# Why Are Replication Rates So Low?

Patrick Vu[†]

**Abstract**

Many explanations have been offered for why replication rates are low in the social sciences, including selective publication, p-hacking, and treatment effect heterogeneity. This article emphasizes that issues with common power calculations in replication studies may also play an important role. Theoretically, I show in a simple model of the publication process that issues with the way that replication power is commonly calculated imply we should always expect replication rates to fall below their intended power targets, even when original studies are unbiased and there is no p-hacking or treatment effect heterogeneity. Empirically, I find that a parsimonious model accounting only for issues with power calculations can fully explain observed replication rates in experimental economics and social science, and two-thirds of the replication gap in psychology. (*JEL* C18, C53, C90)

## 1. Introduction

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 high-profile replication projects which find that the replication rate – i.e. the fraction of replications that are significant with the same sign as the original study – is just 36% in psychology, 61% in experimental economics, and 62% in experimental social science (Open Science Collaboration, 2015; Camerer et al., 2016, 2018).

Understanding the underlying cause of low replication rates is important for researchers and reformers aiming to improve the credibility of published research. There is a large literature examining a wide range of explanations, including selective publication against null results (Franco et al., 2014; Open Science Collaboration, 2015; Camerer et al., 2016, 2018);

[†] *This version:* November 26, 2023. Brown University. Email: 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, University College London and the AIMOS 2022 conference.

$p$-hacking and other questionable research practices (Ioannidis, 2005, 2008; Simonsohn et al., 2014; Brodeur et al., 2016, 2020, 2022; Elliott et al., 2022); and 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).

In this article, the main theoretical result shows that we should expect the replication rate to fall short of its intended target, owing to issues with the common approach of setting power in replications. This is true regardless of whether or not there is selective publication, and even in 'ideal' conditions with no $p$-hacking, no heterogeneity, and relatively high statistical power in original studies (e.g. 80%). Let $RP(x, \sigma_r|\theta)$ be the probability of successfully replicating a study with observed original effect size $x$ and replication standard error $\sigma_r$ conditional on unobserved true effect $\theta$. Replicators commonly set the replication standard error (or equivalently the replication sample size) as a function of the observed effect size $x$, such that $RP(x, \sigma_r(x)|\theta)$ equals a pre-specified intended power target $1-\beta$ when $x = \theta$ (e.g. $1-\beta = 0.9$ would correspond to 90% intended power target, where $\beta$ is the target probability of Type II error). This approach was used, for example, in large-scale replication studies in psychology and economics (Open Science Collaboration, 2015; Camerer et al., 2016), and a survey of replications in the psychology literature by Anderson and Maxwell (2017) shows that it is the most commonly implemented approach. In practice, replication rates consistently fall below the intended power target $1-\beta$, which is commonly interpreted as an indicator that original effects are biased due to factors such as selective publication, $p$-hacking, or treatment effect heterogeneity. However, this article highlights that the replication function $RP(\cdot|\theta)$ is a non-linear, locally concave function. Thus, even if original estimates were unbiased, with $\mathbb{E}_{X|\Theta}[X|\theta] = \theta$, by Jensen's inequality we have that $\mathbb{E}_{X|\Theta}[RP(X, \sigma_r(X)|\theta)|\theta] < RP(\theta, \sigma_r(\theta)|\theta) = 1 - \beta$. That is, stated replication rate targets in large-scale replication studies using the approach described above do not provide an attainable benchmark against which to judge replication rates observed in practice; even if original studies were unbiased, such targets are not in fact reachable in expectation. I also show that the gap between the expected replication rate and its intended power target is larger when the original published studies have low power, a problem that we expect to be severe in practice given evidence of low power in various empirical literatures from (Button et al., 2013; Ioannidis et al., 2017; Stanley et al., 2018; Arel-Bundock et al., 2023).

The main theoretical result applies to studies using what I refer to as the common power rule, which sets replication power to detect the original estimated effect size. More recently, some studies have begun to use a higher-power variant which I refer to as the fractional power rule, wherein replication power is set to detect some fraction of the estimated effect size. Building on results in Andrews and Kasy (2019), I show that the expected replication rate using the fractional power rule can be either above or below the stated power target.

To what extent can these theoretical insights explain the low replication rates actually observed in large-scale replication studies? Although the theory predicts that the actual replication rate will always fall below the target when using the common power rule, the magnitude of this gap is an empirical question. Likewise, for replication studies using the fractional power rule, both the sign and the magnitude of the gap is an empirical question.

To evaluate the importance of power issues in practice, I therefore empirically investigate the results of three replication studies, two of which use the common power rule (Open Science Collaboration, 2015; Camerer et al., 2016) and one of which uses the fractional power rule (Camerer et al., 2018). In each application, I estimate the empirical model in Andrews and Kasy (2019) using a 'metastudy approach' that corrects for publication bias to obtain the underlying distribution of latent studies prior to screening by the publication process.[1] I then use the estimated latent distribution of studies to simulate what we should expect the replication rate to be based on the power calculations actually implemented in replications. Importantly, the model and its predictions are based only on data from original studies and assume away researcher manipulation and heterogeneous treatment effects. The empirical exercise asks, in effect, whether observed replication rates could have been predicted by issues with power alone, before the replication studies themselves were actually undertaken and in a parsimonious model without treatment effect heterogeneity or $p$-hacking.

I find that the predicted replication rate is almost identical to observed replication rates in experimental economics (60% vs. 61%) and experimental social science (54% vs. 57%). Replications in experimental economics implemented the common power rule, while those in experimental social science used a fractional power rule.[2] These empirical results are consistent with the null hypothesis that observed replication rates in these studies are driven entirely by issues with power calculations, rather than other issues such as $p$-hacking or treatment effect heterogeneity. Of course, failure to reject a hypothesis does not mean that it is true, and thus we should not necessarily conclude that these other factors are not present in these settings. Nevertheless, other evidence has also suggested a relatively limited role for $p$-hacking in the context of lab experiments studied here (Brodeur et al., 2016, 2020; Imai et al., 2020).

In psychology, the predicted replication rate is 55%, whereas the observed replication rate is 35%. Since the intended power target was 92%, issues with power calculations explain

[1] It is necessary to model publication bias to estimate the latent distribution of studies. However, for a given latent distribution of studies, the replication rate itself does not depend on the degree to which selective publication suppresses insignificant results (Andrews and Kasy, 2019; Kasy, 2021). This is for the simple reason that the replication rate only includes significant results in its definition. See Section I.B below for additional discussion.

[2] In the experimental social science replications (Camerer et al., 2018), replicators used a fractional power rule in the first stage of replications predicted here, where replication power was set to detect 75% of the original effect size with 90% intended power.

only two-thirds of the gap in psychology. In the case of psychology, we can therefore reject the null that the replication gap is entirely explained by issues with power calculations. This provides strong evidence that some other factors are important in psychology. Some possibilities discussed in the literature include heterogeneous treatment effects, $p$-hacking, and differences across subfields.

In an extension, I examine the relative effect size (defined as the mean of the ratio of the replication effect size and the original effect size), a common complementary continuous measure of replication. I generate relative effect size predictions in each field using a similar method as for the replication rate. I once again find that the predictions are quite similar to observed outcomes in economics (0.70 vs. 0.66). The model is somewhat farther off for social sciences (0.53 vs. 0.44), perhaps suggesting some role for other factors, although the difference is not statistically distinguishable from zero. In psychology, predictions are quite far off (0.64 vs. 0.37), again providing strong evidence for alternative factors.

This article contributes to the large metascience literature and the growing literature on predicting research outcomes (Ioannidis, 2005; Franco et al., 2014; Gelman and Carlin, 2014; Dreber et al., 2015; Maxwell et al., 2015; Anderson and Maxwell, 2017; Stanley et al., 2018; Miguel and Christensen, 2018; Altmejd et al., 2019; Amrhein et al., 2019; DellaVigna et al., 2020; Gordon et al., 2020; Frankel and Kasy, 2022; DellaVigna and Linos, 2022; Nosek et al., 2022). 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). Theoretically, this article builds on this observation by establishing that the expected replication rate is bounded above by its nominal target owing to issues with common power calculations in replication studies. This result holds for any distribution of latent studies. Empirically, I provide evidence that among the profusion of explanations for low replication rates, a parsimonious model accounting only for issues with replication power calculations and low power in original studies can adequately account for observed replication rates in experimental economics and social science.

## 2. Theory

### *2.1. Model of Large-Scale Replication Studies*

I consider the model in Andrews and Kasy (2019). Suppose a large-scale replication study is conducted in an empirical literature of interest and we observe the estimated effect sizes and standard errors for original studies and their replications. Let upper case letters denote random variables, lower case letters realizations. Latent studies (published or unpublished)

have a superscript * and published studies have no superscript. The model of the DGP has five steps:

1. **Draw a population parameter and standard error:** Draw a research question with population parameter ($\Theta^*$) and standard error ($\Sigma^*$):

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

where $\mu_{\Theta,\Sigma}$ is the joint distribution of latent true effects and latent standard errors.

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 $p(\cdot)$, 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:

$$\mathbb{P}(D = 1|X^*/\Sigma^*) = p\left(\frac{X^*}{\Sigma^*}\right) \tag{1}$$

4. **Replication selection:** Replications are sampled from published studies $(X, \Sigma, \Theta)$ (i.e. latent studies $(X^*, \Sigma^*, \Theta^*)$ conditional on publication $(D = 1)$). Replication selection is modelled by the function $r(\cdot)$, 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:

$$\mathbb{P}(R = 1|X/\Sigma) = r\left(\frac{X}{\Sigma}\right) \tag{2}$$

5. **Replication:** A replication draw is made with:

$$X_r|\Theta, X, \Sigma, \Sigma_r, D = 1, R = 1 \sim N\left(\Theta, \Sigma_r^2\right)$$

We observe i.i.d draws of $\left(X, \Sigma, X_r, \Sigma_r\right)$ from the conditional distribution of $\left(X^*, \Sigma^*, X_r, \Sigma_r\right)$ given $D = 1$ and $R = 1$. I consider what happens in the Andrews and Kasy (2019) model outlined above when the replication standard error, $\Sigma_r$, is set to detect the original estimate $X$ with a pre-specified power level $1 - \beta$, where $\beta$ is the target probability of Type II error. This approach is implemented, for example, in Open Science Collaboration (2015) and Camerer et al.

(2016), and a survey of replications the psychology literature by Anderson and Maxwell (2017) shows that it is the most commonly implemented approach. I refer to this as the common power rule, which is formalized as follows:

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

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

*This is equivalent to setting the replication sample size to $N \times \left[\frac{\sigma}{|x|}\left(1.96 - \Phi^{-1}(\beta)\right)\right]^2$, where $N$ and $\sigma$ are the original study's sample size and standard deviation, respectively.*

The justification for the common power rule is that the power in any given replication study will equal its intended power target of $1-\beta$ when $x = \theta$.[3] In practice, replication rates consistently fall below this benchmark, which is typically taken as evidence that original estimates are biased because of selective publication or $p$-hacking. While this argument has intuitive appeal, it does not account for the fact that replication power is a non-linear function of the random original estimate $X$; thus, even if $\mathbb{E}[X|\Theta = \theta] = \theta$, the replication probability evaluated at the expectation (which equals the intended target) will not, in general, be equal to the expected replication rate.

This argument is developed more formally in the following section, under a number of regularity conditions and assumptions imposed on the DGP. First, following Andrews and Kasy (2019), we impose the normalization that true effects are positive:[4]

ASSUMPTION 1 (True effect normalization). *The support of $\Theta$ is a subset of the non-negative real line.*

Second, we impose that the publication probability $p(\cdot)$ is weakly increasing in the $t$-ratio and symmetric around zero:

ASSUMPTION 2 (Publication selection function). *Let $p(t) \neq 0$ for all $|t| \geqslant 1.96$, $p(t)$ be weakly increasing for all $t \geqslant 1.96$, and $p(t) = p(-t)$ for all $t \geqslant 1.96$. Allow $p(\cdot)$ to take any form when $t \in (-1.96, 1.96)$.*

This allows for very general forms of publication bias (or lack thereof). Third, in step 4, which models the replication selection mechanism, we assume that the set of significant results chosen for replication is a random sample from published, significant results:

[3] For a formal statement and proof, see Lemma B1.

[4] 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.

