Estimating statistical power, posterior probability and publication bias of psychological research using the observed replication rate

Michael Ingre¹ and Gustav Nilsonne¹,²,³

1: Karolinska Institutet, Department of Clinical Neuroscience
2: Stockholm University, Stress Research Institute
3: Stanford University, Department of Psychology

Contact: michael.ingre@gmail.com

Abstract

In the present paper, we show how Bayes' theorem can be used to better understand the implications of the 36% reproducibility rate of published psychological findings that was reported by the OSC (Open Science Collaboration, 2015). We demonstrate a method to assess publication bias, and show that the observed reproducibility rate was not consistent with an unbiased literature. We estimate a plausible range for the prior probability of this research, suggesting that statistical power in the original studies was 42%—73% and producing (positive) findings that were likely to be true 43% to 69% of the time. Publication bias was large, assuming a literature with 90% positive findings, indicating that negative evidence was observed 60—100 times before one negative result was published. These findings imply that even when studied associations are truly NULL, we expect the literature to be dominated by statistically significant findings.

Introduction

The Open Science Collaboration (OSC) reported that 36% of published positive findings in experimental psychology were replicated in independent attempts (Open Science Collaboration, 2015). This finding is interesting in itself as an indicator of the reproducibility of published findings in psychology; however, it is also an important data point that can be used together with other information to assess publication bias, statistical power and even the posterior probability of findings published in the psychological literature.

Another important set of observations concerns the proportion of positive findings in the literature. A series of observations spanning five decades has indicated that >90% published studies in psychology reported positive findings, that is, where the authors' hypothesis was
supported by data (Fanelli, 2010; T. D. Sterling, Rosenbaum, & Weinkam, 1995; Theodore D. Sterling, 1959). A similar observation was made by the OSC, where 97% of the original studies sampled supported the proposed hypothesis with a “statistically significant” association (Open Science Collaboration, 2015).

In the present paper, we will show how this information can be used to better understand meta-properties of published psychological research. We demonstrate a mathematical solution, based on Bayes’ theorem, that can be used to assess statistical power, posterior probability, and publication bias of published psychological research. We aim to answer the following questions:

- What are the properties of research that leads to 90% positive findings?
- What is the expected reproducibility of research with 90% positive findings?
- Is the observed 36% reproducibility rate consistent with an unbiased literature?
- What does the observed reproducibility suggest about the prior probability of the tested hypotheses, statistical power of the studies, the posterior probability of the original findings and of publication bias?

**What are the properties of research that leads to 90% positive findings?**

To answer this question we introduce the concept of prior probability from Bayesian theory; it describes the probability that a hypothesis is true before it has been tested on data. When considering a large number of hypotheses, prior probability can also be understood as the proportion of hypotheses that are true a priori, that is, before they have been tested on data. This prior probability can be small and close to zero, for example, in massively exploratory studies where vast amounts of data are searched to try to find the few true associations that may exist; or it can be large and close to one, in theoretically motivated confirmatory research with prior empirical support. We will use theta ($\theta$) to denote a prior probability.

We also need to consider the probability that a study testing a true hypothesis will produce positive evidence. This is generally known as statistical power within a NULL hypothesis significance testing (NHST) paradigm, and is calculated from the type-2 error rate: $1 - \beta$. Finally, we need to consider the test’s type-1 error rate, that describes the probability of observing positive evidence when the hypothesis is false, which we will assume to be $\alpha = .05$ in this text unless stated otherwise.
The probability of observing true positive evidence is calculated by multiplying the prior probability with the statistical power (equation 1) and the probability to observe false positive evidence is the type-1 error rate multiplied by the prior probability that the hypothesis is false (equation 2). Added together, they describe the total probability of observing positive evidence (equation 3).

\[
E_{\text{true}} = \theta (1 - \beta) \quad \text{Eq (1)}
\]

\[
E_{\text{false}} = \alpha (1 - \theta) \quad \text{Eq (2)}
\]

\[
E_{\text{total}} = \theta (1 - \beta) + \alpha (1 - \theta) \quad \text{Eq (3)}
\]

If all hypotheses were true a priori, we could not observe false positive evidence, and the probability of observing positive evidence would reduce to the statistical power \((1 - \beta)\). This shows that one way to produce 90\% positive findings is to only test true hypotheses with 90\% power. Another way to produce close to 90\% positive findings is to run studies with perfect power on hypotheses of which 90\% are true a priori. It should be noted that in such a situation we would actually observe 90.5\% positive evidence, because we would also observe a small number of type-1 errors when the hypothesis is false, as described by equation 2.

