A Bayesian perspective on the Reproducibility Project: Psychology

Alexander Etz[a] **and Joachim Vandekerckhove**[b,†]

This is the author final version.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS ONE, 11*, e0149794.

We revisit the results of the recent Reproducibility Project: Psychology by the Open Science Collaboration. We compute Bayes factors—a quantity that can be used to express comparative evidence for an hypothesis but also for the null hypothesis—for a large subset ($N = 72$) of the original papers and their corresponding replication attempts. In our computation, we take into account the likely scenario that publication bias had distorted the originally published results. Overall, 75% of studies gave qualitatively similar results in terms of the amount of evidence provided. However, the evidence was often weak (i.e., Bayes factor < 10). The majority of the studies (64%) did not provide strong evidence for either the null or the alternative hypothesis in either the original or the replication, and no replication attempts provided strong evidence in favor of the null. In all cases where the original paper provided strong evidence but the replication did not (15%), the sample size in the replication was smaller than the original. Where the replication provided strong evidence but the original did not (10%), the replication sample size was larger. We conclude that the apparent failure of the Reproducibility Project to replicate many target effects can be adequately explained by overestimation of effect sizes (or overestimation of evidence against the null hypothesis) due to small sample sizes and publication bias in the psychological literature. We further conclude that traditional sample sizes are insufficient and that a more widespread adoption of Bayesian methods is desirable.

Bayesian inference | Reproducibility project

The summer of 2015 saw the first published results of the long-awaited Reproducibility Project: Psychology by the Open Science Collaboration (Open Science Collaboration, 2015) (henceforth OSC). In an attempt to closely replicate 100 studies published in leading journals, fewer than half were judged to successfully replicate. The replications were pre-registered in order to avoid selection and publication bias and were evaluated using multiple criteria. When a replication was judged to be successful if it reached statistical significance (i.e., $p < .05$), only 39% were judged to have been successfully reproduced. Nevertheless, the paper reports a $.51$ correlation between original and replication effect sizes, indicating some degree of robustness of results (see their Fig. 3).

Much like the results of the project, the reactions in media and social media have been mixed. In a first wave of reactions, headlines ranged from the dryly descriptive "Scientists replicated 100 psychology studies, and fewer than half got the same results" (Handwerk, 2015) and "More than half of psychology papers are not reproducible" (Jump, 2015) to the crass "Study reveals that a lot of psychology research really is just 'psycho-babble' " (Connor, 2015). A second wave of reactions shortly followed. Editorials with titles such as "Psychology is not in crisis" (Feldman Barrett, 2015) and a statement by the American Psychological Association (American Psychological Association, 2015) were quick to emphasize the possibility of many hidden moderators that rendered the replications ineffective. OSC acknowledges this: "unanticipated factors in the sample, setting, or procedure could still have altered the observed effect magnitudes," but it is unclear what, if any, bearing this has on the robustness of the theories that the original publications supported.

In addition to the unresolved possibility of hidden moderators, there is the issue of lacking statistical power. The statistical power of an experiment is the frequency with which it will yield a statistically significant effect in repeated sampling, assuming that the underlying effect is of a given size. All other things—such as the design of the study and the true size of the effect—being equal, statistical power is determined by an experiment's sample size. Low-powered research designs undermine the credibility of statistically significant results in addition to increasing the probability of nonsignificant ones (see (Button et al., 2013) and the references therein for a detailed argument); furthermore, low-powered studies generally provide only small amounts of evidence (in the form of weak Bayes factors; see below).

Among the insights reported in OSC is that "low-power research designs combined with publication bias favoring positive results together produce a literature with upwardly biased effect sizes," and that this may explain why replications—unaffected by publication bias—show smaller effect sizes. Here, we formally evaluate that insight, and use the results of the Reproducibility Project: Psychology to conclude that publication bias and low-powered designs indeed contribute to the poor reproducibility, but also that many of the replication attempts in OSC were themselves underpowered. While the OSC aimed for a minimum of 80% power (with an average of 92%) in all replications, this estimate was based on the observed effect size in the original studies. In the likely event that these observed effect sizes were inflated (see next section), the

[a]University of Amsterdam; [b]University of California, Irvine
[†]To whom correspondence should be addressed. E-mail: joachim@uci.edu.
This work was partly funded by the National Science Foundation grants #1230118 and #1534472 from the Methods, Measurements, and Statistics panel (www.nsf.gov) and the John Templeton Foundation grant #48192 (www.templeton.org). This publication was made possible through the support of a grant from the John Templeton Foundation. The opinions expressed in this publication are those of the authors and do not necessarily reflect the views of the John Templeton Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

PEER REVIEW AUTHOR VERSION

sample size recommendations from prospective power analysis will have been underestimates, and thus replication studies will tend to find mostly weak evidence as well.