ASSUMPTION 3 (Replication selection function). *For all* $|t| \geqslant 1.96$*, let* $r(t) = c \in (0, 1]$ *and allow* $r(\cdot)$ *to take any form when* $t \in (-1.96, 1.96)$*.*

Finally, note that the article uses three distinct concepts of statistical power. First, power in an original study is defined as the probability of obtaining a statistically significant estimate in the same direction as the true effect.[5] Second, power in a replication study (or the 'replication probability') is defined as the probability of obtaining a significant effect with the same sign as the original study (Definition 2 below), and will depend on the rule for setting replication power (e.g. the common power rule). Finally, the intended power target of a given rule for setting replication power, which we denote by $1 - \beta$.

### *2.2. Common Power Calculations and Low Replication Rates*

This section defines the replication rate and then discusses the main result. First, we define the replication probability of a single study and then use this to define the expected replication rate over multiple studies.

DEFINITION 2 (Replication probability of a single study). *The replication probability of a published study* $(X, \Sigma, \Theta)$ *chosen for replication* $(R = 1)$ *is*

$$RP\Big(X, \Theta, \sigma_r(X, \beta)\Big) = \mathbb{P}\left(\frac{|X_r|}{\sigma_r(X, \Sigma, \beta)} \geqslant 1.96, \text{sign}(X_r) = \text{sign}(X)\Big|X, \Theta, \beta, R = 1\right) \tag{4}$$

This definition captures the dual requirement that the replication estimate is statistically significant and has the same sign as the original study.

DEFINITION 3 (Expected replication rate). *The expected replication probability is defined over published studies* $(X, \Sigma, \Theta)$ *which are chosen for replication* $(R = 1)$ *and statistically significant* $(S_X = 1)$*. It is equal to*

$$\mathbb{E}\Big[RP\big(X, \Theta, \sigma_r(X, \beta)\big)\Big|R = 1, S_X = 1\Big] \tag{5}$$

Substituting the common power rule in Definition 1 for the replication standard error gives the expected replication rate under the common power rule. Note that while insignificant results may be replicated, they are not included in the replication rate in Definition 3, in line with the main definition reported in most large-scale replication studies (Klein et al., 2014;

[5]The arguments made throughout are essentially unchanged if we consider the alternative definition of obtaining a statistically significant estimate irrespective of the sign.

Open Science Collaboration, 2015; Camerer et al., 2016, 2018; Klein et al., 2018).[6] With this, we can state the main theoretical result:

PROPOSITION 1 (The common power rule implies the expected replication rate is below its target.) *Consider the model in I.A. Under assumptions 1, 2, and 3, if replication standard errors are set by the common power rule to detect original estimates with intended power* $1-\beta \geqslant 0.8314$*, then*

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

From a practical perspective, Proposition 1 means that replicators who set the replication sample size to detect original effect sizes should not expect the replication rate to reach its intended target, regardless of whether or not there is selective publication, and even under 'ideal' conditions with 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. That the intended target is not in fact attainable in expectation underscores fundamental difficulties in interpreting replication rate gaps observed in large-scale replication studies.

Figure 1 provides intuition for this result. It plots the replication probability of a single study in Definition 2 as a function of the original effect $X$, for a fixed true effect $\theta$ and assuming that the common power rule is applied with an intended power target of $1-\beta = 0.9$. Denote this conditional replication probability function as $RP\big(X, \sigma_r(X,\beta)\big|\theta\big)$. It is clear that $RP\big(X, \sigma_r(X,\beta)\big|\theta\big)$ is non-linear in $X$, which implies that $\mathbb{E}_{X|\Theta}\big[RP\big(X, \sigma_r(X,\beta)|\theta\big)\big] \neq RP\big(\mathbb{E}_{X|\Theta}[X|\theta], \sigma_r(\mathbb{E}_{X|\Theta}[X|\theta],\beta)|\theta\big)$, even if $X$ is unbiased. If $RP(\cdot|\theta)$ were globally concave, Proposition 1 would immediately follow from Jensen's inequality. However, it is only locally concave around the true effect $\theta$. The proof of Proposition 1 shows that when $1-\beta > 0.8314$, local concavity is sufficient to arrive at the same result for any distribution of latent studies.

The difference between the expected replication rate and its intended target is larger when power in original studies is low. This is because the concavity of $RP(\cdot|\theta)$ is more pronounced when power in original studies is low. As an illustration, Figure 2 plots the relationship between the expected replication rate and power in original studies, again assuming the intended power target in replications is set to 90%, close to mean reported intended replication power in Open

[6] Replication power calculations themselves are typically designed with this definition in mind. Complementary replication measures include: the relative effect size; whether the 95% confidence interval of the replication covers the original estimate; 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).

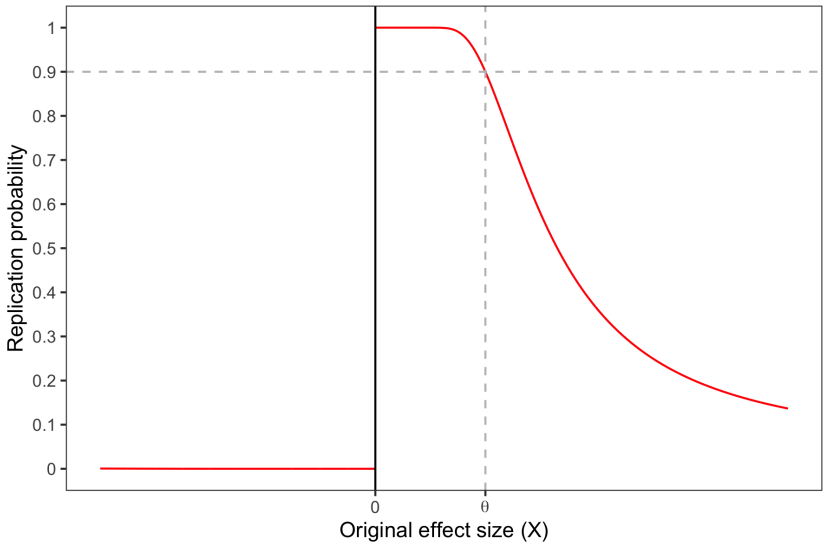


FIGURE 1. REPLICATION PROBABILITY FUNCTION CONDITIONAL ON Θ

*Notes:* Replication probability function in Definition 2 conditional on a fixed $\theta$. The replication standard error is calculated using the common power rule in Definition 1 to detect original effect sizes with 90% power (i.e. $\sigma_r(X, \beta) = |X|/3.242$).

Science Collaboration (2015) and Camerer et al. (2016). To highlight the impact of power in original studies, the relationship is derived assuming no $p$-hacking, no selective publication, and no heterogeneity (i.e. assuming exact replications). The plot shows that the expected replication rate is bounded above by its intended target of 90%, in line with Proposition 1, and is especially low when power in original studies is low. For instance, the expected probability of replicating an original study with 33% power is around 50%. With relatively low estimates of power across various empirical literatures, this provides strong theoretical grounds for expecting low replication rates in practice, even in the absence of issues with $p$-hacking or treatment effect heterogeneity. For intuition, note that if the true effect is zero, the replication probability is 0.025 (regardless of the how the replication standard error is chosen). Continuity implies that when original studies have true effects close to zero (and therefore power in original studies is low), replication probabilities will also be very low.

Two other factors affect the replication rate, although empirically their impact turns out to be relatively small. First, as can be seen in Figure 1, when original estimates are significant but with the 'wrong' sign, the probability of replication is very low (below 0.025) because it requires the highly unlikely event that the replication estimate also has the wrong sign and is statistically significant. Second, the replication rate induces upward bias in original estimates because it is, by definition, calculated on a selected sample of significant findings. Replication estimates will

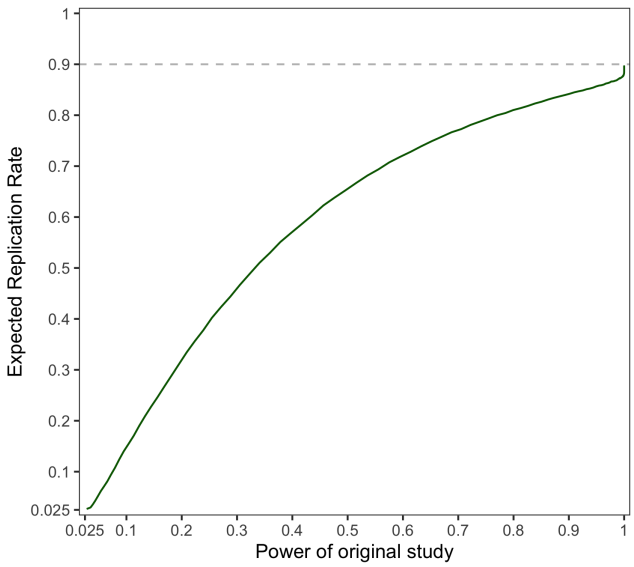


FIGURE 2. ORIGINAL POWER AND THE EXPECTED REPLICATION RATE UNDER THE COMMON POWER RULE

*Notes:* Power of original studies and the expected replication rate under the common power rule are both functions of $\omega = \theta/\sigma$ (normalized to be positive). Power in original studies 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 $10^{-6}\sum_{i=1}^{10^6}\left[1-\Phi\left(1.96-\text{sign}(z_i)\frac{\omega}{\sigma_r(z_i,\beta)}\right)\right]$, with intended power equal to $1-\beta = 0.9$ and depicted by the horizontal dashed line. The replication standard error is calculated using the (normalized )common power rule in Definition 1 to detect original effect sizes with 90% power, which is given by $\sigma_r(z_i,\beta) = |z_i|/3.242$. This figure assumes no $p$-hacking, no heterogeneity in true effects, no selective publication and random replication selection. Further details are provided in Section 2.

regress to the mean (Galton, 1886; Hotelling, 1933; Barnett et al., 2004; Kahneman, 2011)[7], although the ultimate impact on the replication rate is ambiguous because conditioning on significance also tends to select larger true effects, which have higher replication probabilities. Appendix C derives and estimates a decomposition of the replication rate gap in economics and psychology, using the empirical methodology described in the next section, and finds that it is almost entirely explained by the concavity of $RP(\cdot)$.

Proposition 1 applies to replications implementing the common power rule. Some more recent replication studies have used a higher-power variant which I refer to as the fractional power rule, wherein replication power is set to detect some fraction $\psi$ of the estimated effect size (Camerer et al., 2018, 2022). In Proposition B2 in Appendix B, I show that the expected replication rate under the fractional power rule can be either above or below the stated power target $1-\beta$. More specifically, the expected replication rate can range anywhere between 0.025 and $\Phi[1.96-\frac{1}{\psi}(1.96-\Phi^{-1}(\beta))] > 1-\beta$ depending on the statistical power of original studies. For instance, if $\psi = \frac{3}{4}$ and $1-\beta = 0.9$, as in the first-stage in Camerer et al. (2018), then the

[7]For a formal statement and proof, see Proposition B1 in Appendix B.

expected replication rate could range anywhere between 0.025 and 0.99. These results build on those in Andrews and Kasy (2019), who argue that replication rates may vary widely depending on the latent distribution of studies. Finally, note that as with Proposition 1, these conclusions hold whether or not there is selective publication, and even in the absence of $p$-hacking or treatment effect heterogeneity.

Finally, a common perception is that selective publication favouring significant results – either by authors or journals – produces more 'false-positives' in the published literature, which are in turn harder to replicate. This theory is important to address because it enjoys substantial support: over 90% of researchers cite 'selective reporting' as a contributing factor to irreproducibility, more than any other factor (Baker, 2016). However, Andrews and Kasy (2019) and Kasy (2021) show that the replication rate in fact tells us very little about selective publication. Both provide examples showing that the replication rate can take on almost any value depending on the latent distribution of true effects, irrespective of how selective publication is. In fact, the replication rate in the Andrews and Kasy (2019) model is completely insensitive to selective publication against null results.[8] This follows from the simple fact that the replication rate definition does not include statistically insignificant results. Thus, even if insignificant results were being widely published, they would not be included in the replication rate.[9,10]

## 3. Empirical Applications

In this section, I test the null hypothesis that observed replication rates can be entirely explained by issues with common power calculations emphasized in Proposition 1, rather than other issues such as $p$-hacking or heterogeneity. To test this hypothesis, the theory requires that we estimate the latent distribution of studies. This can then be used to generate replication rate predictions which can be compared to observed replication rates. The procedure is as follows:

1. Estimate the latent distribution of studies, $\mu_{\Theta,\Sigma}$ using an augmented version of the Andrews and Kasy (2019) model applied to three large-scale replications.[11] Estimation does

[8] For a formal statement, see Proposition B3 in Appendix B, which proves this more generally for measures $g(\cdot)$ that condition on statistical significance. Setting $g(x,\sigma,x_r,\beta) = \mathbb{1}\Big[\frac{|x_r|}{\sigma_r(x,\sigma,\beta)} \geqslant 1.96, \text{sign}(x_r) = \text{sign}(x)\Big]$ gives the result for the replication rate measure.

