



Scientific self-correction: the Bayesian way

Felipe Romero¹ · Jan Sprenger²

Received: 11 March 2019 / Accepted: 12 May 2020
© The Author(s) 2020

Abstract

The enduring replication crisis in many scientific disciplines casts doubt on the ability of science to estimate effect sizes accurately, and in a wider sense, to self-correct its findings and to produce reliable knowledge. We investigate the merits of a particular countermeasure—replacing null hypothesis significance testing (NHST) with Bayesian inference—in the context of the meta-analytic aggregation of effect sizes. In particular, we elaborate on the advantages of this Bayesian reform proposal under conditions of publication bias and other methodological imperfections that are typical of experimental research in the behavioral sciences. Moving to Bayesian statistics would not solve the replication crisis single-handedly. However, the move would eliminate important sources of effect size overestimation for the conditions we study.

Keywords Statistical inference · Replication crisis · Self-corrective thesis · Bayesian statistics · Null hypothesis significance testing (NHST) · Statistical reform

1 Introduction

In recent years, several scientific disciplines have been facing a **replication crisis**: researchers fail to reproduce the results of previous experiments when copying the original experimental design. By investigating replication rates for the main reported effect in a representative sample of published papers, scientists have tried to assess the seriousness of the crisis in a systematic way. The outcome of these studies is sobering: the number of statistically significant findings and the observed effect

✉ Felipe Romero
c.f.romero@rug.nl

Jan Sprenger
jan.sprenger@unito.it

¹ Faculty of Philosophy, University of Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands

² Center for Logic, Language and Cognition (LLC) Department of Philosophy and Education Sciences, Università degli Studi di Torino, Palazzo Nuovo, Via Sant’Ottavio 20, 10124 Torino, Italy

sizes are often much lower than the theoretical expectation (for the fields of psychology, experimental economics and cancer biology, respectively: Open Science Collaboration 2015; Camerer et al. 2016; Nosek and Errington 2017). While the appropriate interpretation of replication failures is debatable (e.g., Maxwell et al. 2015), there is a shared sentiment that science is not as reliable as it is supposed to be and that something needs to change.

There are several causes of low replicability and hence a wide range of possible reforms to address the crisis. We identify three types of reforms that can be regarded as complementary rather than mutually exclusive. First, **social reforms**, which are inspired by the prevalence of questionable research practices (“QRPs”: Simmons et al. 2011) and more generally, the adverse effects of social and structural factors in science (Bakker et al. 2012; Nuijten et al. 2016; Romero 2017). Social reforms include educating researchers about statistical cognition and methodology (Schmidt 1996; Lakens 2019), but also creating greater incentives for replication work—for example by publishing and co-citing replications alongside original studies (Koole and Lakens 2012) or establishing a separate reward system for confirmatory research (Romero 2018). Second, there are **methodological reforms** such as pre-registering studies and their data analysis plan (Quintana 2015), sharing experimental data for “successful” as well as “failed” studies (van Assen et al. 2014; Munafò et al. 2017) and promoting multi-site experiments (Klein et al. 2014). By “front-staging” important decisions about experimental design and data analysis, these reforms address various forms of post-hoc bias (e.g., selective reporting, adding covariates) and increase the transparency and reliability of published research (see also Freese and Peterson 2018). Third, numerous authors identify “classical” statistical inference based on Null Hypothesis Significance Testing (NHST) as a major cause of the replication crisis (Cohen 1994; Goodman 1999a; Ioannidis 2005; Ziliak and McCloskey 2008) and suggest **statistical reforms**. Some of them remain within the frequentist paradigm and promote novel tools for hypothesis testing (Lakens et al. 2018b) or focus on effect sizes and confidence intervals instead of p -values (Fidler 2005; Cumming 2012, 2014). Others are more radical and propose to replace NHST by Bayesian inference (Goodman 1999b; Rouder et al. 2009; Lee and Wagenmakers 2014), likelihood-based inference (Royall 1997), or even purely descriptive data summaries (Trafimow and Marks 2015).