It is not realistic to assume perfect power, nor that all hypotheses are true a priori. However, a middle ground of 95\% power and 95\% true hypotheses a priori would produce a literature close to what has been observed, with 90.5\% positive evidence (equation 3). Thus, it is possible to produce an unbiased literature with \sim 90\% positive findings when the underlying research tests hypotheses that are 95\% true a priori in studies with 95\% statistical power.

**What is the expected reproducibility of research with 90\% positive findings?**

To calculate the expected reproducibility of published research, we first need to calculate the probability of a positive finding to be true (rather than a type-1 error), which we call the posterior probability. The components needed to calculate this quantity are described by equation 1-3, and we merely need to take the ratio \(E_{\text{true}} / E_{\text{total}}\) to determine the probability of positive findings to be true. This completes a formulation of Bayes’ theorem described by equation 4 below, which can be used to calculate the posterior probability of a hypothesis after observing positive evidence from NHST:

\[
\hat{\theta} = \frac{\theta (1 - \beta)}{\theta (1 - \beta) + \alpha (1 - \theta)} \quad \text{Eq (4)}
\]
When we know the posterior probability and statistical power of the research, it is easy to calculate the expected reproducibility \( R \). Equation 3 above already showed how to calculate the probability of observing positive evidence, but in this case we substitute the assumed prior probability \( (\theta) \) with the actual posterior probability \( (\hat{\theta}) \) of the finding:

\[
R = \hat{\theta}(1 - \beta) + \alpha(1 - \hat{\theta})
\]

Eq (5)

If we assume a middle ground on the prior and statistical power needed to produce ~90% positive findings that we discussed above (i.e. \( \theta = .95 \) and \( 1 - \beta = .95 \)), we can apply equation 4 and find that the posterior probability of such research is \( \hat{\theta} = .997 \); and entering that number into equation 5 shows that the expected reproducibility in an identical study is 95%. Thus, in an unbiased literature with 90% positive findings we would expect positive results to be reproduced ~95% of the time by independent researchers.

**Is the observed 36% reproducibility consistent with an unbiased literature?**

We now know the fundamental properties needed for research to produce 90% positive findings; we need to have close to 95% statistical power and test hypotheses that are close to 95% true *a priori*. We also know that such research would be reproducible 95% of the time in identical studies, and we can use this information to create a formal test of bias of the published literature.

A binomial test on the observed reproducibility of 36% (95% CI: 27%-46%; \( n = 97 \)) reported by the OSC, indicates strong evidence (\( p < 10^{-15} \)) that the replication studies were not drawn from an unbiased literature with 95% reproducibility. This test shows publication bias in the OSC sample, supporting the observation made in their original report of a right-skewed funnel plot (Open Science Collaboration, 2015).

**What does the observed reproducibility suggest about the prior probability of the tested hypotheses, statistical power of the studies, the posterior probability of the original findings and of publication bias?**

One complication with applying Bayes’ theorem (equation 4) is that it is based on several unknown variables. We usually have a good idea of the type-1 error rate that is applied in research, but prior probability and statistical power are often elusive. We can sometimes make informed guesses (Dreber et al., 2015) and calculate the posterior probability, as
illustrated above, but with three unknown variables, statistical power \((1 - \beta)\), prior \((\theta)\), and posterior \((\hat{\theta})\), there is only a limited amount of information we can extract from data. We next attempt to reduce the number of unknown variables to only two, so that we can learn more useful information.

A first step in this process is to form a system of equations based on equation 4 and 5 above so that we can incorporate the observed reproducibility into our calculations. The system of equations defined by equation 6 below uses subscripts to indicate quantities that are specific to original studies \((o)\) and replication studies \((r)\):

\[
\begin{align*}
\hat{\theta} &= \frac{\theta(1-\beta_o)}{\theta(1-\beta_o) + \alpha_o(1-\theta)} \\
R &= \hat{\theta}(1 - \beta_r) + \alpha_r(1 - \hat{\theta})
\end{align*}
\]

Eq (6)

We use two different type-1 error rates; the nominal \(\alpha_o = .05\) is assumed for original studies, and because the OSC applied two-tailed tests of directional hypotheses we assume \(\alpha_r = .025\) in the replication studies. We also assume the replication rate to be \(R = .36\) as reported by the OSC. If we could assume statistical power to be identical in the original and replication studies \((i.e. \beta_r = \beta_o)\), equation 6 would then have only two unknown variables \((\beta_c\ and\ \hat{\theta})\), which is what we are trying to achieve.