[9] A caveat is that the model assumes a fixed distribution of latent studies, whereas in practice it may be endogenous, for example, if researchers engage in more specification searching when publication bias against null results is high (Simonsohn et al., 2014; Brodeur et al., 2016, 2020, 2022).

[10] Appendix D examines measures of replication which may be more sensitive to changes in selective publication than the replication rate. For evaluating efforts to reduce selective publication, simulation results show that the prediction interval approach (Patil et al., 2016), when calculated over both significant and insignificant results, may provide a useful alternative to the replication rate, the confidence interval measure, and the meta-analysis approach.

[11] Note that estimating the latent distribution of studies requires modelling selective publication, as discussed

not use any data from replications.

2. Use the estimated model to simulate replications and predict what fraction of significant results would replicate, absent any other issues such as $p$-hacking or heterogeneity.

3. Compare these predictions (which do not use any data from the replications) to actual replication outcomes.

### *3.1. 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 used the common power rule to detect original effects with at least 90% power. In Open Science Collaboration (2015), replication teams were instructed to achieve at least 80% power using the common power rule, 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 fractional power rule consisting of two stages. In the first stage, replicators aimed 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.[12]

Note that the theoretical result in Proposition 1 showing that the expected replication rate is bounded above by its intended target applies to the common power rule and not to the fractional power rule. For the fractional power rule, the expected replication rate can either above or below the stated power target. In both cases, the magnitude of the gap is an empirical question.

---

in the model in Section I.A. However, with estimates of the latent distribution in hand, replication rate predictions in step 2 will not depend on the degree to which null results are suppressed, since the replication rate is defined only over significant results.

[12] 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.

### *3.2. 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 $\Theta^*$, but not of standard errors $\Sigma^*$. Since predictions of the replication rate also require knowledge of the distribution of $\Sigma^*$, I augment the model to estimate the joint distribution of $(\Theta^*, \Sigma^*)$. 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 assume that $\Sigma^*$ follows a gamma distribution with shape and scale parameters denoted by $\kappa_\sigma$ and $\lambda_\sigma$, respectively.

Table 1 – Maximum Likelihood Estimates

| | Latent true effects $\Theta^*$ | | Latent standard errors $\Sigma^*$ | | Selection parameters | | |
|---|---|---|---|---|---|---|---|
| | $\kappa_\theta$ | $\lambda_\theta$ | $\kappa_\sigma$ | $\lambda_\sigma$ | $\beta_{p1}$ | $\beta_{p2}$ | $\beta_{p3}$ |
| *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 | – |
| | (0.423) | (0.055) | (0.605) | (0.008) | (0.007) | (0.134) | – |
| Andrews and Kasy (2019) | 0.734 | 0.185 | – | – | 0.012 | 0.300 | – |
| | (0.405) | (0.056) | – | – | (0.007) | (0.134) | – |
| *Social science experiments* | | | | | | | |
| Augmented model | 0.077 | 0.644 | 6.249 | 0.028 | 0.000 | 0.000 | 0.611 |
| | (0.106) | (0.333) | (1.762) | (0.009) | (0.000) | (0.000) | (0.427) |
| | (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 $(\kappa, \lambda)$ 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. Parameters $\beta_{p1}$, $\beta_{p2}$ in this case are the relative publication probabilities of studies that are insignificant at the 10% level; and significant at the 10% level but not at the 5% level. For example, in experimental economics, an estimate of $\beta_{p2} = 0.039$ implies that results which are significant at the 5% level are about 26 times more likely to be published and chosen for replication than results that are significant at the 5% level. Note that in economics, results which were insignificant at thew 10% level were not selected for replication and hence $\beta_{p1} = 0$. In social sciences, the omitted category is studies significant at the 1% level. Results below the 5% significance level were not chosen for replication so that $\beta_{p1} = \beta_{p2} = 0$, and $\beta_{p3}$ measures the publication probability of a result that is significant at the 5% level but not at the 1% level, relative to that of a a significant result at the 1% level. Andrews and Kasy (2019) estimates are reproduced from accessible data and code from their analysis.

For all other aspects of the model, I implement identical model specifications as Andrews and Kasy (2019), whose focus is on estimating publication bias. Matching their specifications, I assume that $|\Theta^*|$ follows a gamma distribution with shape and scale parameters $(\kappa_\theta, \lambda_\theta)$; and that the joint probability of being published and chosen for replication, $p(X/\Sigma) \times r(X/\Sigma)$, 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.[13] Table 1 presents the maximum likelihood estimates together with reproduced estimates from Andrews and Kasy (2019) for comparison.[14] For common parameters, estimates are very close.

### *3.3. The Predicted Replication Rate*

Model parameters estimates in Table 1 can be used to generate replication rate predictions by simulating replications using the following procedure:

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 $\hat{\beta}_{\mathbf{p}}$ to determine the subset of studies that are published and chosen for replication.

4. For studies chosen for replication, calculate the replication standard error $\sigma_r^{sim}$ according to the following rule

$$\sigma_r^{sim}(x^{sim}, \beta, \psi) = \frac{\psi \cdot |x^{sim}|}{1.96 - \Phi^{-1}(\beta)} \tag{7}$$

   where $\psi = 1$ and $1 - \beta = 0.92$ in economics and psychology, which corresponds to the common power rule; and $\psi = \frac{3}{4}$ and $1 - \beta = 0.9$ in social science experiments, which corresponds to a fractional power rule.[15]

[13] Details on mechanisms for replication selection are outlined in Appendix E. 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 $r()$.

[14] 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 with the corrected data. Excluding this study in the augmented model leads to very similar replication rate predictions.

[15] This assumes 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 F reports

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

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.[16] $M_{sig}$ is the number of replicated originally-significant studies. The predicted replication rate is equal to

$$\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \mathbb{1}\Big(|x_{r,i}| \geqslant 1.96\sigma_{r,i}, \text{sign}(x_{r,i}) = \text{sign}(x_i)\Big) \tag{8}$$

### *3.4. Results*

In experimental economics, the predicted replication rate is 60%, which is very close to the observed rate of 61.1% (Table 2). This is an "out-of-sample" prediction in the sense that the model is estimated only using information from the original studies, and does not incorporate any information from the replications. The accuracy of this prediction is consistent with the null hypothesis that the observed replication rate in economics can be explained entirely by a parsimonious model accounting only for issues with power calculations, and not other issues such as *p*-hacking or treatment effect heterogeneity. Failure to reject the null hypothesis does not, of course, imply that it is true, and thus we should not necessarily conclude that these other factors are not present. Nonetheless, other evidence points to a relatively limited role for *p*-hacking in the context of lab experiments studied here, perhaps due to fewer researcher degrees of freedom as compared with observational settings (Brodeur et al., 2016, 2020; Imai et al., 2020). Note that despite the very accurate point estimate, standard errors are relatively large, which implies limited power to reject the model's prediction (perhaps owing to the fact that there are only 18 replicated studies).

In psychology, 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 accounts for around two-thirds of the replication rate gap, and we can reject the null hypothesis that the replication gap is entirely explained by issues with common power calculations. The unexplained portion of the gap in psychology provides evidence that 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

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; 52.2% in psychology; and 55.5% in social science).

[16] 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. Predictions are almost identical with a strict 0.05 significance threshold.

the model should account for differences in replicating main effects and interaction effects, and differences across subfields (Open Science Collaboration, 2015; Altmejd et al., 2019).

A popular variant for the common power rule 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)). Theoretically, under the specific rule applied in Camerer et al. (2018), the expected replication rate can range anywhere between 0.025 and 0.99 depending on the power in original studies.[17] Empirically, the predicted replication rate for the experimental social sciences is 54.3%, which is very close to the observed rate of 57.1%. The difference is statistically indistinguishable from zero, although the standard error of the prediction is quite large. Similarly to experimental economics, the accuracy of the point estimate of the prediction implies that we cannot reject the null hypothesis that the observed replication rate can be explained by a parsimonious model accounting only for issues with power calculations.

| | Economics experiments | Psychology | Social sciences |
|---|---|---|---|
| Nominal target (intended power) | 0.92 | 0.92 | – |
| Observed replication rate | 0.611 | 0.348 | 0.571 |
| Predicted replication rate | 0.600 | 0.545 | 0.543 |
| | (0.122) | (0.054) | (0.134) |

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 (8) calculated using parameter estimates Table 1. The fourth row shows standard errors for the predicted replication rate which are calculated using the delta method. 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.

### *3.4.1. Extensions*

I examine three extensions. In Appendix G, I use the empirical models estimated in Table 1 to generate predicted average relative effect sizes, using a similar procedure to the replication rate predictions. I find that the predicted relative effect size is quite similar to the observed value in economics (0.70 vs. 0.66). In the social sciences, the model is somewhat farther off (0.53 vs. 0.44), which may suggest a role for other factors, although the difference is not statistically distinguishable from zero. Finally, in psychology, the prediction is quite far off

[17] Proposition B2 shows that the expected replication rate can range between 0.025 and $1-\Phi[1.96-\frac{1}{\psi}(1.96-\Phi^{-1}(\beta))]$. With the fraction of original effect size to detect equal to $\psi = 3/4$, and intended power set to $1-\beta = 0.9$, the upper range equals 0.99.

(0.64 vs. 0.37), again providing strong evidence for alternative factors. Note that relative effect sizes are affected both by selection of significant results for replication and the level of statistical power in original studies.[18]

A second extension considers the proposed rule of setting replication power equal to original power in Appendix F. 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 issues that stem from conditioning on statistical significance, the third extension in Appendix H examines the suggestion of extending the replication rate definition 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.

## 4. Conclusion

The prominence of the replication rate stems in part from its apparent transparency and ease of interpretation. However, caution should be applied when interpreting the replication rate from large-scale replication studies using the common power rule for setting replication power. In general, intended replication targets are not attainable in expectation. Moreover, the replication rate gap will be particularly large when original power is low. Empirical evidence supports the importance of these theoretical insights. In a parsimonious model with neither heterogeneity nor $p$-hacking, predicted replication rates in experimental economics and social science are very close to observed values. This is consistent with the null hypothesis that problems with power calculations alone are sufficient to explain observed replication rates in these fields.

## References

Altmejd, A., A. Dreber, E. Forsell, et al. (2019). Predicting the Replicability of Social Science Lab Experiments. *PLoS ONE 14*(12).

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

[18] Figure G2 in Appendix G shows that the expected relative effect size is an increasing function of power in original studies and approaches one as original power approach 100%.

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

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

Arel-Bundock, V., R. C. Briggs, H. Doucouliagos, et al. (2023). Quantitative Political Science Research is Greatly Underpowered. *OSF Preprint*.

Baker, M. (2016). 1,500 Scientists Lift the Lid on Reproducibility. *Nature 533*, 452–454.

Barnett, A. G., J. C. Van Der Pols, and A. J. Dobson (2004). Regression To The Mean: What It Is and How To Deal with It. *Journal of Business and Psychology 34*(1), 215–220.

Brodeur, A., N. Cook, and A. Heyes (2020). Methods Matter: P-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review 110*(11), 3634–3660.

Brodeur, A., N. Cook, and A. Heyes (2022). 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*.

Brodeur, A., M. Lé, M. Sangnier, and Y. Zylberberg (2016). Star Wars: The Empirics Strike Back. *American Economic Journal: Applied Economics 8*(1), 1–32.

Bryan, C. J., D. S. Yeager, and J. M. O'Brien (2019). Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate. *Proceedings of the National Academy of Sciences of the United States of America 116*(51), 25535–25545.

Button, K. S., J. P. Ioannidis, C. Mokrysz, et al. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience 14*(5), 365–376.

Camerer, C., Y. Chen, A. Dreber, et al. (2022). Mechanical Turk Replication Project.

Camerer, C. F., A. Dreber, E. Forsell, et al. (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science 351*(6280), 1433–1437.