While science most likely needs a combination of these reforms to improve (e.g., Ioannidis 2005; Romero 2019), we study in this paper the case for statistical reform, and its interaction with various limitations in scientific research (e.g., insufficient sample size, selective reporting of results). In other words, we ask whether the replicability of published research would change if we replaced the conventional NHST method by Bayesian inference.

To address this question, we conduct a systematic computer simulation study that investigates the **self-corrective nature of science** in the context of statistical inference. A strong version of the self-corrective thesis (SCT, Laudan 1981) asserts that scientific method guarantees convergence to true theories in the long run: by staying on the path of scientific method, errors in published research will eventually be discovered, corrected and weeded out (see also Peirce 1931; Mayo 1996). SCT can be operationalized in the context of statistical inference and the

replication crisis in the sense that sequential replications of an experiment will eventually “reveal the truth” (Romero 2016).

SCT* Given a series of exact replications of an experiment, the meta-analytical aggregation of their effect sizes will converge on the true effect size as the length of the series of replications increases.

Arguably, validating SCT* in the precisely defined context of exact replications (i.e., experiments that copy the original design) would be a minimal condition for any of the more far-reaching claims that science eventually corrects errors and converges to the truth. Conversely, if SCT* fails—and the replication crisis provides some preliminary evidence that we should not take SCT* for granted—then claims to the general truth of SCT, and to science as a reliable source of knowledge, are highly implausible.

The truth or falsity of SCT* strongly depends on the conditions in which experimental research operates—in particular on the prevalent kind of **publication bias**, that is, the bias in the process of publishing scientific evidence and disseminating it to the scientific community. Since different statistical frameworks (e.g., NHST and Bayesian inference) classify the same set of experimental results in different qualitative categories, e.g., “strong evidence for the hypothesis”, “moderate evidence”, “inconclusive evidence”, etc., the dominant statistical framework will affect the form and extent of publication bias. This affects, in turn, the accuracy of the meta-analytic effect size estimates and the validity of SCT*.

Our paper studies the validity of SCT* in both statistical frameworks under various conditions that relate to the social dimension of science: in particular, the conventions and biases that affect experimental design and data reporting. We model publication bias in NHST as suppressing (a large percentage of) statistically non-significant results, and in Bayesian inference, as suppressing inconclusive evidence—that is, outcomes that yield Bayes factors in the interval $(\frac{1}{3}; 3)$. Then, under various imperfections that are typical of scientific practice, Bayesian inference yields more accurate effect size estimates than NHST, sometimes significantly so. This makes the long-run estimation of unknown effects more reliable. The results do not imply that Bayesian inference also outperforms other forms of frequentist inference, such as equivalence testing (Lakens et al. 2018b) or pure estimation-based inference (Cumming 2012, 2014)—they just highlight its advantages with respect to the traditional, and still widely endorsed, method of NHST.

The paper is structured as follows: Sect. 2 briefly explains the two competing statistical paradigms—frequentist inference with NHST and Bayesian inference. Section 3 describes the simulation model and the statistical and social factors it includes. Sections 4–6 present the results of multiple simulation scenarios that allow us to evaluate and contrast NHST and Bayesian inference in a variety of practically important circumstances. Finally, Sect. 7 discusses the general implications of the study and suggests projects for further research.

2 NHST and Bayesian inference

Suppose we would like to measure the efficacy of an experimental intervention—for example, whether on-site classes lead to higher student performance than remote teaching. In frequentist statistics, the predominant technique for addressing such a question is **Null Hypothesis Significance Testing (NHST)**. At the basis stands a default or **null hypothesis** H_0 about an unknown parameter of interest. Typically, this hypothesis makes a precise statement about this parameter (e.g., $\mu = 0$), or it claims that the parameter has the same value in two different experimental groups (e.g., $\mu_1 = \mu_2$). For example, the null hypothesis may claim that classroom and remote teaching do not differ in their effect on student grades. Opposed to the null hypothesis is the **alternative hypothesis** H_1 which corresponds, in most practical applications, to the logical negation of the null hypothesis (e.g., $\mu \neq 0$ or $\mu_1 \neq \mu_2$). To test such hypotheses against each other, researchers conduct a **two-sided hypothesis test**: an experimental design where large deviations in either direction from the “null value” count as evidence against the null hypothesis, and in favor of the alternative.¹