However, statistical power was not identical in original and replication studies; the OSC determined the replication sample sizes from power analyses based on the reported effect sizes of the original studies. Such estimates are known to be inflated in the presence of publication bias (Yarkoni, 2009), and since we have determined that the literature is biased, we cannot use these estimates. Data downloaded from the OSC github repository \((rrp, n.d.)\) show that 70% of replications were designed with a larger sample than the original study, 10% had the same sample size, and 20% were smaller than the original study, indicating that statistical power was generally higher in the replication studies. This information can be used to calculate upper and lower bounds of statistical power. The lower bound assumes identical power in original and replication studies \((\beta_r = \beta_o)\), and the upper bound assumes perfect power in replication studies \((\beta_r = 0)\). This reduces the number of unknowns in equation 6 to two variables \((\beta_c\ and\ \hat{\theta})\), and when statistical power is higher in the replication studies, it defines boundaries inside of which which the true value must fall.
We can attempt a more precise approximation of power based on the median degrees of freedom of original studies (df=54) and replication studies (df=68) reported by the OSC. The observed median effect size in replication studies ($r=.2$) is likely to be attenuated by the presence of NULL associations in data, and the observed effect size in the original studies ($r=.4$) is likely to be inflated by publication bias; thus, the true effect size is likely between these two estimates. Calculating statistical power for the range $0.2 < r < 0.4$ shows that a median sized replication study added $\sim 6\%—10\%$ statistical power ($10\%$ at midpoint: $r=.3$) compared to the original study. This estimate gives an approximation of the average increase in statistical power for the replication studies, and allows for a likely range of power to be defined using $\beta_r = \beta_0 - .06$ and $\beta_r = \beta_0 - .10$. This gives two additional applications of equation 6 with only two unknown variables, that define a range in which the true value is likely to fall.

With only two unknown variables left in equation 6, we can solve the equation to calculate statistical power from an assumed range of prior probabilities; and since we then know the statistical power, we can also calculate the posterior probability of the original findings using equation 4. In addition, we can use equation 3 to calculate the expected proportion of positive findings that was observed before publication, and we can compare that estimate to the $>90\%$ positive findings that has been observed in the literature to assess publication bias. The complete solution to equation 6 is presented in the supplemental material and the results are plotted in figure 1 below.
Figure 1. Calculated statistical power and posterior probability of the original research replicated by the OSC (top) together with the expected proportion of positive evidence observed and the estimated publication bias of the research (bottom), based on equation 6. The calculations were performed for the range $0.25 < \theta < 0.975$ of assumed prior probabilities. The plots assume $\alpha_o = 0.05$ in the original studies, $\alpha_r = 0.025$ in replication studies, a reproducibility rate of $R = 0.36$ and a published literature with 90% positive findings. The likely range assumes that the replication studies performed by the OSC had ~6%—10% more statistical power than the original studies. Outer boundaries were calculated assuming that statistical power in the replication studies fell somewhere between the power of the original studies and perfect power. X-axes of all plots and the y-axis of the publication bias plot (bottom, right) were log transformed.
The findings presented in figure 1 give insight into a plausible range of prior probabilities of tested hypotheses in psychology. The top left panel shows that the prior probability of the underlying research cannot be $\theta < .025$, because that would imply better than perfect power in the original studies. Lack of support for small priors is also visible in both top panels when the likely range, assuming 6%—10% better power in the replication studies, is pushed outside of the outer boundaries, indicating the need for better than perfect power in the replications to fit the observed data.

The prior was also unlikely to be smaller than $\theta < .05$; while the lower bound of the power estimate fell at 50%, it is based on the implausible assumption of perfect power in the replications. The likely range suggests 71%—73% power, which seems optimistic for this research; such large statistical power has been indicated only for larger than medium effect sizes in psychological research (Rossi, 1990; Szucs & Ioannidis, 2017). A restricted range of priors was defined as $0.05 < \theta < 0.20$ that indicated power between 42% and 73%, and we assume this to be a plausible range in which the true prior is likely to fall.

With higher assumed prior probabilities, the posterior probability of the original research goes up, and statistical power has to come down to fit the observed data. If we would assume that one out of ten tested hypotheses were true a priori ($\theta = .1$), the posterior of the original research was ~ 55% and the reason the OSC could only replicate 36% of the findings is explained by statistical power in the replication studies just over 60%. For the full range of plausible priors $0.05 < \theta < 0.20$, the proportion of of true positive findings in the original studies fell between 43% and 69%.

The most striking observation in figure 1 was the estimate of publication bias, showing that even the lower bound, above which the true estimate must fall if our assumptions hold, indicated that negative evidence was observed $> 16$ times before one was published over the whole range of priors. For a more plausible range of priors $0.05 < \theta < 0.20$, we see an even more pronounced bias, suggesting that negative evidence was likely to have been observed 60 to 100 times before publication.