Publication bias. Reviewers and editors in psychology journals are known to put a premium on 'positive' results. That is, they prefer studies in which a statistically significant result is used to support the existence of an effect. Nearly six decades ago, Sterling (1959) noted this anomaly in the public record: In four prominent psychology journals, 95 to 99% of studies that performed a significance test rejected the null hypothesis (i.e., $\mathcal{H}_0$). Sterling concludes by noting two key findings, "Experimental results will be printed with a greater probability if the relevant test of significance rejects $\mathcal{H}_0$," and, "The probability that an experimental design will be replicated becomes very small once such an experiment appears in print" (p. 33).

Moreover, it is a truism that studies published in the psychology literature are only a subset of the studies psychologists conduct, and various criteria are used to determine if a study should be published in a given journal. Studies that do not meet the criteria are relegated to lab file drawers (Rosenthal, 1979). A selective preference for publishing studies that reject $\mathcal{H}_0$ is now known as *publication bias*, and is recognized as one cause of the current crisis of confidence in psychology (Pashler & Wagenmakers, 2012).

When journals selectively publish only those studies that achieve statistical significance, average published effect sizes inevitably inflate because the significance threshold acts as a filter; only the studies with the largest effect sizes have sufficiently low *p*-values to make it through to publication. Studies with smaller, non-significant effects are rarely published, driving up the average effect size (Ioannidis, 2008). Readers who wish to evaluate original findings and replications alike must take into account the fact that our "very publication practices themselves are part and parcel of the probabilistic processes on which we base our conclusions concerning the nature of psychological phenomena" (Bakan, 1966, p. 427). Differently put, the publication criteria should be considered part of the experimental design (Walster & Cleary, 1970). For the current project, we choose to account for publication bias by modeling the publication process as a part of the data collection procedure, using a Bayesian model averaging method proposed by Guan and Vandekerckhove (2016) and detailed in Section 2.2 below.

Methods

The Bayes factor. To evaluate replication success we will make use of *Bayes factors* (Etz & Wagenmakers, 2015; Kass & Raftery, 1995). The Bayes factor (B) is a tool from Bayesian statistics that expresses how much a data set shifts the balance of evidence from one hypothesis (e.g., the null hypothesis $\mathcal{H}_0$) to another (e.g., the alternative hypothesis $\mathcal{H}_A$). Bayes factors require researchers to explicitly define the models under comparison.

In this report we compare the null hypothesis of no difference against an alternative hypothesis with a potentially nonzero effect size. Our prior expectation regarding the effect size under $\mathcal{H}_A$ is represented by a normal distribution centered on zero with variance equal to 1 (this is a *unit information prior*, which carries a weight equivalent to approximately one observation; Rouder, Speckman, Sun, Morey, & Iverson, 2009).

Other analysts could reasonably choose different prior distributions when assessing these data, and it is possible they would come to different conclusions. For example, in the case of a replication study specifically, a reasonable choice for the prior distribution of $\mathcal{H}_A$ is the posterior distribution of the originally reported effects (Verhagen & Wagenmakers, 2014). Using the original study's posterior as the replication's prior asks the question, "Does the result from the replication study fit better with predictions made by a null effect or by the originally reported effect?" A prior such as this would lend itself to more extreme values of the Bayes factor because the two hypotheses make very different predictions; the null hypothesis predicts replication effect sizes close to zero, whereas the original studies' posterior distributions will typically be centered on relatively large effect sizes and hence predict large replication effect sizes. As such, Bayes factors for replications that find small-to-medium effect sizes will often favor $\mathcal{H}_0$ (δ=0) over the alternative model that uses the sequential prior because the replication result poorly fits the predictions made by the original posterior distribution, whereas small-to-medium effects will yield less forceful evidence in favor of $\mathcal{H}_0$ over the alternative model using the unit information prior that we apply in this analysis.

There are two main reasons why, in the present paper, we choose to use the unit information prior over this sequential prior. First, our goal is not to evaluate how well empirical results reproduce, but rather to see how the *amount of evidence* gathered in an original study compares to that found in an *independent* replication attempt. This question is uniquely addressed by computing Bayes factors on two data sets, using identical priors. Compared to the sequential prior, the unit information prior we have chosen for our analysis is somewhat conservative, meaning that it requires more evidence before strongly favoring $\mathcal{H}_0$ in a replication study. Indeed, results presented in a blog post by the first author (Etz, 2015) suggest that when a sequential prior is used approximately 20% of replications show strong evidence favoring $\mathcal{H}_0$, as opposed to no replications strongly favoring $\mathcal{H}_0$ with the unit information prior used in this report. Of course, it is to be expected that different analysts obtain different answers with different priors, because they are asking different questions (as Sir Harold Jeffries, 1939, famously quipped: "It is sometimes considered a paradox that the answer depends not only on the observations but on the question; it should be a platitude," p. vi).