Suppose that the data in both experimental conditions (e.g., student grades for on-site and remote teaching) are Normally distributed with unknown variance. Then it is common to analyze them by a ***t*-statistic**, that is, a standardized difference between the sample mean in both groups. This statistic measures the divergence of the data from the null hypothesis $H_0 : \mu_1 = \mu_2$. If the value of t diverges largely from zero—and more precisely, if it falls into the most extreme 5% of the distribution—, we *reject the null hypothesis* and call the result “statistically significant” at the 5% level ($p < .05$). In the above example, such a result means evidence for the alternative hypothesis that classroom and remote teaching differ in their effect on student grades. Otherwise we state a “non-significant result” or a “non-effect” ($p > .05$). Similarly, a result in the 1%-tail of the distribution of the t -statistic is called “highly significant” ($p < .01$).

The implicit logic of NHST—to “reject” the null hypothesis and to declare a result statistically significant evidence if it deviates largely from the null value—has been criticized for a long time in philosophy, statistics and beyond. Critics claim, for example, that it conflates statistical and scientific significance, uses a highly counterintuitive and frequently misinterpreted measure of evidence (p -values) and makes it impossible to express support for the null hypothesis (e.g., Edwards et al. 1963; Hacking 1965; Spielman 1974; Ziliak and McCloskey 2008).

The shortcomings of NHST have motivated the pursuit of alternative models of statistical inference. The most prominent of them is **Bayesian inference**: probabilities express subjective degrees of belief in a scientific hypothesis (Bernardo and Smith 1994; Howson and Urbach 2006). $p(H)$ quantifies prior degree of belief in hypothesis H whereas $p(H|D)$, the conditional probability of H given D , quantifies posterior degree of belief in H —that is, the degree of belief in H after learning data

¹ One-sided, that is, directional, tests also exist, but they are used much less frequently than two-sided tests. For a discussion of their use in the context of behavioral research, see Wagenmakers et al. (2011).

D. While the posterior probability $p(H|D)$ serves as a basis for inference and decision-making, the evidential import of a dataset *D* on two competing hypotheses is standardly described by the **Bayes factor**

$$BF_{10}(D) := \frac{p(H_1|D)/p(H_0|D)}{p(H_1)/p(H_0)} = \frac{p(D|H_1)}{p(D|H_0)}.$$

The Bayes factor is defined as the ratio between posterior and prior odds of H_1 over H_0 (Kass and Raftery 1995). Equivalently, it can be interpreted as the likelihood ratio of H_1 and H_0 with respect to data *D*—that is, as a measure of how much the data discriminate between the two hypotheses, and which hypothesis explains them better. Bayes factors $BF_{10} > 1$ favor the alternative hypothesis H_1 , and Bayes factors in the range $0 < BF_{10} < 1$ favor the null hypothesis H_0 . Finally, note that the Bayes factors for the null and the alternative are each other's inverse: $BF_{01} = 1/BF_{10}$.

In this paper, we shall not enter into the foundational debate between Bayesians and frequentists (for surveys of arguments, see, e.g., Romeijn 2014; Sprenger 2006; Mayo 2018; van Dongen et al. 2019). We just note that while Bayesian inference avoids the typical problems of frequentist inference with NHST, it is not exempt from limitations. These include misinterpretation of Bayes factors, mindless use of “objective” or “default” priors (e.g., exclusive reliance on fat-tailed Cauchy priors in statistical packages), bias in favor of the null hypothesis, and potential mismatch between inference with Bayes factors and estimation based on the posterior distribution (e.g., Sprenger 2013; Kruschke 2018; Lakens et al. 2018a; Mayo 2018; Tenreiro and Kiers 2019).

