'$$

$$+ \frac{1}{C}\int_{2\theta-1.96\sigma}^{2\theta+1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' = \mathbb{P}\Big(X_{sig} \geqslant (1+r^*)\theta\Big) \tag{52}$$

□

The inequality in equation (50) can be further simplified by placing restrictions on intended power. In particular, if intended power satisfies $1-\beta^n \geqslant 0.8314$, then

$$\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big] < \Big(1-\mathbb{P}\big(l^* \leqslant X_{sig} \leqslant u^*\big)\Big)\big(1-\beta^n\big)$$

$$+ \mathbb{P}\big(l^* \leqslant X_{sig} \leqslant u^*\big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] \tag{53}$$

This follows because with $u^* = (1 + r^*)\theta$, we have

$$\frac{1}{2}\Big(1 + RP\big(u^*\big|\theta, \sigma, \beta^n\big)\Big) = \frac{1}{2}\left(1 + \left(1 - \Phi\left(1.96 - \frac{1.96 - \Phi^{-1}(\beta^n)}{1 + r^*(\beta^n)}\right)\right)\right)$$
$$\leqslant 1 - \beta^n \iff 1 - \beta^n \geqslant 0.8314 \tag{54}$$

From equation (53), we can see that $\mathbb{E}\big[RP(X_{sig}|\theta, \sigma, \beta^n)\big|l^* \leqslant X_{sig} \leqslant u^*\big] < 1 - \beta^n$ is a sufficient condition for $\mathbb{E}\big[RP(X_{sig}|\theta, \sigma, \beta^n)\big] < 1 - \beta^n$. Thus, in both cases, the sufficient condition for the desired result is the same.

This sufficient condition is shown in two steps. In the first, I show that this inequality holds even in the case where there is no selective publication and all published results are replicated (i.e. when $X \sim N(\Theta, \Sigma^2)$). In the second, I show that this inequality remains true once we allow for selective publication and truncation of the distribution due to conditioning on statistical significance.

Lemma B2 states the first intermediate step. Its implications are of independent interest and discussed in the main text. It shows that even in the optimistic scenario where original estimates are unbiased, there is no selective publication, and all results are published and replicated, that the expected replication probability still falls below intended power.

Lemma B2. *Let published effects be distributed according to $X|\theta, \sigma \sim N(\theta, \sigma^2)$. Suppose $p(t) = 1$ and $r(t) = 1$ for all $t \in \mathbb{R}$. Assume all results are included in the replication rate calculation. Let power in replications is set according to the common power rule with intended power $1 - \beta^n \geqslant 0.8314$. Then $\mathbb{E}\big[RP(X|\theta, \sigma, \beta^n)\big] < 1 - \beta^n$.*

Proof. Recall that $RP(x|\theta, \sigma, \beta^n)$ is strictly concave with respect to x over the interval (l^*, u^*), where $(l^*, u^*) = \big((1 - r^*)\theta, (1 + r^*)\big)$ when $r^* \in (0, 1)$ and $(l^*, u^*) = \big(0, 2\theta\big)$; in both cases, the mid-point of the interval is θ. We have that

$$\mathbb{E}\Big[RP\big(X|\theta, \sigma, \beta^n\big)\Big|l^* \leqslant X \leqslant u^*\Big] = \int_{l^*}^{u^*} RP\big(x|\theta, \sigma, \beta^n\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{l^*}^{u^*}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} < RP\Big(\theta\Big|\theta, \sigma, \beta^n\big)\Big) = 1 - \beta^n \tag{55}$$

where the strict inequality follows from Jensen's inequality and the fact that $\mathbb{E}[X|l^* \leqslant X \leqslant u^*] = \theta$. The final equality is a property of the replication probability function shown in Lemma 1. This is the sufficient condition required for the desired result.

Note that the inequalities in equations (50) (for when $r^* \geqslant 1$) and (53) (for when $r^* \in (0, 1)$) were derived under more general conditions, where the normal distribution may we reweighted

by $p()$ and truncated based on significance. This setting is a special case with no selective publication (i.e. $p(t) = 1$ for all t), and no truncation such that all results are included in the replication rate irrespective of statistical significance. □

The same conclusions hold when we introduce selective publication (which reweights the normal distribution) and condition on statistical significance (which truncates the 'insignificant' regions of the density). Consider three cases. First, suppose that $u^* \leqslant 1.96\sigma$. Then $\mathbb{E}\big(RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\big) = 0 < 1-\beta^n$ because of truncation. Second, suppose that $l^* \geqslant 1.96\sigma$. Then

$$\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] = \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta^n\big)\frac{p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{l^*}^{u^*} p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'}$$

$$\leqslant \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta^n\big)\frac{\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{l^*}^{u^*}\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'} < RP\Big(\theta\Big|\theta,\sigma,\beta^n\Big)\Big) = 1-\beta^n \tag{56}$$

Note that the distribution is invariant to the scale of $p_{sig}()$. Consider first the weak inequality. This follows because $p_{sig}()$ is assumed to be weakly increasing over (l^*, u^*). When it is a constant function over the interval, the equality holds. If $p_{sig}(x/\sigma) > 0$ for some $x \in (l^*, u^*)$ then the function redistributes weight to larger values of x. Since $RP(x|\theta,\sigma,\beta^n)$ is strictly decreasing over positive values of x (Lemma A1.2), placing higher relative weight on lower values implies that the weak inequality becomes strict. As in the proof to Lemma B2, the strict inequality follows from Jensen's inequality, since $RP(x|\theta,\sigma,\beta^n)$ is strictly concave over (l^*, u^*), and the fact that the expected value of X over this interval is equal to the true value θ. The last equality follows from Lemma 1 in the text.

Finally, consider the case where $l^* < 1.96\sigma < u^*$. Then

$$\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta^n\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] = \int_{1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta^n\big)\frac{p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{1.96\sigma}^{u^*} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'}$$

$$= \int_{1.96\sigma}^{2\theta-1.96\sigma} RP\big(x|\theta,\sigma,\beta^n\big)\frac{p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{1.96\sigma}^{u^*} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'} + \int_{2\theta-1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta^n\big)\frac{p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{1.96\sigma}^{u^*} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'}$$

$$= \omega\int_{1.96\sigma}^{2\theta-1.96\sigma} RP\big(x|\theta,\sigma,\beta^n\big)\frac{p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{1.96\sigma}^{2\theta-1.96\sigma} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'} + (1-\omega)\int_{2\theta-1.96\sigma}^{u^*} RP\big(x|\theta,\sigma,\beta^n\big)\frac{p_{sig}\left(\frac{x}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{2\theta-1.96\sigma}^{u^*} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'}$$

$$= \omega \int_{1.96\sigma}^{2\theta-1.96\sigma} RP(x|\theta,\sigma,\beta^n) \frac{\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{1.96\sigma}^{2\theta-1.96\sigma}\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'} + (1-\omega)\int_{2\theta-1.96\sigma}^{u^*} RP(x|\theta,\sigma,\beta^n)\frac{\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx}{\int_{2\theta-1.96\sigma}^{u^*}\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'}$$

$$< \omega RP\Big(\theta\Big|\theta,\sigma,\beta^n)\Big) + (1-\omega).RP\Big(2\theta-1.96\sigma\Big|\theta,\sigma,\beta^n)\Big) < 1-\beta^n \tag{57}$$

with

$$\omega = \frac{\int_{1.96\sigma}^{2\theta-1.96\sigma} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'}{\int_{1.96\sigma}^{u^*} p_{sig}\left(\frac{x'}{\sigma}\right)\frac{1}{\sigma}\phi\left(\frac{x'-\theta}{\sigma}\right)dx'} \tag{58}$$