A second reason we do not use the sequential prior in this report is that it does not take into account publication bias. Assuming that publication bias has a greater effect on the original studies than it did on the (pre-registered, certain to be published regardless of outcome) replications, the observed effect sizes in original and replicate studies are not expected to be equal. Using the original posterior distribution as a prior in the replication study would penalize bias in the original result; since the replication attempts will nearly always show smaller effect sizes than the biased originals, it will be more common to 'fail to replicate' these original findings (by accumulating evidence in favor of $\mathcal{H}_0$ in the replication). However, here we are interested in evaluating the evidential support for the effects in the replication, rather than using them to quantify the effect of publication bias. In other words, we are interested in answering the following question: If we treat the two results as independent, do they provide similar degrees of evidence?

Interpretation of the Bayes factor. The Bayes factor is most conveniently interpreted as the degree to which the data sway our belief from one to the other hypothesis. In a typical situation, assuming that the reader has no reason to prefer the null hypothesis over the alternative before the study (i.e., $1:1$ odds, or both have a prior probability of $.50$), a Bayes factor of 3 in favor of the alter-

Table 1. Descriptive labels for certain Bayes factors.

Label	B	$p(\mathcal{H}_A\|\text{data})^a$
Strongly support $\mathcal{H}_A$	10	91%
Weakly support $\mathcal{H}_A$	3	75%
Ambiguous information	1	50%
Weakly support $\mathcal{H}_0$	$1/3$	25%
Strongly support $\mathcal{H}_0$	$1/10$	9%

^{a}: $p(\mathcal{H}_A|\text{data})$ is the posterior probability of $\mathcal{H}_A$ assuming prior equiprobability between $\mathcal{H}_0$ and $\mathcal{H}_A$.

native will change their odds to $3:1$ or a posterior probability of .75 for $\mathcal{H}_A$. Since a Bayes factor of 3 would carry a reader from equipoise only to a 75% confidence level, we take this value to represent only weak evidence. Put another way, accepting a 75% posterior probability for $\mathcal{H}_A$ means that the reader would accept a one-in-four chance of being wrong. To put that in a context: that is the probability of correctly guessing the suit of a randomly-drawn card; and the researcher would reasonably prefer to bet on being wrong than on a fair die coming up six. That is to say, it is evidence that would not even be convincing to an uninvested reader, let alone a skeptic (who might hold, say, $10:1$ prior odds against $\mathcal{H}_A$). Table 1 provides posterior probabilities associated with certain Bayes factors B assuming prior odds of $1:1$. In that table, we have also added some descriptive labels for Bayes factors of these magnitudes (these labels are similar in spirit to those suggested by Jeffreys, 1939). Finally, it bears pointing out that if a researcher wants to move the probability of $\mathcal{H}_0$ from 50% to below 5%, a Bayes factor of at least 19 is needed.

It is important to keep in mind that the Bayes factor as a measure of evidence must always be interpreted in the light of the substantive issue at hand: For extraordinary claims, we may reasonably require more evidence, while for certain situations—when data collection is very hard or the stakes are low—we may satisfy ourselves with smaller amounts of evidence. For our purposes, we will only consider Bayes factors of 10 or more as evidential—a value that would take an uninvested reader from equipoise to a 91% confidence level (a level at which an unbiased, rational reader is willing to bet up to ten cents on $\mathcal{H}_A$ to win back one cent if they are right). Since the Bayes factor represents the evidence from the sample, readers can take these Bayes factors and combine them with their own personal prior odds to come to their own conclusions.

Mitigation of publication bias. The academic literature is unfortunately biased. Since studies in which the null hypothesis is confidently rejected are published at a higher rate than those in which it is not, the literature is "unrepresentative of scientists' repeated samplings of the real world" (Young, Ioannidis, & Al-Ubaydli, 2008). A retrospective analysis of published studies must therefore take into account the fact that these studies are somewhat exceptional in having passed the so-called *statistical significance filter* (Ioannidis, 2008).

Guan and Vandekerckhove (2016) define four censoring functions that serve as models of the publication process. Each of these censoring functions formalizes a statistical significance filter, and each implies a particular *expected distribution of test statistics that make it to the literature*. The first, a *no-bias model*, where significant and non-significant results are published with equal probability, implies the typical central and non-central t distributions (for null and non-null effects, respectively). The second, an *extreme-bias model*, indexes a process where non-significant results are never published. This model assigns nonzero density only to regions where significant results occur (i.e., $p < .05$) and nowhere else. The third, a *constant-bias model*, indexes a process where non-significant results are published at a rate that is some constant π ($0 \leq \pi \leq 1$) times the rate at which significant results are published. These distributions look like typical *t* distributions but with the central (non-significant) region weighted down, creating large spikes in density over critical regions in the t-distribution. The fourth, an *exponential-bias model*, indexes a process where the probability that non-significant results are published decreases exponentially as $(p - \alpha)$ increases (i.e., "marginally significant" results have a moderately high chance of being published). These distributions have spikes in density around critical t-values. Fig. 1 shows the predicted distribution of published t values under each of the four possible censoring functions, with and without a true effect.