3 Model description and simulation design

Romero (2016) presents a simulation model to study whether SCT* holds when relaxing ideal, utopian conditions for scientific inquiry in the context of frequentist statistics. This paper follows Romero's simulation model, but we add the choice of the statistical framework (i.e., Bayesian vs. frequentist/NHST inference) as an exogenous variable to study how the validity of SCT* is affected by the statistical framework.

To examine the self-corrective abilities of Bayesian and frequentist inference, we first need to agree on a statistical model. In the behavioral sciences—arguably the disciplines hit most by the replication crisis—, many experiments collect data on a continuous scale and measure how the sample means \bar{X}_1 and \bar{X}_2 differ across two independent experimental conditions (e.g., treatment and control group). The means in each condition are assumed to follow a Normal distribution $N(\mu_{1,2}, \sigma^2)$, and the **true effect size** is described by the standardized difference of the unknown means: $\delta = (\mu_1 - \mu_2)/\sigma$. Conventionally, a δ around 0.2 is considered small, around 0.5 is considered medium, and around 0.8 is considered large. For both Bayesians and frequentists, the natural null hypothesis is $H_0 : \delta = 0$, stating equal means in both groups. Frequentists leave the alternative hypothesis $H_1 : \delta \neq 0$ unspecified whereas

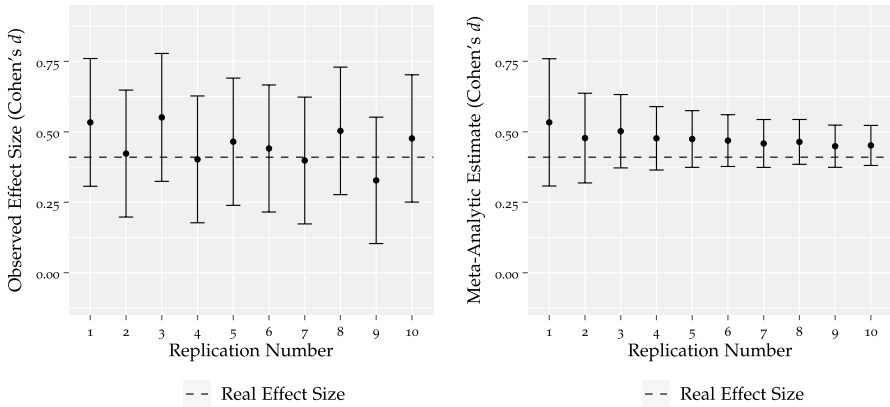


Fig. 1 Observed effect sizes with 95% confidence intervals in the exact replication of an experiment (left figure), and the corresponding aggregated effect size estimates (right figure). Data generated under the assumption $\delta = 0.41$

Bayesians put a diffuse prior over the various values of δ , typically a Cauchy distribution such as $H_1 : \delta \sim \text{Cauchy}(0, \frac{1}{\sqrt{2}})$ (Rouder et al. 2009).

The value of δ can be adequately estimated by Cohen's d , which summarizes **observed effect size** by means of the standardized difference in sample means:

$$\text{Cohen's } d = \frac{\overline{X_1} - \overline{X_2}}{S_p}$$

where S_p denotes the pooled standard deviation of the data.²

Using the statistics software *R*, we randomly generate Normally distributed data for two independent groups. We study two conditions, one where the null hypothesis is (clearly) false and one where it is literally true. As a representative of a positive effect, we choose $\delta = 0.41$, in agreement with meta-studies that consider this value typical of effect sizes in behavioral research (Richard et al. 2003; Fraley and Vazire 2014). The data are randomly generated with standard deviation $\sigma = 1$ in each group. For the first group, the mean is zero while for the second, the mean corresponds to the hypothesized effect size (either $\delta = 0$ or $\delta = 0.41$). The sample size of each group is set to $N = 156$ since this corresponds to a statistical power of 95% (=5% type II error rate) for a true effect of $\delta = 0.41$. We then compute the observed effect size and repeat this procedure to simulate multiple replications of a single experiment. At the same time, we simulate a cumulative meta-analysis of the