The second row simply breaks up the integral. The third row rearranges the sum so that the conditional expectation of the replication probability appears in both terms. The third line follows because, as in the previous case, the p_{sig} function redistributes weight to large values of x and hence lower values of $RP(x|\theta,\sigma,\beta^n)$. In the last line, the first term uses the concavity of $RP(x|\theta,\sigma,\beta^n)$ over $(1.96\sigma, 2\theta-1.96\sigma) \subset (l^*, u*)$, Jensen's inequality, and the fact that the expected value of X over this interval is equal to θ. The second term follows because $2\theta-1.96\sigma$ is the maximum value the function can take because $RP(x|\theta,\sigma,\beta^n)$ is strictly decreasing in x over positive values. The final inequality follows because $RP\big(\theta\big|\theta,\sigma,\beta^n)\big) = 1-\beta^n$ (Lemma 1) and $RP\big(2\theta-1.96\sigma\big|\theta,\sigma,\beta^n)\big) < 1-\beta^n$ because $2\theta-1.96\sigma > \theta$ and the function is strictly decreasing over positive values.

This covers all cases, proving the proposition.

C. Selective Publication Above 1.96 and the Replication Rate

Proposition 1 shows that the replication rate does not depend on the probability of publishing insignificant results relative to significant results. However, it may vary with changes in $p_{sig}()$ i.e when the absolute value of the t-ratio is above 1.96. This section presents a simple example showing how the replication rate varies with the relative probability of publishing 'moderately significant' results to 'highly significant' results.

Suppose that the probability function $p()$ is a stepwise function that distinguishes between insignificant findings, moderately statistically significant findings, and highly statistically significant findings. Specifically, let $\kappa > 1.96$ be a value such that $|z| \in (1.96, \kappa)$ is defined as moderately significant and $|z| > \kappa$ as highly significant. Let β_{p1} refer to the constant probability of publishing insignificant findings, and β_{p2} to the constant probability of publishing moderately significant findings. Both these probabilities are defined relative to the probability of publishing a highly significant finding, which we normalize to 1 (since only the ratio of probabilities are identified). The top-left panel of Figure C1 provides an illustration with $\kappa = 3$, $\beta_{p2} = 0.7$ and $\beta_{p1} = 0.2$.

The top-right panel of Figure C1 shows how the replication rate in economics experiments

varies with β_{p2}. The results show that as β_{p2} decreases – that is, as highly significant results become more favoured for publication relative to moderately significant results – the replication rate increases. The size of the changes in the replication rate as we vary β_{p2} can be relatively large, increasing, for example, by more than 10 percentage points when we move from $\beta_{p2} = 1$ to $\beta_{p2} = 0$.

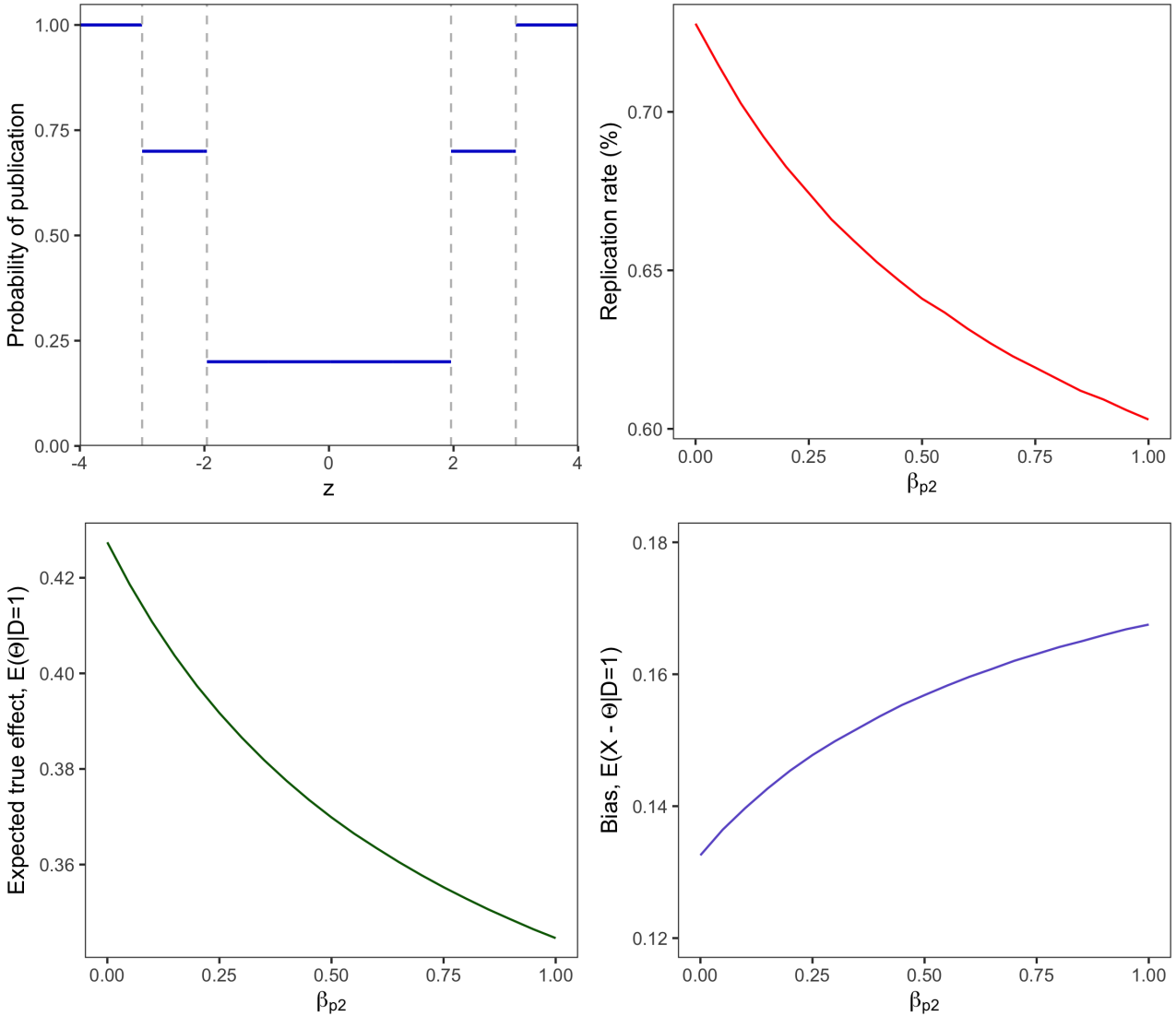


FIGURE C1. The replication rate, true effect, and bias changing the probability of publishing moderately significant results relative to highly significant results (β_{p2}). Results are based on estimated parameters for economics experiments in Table 1, setting $\kappa = 3$ and varying β_{p2}. Power in replication studies is set detect the original estimate with 92% power. The top left panel is an illustration of a stepwise publication probability function which distinguishes between insignificant findings $z \in (-1.96.1.96)$, moderately statistically significant findings $|z| \geqslant 1.96$ and $|z| \leqslant 3$, and highly statistically significant findings $|z| > 3$. The top-right panel shows the replication rate, which is defined as the share of significant results that obtain a statistically significant results with the same sign in replication. The bottom-left panel plots the expected true value of published results and the bottom-right mean bias.