**Discussion**

In the present paper we showed how Bayes’ theorem can be used to better understand the implications of the observed 36% reproducibility rate of published psychological findings that was reported by the OSC (Open Science Collaboration, 2015). We demonstrated a method to assess publication bias, and performed a formal test indicating that the observed reproducibility rate was not consistent with an unbiased literature. The findings also showed
that the prior probability of this research cannot be smaller than $\theta < .025$, and is unlikely to be smaller than $\theta < .05$; we suggest a plausible prior for this research somewhere in the range $.05 < \theta < .20$, where statistical power of the original studies was estimated to be 42%—73%. We found that 43% to 69% of the original findings were likely to be true, and that the reproducibility rate observed by the OSC was lower due to less than perfect power in the replications. Publication bias was large, assuming a literature with 90% positive findings, indicating that negative evidence was observed 60—100 times before one negative result was published.

Another recent Bayesian reanalysis (Johnson, Payne, Wang, Asher, & Mandal, 2017) of the data observed by OSC focused on observed effect sizes, and was restricted to the subsample for which a correlation ($r$) with standard errors could be derived (73/100 studies). This subsample had 71 positive findings and the observed reproducibility rate was 41%. The authors estimated ~93% true NULL hypotheses in this research, i.e. $\theta = .07$. Furthermore, they assumed $\alpha_o = .052$, and that both original studies and replication studies had 75% power, to arrive at an estimated posterior of $37/71=52\%$ for the original positive findings. They also suggest that ~700 hypothesis tests were performed to produce the 73 published findings. We can verify these estimates under the above assumptions using equation 3 & 4.

However, the estimates presented in the present paper pertain to the full sample replicated by the OSC, with 36% observed reproducibility. Also, as discussed above, we do not assume identical power in replication studies and original studies, since the replications were designed with larger sample sizes; thus, we only use that assumption for the lower bound of power. If we were to accept the prior suggested by Johnson et al (Johnson et al., 2017) ($\theta = .07$), statistical power would be in the range 62%—65% in the original and 71%—72% in the replication studies. The posterior of the original findings would be 48%—49%. In addition, the expected proportion of positive evidence observed in the original studies was ~9%, indicating that ~97/0.9=1077 studies were performed to produce the 97 positive and 3 negative findings that were published and subsequently replicated by the OSC; this means that negative evidence would have been observed ~ 90 times before a negative finding was published, assuming a literature with 90% positive findings.

The prior ($\theta = .07$) suggested by Johnson et al (Johnson et al., 2017) implies > 62% power in the original studies; such high power has been indicated for medium ($r=.3$) and larger effect sizes in psychological research (Rossi, 1990; Szucs & Ioannidis, 2017), and is larger than empirical estimates of median power observed in other fields (Button et al., 2013; Dumas-Mallet, Button, Boradu, Gonon, & Munafò, 2017). Considering that the the median
effect size observed in the replication studies by OSC was only $r=.2$, assuming such high power is optimistic, but not implausible due to the likely attenuation of this estimate from the presence of NULL associations in data. We propose a plausible prior somewhere in the range $0.05 < \theta < 0.20$; corresponding to statistical power in the range 42%—73% of the original studies and suggesting that 43%—69% of the positive findings observed in the original studies were likely to be true.

Applying Bayes’ theorem in this way has important implications; it assumes that hypotheses are either true or false, and such binary hypothesis testing has been criticized (Cohen, 1994). Indeed, it can be argued that there are no truly non-zero associations in observational data. If we assume that no associations are truly zero, but we are not interested in making inferences from very small true effect sizes, $p$-values from NULL hypothesis significance testing (NHST) would be biased with inflated type-1 errors. In addition, we may conclude that any (non-directional) hypothesis is necessarily true, giving a trivial prior probability of $\theta = 1$. However, we should recognize that these are not limitations of binary hypothesis testing per se, but rather limitations of how specific hypotheses are formulated and tested. It is possible to define a different “NULL” hypothesis, with a mean other than zero, to protect inferences from true effect sizes of “trivial” magnitudes (Ingre, 2013) and make the prior more informative in observational studies at $\theta < 1$. Also, binary NHST is not inherently problematic in true experimental designs (with randomisation), since we can then assume associations in data that are truly NULL. In the present analysis we have assumed the same position on binary NHST as the publishing authors of the original studies that were replicated by the OSC, and the limitations discussed above apply similarly to how they would apply to the original studies.