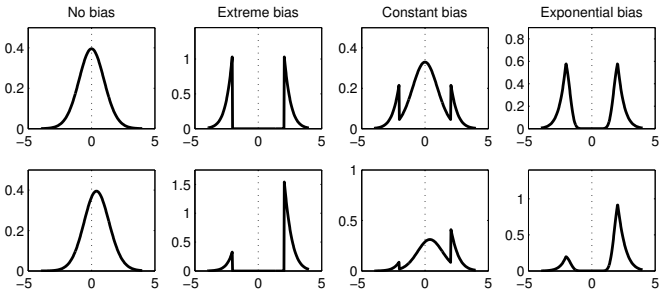

Fig. 1. Predicted distributions of t statistics in the literature. Predicted distributions are shown under the four censoring mechanisms we consider (columns) and two possible states of nature (top row: $\mathcal{H}_0$ true ($\delta = 0$); bottom row: $\mathcal{H}_0$ false ($\delta \neq 0$)).

None of these censoring functions are likely to capture the exact nature of the publication process in all of the cases we consider, but we believe they span a reasonable range of possible processes. Assuming that these four models reasonably represent possible statistical significance filters, we can use a Bayesian model averaging method to compute a single *mitigated Bayes factor* (B^M) that takes into account that a biased process may have led to the published effect. The procedure essentially serves to raise the evidentiary bar for published studies if publication bias was not somehow prevented (e.g., through pre-registration). A unique feature of this method (compared to other bias mitigation methods such as PET-PEESE; Stanley & Doucouliagos, 2014) is that it allows us to quantify mitigated evidence for or against the null hypothesis on a continuous scale—a feature that will become useful when we compare original and replicated studies, below.

Calculation of the mitigated Bayes factor. To calculate B^M, we first define a likelihood function in which the t distribution is multiplied by a weighting function w, so that

$$p_w^+(x|n,\delta,\theta) \propto t_n(x|\delta)\, w(x|\theta)\,. \qquad [1]$$

Here, x is the reported t-value, n stands for the associated degrees of freedom, δ is the effect size parameter of the noncentral t distribution, and w is one of the four censoring functions which has optional parameters θ (see Table 2 for details regarding weighting functions).

Equation 1 describes four possible models, each with some effect size δ. Together, these four models form the alternative hypothesis $\mathcal{H}_A$. We construct four additional models in which $\delta = 0$ (i.e., there is no underlying effect): $p_w^-(x|n,\theta) = p_w^+(x|n,\delta=0,\theta)$.

Here the t distribution reduces to the central t, and these four models together form the null hypothesis $\mathcal{H}_0$.

Second, we obtain the Bayesian evidences E_w^+ and E_w^- by integrating the likelihood for each model over the prior:

$$\begin{aligned} E_w^+ &= \int_\Theta \int_\Delta p_w^+ (x|n, \delta, \theta)\, p(\delta) p(\theta) d\delta d\theta \\ E_w^- &= \int_\Theta p_w^- (x|n, \theta)\, p(\theta) d\theta. \end{aligned}$$

E_w^+ and E_w^- are also known as the *marginal likelihoods* of these models (i.e., the probability density of the data under the model, as a prior-weighted average over all possible parameter constellations), and they can be conveniently approximated with Gaussian quadrature methods (Abramowitz, Stegun, et al., 1972).

Finally, the posterior probability of each hypothesis can be calculated by (1) multiplying each evidence value with the corresponding model prior (where a 'model' is any one of the eight possible combinations of weighting function w and the null or alternative hypothesis; see Fig. 1); (2) dividing each of those products with the sum of all such products for all models; and (3) summing the posterior probabilities for all models within an hypothesis. This can be rearranged to yield the following expression for the posterior:

$$Pr\left(\mathcal{H}_A|x\right) = \frac{Pr\left(\mathcal{H}_A\right) \times \sum_w Pr\left(w\right) E_w^+}{\sum_k Pr\left(k\right) \left[Pr\left(\mathcal{H}_A\right) E_k^+ + Pr\left(\mathcal{H}_0\right) E_k^-\right]},$$

where $Pr\left(w\right)$ is the prior probability of censoring function w and $Pr\left(\mathcal{H}_A\right)$ is the prior probability that there is a nonzero effect. To obtain the Bayes factor, we restate in terms of posterior and prior ratios to obtain the simple expression:

$$\underbrace{\frac{Pr\left(\mathcal{H}_A|x\right)}{Pr\left(\mathcal{H}_0|x\right)}}_{\text{Posterior odds}} = \underbrace{\frac{Pr\left(\mathcal{H}_A\right)}{Pr\left(\mathcal{H}_0\right)}}_{\text{Prior odds}} \times \underbrace{\frac{\sum_w Pr(w) E_w^+}{\sum_w Pr(w) E_w^-}}_{\text{Mitigated Bayes factor}},$$

where the second factor on the right hand side now represents the mitigated Bayes factor B^M. Full details and MATLAB/Octave code to implement the procedure can be found here: `http://bit.ly/1Nph9xQ`.