The intuition is that increasing the relative probability of publishing highly significant results compared to moderately significant results (i.e. decreasing β_{p2}) has the effect of increasing the mean true effect in original published studies (bottom-left panel of Figure C1. All else equal, increasing the mean true effect will increase power, and therefore the replication rate. This is because the larger the true effect, the smaller is the bias of the original estimate (bottom-right panel of Figure C1). Power will therefore increase with larger true effects based on the rule. While this may be counter-intuitive from the perspective of selective publication – if we think of favouring highly significant results over moderately significant results as a 'worsening' of selective publication – it is in fact very similar to what Benjamin et al. (2018) propose for setting a new standard of statistical significance for novel findings at $p < 0.005$. McShane et al. (2019) provide arguments against this proposal.

D. Replication Selection in Empirical Applications

Replication selection is a multi-step mechanism that first selects studies, and then selects results within those studies to replicate (since studies typically report multiple results). It consists of three steps:

1. **Eligibility**: define the set of eligible studies (e.g. journals, time-frame, study designs).
2. **Study selection:** on the set of eligible studies, a mechanism that select which studies will be included in the replication study.
3. **Within-study replication selection:** for selected studies, a mechanism for selecting which result(s) to replicate.

These three features of the replication selection mechanism determine: (i) the latent distribution estimated in the empirical exercise; and (ii) the interpretation of the selection parameters ($\beta_{p1}, \beta_{p2}, \beta_{p3}$).

Economics experiments.—Consider these three steps in Camerer et al. (2016):

1. **Eligibility**: Between-study laboratory experiments in *American Economic Review* and *Quarterly Journal of Economics* published between 2011 and 2014.
2. **Study selection:** Camerer et al. (2016) select for publication all eligible studies that had 'at least one significant between subject treatment effect that was referred to as statistically significant in the paper.' Andrews and Kasy (2019) review eligible studies and conclude that no studies were excluded by this restriction. Thus, the complete set of eligible studies was selected for replication.

3. **Within-study replication selection:** the most important *statistically significant* result within a study, as emphasized by the authors, was chosen for replication. Further details are in the supplementary materials in Camerer et al. (2016). Of the 18 replication studies, 16 were significant at the 5% level and two had p-values slightly above 0.05 but were treated as 'positive' results for replication and included in the replication rate calculation.

I assume replication selection is random with respect to the t-ratio for results whose p-values are below or only slightly above 0.05. This implies that β_{p2} measures the relative probability of being published and chosen for replication for a result whose p-value is slightly above 0.05, compared to if it were strictly below 0.05. Overall, the empirical results are valid for the population of 'most important' significant (or 'almost significant') results, as emphasized by authors, in experimental economics papers published in top economics journals between 2011 and 2014.

Psychology.—Next, consider replication selection in Open Science Collaboration (2015):

1. **Eligibility**: Studies published in 2008 in one of the following journals: *Psychological Science*, *Journal of Personality and Social Psychology*, and *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

2. **Study selection:** Open Science Collaboration (2015) write: 'The first replication teams could select from a pool of the first 20 articles from each journal, starting with the first article published in the first 2008 issue. Project coordinators facilitated matching articles with replication teams by interests and expertise until the remaining articles were difficult to match. If there were still interested teams, then another 10 articles from one or more of the three journals were made available from the sampling frame.' Importantly, the most common reason why an article was not matched was due to feasibility constraints (e.g. time, resources, instrumentation, dependence on historical events, or hard-to-access samples).

3. **Within-study replication selection:** the last experiment reported in each article was chosen for replication. Open Science Collaboration (2015) write that, 'Deviations from selecting the last experiment were made occasionally on the basis of feasibility or recommendations of the original authors.' A small number of results had p-values just above 0.05 but were treated as 'positive' results for replication, as in Camerer et al. (2016).

This selection mechanism implies that the empirical results are valid for the distribution of last experiments in the set of eligible journals. Since neither studies nor results were selected

based on statistical significance, it is reasonable to treat the 'last experiment' rule as effectively random. In this case, we can interpret the results are being valid for all results in the eligible set of journals.

Social science experiments.—Finally, consider replication selection in Camerer et al. (2018):

1. **Eligibility**: Experimental studies in the social sciences published in *Nature* or *Science* between 2010 and 2015.

2. **Study selection:** Camerer et al. (2018) include all studies that: '(1) test for an experimental treatment effect between or within subjects, (2) test at least one clear hypothesis with a statistically significant finding, and (3) were performed on students or other accessible subject pools. Twenty-one studies were identified to meet these criteria.'

3. **Within-study replication selection:** Camerer et al. (2018) write, 'We used the following three criteria in descending order to determine which treatment effect to replicate within each of these 21 papers: (a) select the first study reporting a significant treatment effect for papers reporting more than one study, (b) from that study, select the statistically significant result identified in the original study as the most important result among all within- and between-subject treatment comparisons, and (c) if there was more than one equally central result, randomly select one of them for replication.' All results selected for replication had p-values strictly below 0.05.

This selection mechanism implies that the empirical results are valid for the population of statistically significant between- or within-subject treatment comparisons in experimental social science, which were identified by authors as the most 'important' and published in *Nature* or *Science* between 2010 and 2015.

E. Predicted Replication Rates Under Alternative Power Calculations

This appendix presents several extensions to the main empirical results on predicting replication rates in experimental economics, psychology and social science. The first extension allows for variation in the application of the common power rule around mean intended power. Results are similar to those in the main text, which assume no variability in the application of the common power rule. The second extension generates replication rate predictions under the rule of setting replication power equal to original power. This delivers lower replication rates than the common power rule.

Alternative power calculation rules.—Consider first the rule used for calculating replication power in the main text, and then two additional approaches. For concreteness, suppose we want to calculate the replication standard error for a simulated original study $(x^{sim}, \sigma^{sim}, \theta^{sim})$.

1. **Common power rule (mean):** This is the rule reported in the results in the main text. It assumes no variability in the application of the common power rule, such that all replications have mean intended power $1 - \beta^n$. This rule implies

$$\sigma_r^{sim}(x^{sim}, \beta^n) = \frac{|x^{sim}|}{1.96 - \Phi^{-1}(\beta^n)} \tag{59}$$

2. **Common power rule (realized):** Intended power for individual replications varied around mean intended power for at least two reasons. First, replication teams were instructed to meet minimum levels of statistical power, and encouraged to obtain higher power if feasible. Second, a number of replication in Open Science Collaboration (2015) did not meet this requirement. Figure E1 shows the distribution of realized intended power in replications for experimental economics and psychology. Realized intended power is right-skewed for psychology. In experimental economics, realized intended power is distributed more tightly around mean.

   To capture variability in the application of the common power rule, take a random draw from the empirical distribution of $|x|/\sigma_r$ and denote it $1.96 - \widehat{\beta}^n$. Then realized intended power for simulated study $(x^{sim}, \sigma^{sim}, \theta^{sim})$ is equal to

$$\sigma_r^{sim}(x^{sim}, \widehat{\beta}^n) = \frac{|x^{sim}|}{1.96 - \Phi^{-1}(\widehat{\beta}^n)} \tag{60}$$

3. **Original power:** Set replication power equal to the power in the original study:

$$\sigma_r^{sim}(\sigma^{sim}) = \sigma^{sim} \tag{61}$$

   This rule has been proposed as a straightforward, intuitive approach for designing replication studies. In a review of replication studies by Anderson and Maxwell (2017), 19 of 108 studies used this approach.