Camerer, C. F., A. Dreber, F. Holzmeister, et al. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour 2*(9), 637–644.

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

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

DellaVigna, S., N. Otis, and E. Vivalt (2020). Forecasting the Results of Experiments: Piloting an Elicitation Strategy. *AEA Papers and Proceedings 110*, 75–79.

Dreber, A., T. Pfeiffer, J. Almenberg, et al. (2015). Using Prediction Markets to Estimate the Reproducibility of Scientific Research. *Proceedings of the National Academy of Sciences of the United States of America 112*(50), 15343–15347.

Elliott, G., N. Kudrin, and K. Wüthrich (2022). Detecting p-Hacking. *Econometrica 90*(2), 887–906.

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

Franco, A., N. Malhotra, and G. Simonovits (2014). Publication Bias in the Social Sciences: Unlocking the File Drawer. *Science 345*(6203), 1502–1505.

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

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

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

Gordon, M., D. Viganola, M. Bishop, et al. (2020). Are Replication Rates the Same Across Academic Fields? Community Forecasts from the DARPA SCORE Programme. *Royal Society Open Science 7*.

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

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

Imai, T., K. Zemlianova, N. Kotecha, et al. (2020). How Common are False Positives in Laboratory Economics Experiments? Evidence from the P-Curve Method. *Working Paper*.

Ioannidis, J. P. (2005). Why Most Published Research Findings Are False. *PLoS Medicine 2*(8).

Ioannidis, J. P. (2008). Why Most Discovered True Associations Are Inflated. *Epidemiology 19*(5), 640–648.

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

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

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

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

Klein, R. A., M. Vianello, F. Hasselman, et al. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science 1*(4), 443–490.

Laird, N. M. and F. Mosteller (1990). Some Statistical Methods for Combining Experimental Results. *International Journal of Technology Assessment in Health Care 6*(1), 5–30.

Maxwell, S. E., M. Y. Lau, and G. S. Howard (2015). Is Psychology Suffering from a Replication Crisis? What Does "Failure to Replicate" Really Mean? . *American Psychologist 70*(6), 487–498.

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

Nosek, B. A., T. E. Hardwicke, H. Moshontz, et al. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology 73*, 719–748.

Open Science Collaboration (2015). Estimating the Reproducibility of Psychological Science. *Science 349*(6251).

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

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

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science 26*(5), 559–69.

Simonsohn, U., L. D. Nelson, and J. P. Simmons (2014). P-Curve: A Key to the File-Drawer. *Journal of Experimental Psychology: General 143*(2), 534–547.

Stanley, T. D., E. C. Carter, and H. Doucouliagos (2018). What Meta-Analyses Reveal About the Replicability of Psychological Research. *Psychological Bulletin 144*(12), 1325–1346.

Vu, P. (2022). Replication data for: Can the Replication Rate Tell Us About Selective Publication? *American Economic Association [publisher], Inter-university Consortium for Political and Social Research [distributor]*.

Wagenmakers, E.-J., J. Verhagen, and A. Ly (2016). How to Quantify the Evidence for the Absence of a Correlation. *Behavior Research Methods 48*(2), 413–26.

## Online Appendix

This appendix contain proofs and supplementary materials for "Can the Replication Rate Tell Us About Selective Publication?" Section A derives properties of the replication probability function. Section B contains proofs for results in the main text, in addition to other theoretical results. Section C presents an illustrative example of how the replication rate can vary with changes in selective publication above the 1.96 significance threshold. Section D details replication selection mechanisms implemented in the three applications. Section E presents extensions of the empirical results using alternative power calculations. Section F builds intuition for the empirical replication rate decomposition results. Section G examine two further extensions to the empirical results: examining the impact of $p$-hacking on the replication rate; and an analysis of the relative effect size measure of replication. Appendix H examines a generalization of the replication rate definition to include insignificant results.

## 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 $1-\beta$ intended power (Definition 3). Recall that the replication probability for original study $(x, \sigma, \theta)$ is equal to

$$RP\big(x, \theta, \sigma_r(x, \sigma, \beta)\big) = \mathbb{P}\left(\frac{|X_r|}{\sigma_r(x, \beta)} \geqslant 1.96, \text{sign}(X_r) = \text{sign}(x)\right) \tag{9}$$

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 $1-\beta = 0.9$ and a fixed value of $\theta$.

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

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

$$RP\big(x, \theta, \sigma_r(x, \sigma, \beta)\big) = 1 - \Phi\left(1.96 - \text{sign}(x)\frac{\theta}{\sigma_r(x, \sigma, \beta)}\right) \tag{10}$$

The remaining properties assume the replication standard error $\sigma_r(x, \beta)$ is set using the

common power rule in Definition 3 with intended power $1-\beta$:

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

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

$$r^*(\beta) = -\big(2+1.96.h(\beta)\big) + \sqrt{\frac{\big(2+1.96.h(\beta)\big)^2 - 4\times\big(1+1.96.h(\beta)-h(\beta)^2\big)}{2}} > 0 \quad (11)$$

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

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

$$\lim_{x\to\infty} RP\big(x,\theta,\sigma_r(x,\beta)\big) = 0.025 \text{ and } \lim_{x\to-\infty} RP\big(x,\theta,\sigma_r(x,\beta)\big) = 0.025 \quad (12)$$

$$\lim_{x\uparrow 0} RP\big(x,\theta,\sigma_r(x,\beta)\big) = 0 \text{ and } \lim_{x\downarrow 0} RP\big(x,\theta,\sigma_r(x,\beta)\big) = 1 \quad (13)$$

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

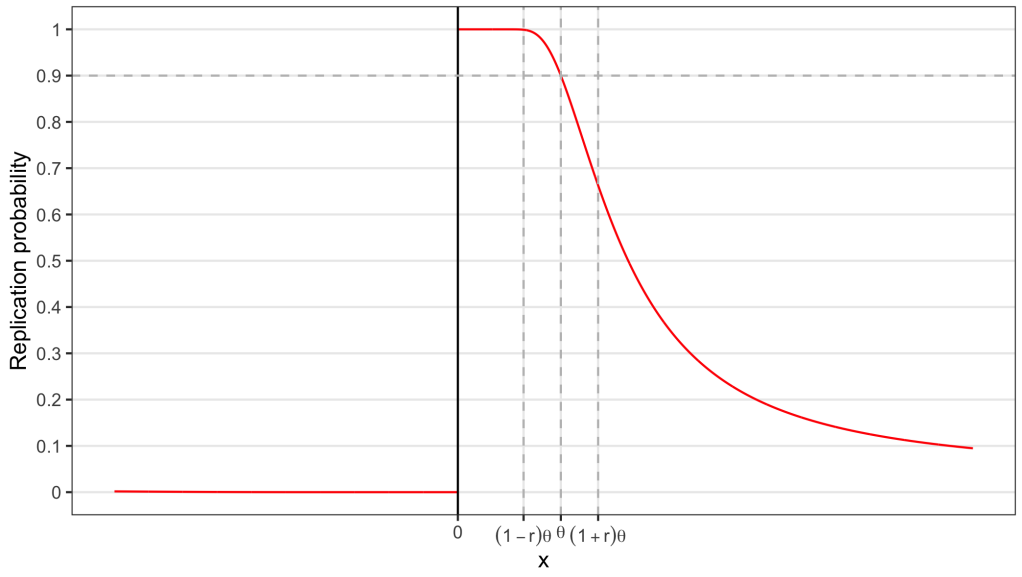


FIGURE A1. Example of the replication probability function under the common power rule with intended power $(1-\beta) = 0.9$. The two vertical lines around $\theta$ marks the open interval over which the replication probability function is strictly concave, where $r^*$ is given by equation (11).

*Proof of 1.*

The probability in equation (9) 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 (10) 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

$$\frac{\partial RP\big(x, \theta, \sigma_r(x, \beta)\big)}{\partial x} = \begin{cases} -\frac{\theta}{x^2}\big(1.96 - \Phi^{-1}(\beta)\big) \times \phi\Big(1.96 - \frac{\theta}{x}\big(1.96 - \Phi^{-1}(\beta)\big)\Big), & x > 0 \\ -\frac{\theta}{x^2}\big(1.96 - \Phi^{-1}(\beta)\big) \times \phi\Big(-1.96 - \frac{\theta}{|x|}\big(1.96 - \Phi^{-1}(\beta)\big)\Big), & x < 0 \end{cases} \tag{14}$$

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

*Proof of 3.*

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

$$\frac{\partial^2 RP\big(x, \theta, \sigma_r(x, \beta)\big)}{\partial x^2} = \left(\frac{h(\beta)\theta}{x^3}\right)\phi\left(1.96 - \frac{h(\beta)\theta}{x}\right)\left[1 + \left(\frac{h(\beta)\theta}{x}\right)\left(1.96 - \frac{h(\beta)\theta}{x}\right)\right] \tag{15}$$

Let $x = (1 + r)\theta$. Substituting this into the previous equation and simplifying shows that equation (15) is strictly negative when the following inequality is satisfied

$$r^2 + \big(2 + 1.96h(\beta)\big).r + \big(1 + 1.96h(\beta) - h(\beta)^2\big) < 0 \tag{16}$$

The solution to the quadratic equation has a unique positive solution $r^*(\beta)$ whenever $(1 - \beta) > 0.6628$. To see this, note that there exists a unique positive solution when $\big(1 + 1.96h(\beta) - h(\beta)^2\big) < 0$. This quadratic equation in $h(\beta)$ must have a unique positive and negative solution in turn, since the parabola opens downwards and equals 1 when $h(\beta) = 0$. 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 $h(\beta) > 2.38014$, we have $\big(1 + 1.96h(\beta) - h(\beta)^2\big) < 0$.

Thus, a unique positive solution to equation (16) exists whenever this condition is satisfied. In particular, a unique positive solution exists whenever

$$\begin{aligned} & h(\beta) = 1.96 - \Phi^{-1}(\beta) > 2.38014 \\ \iff & \Phi(1.96 - 2.38014) > \beta \\ \iff & (1 - \beta) > 0.6628 \end{aligned} \tag{17}$$

The unique positive solution for equation (16) can again be obtained by the quadratic formula, which gives equation (11). Note that for any $r > 0$ where the inequality for concavity in equation (16) 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 $(\max\{0, [1 - r^*(\beta)]\theta\}, [1 + r^*(\beta)]\theta)$, 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

$$RP\big(x, \theta, \sigma_r(x, \beta)\big) = 1 - \Phi\left(1.96 - \frac{\theta}{x}\big(1.96 - \Phi^{-1}(\beta)\big)\right) \tag{18}$$

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 $\theta$, as $\theta \to \infty$ for fixed $\sigma$. 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

$$\left.\frac{\partial^2 RP\big(x, \theta, \sigma_r(x, \beta)\big)}{\partial x^2}\right|_{x = c\theta} = \left(\frac{h(\beta)}{c^3\theta^2}\right)\phi\left(1.96 - \frac{h(\beta)}{c}\right)\left[1 + \left(\frac{h(\beta)}{c}\right)\left(1.96 - \frac{h(\beta)}{c}\right)\right] \tag{19}$$

This approaches zero as $\theta \to \infty$, which implies that $RP\big(x, \theta, \sigma_r(x, \beta)\big)$ 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 $\sigma$, we have that

$$\mathbb{P}\big[X^* \in \big(r\theta, (1+r)\theta\big)|\theta, \sigma\big] = \Phi\left(\frac{(1+r)\theta - \theta}{\sigma}\right) - \Phi\left(\frac{r\theta - \theta}{\sigma}\right) \to 1 \tag{20}$$

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 $\sigma$. This implies in the limit that $\mathbb{E}\big[RP\big(X, \theta, \sigma_r(X, \beta)\big)\big] = RP\big(\mathbb{E}[X], \theta, \sigma_r(X, \beta)\big) = RP\big(\theta, \theta, \sigma_r(X, \beta)\big) = 1 - \beta$, as shown in Lemma 1 in the main text.