Sample. We limited our analysis to studies that relied on univariate tests in order to apply the statistical mitigation method developed by Guan and Vandekerckhove (2016). A total of $N = 72$ studies were eligible. This includes all studies that relied on t-tests, univariate F-tests, and univariate regression analyses. This limits the generality of our conclusions to these cases, which fortunately constitute the bulk of studies in the Reproducibility Project: Psychology. A list of included studies and their inferential statistics is provided in the Supporting Information. Additionally, we conducted a sensitivity analysis varying the scale of the prior distribution among reasonable values ($.5$ to 2.0); this revealed no concerns that affect the conclusions or recommendations of the present analysis.

Table 2. The four weighting functions.

Model	Weight w if $p > .05$	Parameters θ
No bias	$w(x) = 1$	None
Extreme bias	$w(x) = 0$	None
Constant bias	$w(x|\pi) = \pi$	π
Exponential bias	$w(x|\lambda) = e^{(-\lambda(p-.05))}$	λ

Note: $w(x)$ is always 1 for results that are statistically significant at the .05-level. The dependency on the design and data properties that determine statistical significance is implied.

Results

Evidence in the original studies, taken at face value. For the original studies, we first computed "face value" Bayes factors that do not take into account the possibility of a biased publication process. By this measure, we find that 31 of the original studies (43%) provide strong support for the alternative hypothesis ($B \geq 10$). No studies provide strong evidence for the null hypothesis. The remaining 57% provide only weak evidence one way or the other.

The small degrees of evidence provided by these published reports, taken at face value, are consistent with observations by Wetzels et al. (2011) as well as the cautionary messages by Johnson (2013) and Maxwell, Lau, and Howard (2015).

Evidence in the original studies, corrected for publication bias. When we apply the statistical mitigation method of Guan and Vandekerckhove (2016), the evidence for effects generally shrinks. After correction for publication bias, only 19 (26%) of the original publications afford strong support for the alternative hypothesis ($B^M \geq 10$). A sizable majority of studies (53, or 74%) provide only ambiguous or weak information, with none finding strong evidence in favor of the null.

Evidence in the replication studies. The set of replication studies was entirely preregistered, with all data sets fully in the open and no opportunity for publication bias to muddy the results. Hence, no mitigation of bias is called for. Of the 72 replication studies, 15 (21%) strongly support the alternative hypothesis ($B^R \geq 10$) and none strongly support the null. Twenty-seven (38%) provide only ambiguous information, and another 25 (35%) provide weak evidence for the null hypothesis.

Consistency of results. One of the stated goals of the Reproducibility Project: Psychology was to test whether previously found effects would obtain in an identical replication of a published study. Focusing on Bayesian evidence, we can now evaluate whether similar studies support similar conclusions. In 46 cases (64%), neither the original study nor the replication attempt yielded strong evidence (i.e., $B \geq 10$). In only 8 cases (11%) did both the original study and the replication strongly support the alternative hypothesis. In 11 cases (15%) the original study strongly supported the alternative but the replication did not, and in 7 cases (10%) the replication provided strong evidence for the alternative whereas the original did not. The frequencies of these Bayes factor transitions are given in Table 3.

Fig. 2 shows (in logarithmic coordinates) the Bayes factor of the replication B^R plotted against the bias-corrected Bayes factor of the original result B^M. The majority of cases in which neither the original nor the replication provided strong evidence are displayed as the cluster of small crosses in the lower left of the figure. Circles represent cases where at least one of the attempts yielded strong evidence.

The observation that there are only 8 cases where both original and replication find strong evidence for an effect, while there are 18 cases in which one does and the other does not, seems at first to indicate a large degree of inconsistency between pairs of otherwise similar studies. To explain this inconsistency, Fig. 2 highlights a major difference between each original and replication:

AUTHOR VERSION

Table 3. Consistency of Bayes factors across original and replicate studies. Columns indicate the magnitude of the mitigated Bayes factor from the original study, and rows indicate the magnitude of the Bayes factor obtained in the replication project.

		Mitigated Bayes factor (original study)					
		$0 - 1/10$	$1/10 - 1/3$	$1/3 - 3$	$3 - 10$	$10 - \infty$	*sum*
Replication	$0 - 1/10$	0	0	0	0	0	0
study	$1/10 - 1/3$	0	0	18	4	3	25
"face-value"	$1/3 - 3$	0	0	16	4	7	27
Bayes	$3 - 10$	0	0	3	1	1	5
factor	$10 - \infty$	0	1	6	0	8	15
	sum	0	1	43	9	19	72

The chosen sample size. The size of the circles indicates the ratio of the replication sample size to the original sample size. In each of the 11 cases where the original study supported the alternative but the replication did not, the original study had the larger sample size. In each of the 7 cases where the replication provided strong evidence for the alternative but the original did not, it was the replication that had the larger sample size.