² S_p is defined as $S_p = \sqrt{\frac{(N_1-1)S_1^2 + (N_2-1)S_2^2}{N_1 + N_2 - 2}}$ where N_1 and N_2 are the sample sizes for both conditions, and S_1^2 and S_2^2 denote the corrected within-sample variance.

effect size estimates. Figure 1 shows the observed effect sizes from 10 replications and how they are aggregated into an overall meta-analytic estimate.^{3,4}

We expect that frequentist and Bayesian inference both validate SCT* under ideal conditions where various biases and imperfections are absent. The big question is whether Bayesian statistics improves upon NHST when we move to more realistic scenarios. In particular: Are the experiments sufficiently powered to detect an effect? Are the researchers biased in a particular direction? Are non-significant results systematically dismissed? The available evidence on published research suggests that the answers to these questions should not always be yes, leaving open whether SCT* will still hold in those cases. We model the relevant factors as binary variables, contrasting an ideal or utopian condition to a less perfect (and more realistic) condition. Let's look at them in more detail.

3.1 Variable 1: sufficient versus limited resources

NHST is justified by its favorable long-run properties, spelled out in terms of error control: a true null hypothesis is rarely “rejected” by NHST and a true alternative hypothesis typically yields a statistically significant result. To achieve these favorable properties, experiments require an adequate sample size. Due to lack of resources and other practical limitations (e.g., availability of participants/patients, costs of trial, time pressure to finish experiments), the sample size is often too small to bound error rates at low levels. Since the type I error level—that is, the rate of rejecting the null hypothesis when it is true—is conventionally fixed at 5%, this means that the power of a test is frequently low and can even fall below 50% (e.g., Ioannidis 2005).

In our simulation study, we compare two cases: first, a condition where the type I error rate in a two-sample *t*-test is bound at the 5% level and power relative to $\delta = 0.41$ equals 95%. This condition of **sufficient resources** corresponds to a sample size of $N = 156$. It is contrasted to a condition of **limited resources** that is typical of many experiments in behavioral research. In that condition, both experimental groups have sample size $N = 36$, resulting in a power of only 40%.

The Bayesian analogue to power analysis is to control the **probability of misleading evidence** (Royall 2000; Schönbrodt and Wagenmakers 2018), and to design an experiment such that the Bayes factor will, with high probability, state evidence

³ The details of the aggregation procedure are as follows (again, we follow Romero 2016): We assume that effect size is fixed across experiments, or in other words, that all single experiments are measuring the same effect size. Then, the aggregated effect size D after M experiments is given by $D = \frac{\sum_{i=1}^M w_i d_i}{\sum_{i=1}^M w_i}$. Here d_i denotes the effect size observed in experiment i , and $w_i = 1/v_i$ denotes the inverse of the variance of observed effect size, approximated by $v_i := \frac{2}{N} + \frac{d_i^2}{2N^2}$ for sample size N . The variance of D , which is necessary to calculate the associated confidence intervals, is given by $v_D = \frac{1}{\sum w_i}$ (Cumming 2012, 210–213; Borenstein et al. 2009, 63–67).

⁴ It would also be possible—and might be an interesting direction for future research—to use Bayesian methods for aggregating the individual estimates. However, that would require making potentially substantive assumptions (e.g., the prior distribution of effect size) that would make a fair comparison of the statistical frameworks difficult. We prefer to model the aggregation procedure as framework independent.

for H_1 when it is true, and *mutatis mutandis* for H_0 . For the parameters in our study, such a “Bayes factor design analysis” yields the sample size $N = 190$. To ensure a level playing field between both approaches, we use the same values ($N = 156$ and $N = 36$) for the frequentist and Bayesian scenarios. The simulation results for $N = 190$ instead of $N = 156$ in the sufficient resources condition are also qualitatively identical.