The most crucial estimate used in our analysis was the observed reproducibility rate of 36% reported by the OSC (Open Science Collaboration, 2015). Reproducibility is a complicated concept with many different facets, in particular in psychology and the social sciences; some “true” findings may not be possible to replicate in a different time, social or cultural context, because the underlying meaning of the constructs used to design the study or define the variables may have changed. The underlying theory may still be valid but needs to be adapted to the new environment, and this has been proposed as an argument against the validity of direct replication of a study’s methods on an independent sample (Stroebe & Strack, 2014). But from a more general scientific perspective, it can be seen as a flaw in the formulated theory and the methods defined to test it; science needs to be verifiable to stand out from other types of claims and should have some generalizability to be a useful source of knowledge and important context need to be included. Another factor to consider is poorly
described methods in the original study that may impact the success rate in replications; but this is essentially the same problem. If the study report did not present sufficient information to accurately replicate the methods: how can it be properly understood and evaluated by the readers?

Reproducibility may have been impaired because of mistakes made by the replicating team of researchers; however, this does not seem to be a major risk in the OSC study. The study was pre-registered and performed by well motivated researchers under more or less public scrutiny; the team were in frequent contact with authors of the original studies to obtain material and information about the design and procedure of their studies; and they employed a system of internal reviews of all studies to ensure quality. However, any potential mistakes that may have lowered the reproducibility rate is part of the overall type-2 error rate ($\hat{\beta}$) in equation 5 and 6, and can be seen as a reduction of “statistical power” in the replication studies below what we have nominally assumed. It seems unlikely that this would pose a problem large enough to invalidate the lower bound of the estimate used in this study, assuming power to be identical in the original and replication studies.

Studies eligible for replication by the OSC were picked from three prestigious journals in experimental psychology. Approximately one third of the total sample was never submitted for replication, mostly because these studies were deemed infeasible to replicate, for example, because they required special samples, knowledge or equipment. This introduces uncertainty and potential bias in the reproducibility estimate; it is possible that the more specialized or complicated designs would have worse (or, less likely, better) reproducibility. Thus, the reproducibility rate estimated by the OSC is an estimate representative of the two thirds most accessible research in three well-regarded journals in experimental psychology; and might not generalize to psychology in general.

Findings presented in this paper refer to studies in experimental psychology; data from other scientific fields suggest a less pronounced focus on positive evidence, with 70-90% significant findings supporting the authors’ hypothesis (Fanelli, 2010; T. D. Sterling et al., 1995), but even worse reproducibility rates in the range 11—24% in certain fields (Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011; Steward, Popovich, Dietrich, & Kleitman, 2012). This suggests that while all estimates may not generalize, publication bias may still be of similar magnitude in other fields; but specific fields with a higher proportion of published negative evidence and/or a higher demonstrated reproducibility (Camerer et al., 2016) are likely to be less affected by publication bias.
One should recognize that most findings suppressed from publication describe NULL effects that many may find uninformative or not interesting (Stroebe & Strack, 2014); but the fact that they are never published makes it more likely that similar studies are performed repeatedly by independent researchers; and eventually one will become “significant” by chance, dramatically increasing its chance of being published. Thus, the fact that such a large portion of negative evidence was suppressed from publication not only represents a serious threat to the veracity of published positive evidence; it also means that false theories that have been published may never become “falsified” in the literature, and that researchers are likely to spend time and resources testing hypotheses that should already have been rejected.

Publication bias may be the single most important problem to solve in order to increase the efficiency of the scientific project and bring the veracity of published research to higher standards. The implications of suppressing > 60 negative observations for each one published should not be underestimated. With $\alpha = .05$, we expect a significant finding by chance for every 20 observations made on random data. Thus, our results suggest that even when studied associations are truly NULL, the literature will be dominated by statistically significant findings.

Authors’ contributions

MI conceived of the study, performed the analysis, interpreted results, and drafted the paper. GN participated in interpreting results and writing the paper.

Acknowledgements

We thank professor Anna Dreber Almenberg for comments on a preprint version of this paper.
References

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience, 14*(5), 365–376.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436.

Cohen, J. (1994). The earth is round (p < .05). *The American Psychologist, 49*(12), 997.

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., … Johannesson, M. (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.

Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: a review of three human research domains. *Royal Society Open Science, 4*(2), 160254.

Fanelli, D. (2010). “Positive” results increase down the Hierarchy of the Sciences. *PloS One, 5*(4), e10068.

Ingre, M. (2013). Why small low-powered studies are worse than large high-powered studies and how to protect against “trivial” findings in research: comment on Friston (2012). *NeuroImage, 81*, 496–498.

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the Reproducibility of Psychological Science. *Journal of the American Statistical Association, 112*(517), 1–10.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). https://doi.org/10.1126/science.aac4716

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews. Drug Discovery, 10*(9), 712.