Results.—Table E1 presents the results for all three applications. Panel *A* shows that allowing intended power to vary across replications ('Realized power') yields similar replication rate prediction to assuming all replications have intended power equal to the report mean ('92% on X'). In fact, in all three applications, the accuracy improves very slightly under the realized

power rule. The biggest differences is in psychology, because the realized power rule accounts for the fact that the distribution of intended power is right skewed.

Panel B examines the proposed rule of setting replication power equal to original power. In all three cases, the expected replication rate is lower than under the common power rule.

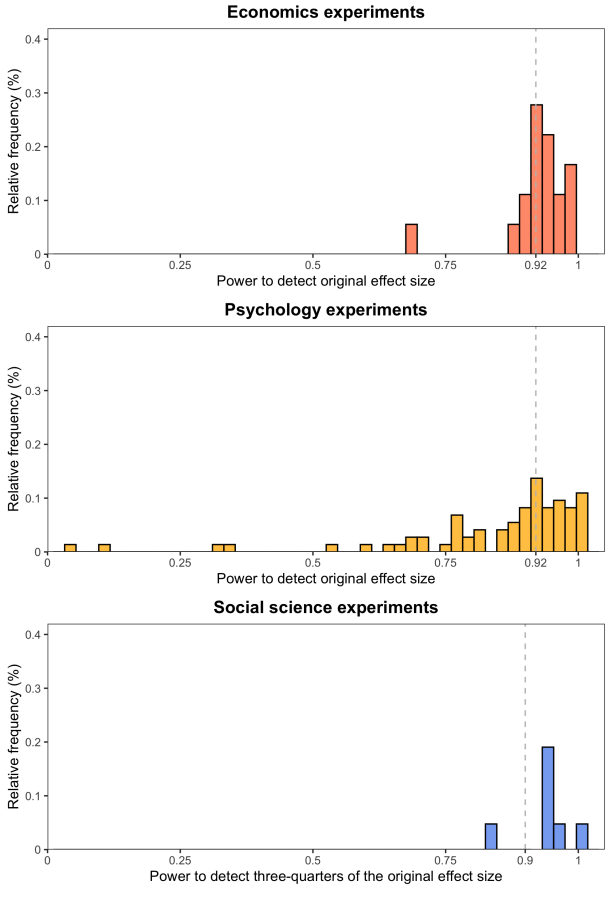


FIGURE E1. Histograms of realized intended power in replication studies in experimental economics, psychology, and social science. Data are from Camerer et al. (2016), Open Science Collaboration (2015), and Camerer et al. (2018), respectively. Realized intended power is defined as $1 - \Phi(1.96 - \psi \cdot \frac{x}{\sigma_r})$ with $\psi = 1$ in economics and psychology and $\psi = 3/4$ in social science. The horizontal dashed line is reported mean power in each application. In economics and psychology, this is 92% to detect the original effect size. In social science, this is 90% to detect three quarters of the effect size.

.

Table E1 – Replication Rate Predictions Under Alternative Replication Power Rules

	Economics	Psychology	Social science
A. Replication rate predictions			
Nominal target (intended power)	0.92	0.92	–
Observed replication rate	0.611	0.348	0.571
92% on X	0.600	0.545	0.553
Realized power	0.615	0.523	0.565
B. Alternative rule			
Same power	0.550	0.486	0.494

Notes: Economics experiments refer to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015), and social science experiments to Camerer et al. (2018). The replication rate is defined as the share of original estimate whose replications have statistically significant findings of the same sign. Figures in the first row are observed outcomes from large-scale replication studies. Remaining rows report predicted replication rates using parameter estimates Table 1 and assuming different rules for calculating replication power.

F. Intuition Behind Empirical Decomposition Results

This Appendix provides intuition behind the empirical decomposition results. For reference, the decomposition of the replication rate gap derived in the main text is restated below. Calculating this decomposition empirically in experimental economics and psychology shows that: (i) failing to account for the non-linearity of the power function explains over 90% of the explained replication rate gap; (ii) attempts to replicate original estimates with the 'wrong' sign account for between 5.7–8% of the gap; and (iii) regression-to-the-mean in replication attempts accounts for a small amount of the replication rate gap in economics and *decreases* the gap in psychology. Below I provide details underlying the intuition behind the empirical results.

$$
\begin{aligned}
&(1-\beta^n) - \mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1\big] \\
&= \underbrace{(1-\beta^n) - \mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|r(t)=1\ \forall t, p(t)=1\ \forall t, X\geqslant 0\big]}_{\text{(i) non-linearity gap}} \\
&+ \underbrace{\mathbb{P}\Big(X<0|S_X=1\Big)\Big(\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1, X\geqslant 0\big] - \mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1, X<0\big]\Big)}_{\text{(ii) 'wrong' sign gap}} \\
&+ \underbrace{\mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|r(t)=1\ \forall t, p(t)=1\ \forall t, X\geqslant 0\big] - \mathbb{E}\big[RP(X,\Theta,\sigma_r(X,\Sigma,\beta^n))\big|S_X=1, X\geqslant 0\big]}_{\text{(iii) regression-to-the-mean gap}} \quad (62)
\end{aligned}
$$

Non-linearity gap.—Figure F1 presents normal simulations showing that the non-linearity gap is largest for standardized true effects $\omega \equiv \theta/\sigma$ which are close to 0, and remains above 0.2 for $\omega \leqslant 1$. It decreases monotonically as the true effect size ω increases and approaches zero in the limit.[21] It follows that the size of the non-linearity gap depends on the distribution of

[21] See Lemma A1.5 in Appendix A for a proof which shows that the non-linearity issue vanishes as true effect

ω. The first row of graphs in Figure F2 plot the distribution of latent studies that have the 'correct' sign (this corresponds to the expression for the 'non-linearity' gap in equation (62)). We see that a high fraction of latent studies have $\omega < 1$, which explains why the non-linearity gap explains such a large role.

Wrong-sign gap.—Random sampling variation means that original estimates will occasionally have the 'wrong' sign. When this occurs, the replication probability is bounded above by 0.025. The extent to which this issue contributes to low replication rates therefore depends on the share of studies that have the wrong sign among significant studies. This share will be higher in settings with small true effects and low statistical power (Gelman and Carlin, 2014; Ioannidis et al., 2017). As power approaches 100%, the 'wrong-sign gap' approaches zero because the probability of drawing an estimate with the 'wrong' sign shrinks to zero.

Table F1 presents figures based on the estimated models, which show that significant results in experimental economics and psychology are relatively low-powered. The share of significant studies with the 'wrong' sign is 3% in economics, and 5% in psychology owing to lower statistical power. As a consequence, the wrong-sign gap is around 1 percentage point higher in psychology compared to economics.

Table F1 – Power and Estimates With the Wrong Sign For Statistically Significant Studies

	Experimental economics	Experimental psychology
Mean normalized true effect	2.835	2.251
Mean power	0.550	0.486
Share with wrong sign	0.030	0.054
Wrong-sign gap	0.022	0.033

Notes: Figures are based on simulated draws from the estimated distribution of latent studies in Table 1. All statistics are calculated on the subset of statistically significant studies. The normalized true effect is defined as θ/σ. Power is defined as the probability of obtaining a statistically significant effect at the 5% level. The wrong-sign gap is defined in (62).