## B. Proofs of Propositions

For convenience, some proofs use notation distinguishing the publication probability function $p(\cdot)$ over significant and insignificant regions:

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

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

**Lemma B1** (Justification of the common power rule). *Consider a published study $(x, \sigma, \theta)$. If $x = \theta$ and a replication uses the common power rule to detect the original effect with intended power $1 - \beta$, then*

$$RP\Big(\theta, \theta, \sigma_r(\theta, \beta)\Big) = 1 - \beta \tag{21}$$

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

$$RP\big(\theta, \theta, \sigma_r(\theta, \beta)\big) = 1 - \Phi\left(1.96 - \text{sign}(\theta)\frac{\theta}{\sigma_r(\theta, \beta)}\right) = 1 - \Phi\left(1.96 - \frac{\theta}{\theta}\big(1.96 - \Phi^{-1}(\beta)\big)\right) = 1 - \beta \tag{22}$$

□

**Proof of Proposition 1:** 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

$$\mathbb{E}_{X^*, \Sigma^*, \Theta^*|D, R, S_X^*}\Big[RP\Big(X^*, \Theta^*, \sigma_r(X^*, \beta)\Big)\Big|D = 1, R = 1, |X^*/\Sigma^*| \geqslant 1.96\Big]$$

$$= \mathbb{E}_{X,\Sigma,\Theta|S_X}\Big[RP\big(X,\Theta,\sigma_r(X,\Sigma,\beta)\big)\big||X/\Sigma| \geqslant 1.96\Big]$$

$$= \mathbb{E}_{X_{sig},\Sigma_{sig},\Theta_{sig}}\Big[RP\Big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta)\Big)\Big]$$

$$= \mathbb{E}_{\Sigma_{sig},\Theta_{sig}}\left[\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\Big[RP\Big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta)\Big)|\Theta_{sig}=\theta,\Sigma_{sig}=\sigma\Big]\right] \tag{23}$$

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 $\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\big[RP\big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta)\big)|\Theta_{sig}=\theta,\Sigma_{sig}=\sigma\big] < 1-\beta$ which implies that the expected replication probability is also less than intended power $1-\beta$. For greater clarity in what follows, let $\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\big]$ be shorthand for $\mathbb{E}_{X_{sig}|\Sigma_{sig},\Theta_{sig}}\big[RP\big(X_{sig},\Theta_{sig},\sigma_r(X_{sig},\beta)\big)|\Theta_{sig}=\theta,\Sigma_{sig}=\sigma\big]$.

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

$$\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\big] = \int\left(1-\Phi\left(1.96-\text{sign}(x)\frac{\theta}{|x|}\big(1.96-\Phi^{-1}(\beta)\big)\right)\right)\frac{p\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}|\geqslant 1.96\big)dx}{\int_{x'}p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)\mathbb{1}\big(|\frac{x}{\sigma}|\geqslant 1.96\big)dx'} \tag{24}$$

where the integrand in equation (24) 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. This density differs from a normal density 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 $(1-\beta) > 0.6628$, then $RP\big(x,|\theta,\sigma,\beta\big)$ is strictly concave over the open interval $(\max\{0,[1-r^*(\beta)]\theta\},[1+r^*(\beta)]\theta)$, where $r^*(\beta)$ is given by equation (11). This Proposition assumes $(1-\beta) > 0.8314$, 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 $\theta$.

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

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

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

$$< \mathbb{P}\Big(X_{sig} < l^*\Big)0.025 + \mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] + \mathbb{P}\Big(X_{sig} > u^*\Big)(1-\beta) \tag{25}$$

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 $RP\big(2\theta|\theta,\sigma,\beta\big)$ 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 $RP\big(2\theta|\theta,\sigma,\beta\big) < RP\big(\theta|\theta,\sigma,\beta\big) = 1-\beta$, where the equality is shown in Lemma 1. From equation (25), we can see that $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta)|l^* \leqslant X_{sig} \leqslant u^*\big] < 1-\beta$ is a sufficient condition for $\mathbb{E}\big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\big] < 1-\beta$.

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

$$\begin{aligned}\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big] &= \mathbb{P}\Big(X_{sig} \leqslant l^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|X_{sig} \leqslant l^*\Big] \\ &+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] + \mathbb{P}\Big(X_{sig} \geqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|X_{sig} \geqslant u^*\Big] \\ &< \mathbb{P}\Big(X_{sig} \leqslant l^*\Big) + \mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] + \mathbb{P}\Big(X_{sig} \geqslant u^*\Big)RP\Big(u^*|\theta,\sigma,\beta\Big)\end{aligned} \tag{26}$$

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, $RP(u^*|\theta,\sigma,\beta)$ 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

$$\begin{aligned}\mathbb{E}\Big[\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big] &< \frac{1}{2}\Big(1-\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\Big)\Big(1+RP\big(u^*|\theta,\sigma,\beta\big)\Big) \\ &+\mathbb{P}\Big(l^* \leqslant X_{sig} \leqslant u^*\Big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big]\end{aligned} \tag{27}$$

This follows because $\mathbb{P}(X_{sig} \leqslant l^*) \leqslant \mathbb{P}(X_{sig} \geqslant u^*)$ and $RP(u^*|\theta,\sigma,\beta) < 1$. 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 B2.** *Suppose $X|\theta,\sigma$ follows the truncated normal pdf in equation (24). 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 $\theta$. 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

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

$$< \frac{1}{C}\int_{2\theta+1.96\sigma}^{\infty} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' + \frac{1}{C}\int_{\max\{1.96\sigma,(1+r^*)\theta\}}^{2\theta+1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' = \mathbb{P}\Big(X_{sig} \geqslant (1+r^*)\theta\Big) \tag{28}$$

Consider the weak inequality. Note that the mid-point between $-1.96\sigma$ and $2\theta + 1.96\sigma$ is $\theta$. 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 $\theta$ 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:

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

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

$$+ \frac{1}{C}\int_{2\theta-1.96\sigma}^{2\theta+1.96\sigma} p_{sig}\Big(\frac{x}{\sigma}\Big)\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx' = \mathbb{P}\Big(X_{sig} \geqslant (1+r^*)\theta\Big) \tag{29}$$

□

The inequality in equation (27) can be further simplified by placing restrictions on intended power. In particular, if intended power satisfies $1-\beta \geqslant 0.8314$, then

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

$$+ \mathbb{P}\big(l^* \leqslant X_{sig} \leqslant u^*\big)\mathbb{E}\Big[RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\Big] \tag{30}$$

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

$$\frac{1}{2}\Big(1 + RP\big(u^*\big|\theta,\sigma,\beta\big)\Big) = \frac{1}{2}\left(1 + \left(1 - \Phi\left(1.96 - \frac{1.96 - \Phi^{-1}(\beta)}{1 + r^*(\beta)}\right)\right)\right)$$
$$\leqslant 1 - \beta \iff 1 - \beta \geqslant 0.8314 \tag{31}$$

From equation (30), we can see that $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta)\big|l^* \leqslant X_{sig} \leqslant u^*\big] < 1-\beta$ is a sufficient condition for $\mathbb{E}\big[RP(X_{sig}|\theta,\sigma,\beta)\big] < 1-\beta$. 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 B3 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 B3.** *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 $1 - \beta \geqslant 0.8314$. Then $\mathbb{E}\big[RP(X|\theta,\sigma,\beta)\big] < 1 - \beta$.*

*Proof.* Recall that $RP(x|\theta,\sigma,\beta)$ 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 $\theta$. We have that

$$\mathbb{E}\Big[RP\big(X|\theta,\sigma,\beta\big)\Big|l^* \leqslant X \leqslant u^*\Big] = \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{l^*}^{u^*}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} < RP\Big(\theta\Big|\theta,\sigma,\beta\big)\Big) = 1 - \beta \tag{32}$$

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 in the main text. This is the sufficient condition required for the desired result.

Note that the inequalities in equations (27) (for when $r^* \geqslant 1$) and (30) (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 $\mathbb{E}\big(RP\big(X_{sig}|\theta,\sigma,\beta\big)\Big|l^* \leqslant X_{sig} \leqslant u^*\big) = 0 < 1-\beta$ because of truncation. Second, suppose that $l^* \geqslant 1.96\sigma$. Then

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

$$\leqslant \int_{l^*}^{u^*} RP\big(x|\theta,\sigma,\beta\big)\frac{\frac{1}{\sigma}\phi\Big(\frac{x-\theta}{\sigma}\Big)dx}{\int_{l^*}^{u^*}\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} < RP\Big(\theta\Big|\theta,\sigma,\beta\big)\Big) = 1-\beta \tag{33}$$

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 $RP(x|\theta,\sigma,\beta)$ 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 B3, the strict inequality follows from Jensen's inequality, since $RP(x|\theta,\sigma,\beta)$ is strictly concave over $(l^*, u^*)$, and the fact that the expected value of $X$ over this interval is equal to the true value $\theta$. The last equality follows from Lemma 1 in the main text.

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

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

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

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

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

$$< \omega RP\Big(\theta\Big|\theta,\sigma,\beta\Big)\Big) + (1-\omega).RP\Big(2\theta - 1.96\sigma\Big|\theta,\sigma,\beta\Big)\Big) < 1-\beta \tag{34}$$

with

$$\omega = \frac{\int_{1.96\sigma}^{2\theta-1.96\sigma} p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'}{\int_{1.96\sigma}^{u^*} p_{sig}\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\Big(\frac{x'-\theta}{\sigma}\Big)dx'} \tag{35}$$

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 $RP(x|\theta,\sigma,\beta)$. In the last line, the first term uses the concavity of $RP(x|\theta,\sigma,\beta)$ 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 $\theta$. The second term follows because $2\theta - 1.96\sigma$ is the maximum value the function can take because $RP(x|\theta,\sigma,\beta)$ is strictly decreasing in $x$ over positive values. The final inequality follows because $RP\big(\theta\big|\theta,\sigma,\beta\big)\big) = 1-\beta$ (Lemma 1) and $RP\big(2\theta - 1.96\sigma\big|\theta,\sigma,\beta\big)\big) < 1-\beta$ because $2\theta - 1.96\sigma > \theta$ and the function is strictly decreasing over positive values.

This covers all cases, proving the proposition.

**Proposition B1** (Regression to the mean in replications). *Suppose $p_{sig}()$ is symmetric about zero, non-zero over all values, 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*

$$\mathbb{E}\big[X\big|\Theta = \theta, S_X = 1\big] > \theta = \mathbb{E}\big[X_r|\Theta = \theta\big] \tag{36}$$

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

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

$$= \mathbb{E}_{\Sigma|\Theta,S_X}\left(\mathbb{E}_{X|\Theta,\Sigma,S_X}\Big(X|\Theta = \theta, \Sigma = \sigma, |X/\sigma| \geqslant 1.96\Big)\right) \tag{37}$$

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 the expression in equation (37) is also greater than $\theta$. Recall that $X|\theta,\sigma$ is the effect size of published studies and follows a truncated normal distribution:

$$\frac{p\big(\frac{x}{\sigma}\big)\frac{1}{\sigma}\phi\big(\frac{x-\theta}{\sigma}\big)\mathbb{1}\big(|\frac{x}{\sigma}| \geqslant 1.96\big)}{\int p\big(\frac{x'}{\sigma}\big)\frac{1}{\sigma}\phi\big(\frac{x'-\theta}{\sigma}\big)\mathbb{1}\big(|\frac{x}{\sigma}| \geqslant 1.96\big)dx'} \tag{38}$$

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

$$\frac{p\big(z + \frac{\theta}{\sigma}\big)\phi\big(z\big)\mathbb{1}\big(|z + \frac{\theta}{\sigma}| \geqslant 1.96\big)}{\int p\big(z' + \frac{\theta}{\sigma}\big)\phi\big(z'\big)\mathbb{1}\big(|z + \frac{\theta}{\sigma}| \geqslant 1.96\big)dz'} \tag{39}$$

For notational convenience, define the following normalization constants:

$$\bar{\eta} = \mathbb{P}(X \leqslant -1.96\sigma) + \mathbb{P}(X \geqslant 1.96\sigma) = \mathbb{P}\left(Z \leqslant -1.96 - \frac{\theta}{\sigma}\right) + \mathbb{P}\left(Z \geqslant 1.96 - \frac{\theta}{\sigma}\right) \tag{40}$$

$$\eta_1 = \mathbb{P}(X \leqslant -1.96\sigma) = \mathbb{P}\left(Z \leqslant -1.96 - \frac{\theta}{\sigma}\right) \tag{41}$$

$$\eta_2 = \mathbb{P}(X \geqslant 2\theta + 1.96\sigma) = \mathbb{P}\left(Z \geqslant \frac{\theta}{\sigma} + 1.96\right) \tag{42}$$

$$\eta_3 = \mathbb{P}(1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma) = \mathbb{P}\left(1.96 - \frac{\theta}{\sigma} \leqslant Z \leqslant \frac{\theta}{\sigma} - 1.96\right) \tag{43}$$