Rossi, J. S. (1990). Statistical power of psychological research: What have we gained in 20 years? *Psycnet.apa.org*. Retrieved from http://psycnet.apa.org/journals/ccp/58/5/646/rpp. (n.d.). Github. Retrieved from https://github.com/CenterForOpenScience/rpp

Sterling, T. D. (1959). Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance--Or Vice Versa. *Journal of the American Statistical Association, 54*(285), 30–34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician, 49*(1), 108–112.

Steward, O., Popovich, P. G., Dietrich, W. D., & Kleitman, N. (2012). Replication and reproducibility in spinal cord injury research. *Experimental Neurology, 233*(2), 597–605.

Stroebbe, W., & Strack, F. (2014). The Alleged Crisis and the Illusion of Exact Replication. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 9*(1), 59–71.

Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797.

Yarkoni, T. (2009). Big Correlations in Little Studies: Inflated fMRI Correlations Reflect Low Statistical Power-Commentary on Vul et al. (2009). *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 4*(3), 294–298.
**Supplemental material**

**Solving the system of equations presented in equation 6 to find $\beta_o$ from $\theta$**

The system of equations defined by equation 6 in the main text (replicated below) needs to be solved for unique values of the reproducibility rate ($R$), the assumed type-2 error rate in the replication studies ($\beta_r$) as well as the type-1 error rate of the original ($\alpha_o$) and replication studies ($\alpha_r$). This can be simplified using a computerized equation solver and cross checking the math of the suggested solution. Syntax for solving the equations using a web based equation solver ([www.wolframalpha.com](http://www.wolframalpha.com)) together with R-code for cross checking the math is given below.

\[
\begin{align*}
\hat{\theta} &= \frac{\theta(1-\beta_o)}{\theta(1-\beta_o)+\alpha_o(1-\theta)} \\
R &= \hat{\theta}(1-\beta_r) + \alpha_r(1-\hat{\theta})
\end{align*}
\]

Eq (6)

Since the picked equation solver was somewhat limited in the choice of symbols, equation 6 was rewritten in plain text like so:

\[
P=(\theta(1-\beta))/((\theta(1-\beta)+\alpha(1-\theta))); \\
R=P*(1-b)+a*(1-P);
\]

To find the solution for the lower bound of the range, in which the true value must fall (discussed in detail in the main text), we assume identical power in the original and replication studies, and added the following constraints:

\[
\alpha=.05; \ R=.36; \ a=.025; \ b=\beta; \ 0<\theta<1;
\]

And to finish the command we added instructions to solve for $\beta_o$ and $\hat{\theta}$:

\[
solve \beta_o \ and \ P
\]

This produced the following solution for the lower bound of the range in which the true value **must** fall (see the main text):
Solving the equation for the upper bound of this range produced the following solution:

\[
\beta = \frac{2527b-67}{2560b} \quad \text{and} \quad P = \frac{67}{105} \quad \text{and} \quad 0 < \theta < 1
\]

\begin{align*}
P &= \frac{(\theta * (1-b))}{(\theta + (1-b) + \alpha * (1-theta))}; \\
R &= P * (1-b) + a * (1-P); \quad \alpha = .05; \\
R &= .36; \quad a = .025; \quad b = 0; \\
0 < \theta < 1; \quad \text{solve beta and P}
\end{align*}

Computed by WolframAlpha

Solving the equation for the lower end of the more narrow likely range (see discussion in the main text) produced the following solution (note, this solution is valid for the whole range \(0 < \theta < 1\) but the solver erroneously produced a constant solution for the specific case of \(\theta = 10/17\)):

\[
\beta = b + \frac{3}{50} \quad \text{and} \quad P = \frac{20(b-1)b}{(20 \beta - 19 \theta - 1)}
\]

\begin{align*}
0 < \theta < \frac{10}{17} \quad \text{and} \quad b = \frac{1}{200} \left(158 - \sqrt{\frac{670}{\theta} + 230}\right) \\
P &= \frac{(\theta * (1-b))}{(\theta + (1-b) + \alpha * (1-theta))}; \\
R &= P * (1-b) + a * (1-P); \quad \alpha = .05; \\
R &= .36; \quad a = .025; \quad b = \text{beta-0.06}; \\
0 < \theta < 1; \quad \text{solve beta and P}
\end{align*}