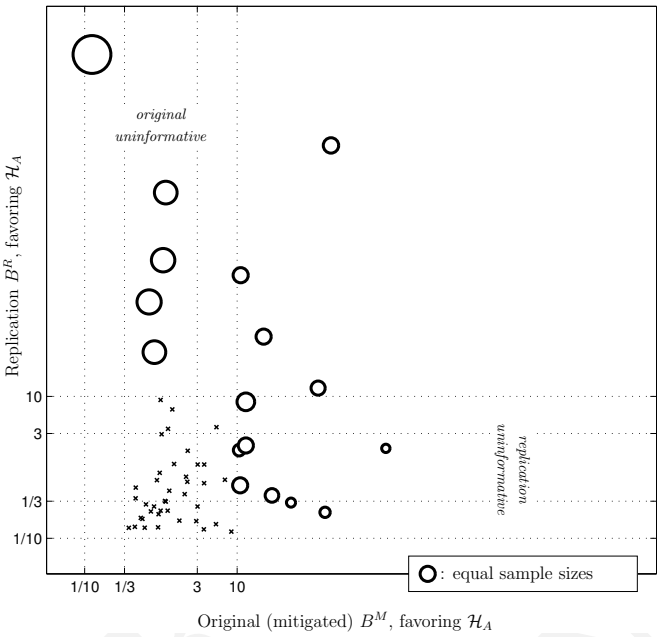

Fig. 2. Evidence resulting from replicated studies plotted against evidence resulting from the original publications. For the original publications, evidence for the alternative hypothesis was calculated taking into account the possibility of publication bias. Small crosses indicate cases where neither the replication nor the original gave strong evidence. Circles indicate cases where one or the other gave strong evidence, with the size of each circle proportional to the ratio of the replication sample size to the original sample size (a reference circle appears in the lower right). The area labeled 'replication uninformative' contains cases where the original provided strong evidence but the replication did not, and the area labeled 'original uninformative' contains cases where the reverse was true. Two studies that fell beyond the limits of the figure in the top right area (i.e., that yielded extremely large Bayes factors both times) and two that fell above the top left area (i.e., large Bayes factors in the replication only) are not shown. The effect that relative sample size has on Bayes factor pairs is shown by the systematic size difference of circles going from the bottom right to the top left. All values in this figure can be found in the Appendix.

Discussion

Small sample sizes and underpowered studies are endemic in psychological science. Publication bias is the law of the land. These two weaknesses of our field have conspired to create a literature that is rife with false alarms (Ioannidis, 2005). From a Bayesian reanalysis of the Reproducibility Project: Psychology, we conclude that one reason many published effects fail to replicate appears to be that the evidence for their existence was unacceptably weak in the first place.

Crucially, our analysis revealed no obvious inconsistencies between the original and replication results. In no case was an hypothesis strongly supported by the data of one team but contradicted by the data of another. In fact, in 75% of cases the replication study found qualitatively similar levels of evidence to the original study, after taking into account the possibility of publication bias. In many cases, one or both teams provided only weak or ambiguous evidence, and whenever it occurred that one team found strong evidence and the other did not, this was easily explained by (sometimes large) differences in sample size. The apparent discrepancy between the original set of results and the outcome of the Reproducibility Project can be adequately explained by the combination of deleterious publication practices and weak standards of evidence, without recourse to hypothetical hidden moderators.

The Reproducibility Project: Psychology is a monumental effort whose preliminary results are already transforming the field. We conclude with the simple recommendation that, whenever possible, empirical investigations in psychology should increase their planned replication sample sizes beyond what is implied by power analyses based on effect sizes in the literature. Our analysis in that sense echoes that of Fraley and Vazire (2014).

Decades of reliance on orthodox statistical inference—which is known to overstate the evidence against a null hypothesis (Berger & Delampady, 1987; Berger & Sellke, 1987; Edwards, Lindman, & Savage, 1963; Goodman, 2008)—have obfuscated the widespread problem of small samples in psychological studies in general and in replication studies specifically. While 92% of the original studies reached the statistical significance threshold ($p < .05$), only 43% met our criteria for strong evidence, with that number shrinking further to 26% when we took publication bias into account. Furthermore, publication bias inflates published effect sizes. If this inflationary bias is ignored in prospective power calculations then replication attempts will systematically tend to be underpowered, and subsequently will systematically obtain only weak or ambiguous evidence. This appears to have been the case in the Reproducibility Project: Psychology.