*Case 1.*

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

$$\begin{aligned}\mathbb{E}(X|1.96\sigma \leqslant |X|) = \frac{1}{\bar{\eta}}\Bigg(&\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) \\ &+ \big(\bar{\eta} - \eta_1 - \eta_2\big)\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta + 1.96\sigma)\Bigg)\end{aligned} \tag{44}$$

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:

$$\mathbb{E}(X|X \leqslant -1.96\sigma) = \mathbb{E}\Big(\theta + \sigma Z | Z \leqslant -1.96 - \frac{\theta}{\sigma}\Big) = \theta + \sigma\mathbb{E}\Big(Z|Z \leqslant -1.96 - \frac{\theta}{\sigma}\Big) \tag{45}$$

Evaluating the expectation in the right-hand-side of equation (45) gives

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

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

$$= -\frac{1}{\eta_1}p_{sig}(-1.96)\phi\Big(-1.96 - \frac{\theta}{\sigma}\Big) + \frac{1}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz \tag{46}$$

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 (45) gives

$$\mathbb{E}(X|X \leqslant -1.96\sigma) = \theta - \frac{\sigma}{\eta_1}p_{sig}(-1.96)\phi\Big(-1.96 - \frac{\theta}{\sigma}\Big) + \frac{\sigma}{\eta_1}\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz \tag{47}$$

Next, note that

$$\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) = \theta + \sigma\mathbb{E}\Big(Z|Z \leqslant \frac{\theta}{\sigma} + 1.96\Big) \tag{48}$$

where

$$\mathbb{E}\Big(Z|Z \leqslant \frac{\theta}{\sigma} + 1.96\Big) = \frac{1}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} z p_{sig}\Big(z + \frac{\theta}{\sigma}\Big)\phi(z)dz \geqslant \frac{1}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} z p_{sig}\Big(z - \frac{\theta}{\sigma}\Big)\phi(z)dz \tag{49}$$

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 (46). Substituting the result into equation (48) gives

$$\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma) \geqslant \theta + \frac{\sigma}{\eta_2}p_{sig}(1.96)\phi\Big(1.96 + \frac{\theta}{\sigma}\Big) + \frac{\sigma}{\eta_2}\int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\Big(z - \frac{\theta}{\sigma}\Big)\phi(z)dz \tag{50}$$

Equations (47) and (50) imply

$$\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma)$$
$$\geqslant (\eta_1+\eta_2)\theta + \sigma\left[p_{sig}(1.96)\phi\left(1.96+\frac{\theta}{\sigma}\right) - p_{sig}(-1.96)\phi\left(-1.96-\frac{\theta}{\sigma}\right)\right]$$
$$+\sigma\left[\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz + \int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\left(z-\frac{\theta}{\sigma}\right)\phi(z)dz\right] = (\eta_1+\eta_2)\theta \tag{51}$$

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

$$\int_{-\infty}^{-1.96-\frac{\theta}{\sigma}} p'_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz = \int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\left(-u+\frac{\theta}{\sigma}\right)\phi(u)du = -\int_{1.96+\frac{\theta}{\sigma}}^{\infty} p'_{sig}\left(u-\frac{\theta}{\sigma}\right)\phi(u)du \tag{52}$$

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 $(\theta, \sigma)$, we have

$$\mathbb{E}(X|1.96\sigma \leqslant |X|) = \frac{1}{\bar{\eta}}\Bigg(\eta_1\mathbb{E}(X|X \leqslant -1.96\sigma) + \eta_2\mathbb{E}(X|X \geqslant 2\theta + 1.96\sigma)$$
$$\eta_3\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma) + \big(\bar{\eta} - \eta_1 - \eta_2 - \eta_3\big)\mathbb{E}(X|2\theta - 1.96\sigma \leqslant X \leqslant 2\theta + 1.96\sigma)\Bigg)$$
$$> \frac{1}{\bar{\eta}}\Bigg(\theta(\eta_1+\eta_2) + \big(\bar{\eta} - \eta_1 - \eta_2 - \eta_3\big)\theta + \eta_3\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma)\Bigg) \tag{53}$$

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

$$\mathbb{E}(X|1.96\sigma \leqslant X \leqslant 2\theta - 1.96\sigma) = \theta + \sigma\mathbb{E}\left(Z\middle|1.96 - \frac{\theta}{\sigma} \leqslant Z \leqslant -1.96 + \frac{\theta}{\sigma}\right) \tag{54}$$

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

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

$$= \frac{1}{\eta_3}\left(\int_{0}^{\frac{\theta}{\sigma}-1.96} z\left[p_{sig}\left(z+\frac{\theta}{\sigma}\right) - p_{sig}\left(-z+\frac{\theta}{\sigma}\right)\right]\phi(z)dz\right) \geqslant 0 \tag{55}$$

The second equality follows because

$$\int_{1.96-\frac{\theta}{\sigma}}^{0} zp_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz = -\int_{0}^{1.96-\frac{\theta}{\sigma}} zp_{sig}\left(z+\frac{\theta}{\sigma}\right)\phi(z)dz = -\int_{0}^{\frac{\theta}{\sigma}-1.96} up_{sig}\left(-u+\frac{\theta}{\sigma}\right)\phi(u)du \tag{56}$$

which uses the substitution $u = -x$ and the symmetry of $\phi()$. The weak inequality in equation (55) 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. □

**Proposition B2** *Under the fractional power rule which sets the replication standard error according to* $\sigma_r(X, \beta, \psi) = \frac{\psi \cdot |X|}{1.96 - \Phi^{-1}(\beta)}$ *with* $\psi < 1$*, the expected replication rate can range between 0.025 and* $1 - \Phi[1.96 - \frac{1}{\psi}\big(1.96 - \Phi^{-1}(\beta)\big)] > 1 - \beta$.

**Proof of Proposition B2**: Under the fractional power rule, the expected replication rate conditional on fixed $(\theta, \sigma)$ is given by

$$\mathbb{E}[RP(X, \Theta, \sigma_r(X, \beta, \psi)|\Theta = \theta, \Sigma = \sigma]$$

$$= \int \left[1 - \Phi\Big(1.96 - \text{sign}(x)\frac{\theta}{\psi \cdot |x|}\big(1.96 - \Phi^{-1}(\beta)\big)\Big)\right]\frac{1}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)dx \tag{57}$$

If $\theta = 0$, then this equals 0.025. Next, suppose that $\theta > 0$ and consider the case where $\sigma \to 0$ such that power in original studies approaches one. See that the integrand is bounded above by one and converges pointwise as $\sigma \to 0$ to

$$1-\Phi\Big(1.96-\text{sign}(x)\frac{\theta}{\psi\cdot|x|}\big(1.96-\Phi^{-1}(\beta)\big)\Big)\mathbb{1}\{x=\theta\} \tag{58}$$

since the normal distribution converges to a degenerate distribution when the variance goes to zero. Thus, by the dominated convergence theorem (and the fact that $\theta>0$), we have that

$$\lim_{\sigma\to 0}\mathbb{E}[RP(X,\Theta,\sigma_r(X,\beta,\psi)|\Theta=\theta,\Sigma=\sigma]=1-\Phi\Big(1.96-\frac{1}{\psi}\big(1.96-\Phi^{-1}(\beta)\big)\Big) \tag{59}$$

When $\psi=1$, this equals $1-\beta$. Since equation (59) is strictly decreasing in $\psi$, it follows that equation (59) is strictly above $1-\beta$ when $\psi<1$.

This shows that the expected replication of an *individual* study can range between 0.025 and $1-\Phi[1.96-\frac{1}{\psi}\big(1.96-\Phi^{-1}(\beta)\big)]>1-\beta$. Integrating over the distribution of latent studies gives the desired result. □

**Proposition B3** *For any function* $g(X,\Sigma,X_r,\beta)$,$\mathbb{E}\big[g(X,\Sigma,X_r,\beta)|D=1,R=1,S_X=1\big]$ *does not depend on* $p_{insig}()$.

**Proof of Proposition B3**: We can write $\mathbb{E}\big[g(X,\Sigma,X_r,\beta)|D=1,R=1,S_X=1\big]$ as

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

$$=\int_{x,\sigma,\theta}\Big(\int_{x_r}g(x,\sigma,x_r,\beta)f_{X_r|X*,\Sigma*,\Theta*}\big(x_r|\theta,\sigma_r(x,\sigma,\beta)\big)dx_r\Big)f_{X*,\Sigma*,\Theta*|D,R,S_X^*}(x,\sigma,\theta|D=1,R=1,S_X^*=1)dxd\sigma d\theta \tag{60}$$

The equality uses the Law of Iterated Expectations and $f_{X_r|X*,\Sigma*,\Theta*,D,R,S_X^*}\big(x_r|\theta,\sigma_r(x,\sigma,\beta)\big)=f_{X_r|X*,\Sigma*,\Theta*}\big(x_r|\theta,\sigma_r(x,\sigma,\beta)\big)$. 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

$$f_{X*,\Sigma*,\Theta*|D,R,S_X^*}(x,\sigma,\theta|D=1,R=1,S_X^*=1)$$

$$=\frac{\mathbb{P}\Big(D=1|X^*=x,\Sigma^*=\sigma,\Theta^*=\theta,R=1,S_X^*=1\Big)}{\mathbb{P}\Big(D=1|R=1,S_X^*=1\Big)}\times\frac{\mathbb{P}\Big(R=1|X^*=x,\Sigma^*=\sigma,\Theta^*=\theta,S_X^*=1\Big)}{\mathbb{P}\Big(R=1|S_X^*=1\Big)}$$

$$\times f_{X*,\Theta,\Sigma*|S_X^*}\Big(x,\theta,\sigma\big|S_X^*=1\Big)$$

$$=\frac{p_{sig}(x/\sigma)}{\mathbb{E}\big(p_{sig}(X*/\Sigma*)\big|S_X^*=1\big)}\cdot\frac{r_{sig}(x/\sigma)}{\mathbb{E}\big(r_{sig}(X*/\Sigma*)\big|S_X^*=1\big)}\cdot f_{X*,\Sigma*,\Theta*|S_X^*}\Big(\theta,x,\sigma\big|S_X^*=1\Big) \tag{61}$$

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. □

## C. Replication Rate Gap Decomposition

How can we measure the relative importance of non-linearities as compared to distortions from selection on significance? To answer this question, I derive a decomposition of the replication rate gap, which I implement in the empirical section.

The decomposition is based on two regimes. Regime 1 ($M1$) assumes use of the standard definition of the replication rate: only significant results are included, and replication selection is a random sample of significant results. Regime 2 ($M2$) is based on a counterfactual scenario where all results are published and replication is random. This implies the distribution of published, replicated studies coincides with the distribution of latent studies. Formally, the expectation operators under both regimes are defined by:

$$\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big]=\int RP(x,\theta,\sigma_r(x,\beta))f_{X*,\Theta*|D,R,S_X^*}(x,\theta|D=1,R=1,S_X^*=1)dxd\theta \tag{62}$$

$$\mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big]=\int RP(x,\theta,\sigma_r(x,\beta))f_{X*,\Theta*}(x,\theta)dxd\theta \tag{63}$$

Using these, we have the following decomposition:

$$\underbrace{(1-\beta)-\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big]}_{\text{replication rate gap}}=\underbrace{(1-\beta)-\mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geqslant 0\big]}_{\text{(i) concavity gap}}$$

$$+\underbrace{\mathbb{P}_{M1}(X<0)\Big(\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geqslant 0\big]-\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X<0\big]\Big)}_{\text{(ii) wrong-sign gap}}$$

$$+\underbrace{\mathbb{E}_{M2}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geqslant 0\big]-\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geqslant 0\big]}_{\text{(iii) selection-on-significance gap}} \tag{64}$$

*Proof.* Write the expected replication probability under model 1 as

$$\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big]=\mathbb{E}_{M1}\big[RP(X,\Theta,\sigma_r(X,\beta))\big|X\geqslant 0\big]$$

$$+ \mathbb{P}_{M1}(X < 0)\Big(\mathbb{E}_{M1}\big[RP(X, \Theta, \sigma_r(X, \beta))|X < 0\big]\Big) - \mathbb{E}_{M1}\big[RP(X, \Theta, \sigma_r(X, \beta))|X \geqslant 0\big]\Big) \tag{65}$$

To arrive at equation (64), substitute equation (65) into the replication rate gap; add and subtract $\mathbb{E}_{M2}\big[RP(X, \Theta, \sigma_r(X, \beta))|X \geqslant 0\big]$; and rearrange the terms. □

Note that the concavity gap and the selection-on-significance gap condition on estimates with the same sign as the underlying true effect. This allows us to determine their contribution separate from the impact of attempting to replicate original estimates with the 'wrong' sign.