Computed by WolframAlpha

Solving the equation for the upper end of the likely range produced the following solution (note, similar to above we got an erroneous constant solution for \(\theta = 2/5\)):

\[
\beta = b + \frac{1}{10} \quad \text{and} \quad P = \frac{20(b-1)b}{(20 \beta - 19 \theta - 1)} \quad \text{and} \quad 0 < \theta < \frac{2}{5}
\]

\begin{align*}
0 < \theta < \frac{2}{5} \quad \text{and} \quad b = \frac{1}{200} \left(154 - \sqrt{\frac{670}{\theta} + 6}\right) \\
P &= \frac{(\theta * (1-b))}{(\theta + (1-b) + \alpha * (1-theta))}; \\
R &= P * (1-b) + a * (1-P); \quad \alpha = .05; \\
R &= .36; \quad a = .025; \\
b = \text{beta-0.10}; \quad 0 < \theta < 1; \quad \text{solve beta and P}
\end{align*}

Computed by WolframAlpha
Cross checking the proposed solutions

We crosschecked the solution for beta that was supplied by the equation solver above, by calculating beta for the full range of priors \(0.25 < \theta < 0.75\), apply equation 4 to calculate the posterior probability and then verify that the expected replication rate was \(R = 0.36\) according to equation 5 for the full range of priors.

R-code to reproduce all findings presented in the paper

The code can be downloaded here: https://github.com/micing/publication_bias_psychology

```r
## This R-code reproduces all findings presented in the paper:
## "Estimating statistical power, posterior probability and publication bias
## of psychological research using the observed replication rate"
## by Michael Ingre & Gustav Nilsonne, 2017.

library("Rcurl")
library("pwr")

### Functions ###

# expected total positive findings: equation 3
exp_total_pos <- function(alpha, beta, theta) { theta*(1-beta)+alpha*(1-theta)}

# Bayes theorem: equation 4
theta_hat <- function(alpha, beta, theta) {
  theta*(1-beta) / (theta*(1-beta) + alpha*(1-theta))
}

# Reproducibility: equation 5
R <- function(alpha, beta, theta_hat) {theta_hat*(1-beta)+alpha*(1-theta_hat)}

## Solution for beta (alpha_o=.05, alpha_r=.025, R=.36): equation 6
beta <- function(theta, i=1){
  switch(i,
    {1/200*(164 - sqrt(670/theta + 626))},       ## lower bound
    {(2627*theta - 67)/2560*theta),            ## upper bound
    (1/200*(158 - sqrt(670/theta + 230)))+3/50, ## lower likely
    {1/200*(154 - sqrt(670/theta + 6))}+1/10    ## upper likely
  )
}

# publication bias
bias <- function(alpha, beta, theta, obs_pos=.90){
  (1-exp_total_pos(alpha,beta,theta)) /exp_total_pos(alpha,beta,theta) /
  ((1-obs_pos)/obs_pos)
}
```
# Function to generate data for plotting 

gen_data <- function(theta, beta){
data <- list(theta=theta)
  for (i in 1:4){
    data$beta[[i]] <- beta(theta, i)
data$power[[i]] <- 1 - data$beta[[i]]
data$theta_hat[[i]] <- mapply(theta_hat, .05, data$beta[[i]], theta)
data$exp_total_pos[[i]] <- mapply(exp_total_pos, .05, data$beta[[i]], theta)
data$bias[[i]] <- mapply(bias, .05, data$beta[[i]], theta)
  }
return(data)
}

# plot function 

plotit <- function(d, theta, ylim=c(0,1), log="x", legpos="topleft", xlab="", ylab="", main="") {
  plot(theta, d[[1]], type="n", ylim=ylim, xlab=xlab, ylab=ylab, log=log, main=main, xaxt="n")
  axis(1, at = c(.025, .05, .10, .20, .50, 1))
  polygon(c(theta, rev(theta)), c(d[[1]], rev(d[[2]]))), col="grey80", border=NA)
polygon(c(theta, rev(theta)), c(d[[3]], rev(d[[4]]))), col="black", border=NA)

  if (!is.na(legpos)) {
    legend(legpos, 
      c("Likely range", "Outer boundary"), 
      col=c("black", "grey80"), 
      lwd=c(5, 5), 
      bty="n",
    }
  }
}

### Summarize RPP data by the OSC ###

rppurl=getURL("https://raw.githubusercontent.com/CenterForOpenScience/rpp/master/data_allformats/RPPdataConverted.csv")
rpp=read.csv(text = rppurl)
d=subset(rpp, !is.na(rpp$Replicate.R), select=c("N.O", "N.R", "Replicate.R"))
d$larger=d$R>d$O
cl
big=3
small=1
for(i in 1:4){
  d$larger=d$R*d$O
d$same=d$R==d$O

  sum(d$larger)
  sum(d$same)
  sum(d$smaller)

  c<-
  theta <- seq(.025, .975, .001)
data <- gen_data(theta, beta)
  op <- par(no.readonly = TRUE)
par(mfrow=c(2,2), mar=c(5, 5.5, 2, 1), cex=1)