A major selling point of Bayesian statistical methods is that sample sizes need not be determined in advance (Rouder, 2014), which allows analysts to monitor the incoming data and stop data collection when the results are deemed adequately informative; see (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) for more detail and see (Matzke et al., 2015) for an implementation of this kind of sampling plan, and also see (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, in press) for a detailed step-by-step guide and discussion of this design. Subsequently, if the planned sample size is reached and the results remain uninformative, more data can be collected or else researchers can stop and simply acknowledge the ambiguity in their results. Free and easy-to-use software now exists that allows this brand of sequential analysis (e.g., JASP; Love et al., 2015).

This is the first of several retrospective analyses of the Reproducibil-

ity Project data. We have focused on a subset of the reproduced studies that are based on univariate tests in order to account for publication bias. Other retrospectives include those that focus on Bayes factors and Bayesian effect size estimates (Marsman et al., 2015).

References

Abramowitz, M., Stegun, I. A., et al. (1972). *Handbook of mathematical functions* (Vol. 1) (No. 5). Dover New York.

American Psychological Association. (2015). *Science paper shows low replicability of psychology studies.* Retrieved from http://bit.ly/1NogtsD (Accessed: 2015-10-31)

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*, 423–437.

Berger, J. O., & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, *2*, 317–352.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*, 112–139.

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

Connor, S. (2015). *Study reveals that a lot of psychology research really is just 'psycho-babble'.* Retrieved from http://ind.pn/1RO7hby (Accessed: 2015-10-31)

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*, 193–242.

Etz, A. (2015). *The bayesian reproducibility project.* Retrieved from http://bit.ly/1U1YO9J (Accessed: 2015-10-31)

Etz, A., & Wagenmakers, E.-J. (2015). Origin of the bayes factor. *Preprint available: arXiv:1511.08180 [stat.OT]*. Retrieved from http://arxiv.org/abs/1511.08180

Feldman Barrett, L. (2015). *Psychology is not in crisis.* Retrieved from http://nyti.ms/1PInTEg (Accessed: 2015-10-31)

Fraley, R. C., & Vazire, S. (2014). The n-pact factor: evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE*, *9*, e109019.

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. In *Seminars in hematology* (Vol. 45, pp. 135–140).

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin & Review*, *23*(1), 74–86. doi:

Handwerk, B. (2015). *Scientists replicated 100 psychology studies, and fewer than half got the same results.* Retrieved from http://bit.ly/1OYZVHY (Accessed: 2015-10-31)

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, 696–701.

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648.

Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*(48), 19313–19317. doi:

Jump, P. (2015). *More than half of psychology papers are not reproducible.* Retrieved from http://bit.ly/1GwLHGh (Accessed: 2015-10-31)

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., ... Wagenmakers, E.-J. (2015). *JASP [computer software].* Retrieved from https://jasp-stats.org/

Marsman, M., Dablander, F., Baribault, B., Etz, A., Gronau, Q. F., Jamil, T., ... Wagenmakers, E.-J. (2015). *A Bayesian reanalysis of the reproducibility project.* Manuscript in preparation.

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487.

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

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225-237.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (in press). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*.

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*, 60–78.

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*, 30–34.

Verhagen, A. J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 627–633.

Walster, G. W., & Cleary, T. A. (1970). A proposal for a new editorial policy in the social sciences. *The American Statistician*, *24*(2), 16–19.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298. doi:

Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, *5*(10), e201.

Appendix

S1 Table. Inferential statistics for each of the 72 studies and their replication attempts.

Relevant statistics for each of the 72 included studies. Note that Bayes factors are presented on the $\log_{10}$ scale, so positive values favor $\mathcal{H}_A$ and negative values favor $\mathcal{H}_0$; $|\log_{10}(BF)| > 1$ indicate strong evidence favoring the respective hypothesis.