Regression-to-the-mean gap.—The regression-to-the-mean gap is 1% in economics and slightly negative for psychology (i.e. conditioning on statistical significance increases the replication rate compared to when there is no conditioning). The sign of the regression-to-the-mean gap is ambiguous because of two opposing effects from conditioning on statistical significance. To see these two effects, consider the figures in Table F2 which are based on the estimated empirical model. For the first effect, note that conditioning on significant findings increases mean bias in both applications.[22] This makes replication more difficult for any fixed level of ω. For

sizes approach infinity.

[22]Bias is positive for latent studies because these statistics condition on original estimates X^* to have the same sign as true effects.

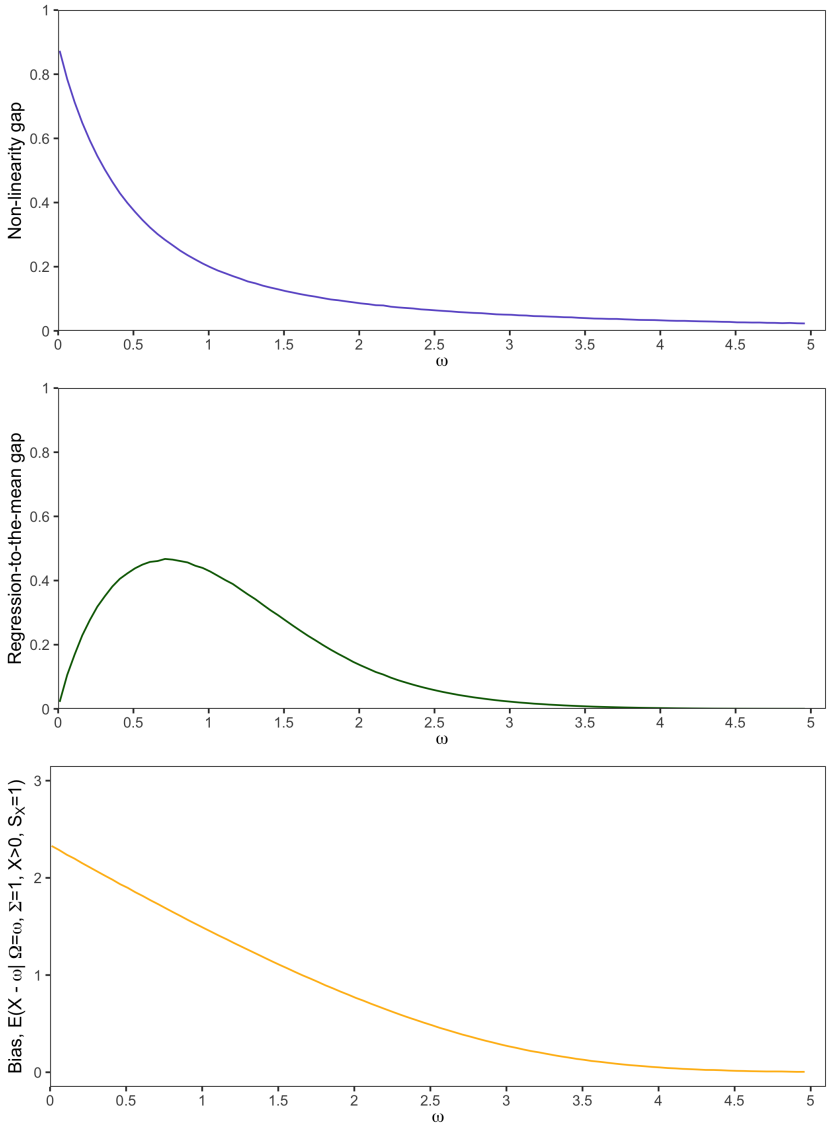


FIGURE F1. REPLICATION RATE GAP DECOMPOSITION: MONTE CARLO SIMULATIONS

Notes: Plots are based on simulating studies from an $N(\omega, 1)$ distribution, for different values of ω. Replication estimates are drawn from a $N(\omega, \sigma_r(x, \beta^n)^2)$, where $\sigma_r(x, \beta^n)$ is set based on the common power rule to detect the original effect x with $1 - \beta^n = 0.92$ intended power. The non-linearity gap and regression-to-the-mean gap are based on equation (62) and calculated using Monte Carlo methods.

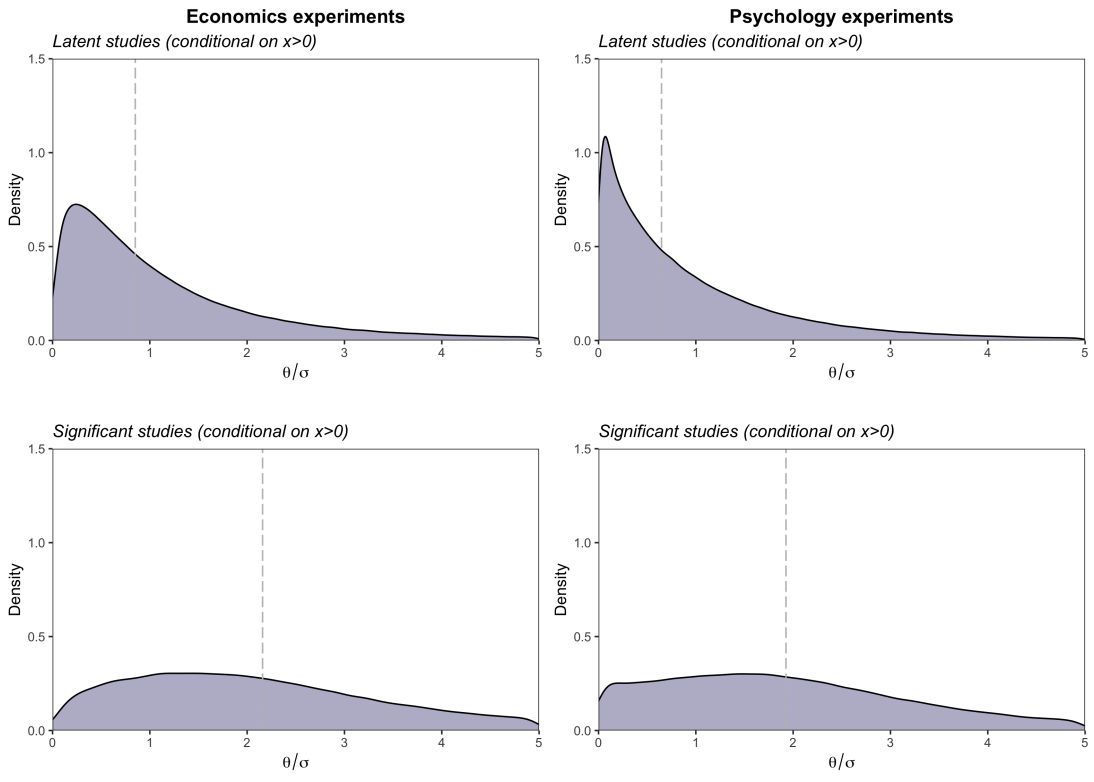


FIGURE F2. DISTRIBUTION OF NORMALIZED TRUE EFFECTS: LATENT STUDIES AND SIGNIFICANT STUDIES

Notes: Economics experiments refers to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). Densities are based on simulated draws from the estimated distribution of latent studies in Table 1. Dashed vertical lines show the median of the distribution.

the second effect, note that conditioning also tends to select studies with larger standardized true effects ω, which have higher replication probabilities.[23] High replication probabilities arise because (i) bias is lower for larger true effects; and (ii) non-linearity effects are more severe for low-powered studies.