Table C1 presents the results. Panel A reproduces the results in the main text, and Panel B present the decomposition results. The empirical results for the decomposition show that failing to account for the concavity of the replication power function explains the overwhelming majority of the explained replication rate gap in both economics and psychology. The selection-on-significance gap in small, explaining only 3.1% of the gap in economics, while actually *decreasing* 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 smaller true effects.

| | 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 | 0.543 |
| *B. Decomposition of explained gap* | | | |
| Predicted replication rate gap | 0.320 (100%) | 0.375 (100%) | – |
| Concavity gap | 0.292 (91.16%) | 0.364 (97.16%) | – |
| Wrong-sign gap | 0.018 (5.72%) | 0.030 (8.03%) | – |
| Selection-on-significance gap | 0.010 (3.12%) | -0.019 (-5.18%) | – |

TABLE C1 – REPLICATION RATE PREDICTIONS AND DECOMPOSITION RESULTS

*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 (8) calculated using parameter estimates Table 1. 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.

Below I provide details underlying the intuition behind the decomposition results.

*Concavity gap.*—Figure C1 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 $\omega$ increases and approaches zero

in the limit.[19] It follows that the size of the non-linearity gap depends on the distribution of $\omega$. 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 (**??**)). 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 C2 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 C2 – 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 | 0.018 | 0.030 |

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

*Selection-on-significance gap.*—The Selection-on-significance 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 this gap is ambiguous because of two opposing effects from conditioning on statistical significance. To see these two effects, consider the figures in Table C2 which are based on the estimated empirical models. For the first effect, note that conditioning on significant findings increases mean bias in both

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

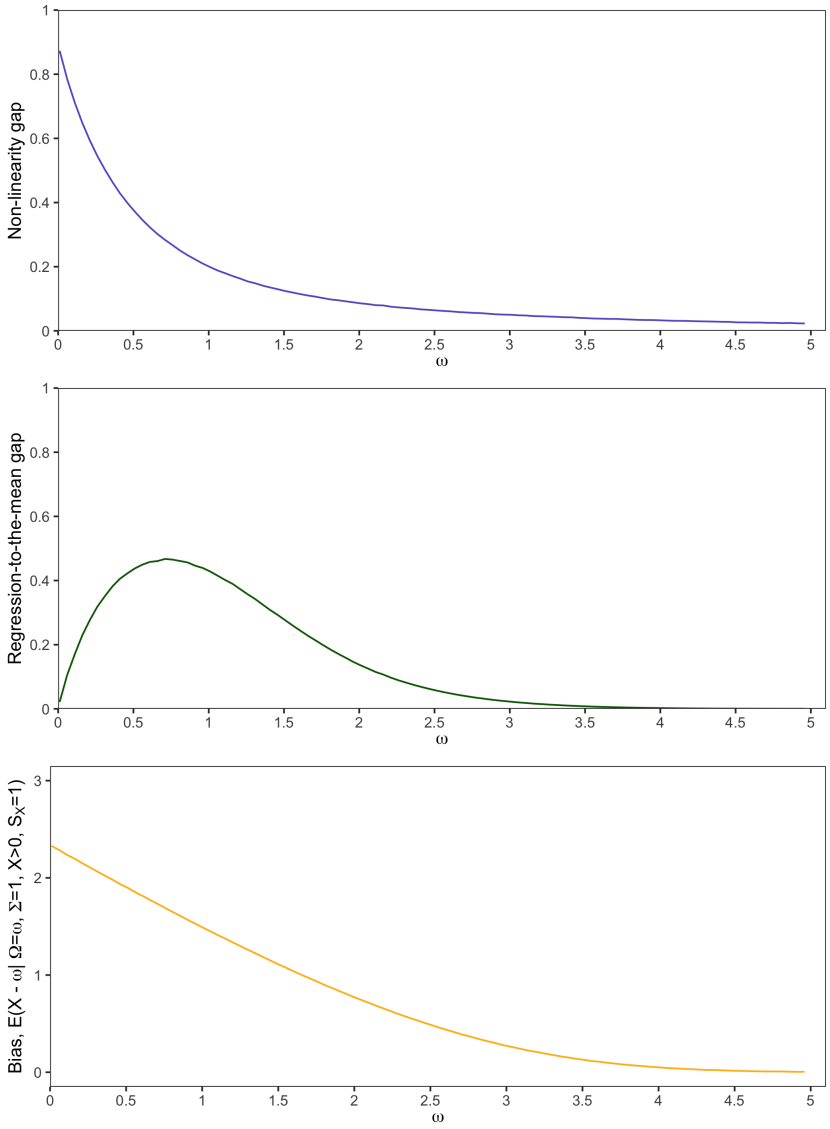


FIGURE C1. REPLICATION RATE GAP DECOMPOSITION: MONTE CARLO SIMULATIONS

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

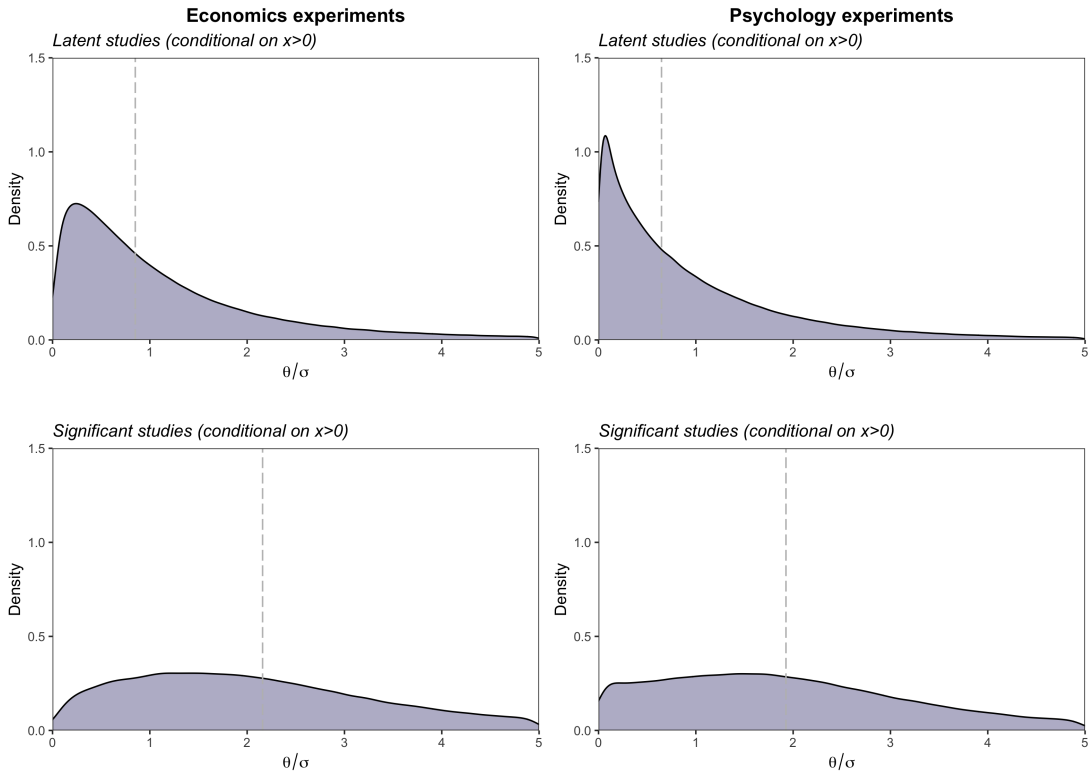


FIGURE C2. DISTRIBUTION OF NORMALIZED TRUE EFFECTS: LATENT STUDIES AND SIGNIFICANT STUDIES

*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 in the main text. Dashed vertical lines show the median of the distribution.

applications.[20] This makes replication more difficult for any fixed level of $\omega$. For the second effect, note that conditioning also tends to select studies with larger standardized true effects $\omega$, which have higher replication probabilities.[21] Higher replication probabilities arise because (i) bias is lower for larger true effects; and (ii) non-linearity effects are less severe for more highly powered studies.

The bottom panel in Figure C1 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.

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

[21]The impact of conditioning on the full distribution of $\omega$ can be seen in Figure C2.

TABLE C3 – 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 from Table 1 in the main text. 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 $D = 1$.

## D. 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: $\mathbb{1}\big[X \in \big(X_r - 1.96\Sigma_r, X_r + 1.96\Sigma_r\big)\big]$.

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 $\Sigma_m$ are the meta-analytic estimate and standard error, respectively.[22]

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: $\mathbb{1}\big[X_r \in \big(X - 1.96\sqrt{\Sigma^2 + \Sigma_r^2}, X + 1.96\sqrt{\Sigma^2 + \Sigma_r^2}\big)\big]$.[23]

These alternative replication measures are frequently reported in large-scale replication studies (Open Science Collaboration, 2015; Camerer et al., 2016, 2018). In simulations, I

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

[23]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).

calculate these measures over significant and insignificant published results, since conditioning on statistical significance makes them unresponsive to selective publication on null results (Proposition B2).

Simulations assume that all results significant at the 5% level are published, and that results insignificant at the 5% level are published with probability $\beta_p$. I then calculate how the various measures change with $\beta_p$ 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 lead to an increase in the policy variable, $\beta_p$. 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 D1 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.[24] 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.

## E. 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).

[24] 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 $X_r$.

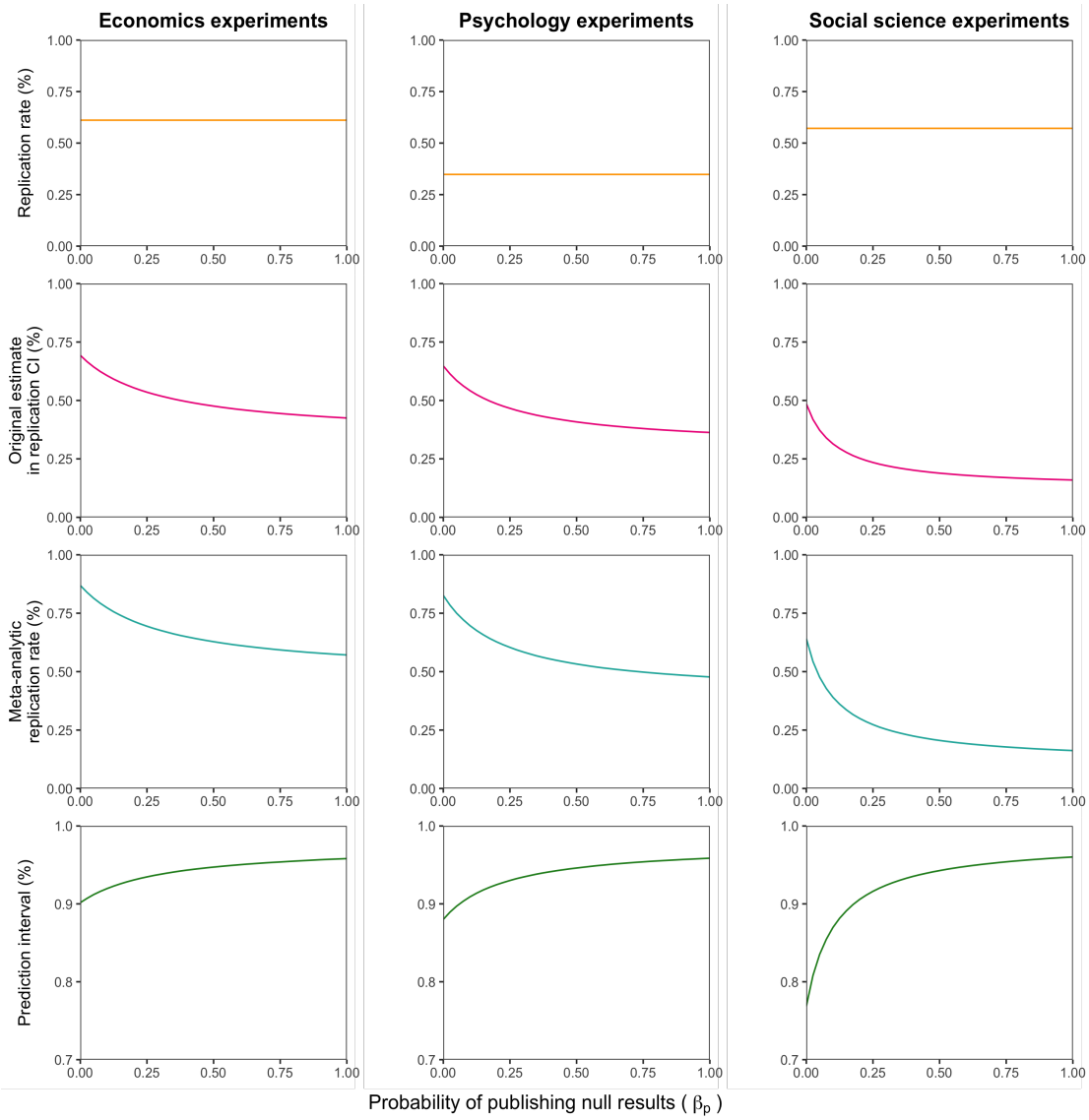


FIGURE D1 – POLICY SIMULATIONS: ALTERNATIVE MEASURES OF REPLICATION AND SELECTIVE PUBLICATION

*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 $\beta_p$. The first column reproduces replication rate predictions in Table 2.