plotit(data$power, theta=data$theta, legpos="bottomleft",
       xlab="Assumed prior probability",
       ylab="Estimated statistical power\nnot the original studies",
       main="Statistical power")

plotit(data$theta_hat, theta=data$theta, legpos="bottomleft",
       xlab="Assumed prior probability",
       ylab="Estimated posterior probability\nnot the original findings",
       main="Posterior probability")

plotit(data$exp_total_pos, theta=data$theta, legpos="topleft", ylim=c(0, 0.4),
       xlab="Assumed prior probability",
       ylab="Expected proportion of\npositive evidence observed",
       main="Positive evidence")

plotit(data$bias, theta=data$theta, legpos="bottomleft", ylim=c(1, 300), log="xy",
       xlab="Assumed prior probability",
       ylab="Odds for observed negative evidence\nnot to be suppressed from publication",
       main="Publication bias")

par(op)

dev.copy(device = jpeg, filename = 'figure1.jpg', width = 700, height = 700)
dev.off()

### Calculations in text ###

# expeted positive findings
.95*.95+(1-.95)*.05

# expected posterior probability
theta_hat(alpha=.05, beta=1-.95, theta=.95)

# expected reproducibility
R(alpha=.025, beta=.05, theta_hat(alpha=.05, beta=.05, theta=.95))

# testing if the observed reproducibility
# is consistent with the expected reproducibility
# assuming a zero bias literature
binom.test(35, 97, R(alpha=.025, beta=.05, theta_hat(alpha=.05, beta=.05, theta=.95)))

### Median difference in power between original and replication

df_o = 54
df_r = 68
r = seq(.2,.4,.01)

delta_power = pwr.r.test(n=35, r=r)$power - pwr.r.test(n=54, r=r)$power
max(delta_power)
min(delta_power)

## comments for figure 1

# power

data$power[[3]][[match(.050, data$theta)]]
data$power[[4]][[match(.050, data$theta)]]
data$power[[3]][[match(.100, data$theta)]]
data$power[[4]][[match(.100, data$theta)]]
data$power[[3]][[match(.200, data$theta)]]
data$power[[4]][[match(.200, data$theta)]]

data$power[[3]][[match(.070, data$theta)]]
data$power[[4]][[match(.070, data$theta)]]

# posterior

data$theta_hat[[3]][[match(.050, data$theta)]]
data$theta_hat[[4]][[match(.050, data$theta)]]
data$theta_hat[[3]][[match(.100, data$theta)]]
data$theta_hat[[4]][[match(.100, data$theta)]]
data$theta_hat[[3]][[match(.200, data$theta)]]
data$theta_hat[[4]][[match(.200, data$theta)]]

data$theta_hat[[3]][[match(.070, data$theta)]]
data$theta_hat[[4]][[match(.070, data$theta)]]

# publication bias

data$bias[[1]][[match(.975, data$theta)]]
data$bias[[3]][[match(.050, data$theta)]]
data$bias[[4]][[match(.050, data$theta)]]
data$bias[[3]][[match(.20, data$theta)]]
data$bias[[4]][[match(.20, data$theta)]]

# theta=.07

data$power[[3]][[match(.070, data$theta)]]
data$power[[4]][[match(.070, data$theta)]]
data$theta_hat[[3]][[match(.070, data$theta)]]
data$theta_hat[[4]][[match(.070, data$theta)]]

data$exp_total_pos[[3]][[match(.070, data$theta)]]
data$exp_total_pos[[4]][[match(.070, data$theta)]]
data$bias[[3]][[match(.070, data$theta)]]
data$bias[[4]][[match(.070, data$theta)]]

#### Cross checking the math ####

# We verified that the data produced by our solution
# for beta implies R=.36 reproducibility

digits = 16
# lower bound: identical power
signif(min(mapply(R, .025, data$beta[[1]], data$theta_hat[[1]])), digits)
signif(max(mapply(R, .025, data$beta[[1]], data$theta_hat[[1]])), digits)

## upper bound: perfect power
signif(min(mapply(R, .025, 0, data$theta_hat[[2]])), digits)
signif(max(mapply(R, .025, 0, data$theta_hat[[2]])), digits)

## lower likely: +6% power
signif(min(mapply(R, .025, data$beta[[3]]-.06, data$theta_hat[[3]])), digits)
signif(max(mapply(R, .025, data$beta[[3]]-.06, data$theta_hat[[3]])), digits)

## upper likely: +10% power
signif(min(mapply(R, .025, data$beta[[4]]-.10, data$theta_hat[[4]])), digits)
signif(max(mapply(R, .025, data$beta[[4]]-.10, data$theta_hat[[4]])), digits)