3.2 Variable 2: direction bias

Scientists sometimes conduct their research in a way that is shaped by selective perception and biased expectations. For example, feminist critiques of primatological research have pointed out that evidence on the mating behavior of monkeys and apes was often neglected when it contradicted scientists’ theoretical expectations (e.g., polyandrous behavior of females: Hrdy 1986; Hubbard 1990). More generally, researchers often exhibit *confirmation bias* (e.g., MacCoun 1998; Douglas 2009): their perception of empirical findings is shaped by the research program to which they are committed. There is also specific evidence that results are more likely to be published if they agree with previously found effects and exhibit positive magnitude (Hopewell et al. 2009; Lee et al. 2013). Effects that contradict one’s theoretical expectations and have a negative magnitude may either be suppressed as an act of self-censuring or be discarded in the peer-review process. Such **direction bias** is obviously detrimental to the impartiality and objectivity of scientific research, and we expect that it affects the accuracy of meta-analytic effect estimation and the validity of SCT*, too.

We model direction bias by a variable that can have two values: either **all results are published**, regardless of whether the effect is positive or negative (=no direction bias), or **all results with negative effect size magnitude are suppressed** (=direction bias present).

3.3 Variable 3: suppressing inconclusive evidence

Statistically non-significant outcomes of NHST ($p > .05$) are in practice often filtered out and end up in the proverbial file drawer (Rosenthal 1979; Ioannidis 2005; Fanelli 2010). An epistemic explanation for this is that non-significant outcomes are ambiguous between supporting the null hypothesis and the study not having enough statistical power to find an effect. Due to this ambiguity, they are hard to package into a clear narrative and published much less frequently. In our model, we distinguish between a non-ideal condition where **only results significant at the 5% level are published** and an ideal condition where **all results are published**, also non-significant ones (i.e., results with a p -value exceeding .05). This dichotomous picture (which we relax when we extend the model) is in line with scientometric evidence for the increasing prevalence of statistically significant over non-significant findings (Fanelli 2012). The choice of 5% as a cutoff level is a well-entrenched convention in the behavioral sciences; that said, also “marginally significant” p -values

(i.e., $.05 \leq p < .10$) are often reported in economics and the biomedical sciences (De Winter and Dodou 2015; Lakens 2015; Bruns et al. 2019).

For the Bayesian, the inconclusiveness of findings is spelled out by means of the Bayes factor instead of the p -value. When the Bayes factor is close to 1, the evidence is inconclusive: the null hypothesis and the alternative are equally likely to explain the observed data. We set up the two conditions analogously to the frequentist case: in the ideal condition, **all observed Bayes factors enter the meta-analysis**, regardless of their value, whereas the non-ideal condition **excludes all Bayes factors reporting weak evidence**, that is, those values where neither the null hypothesis nor the alternative are clearly favored by the data.

Specifically, we use the range $\frac{1}{3} < BF_{10} < 3$ for denoting inconclusive or weak evidence. This range is appropriate for two reasons. First, the qualitative meaning of the $p < .05$ significance threshold corresponds to the Bayesian threshold $\frac{1}{3} < BF_{10} < 3$. Frequentists consider p -values between .05 and .10 as weak or anecdotal evidence, as witnessed by formulations such as “marginally significant” and “trend”. Similarly, Bayesian researchers use a scale where the interval 1–3 corresponds to anecdotal or weak evidence for H_1 , 3–10 to moderate evidence, 10–30 to strong evidence, and so on (Jeffreys 1961; Lee and Wagenmakers 2014). Reversely for the ranges $1/3$ to 1, $1/10$ to $1/3$, and so on. Second, Bayesian re-analysis of data with an observed significance level of $p \approx .05$ often corresponds to a Bayes factor around $BF_{10} = 3$.⁵ Wider ranges for inconclusive evidence, such as $\frac{1}{6} < BF_{10} < 6$ (Schönbrodt and Wagenmakers 2018), are possible, but such proposals do not correspond to an interpretation of Bayes factors anchored in existing conventions.