The bottom panel in Figure F1 present normal simulations which show that mean bias decreases as the standard effect size increases, and approaches zero in the limit. The intuition is that censoring insignificant original estimates has little 'bite' when the true effect is very large, since the probability of drawing an insignificant estimate is very small. Thus, as true effects become very large, the regression-to-the-mean gap approaches zero because the expected replication probability of statistically significant findings with the 'correct' sign converges to the expected replication probability of latent studies with the 'correct' sign.

[23]The impact of conditioning on the full distribution of ω can be seen in Figure F2.

Table F2 – True Effect Sizes and Bias For Studies with the 'Correct' Sign

	Economics experiments		Psychology experiments	
	Latent	Published & significant	Latent	Published & significant
Mean bias	0.113	0.200	0.091	0.173
Mean standardized true effect	1.415	2.915	1.084	2.367

Notes: Economics experiments refers to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). Figures are based on simulated draws from the estimated distribution of latent studies in Table 1. The mean of the standardized true effect is equal to $\mathbb{E}[\Omega^*|S^*_X, X^* > 0, D]$. Mean Bias is equal to $\mathbb{E}[X^* - \Omega^*|S^*_X, X^* > 0, D]$. 'Latent studies' allow S^*_X and D to be either 0 or 1. 'Published & significant studies' set $S^*_X = 1$ and $D = 1$.

G. Extensions to the Empirical Model

This appendix presents results from two extensions. The first extension incorporates p-hacking and manipulation into the empirical model. In the second extension, I use the model (without p-hacking) to predict relative effect sizes, a continuous measure of the replication that is complementary to the replication rate.

A. Augmented Model with p-Hacking and Manipulation

The augmented model.—p-hacking captures a wide range of researcher behaviors to obtain 'more favorable' p-values e.g. the way data are cleaned, adopting convenient variable definitions, reporting favorable specifications, dropping inconvenient observations etc. I augment the empirical model to incorporate a particularly egregious form of manipulation, where researchers who obtain a marginally insignificant standardized estimates misreport their result as significant with some prespecified probability.

Formally, the augmented model adds one step in the model outlined in the General Theory section. The p-hacking step occurs after the first stage where researchers draw a latent estimate X^*. The model of p-hacking assumes that researchers misreport marginally insignificant results, defined as any $X^* \in (1.46\Sigma^*, 1.96\Sigma^*)$, with probability β_h. Misreported results $\widetilde{X^*}$ are drawn from a uniform distribution over $[1.96\Sigma^*, 2.46\Sigma^*]$. Publication and replication selection depend on the reported estimate $\widetilde{X^*}$. When $\beta_h = 0$ we have the original model. This augmented model allows us to show how the predicted replication rate decreases as we vary β_h from zero (no p-hacking) to one (all marginally insignificant results are p-hacked). Note that the replication rate is defined as the share of *reported* significant results that are replicated with the same sign.

Remarks.—Predictions are based on latent distribution of studies estimated in the standard model with no p-hacking. Thus, the augmented model assumes that the estimated standard

models reflect the true DGP. A close model fit in economics and social science is consistent with no manipulation in these fields. This suggests that it may be reasonable to use estimates of the latent distribution of studies as a basis for the augmented p-hacking model, since the assumed p-hacking step occurs *after* latent studies are drawn. More caution should be applied when interpreting the results in psychology, where the model can only explain two-thirds of the replication rate gap.

Additionally, note that the accompanying code is available. It allows any one interested to test alternative specifications e.g. modifying the ranges over which marginally insignificant results are defined and misreported results are drawn from.

Results.—Figure G1 shows the results for specifications with $\beta_h = 0, 0.1, 0.2, ..., 1$. Across all three applications, the replication rate decreases and the share of misreported results increases from no p-hacking ($\beta_h = 0$) to complete p-hacking ($\beta_h = 1$). Figure G1 shows that the magnitude of the decline in the replication rate is relatively small for most values of β_h when compared to the overall distance between the replication rate and intended power in economics and psychology. If all marginally significant results are manipulated ($\beta_h = 1$), then the share of significant results that are p-hacked is around 30%.

A useful benchmark for what might constitute a realistic value for β_h comes from Brodeur et al. (2016) and Brodeur et al. (2022), who estimate that the proportion of 'wrongfully claimed significant results' is around 10%. In our model, this implies that between 22–26% of significant results are p-hacked (Table G1). Under these specifications, the replication rate falls by between 2-4 percentage points. In economics and psychology, this implies that p-hacking accounts for between 5-6% of the total gap between the predicted replication rate and mean intended power of 92%.

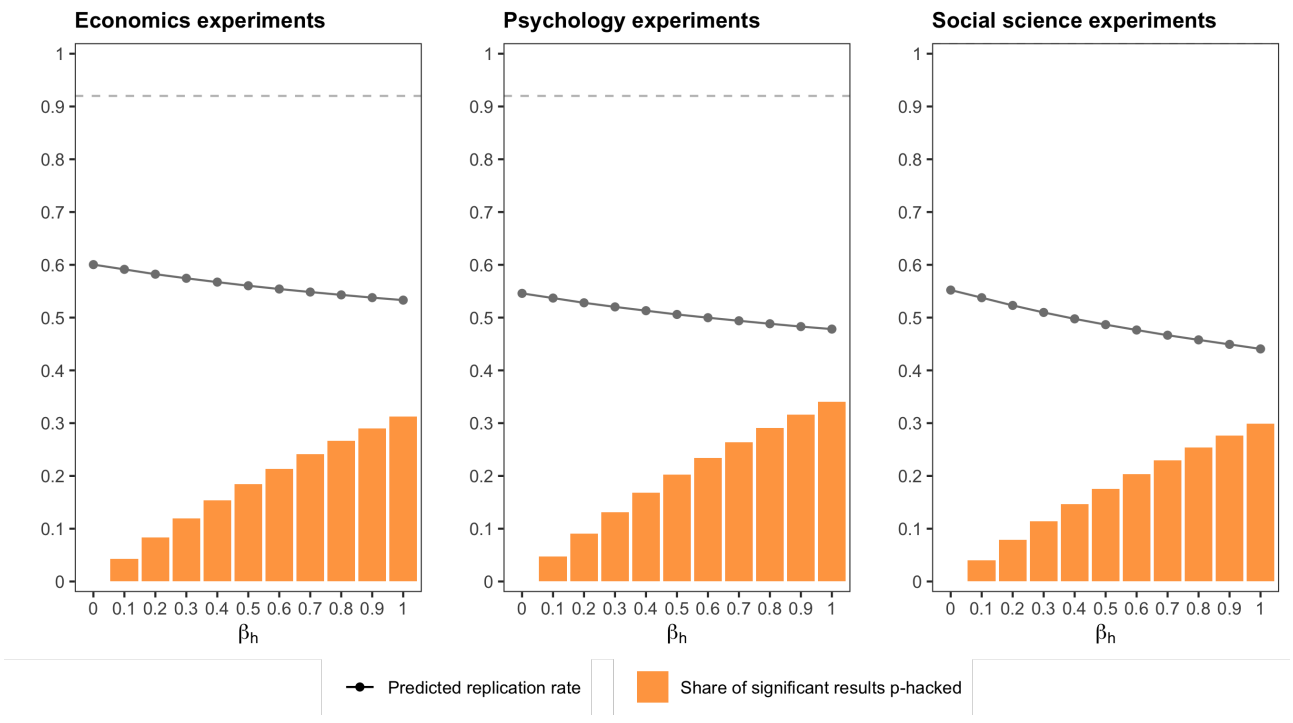


FIGURE G1. PREDICTED REPLICATION RATE WITH *p*-HACKING

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social sciences to Camerer et al. (2018). See text for details on the augmented model. The horizontal dashed line denotes mean intended power in economics and psychology. Replications for social science experiments implemented the fractional power rule to detect three-quarters of original effects with 90% power in its first stage. This rule for setting power does not have a well-defined replication target.