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 influence 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 $\beta_{p2}$ 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.

## F. 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 $1-\beta$. This rule implies

$$\sigma_r^{sim}(x^{sim}, \beta) = \frac{|x^{sim}|}{1.96 - \Phi^{-1}(\beta)} \tag{66}$$

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 F1 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 and social science, 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

$$\sigma_r^{sim}(x^{sim}, \widehat{\beta}^n) = \frac{|x^{sim}|}{1.96 - \Phi^{-1}(\widehat{\beta}^n)} \tag{67}$$

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

$$\sigma_r^{sim}(\sigma^{sim}) = \sigma^{sim} \tag{68}$$

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.

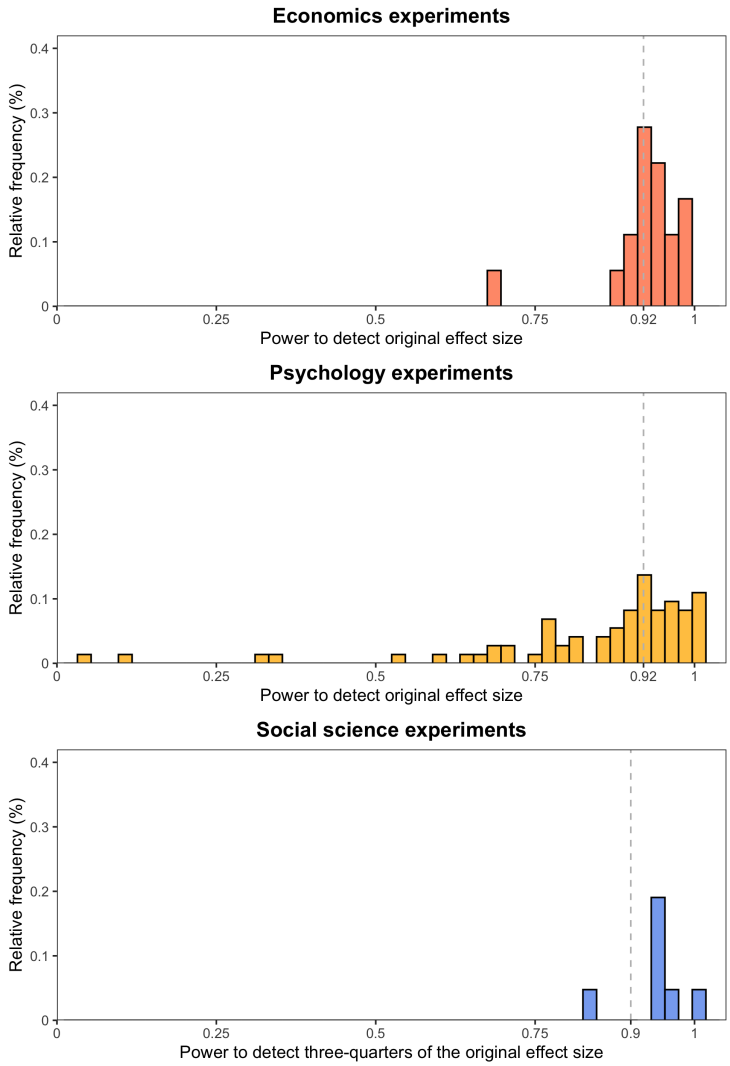


FIGURE F1. Histograms of realized intended power in replication studies in experimental economics, psychology, and social science. Data are from Camerer et al. (2016), Open Science Collaboration (2015), and Camerer et al. (2018), respectively. Realized intended power is defined as $1 - \Phi(1.96 - \psi \cdot \frac{x}{\sigma_r})$ 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.

.

*Results.*—Table F1 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.

Table F1 – 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 |
| Mean power | 0.600 | 0.545 | 0.543 |
| Realized power | 0.615 | 0.522 | 0.555 |
| *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 in the main text and assuming different rules for calculating replication power.

## G. Relative Effect Size Predictions

The main focus of this article is the binary measure of replication based on the statistical significance criterion. This is because of its status as the primary replication indicator in the large-scale replication studies.[25] 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 using the estimated models.

*Theoretical discussion.*—The relative effect size 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

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

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 B2).[26] 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 G1 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%.

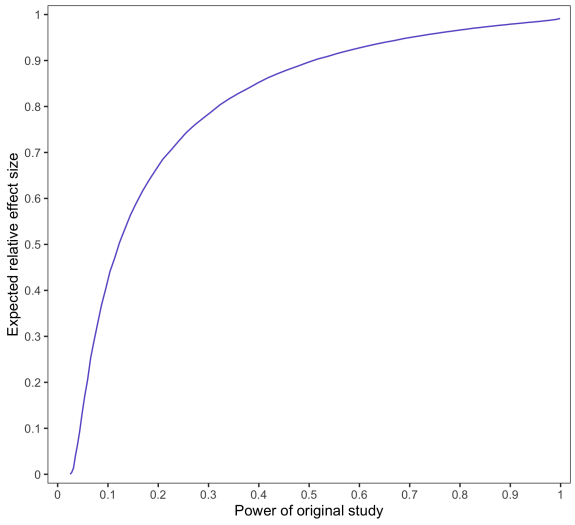


FIGURE G1. EXPECTED RELATIVE EFFECT SIZE OF SIGNIFICANT ORIGINAL STUDIES AND THEIR STATISTICAL POWER

*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 $N(\omega, \sigma_{r,i}(z_i, \beta)^2)$ distribution. The replication standard error is calculated using the common power rule to detect original effect sizes with 90% power (i.e. $1-\beta = 0.9$), which is given by $\sigma_r(z_i, \beta) = |z_i|/[1.96 - \Phi^{-1}(\beta)] = |z_i|/3.242$.

[26] 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 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.

*Empirical results.*—The estimated models in Table 1 in the main text can be used to generate predictions of the average relative effect sizes. To procedure for simulating replications is identical to the procedure outlined in the main text for the replication rate case. 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 their corresponding replication results; $M_{sig}$ is the size of the set. The predicted relative effect size is equal to

$$\frac{1}{M_{sig}} \sum_{i=1}^{M_{sig}} \frac{\rho_{i,r}^{sig}}{\rho_i^{sig}} \tag{69}$$

where $\rho = \tanh z$ denotes the Pearson correlation coefficient which is obtained by transforming the Fisher-transformed correlation coefficient (Fisher, 1915). I also present results for the median relative effect size. Results are presented in Table G2. The predicted average relative effect size is relatively close to observed average relative effect size in economics, somewhat further off in social science, and quite far off in psychology. In each case, the predicted average relative effect size is optimistic compared to the observed value. In economics and psychology, the difference in predicted and observed relative effect sizes is not statistically different from zero, while in psychology it is. Predictions for median relative effect sizes show qualitatively similar results.

| | Economics | Psychology | Social Sciences |
|---|---|---|---|
| Observed relative effect size (mean) | 0.657 | 0.374 | 0.443 |
| Predicted relative effect size (mean) | 0.703 | 0.637 | 0.533 |
| | (0.135) | (0.060) | (0.141) |
| Observed relative effect size (median) | 0.691 | 0.292 | 0.527 |
| Predicted relative effect size (median) | 0.747 | 0.674 | 0.595 |
| | (0.129) | (0.063) | (0.240) |

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 (69) and the procedure outlined in the text. Standard errors are calculated using the delta method.

## 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* $(X, \Sigma, \Theta)$ *which is published* $(D = 1)$ *and chosen for replication* $(R = 1)$ *is*

$$\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Big) = \begin{cases} \mathbb{P}\Big(\frac{|X_r|}{\sigma_r(X,\Sigma,\beta)} \geqslant 1.96, \text{sign}(X) = \text{sign}(X_r)\Big| X, \Theta, \sigma_r(X, \Sigma, \beta)\Big) & \text{if } 1.96.\Sigma \leqslant |X| \\ \mathbb{P}\Big(\frac{|X_r|}{\sigma_r(X,\Sigma,\beta)} < 1.96\Big| X, \Theta, \sigma_r(X, \Sigma, \beta)\Big) & \text{if } 1.96.\Sigma > |X| \end{cases} \tag{70}$$

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

$$\begin{aligned} \mathbb{E}\Big[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Big)\Big] = \mathbb{P}\Big(1.96.\Sigma \leqslant |X|\Big)\mathbb{E}\Big[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Big| X, \Theta, \sigma_r(X, \Sigma, \beta), 1.96.\Sigma \leqslant |X|\Big] \\ + \Big(1 - \mathbb{P}\Big(1.96.\Sigma \leqslant |X|\Big)\Big)\mathbb{E}\Big[\widetilde{RP}\Big(X, \Theta, \sigma_r(X, \Sigma, \beta)\Big| X, \Theta, \sigma_r(X, \Sigma, \beta), 1.96.\Sigma > |X|\Big] \end{aligned} \tag{71}$$

First, note that Definition H2 equals the standard replication rate definition when the expectation is taken only over significant studies because, in this case, $\mathbb{P}\big(|X| \leqslant 1.96.\Sigma\big) = 0$. Thus, 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.[27]

*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 in the main text. 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.

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

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 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 | 0.553 |
| $\mathbb{P}(\text{Replicated}\|S_X = 1)$ | 0.600 | 0.551 |
| $\mathbb{P}(\text{Replicated}\|S_X = 0)$ | 0.574 | 0.789 |
| $\mathbb{P}(S_X = 1)$ | 0.993 | 0.993 |
| $\mathbb{P}(S_X = 0)$ | 0.007 | 0.007 |
| *No selective publication* | | |
| Generalized replication rate | 0.432 | 0.773 |
| $\mathbb{P}(\text{Replicated}\|S_X = 1)$ | 0.582 | 0.515 |
| $\mathbb{P}(\text{Replicated}\|S_X = 0)$ | 0.378 | 0.867 |
| $\mathbb{P}(S_X = 1)$ | 0.268 | 0.268 |
| $\mathbb{P}(S_X = 0)$ | 0.732 | 0.732 |
| **B Psychology experiments** | | |
| *Selective Publication* | | |
| Generalized replication rate | 0.546 | 0.526 |
| $\mathbb{P}(\text{Replicated}\|S_X = 1)$ | 0.544 | 0.487 |
| $\mathbb{P}(\text{Replicated}\|S_X = 0)$ | 0.563 | 0.839 |
| $\mathbb{P}(S_X = 1)$ | 0.890 | 0.890 |
| $\mathbb{P}(S_X = 0)$ | 0.110 | 0.110 |
| *No selective publication* | | |
| Generalized replication rate | 0.490 | 0.798 |
| $\mathbb{P}(\text{Replicated}\|S_X = 1)$ | 0.535 | 0.469 |
| $\mathbb{P}(\text{Replicated}\|S_X = 0)$ | 0.478 | 0.886 |
| $\mathbb{P}(S_X = 1)$ | 0.209 | 0.209 |
| $\mathbb{P}(S_X = 0)$ | 0.791 | 0.791 |

*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 in the main text. Different column represent different rules for calculating power in replications.