study	original		replicate		$\log_{10}$ Bayes factor		
number	df	t-value	df	t-value	original	mitigated	replicate
1	13	2.6665	28	0.7937	0.6441	0.1108	-0.3291
2	23	3.7027	23	1.1314	1.5810	0.8406	-0.1752
3	24	2.3000	31	1.2272	0.4905	-0.0513	-0.1799
4	190	3.2388	268	0.1000	1.3270	0.5640	-0.8735
5	31	2.8948	47	0.9327	0.9802	0.3126	-0.3767
6	23	3.5500	31	2.4000	1.4556	0.7298	0.5672
7	99	10.1800	14	0.4960	13.9830	13.1400	-0.2826
8	37	4.1267	31	0.6197	2.2564	1.4573	-0.3953
10	28	5.1662	29	6.7283	3.0554	2.2314	4.5391
11	21	4.1593	29	2.8397	1.8813	1.1144	0.9241
15	94	1.9290	241	3.9550	0.0730	-0.1539	2.3326
19	31	3.7683	19	1.9134	1.7949	1.0271	0.2423
20	94	2.2294	106	0.2000	0.3236	-0.2641	-0.7108
24	152	4.8141	48	2.0543	3.7866	2.9503	0.2679
26	94	1.5811	92	1.3964	-0.1753	-0.3290	-0.2842
27	31	2.2738	70	3.4326	0.4696	-0.0854	1.6253
28	31	2.0248	90	0.9849	0.2879	0.0638	-0.4829
29	7	2.8920	14	3.7080	0.5192	0.0899	1.2588
32	36	4.7833	37	3.3347	2.9577	2.1352	1.4245
33	39	3.7700	39	2.0800	1.8938	1.1147	0.3089
36	20	4.5596	20	4.1653	2.1323	1.3475	1.8434
37	11	2.1909	17	1.5395	0.3697	0.1730	0.0476
44	67	3.0800	176	2.0160	1.2134	0.4831	0.0398
48	92	-2.2200	192	-0.7255	0.3186	-0.2666	-0.7393
49	34	2.3833	86	0.2828	0.5528	-0.0301	-0.6593
52	131	2.4062	111	0.9950	0.4373	-0.1962	-0.5215
53	31	2.2672	73	0.6573	0.4646	-0.0891	-0.5524
56	99	4.0768	38	-0.2600	2.5232	1.7072	-0.4970
58	182	2.2891	278	0.6132	0.2790	-0.3413	-0.8382
61	108	-2.3400	220	0.0700	0.4038	-0.2116	-0.8509
63	68	2.3495	145	0.8911	0.4744	-0.1317	-0.6207
65	41	3.0659	131	0.1342	1.1730	0.4637	-0.7584
68	116	2.0372	222	0.0447	0.1246	-0.4201	-0.8525
71	373	4.4000	175	0.9730	2.9768	2.1537	-0.6328
72	257	3.4029	247	0.7000	1.5031	0.7229	-0.8005
81	90	2.6420	137	1.1958	0.7253	0.0539	-0.4730

study	original		replicate		$\log_{10}$ Bayes factor		
number	df	t-value	df	t-value	original	mitigated	replicate
87	51	3.0757	47	0.0894	1.2009	0.4805	-0.5511
89	26	0.7200	26	0.1500	-0.3374	-0.3756	-0.4331
93	83	3.0500	68	-1.1240	1.1752	0.4430	-0.3673
94	26	1.8700	59	2.3250	0.1981	0.0087	0.4679
97	73	3.4914	1486	1.4248	1.7004	0.9244	-0.9051
106	34	2.4083	45	1.5340	0.5730	-0.0149	-0.0775
107	84	2.0900	156	1.3180	0.2209	-0.3317	-0.4358
110	278	11.1077	142	1.0909	20.9560	20.1120	-0.5321
111	55	2.6230	116	2.4960	0.7462	0.0937	0.5443
112	9	2.9496	9	3.4059	0.6473	0.1493	0.8169
113	124	10.3600	175	15.6400	15.4490	14.6050	31.4220
114	30	3.8066	30	4.7191	1.8159	1.0472	2.7092
115	31	3.2300	8	-1.4260	1.2825	0.5684	0.0489
116	172	3.9400	139	4.0200	2.3526	1.5385	2.4698
118	111	2.3046	158	0.6156	0.3665	-0.2416	-0.7251
120	29	2.2123	41	1.6533	0.4258	-0.1123	0.0091
122	7	2.7600	16	-9.5900	0.4803	0.0636	3.8735
124	34	2.4269	68	0.2828	0.5880	-0.0035	-0.6110
127	28	4.9800	25	-3.1030	2.8817	2.0618	1.1170
129	26	2.0421	64	0.1414	0.3105	0.0892	-0.6110
133	23	2.3875	37	2.8425	0.5513	-0.0038	0.9505
134	115	2.3030	234	8.8360	0.3596	-0.2489	13.5090
135	562	-0.1100	3511	-6.3100	-0.9042	-0.9044	5.8207
136	28	3.0400	56	-0.7700	1.0900	0.4085	-0.4664
145	76	10.4757	36	5.1730	13.1250	12.2820	3.3895
146	14	3.2000	11	1.9000	0.9709	0.3512	0.2339
148	194	2.6758	259	0.4858	0.6592	-0.0370	-0.8442
149	194	2.6758	314	0.3240	0.6592	-0.0370	-0.8799
150	13	3.7683	18	0.9000	1.2348	0.5677	-0.2222
151	41	2.7946	124	0.0316	0.9115	0.2428	-0.7509
153	7	4.4500	7	0.3200	0.8838	0.3464	-0.2037
154	68	3.9275	14	0.4141	2.2479	1.4437	-0.2958
155	51	2.3286	70	0.2846	0.4848	-0.1069	-0.6167
158	38	2.4920	93	4.3520	0.6405	0.0299	2.9206
161	44	3.6633	44	1.1987	1.8164	1.0407	-0.2521
167	17	3.0545	21	1.2042	0.9613	0.3306	-0.1315

IN REVIEW VERSION