TABLE G1 – SPECIFICATIONS WHERE 10% OF SIGNIFICANT RESULTS ARE *p*-HACKED

	Implied β_h	Replication rate (*p*-hacking)	Replication rate (no *p*-hacking)	Share of gap explained by *p*-hacking
Economics	0.25	0.578	0.600	0.063
Psychology	0.22	0.526	0.545	0.047
Social sciences	0.26	0.515	0.553	–

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social science experiments to Camerer et al. (2018). The implied β_h is the probability of *p*-hacking which is consistent with having 10% of significant results wrongfully claimed as *p*-hacked. The share of the gap explained by *p*-hacking is defined as the difference between the non-*p*-hacked replication rate and the *p*-hacked replication rate divided by the difference between intended power and the *p*-hacked replication rate.

B. Relative Effect Size

The main focus of this article is the binary measure of replication because of its status as the primary replication indicator in the large-scale replication studies.[24] However, complementary measures are frequently presented alongside the replication rate. Perhaps the most common is the relative effect size, a continuous measure of replication defined as the ratio of replication effect size and original effect size. Relative effect sizes typically range between 0.35 and 0.7. Below, I include a brief theoretical discussion of the relative effect size and then present predictions of this measure based on the estimated models.

Theoretical discussion.—The relative effect size measure for individual studies may be informative about biases affecting original studies, especially when original studies are well-powered. However, as an *aggregate* measure of reproducibility, the relative effect size measure may be subject to similar issues to the replication rate, at least in the case where it is defined exclusively over significant findings.

First, if the relative effect size is defined over significant original results, then it will be largely uninformative about the 'file-drawer' problem (Proposition B1).[25] Second, non-random sampling of significant results for replication mechanically induces inflationary bias in original estimates and regression to the mean in replication estimates, such that relative effect sizes are below one in expectation. Thus, similar to the replication rate, it has no natural benchmark against which to judge deviations, making it challenging to interpret. Relatedly, the average relative effect size is also very sensitive to power in original studies, which is unobserved. Figure G2 provides an illustration with intended power set to 0.9, which shows that the expected relative effect size for significant results is increasing in the power of original studies, and approaches one only as statistical power approaches 100%.

[24] Power calculations in replications are themselves typically designed to measure a binary notation of replication 'success' or 'failure'.

[25] Defining it over null results may present its own difficulties. For a perfectly measured null effect, the denominator in the statistic is equal to zero and the statistic is not well defined. On the other hand, if it is close to but not equal to zero, then the statistic is highly sensitive to the precision of replication estimates; this raises questions about how one should set replication power when replicating a null effect.

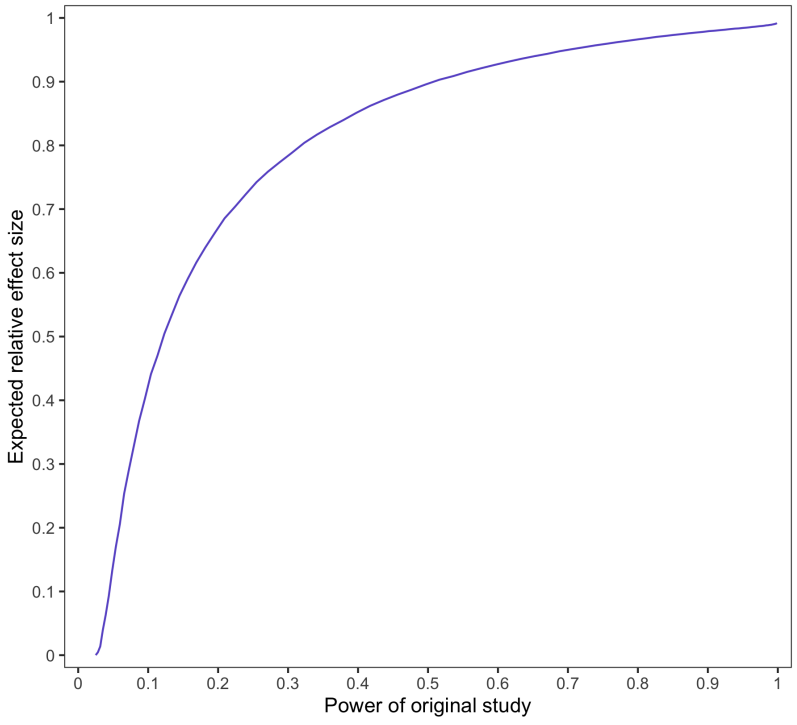


FIGURE G2. EXPECTED RELATIVE EFFECT SIZE OF SIGNIFICANT ORIGINAL STUDIES AND THEIR STATISTICAL POWER

Notes: Illustration for the relationship between original power and the expected relative effect size of significant findings under the common power rule are both functions of $\omega = \theta/\sigma$ (normalized to be positive). Original power to obtain a significant effect with the same sign as the true effect is equal to $1-\Phi(1.96-\omega)$. The expected relative effect size is calculated by taking 10^6 draws of Z from $N(\omega, 1)$ and then calculating $\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \rho_{i,r}^{sig}/\rho_i^{sig}$, where $\rho = \tanh z$ denotes the Pearson correlation coefficient obtained by transforming the Fisher-transformed correlation coefficient (Fisher, 1915); and M_{sig} is the number of significant latent studies. The superscript *sig* reflects the fact that only statistically significant original results at the 5% level and their replications are included in the calculation. Replication estimates $z_{i,r}$ are drawn from an $N(\omega, \sigma_{r,i}(z_i, \beta^n)^2)$ distribution. The replication standard error is calculated using the common power rule to detect original effect sizes with 90% power (i.e. $1-\beta^n = 0.9$), which is given by $\sigma_r(z_i, \beta^n) = |z_i|/[1.96 - \Phi^{-1}(\beta^n)] = |z_i|/3.242$.

Empirical results.—The estimated models in Table 1 can be used to generate predictions of the average relative effect sizes. To procedure for simulating replications is identical as presented in the main text for the replication rate. Let $\{x_i, \sigma_i, x_{r,i}, \sigma_{r,i}\}_{i=1}^{M_{sig}}$ be the set of simulated original studies that are published and significant, and corresponding replication results; M_{sig} is the size of the set. The predicted relative effect size is equal to

$$\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \frac{\rho_{i,r}^{sig}}{\rho_i^{sig}} \tag{63}$$

where $\rho = \tanh z$ denotes the Pearson correlation coefficient which is obtained by transforming the Fisher-transformed correlation coefficient (Fisher, 1915). Results are presented in Table G2.

The predicted average relative effect size is relatively close to observed average relative effect size in economics and social science, although somewhat optimistic in both cases. In psychology, the predicted average relative effect size is very optimistic compared to the observed value.

	Economics	Psychology	Social Sciences
Observed average relative effect size	0.657	0.374	0.443
Predicted average relative effect size	0.703	0.637	0.542

TABLE G2. AVERAGE RELATIVE EFFECT SIZE PREDICTIONS

Notes: Economics experiments refers to Camerer et al. (2016), psychology experiments to Open Science Collaboration (2015) and social science experiments to Camerer et al. (2018). Observed relative effect sizes are based on data from large-scale replication studies. Predicted average relative effect sizes are calculated using equation (63) and the procedure outlined in the text.

H. Extending the Replication Rate Definition

This appendix analyzes a generalization of the replication rate definition that extends to insignificant results. It outlines a number of issues with this proposal.

The Generalized Replication Rate.—Suppose we extend the definition of the replication rate such that insignificant original results are counted as 'successfully replicated' if they are also insignificant in replications. Assume replication selection is a random sample of published results. Then we have the following definitions:

Definition H1 (Generalized replication probability of a single study). *The replication probability of a study* (X, Σ, Θ) *which is published* $(D = 1)$ *and chosen for replication* $(R = 1)$ *is*

$$\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta^n)\Big) = \begin{cases} \mathbb{P}\left(\frac{|X_r|}{\sigma_r(X,\Sigma,\beta^n)} \geqslant 1.96, \operatorname{sign}(X) = \operatorname{sign}(X_r) \middle| X, \Theta, \sigma_r(X, \Sigma, \beta^n)\right) & \text{if } 1.96.\Sigma \leqslant |X| \\ \mathbb{P}\left(\frac{|X_r|}{\sigma_r(X,\Sigma,\beta^n)} < 1.96 \middle| X, \Theta, \sigma_r(X, \Sigma, \beta^n)\right) & \text{if } 1.96.\Sigma > |X| \end{cases} \tag{64}$$

Definition H2 (Expected generalized replication probability). *The expected generalized replication probability equals*

$$\begin{aligned} \mathbb{E}\Big[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta^n)\Big)\Big] &= \mathbb{P}\Big(1.96.\Sigma \leqslant |X|\Big)\mathbb{E}\left[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta^n)\Big) \middle| X, \Theta, \sigma_r(X, \Sigma, \beta^n), 1.96.\Sigma \leqslant |X|\right] \\ &+ \Big(1 - \mathbb{P}\Big(1.96.\Sigma \leqslant |X|\Big)\Big)\mathbb{E}\left[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta^n)\Big) \middle| X, \Theta, \sigma_r(X, \Sigma, \beta^n), 1.96.\Sigma > |X|\right] \end{aligned} \tag{65}$$

First, note that Definition H2 equals the standard replication rate definition when the expectation is taken only over significant studies because, in this case, $\mathbb{P}\big(|X| \geqslant 1.96.\Sigma\big) = 1$. The degree to which the expected generalized replication probability differs from the standard

expected replication probability depends on two factors. First, the share of published results that are insignificant. Second, the expected probability that replications will be insignificant conditional on original estimates being insignificant.[26]

Empirical Results.—To analyze the generalized replication rate, we can apply the empirical approach outlined in the main text, but using the generalized definition in place of the original definition. Recall that the original replication rate is invariant to publication bias against null results. The generalized replication rate, by contrast, does vary as the degree of selective publication against null results changes. Thus, two sets of results are presented for comparison. The first set assumes selective publication using estimated selection parameters in Table 1. The second set assumes no selective publication (i.e. that all results are published with equal probability). We examine two rules for calculating replication power: the common power rule and the original power rule (where the replication standard error is set equal to the original standard error). For more details on different rules for calculation replication power, see Appendix E.

Table H1 reports the results for both applications. Under the common power rule, the simulated generalized replication rate remains below intended power in both publication regimes. Under the original power rule, it is relatively low when there is selective publication and around 80% when there is no selective publication.

These generalized replication rate predictions differs from the standard replication rate predictions for two reasons: (i) the share of insignificant results in the published literature and (ii) the replication probability when results are insignificant, which depends on the power rule used in replication studies. On the first point, moving from the selective publication regime to the no selective publication regime implies a dramatic increase in the share of insignificant published results; in both applications, null results change from a minority of published results to a majority. On the second point, the results show that the replication power rules considered here have some undesirable properties. First, note that the common power rule is designed to detect original estimates with high statistical power. This implies that low-powered, insignificant original results will be high-powered in replications, which increases the probability that they are significant and thus counted as replication 'failures' under the generalized definition. The original power rule has the reverse problem. On the one hand, low-powered, insignificant original studies are likely to be insignificant in replications, which counts as a 'successful' replication under the generalized definition. However, on the other hand, low-powered, significant original studies will have low replication probabilities when the same low-powered design is

[26] Additionally, note that this definition implies that if $\theta = 0$, then $\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta^n) | \Theta = 0\Big) = 0.90375$. That is, the replication probability of null results is constant and independent of power in original studies and replication studies.

repeated in replications. The generalized replication rate therefore depends crucially on the share of significant and insignificant findings in the published literature, and the distribution of standard errors. Under the original power rule with no selective publication, the generalized replication rate is around 80% in both applications; however, with greater power in original studies, the replication rate would fall.

While the generalized replication rate changes as selective publication is reduced, the direction of this change depends on which replication power rule is used: with the original power rule the replication rate increases, while with the common power rule it decreases.

Overall, generalizing the replication rate with Definition H2 does not deliver replication rates close to intended power under the common power rule. For the original power rule, it is higher when there is no selective publication because replications repeat low-power designs for low-powered original studies with insignificant results. The generalized replication rate under this original power rule will therefore be sensitive to the distribution of power in original studies.

Table H1 – Predicted Generalized Replication Rate Results

	Simulated statistics	
A Economics experiments	92% for X	Original power
Selective publication		
Generalized replication rate	0.600	0.555
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.601	0.552
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.542	0.774
$\mathbb{P}(S_X = 1)$	0.988	0.988
$\mathbb{P}(S_X = 0)$	0.016	0.016
No selective publication		
Generalized replication rate	0.436	0.789
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.601	0.551
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.385	0.862
$\mathbb{P}(S_X = 1)$	0.236	0.236
$\mathbb{P}(S_X = 0)$	0.764	0.764
B Psychology experiments		
Selective Publication		
Generalized replication rate	0.541	0.526
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.539	0.478
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.554	0.824
$\mathbb{P}(S_X = 1)$	0.861	0.861
$\mathbb{P}(S_X = 0)$	0.139	0.139
No selective publication		
Generalized replication rate	0.474	0.805
$\mathbb{P}(\text{Replicated}\|S_X = 1)$	0.537	0.479
$\mathbb{P}(\text{Replicated}\|S_X = 0)$	0.460	0.88
$\mathbb{P}(S_X = 1)$	0.188	0.188
$\mathbb{P}(S_X = 0)$	0.812	0.812

Notes: Economics experiments refer to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). The generalized replication rate is defined in the text. The indicator variable S_X equals one for significant results and zero otherwise. Economics experiments refers to Camerer et al. (2016) and psychology experiments to Open Science Collaboration (2015). Simulated statistics are based on parameter estimates in Table 1. Different column represent different rules for calculating power in replications.