To date, there has not yet been a systematic study of evidence filtering in Bayesian statistics. Hence, it is an open question whether in practice researchers would publish evidence for the null hypothesis when they have the necessary statistical tools to express it, e.g., Bayes factors. We return to this question in the discussion section.

4 Results: the baseline condition

Our simulations compare the performance of NHST and Bayesian inference in two types of situations: the baseline conditions and extensions of the model. The **baseline conditions**, numbered S1–S16, take the three variables described in Sect. 3 and the true effect size as independent variables. Table 1 explains which scenario corresponds to which combination of values of these variables. The **model extensions** explore a wider range of situations: we examine conditions where some, but not all negative results are published, and we contrast Bayesian and frequentist inference

⁵ For example, Benjamin et al. (2018) compare Bayesian and frequentist analysis for testing the mean of a Normal distribution with known variance. They define the prior over the alternative H_1 according to various constraints on experimental designs (75% power, uniformly most powerful Bayesian test, upper bounds on the Bayes factor, etc.). For all their designs, the Bayes factor corresponding to $p = .05$ falls into the interval [2.5; 3.4].

Table 1 The 16 possible simulation scenarios

	$\delta = 0.41$				$\delta = 0$				$\delta = 0.41$				$\delta = 0$			
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
SUFFICIENT RESOURCES	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
NO DIRECTION BIAS	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗	✓	✓	✗	✗
INCONCLUSIVE EVIDENCE IS PUBLISHED	✓	✓	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗

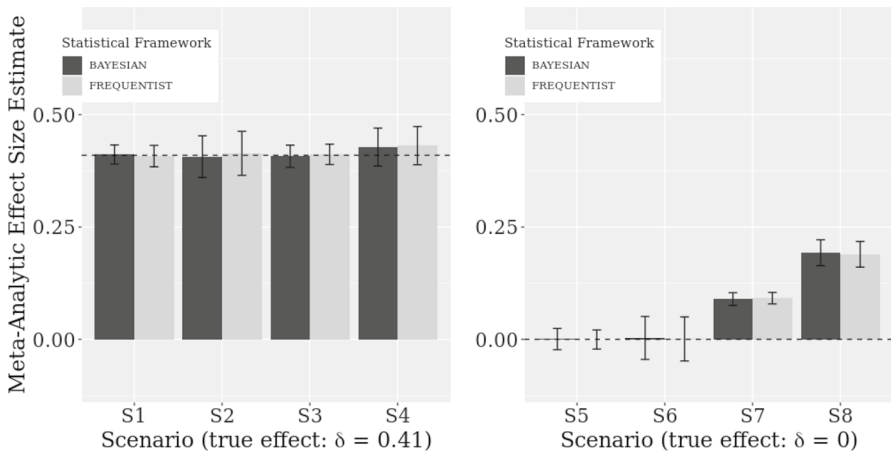


Fig. 2 Meta-analytic effect size estimates for Bayesian inference (dark bars) and frequentist inference (light bars) in conditions S1–S8 after 25 reported experiments. Upper graph: scenarios S1–S4 where $\delta = 0.41$, lower graph: scenarios S5–S8 where $\delta = 0$. All inconclusive evidence is published. The dashed line represents the true effect size, the error bars show one standard deviation

for a wider range of effect sizes (e.g., small effects such as $\delta \approx 0.2$ or large effects such as $\delta \approx 1$).

As revealed by Fig. 2, there is no difference between Bayesian and frequentist inference as long as “negative results” (i.e., results with inconclusive evidence) are published. This is to be expected since the difference between Bayesian and frequentist analysis in our study consists in the way inconclusive evidence is explicated and filtered. Thus both frameworks yield the same result in S1–S8: when the alternative hypothesis is true, meta-analytic estimates are accurate (scenarios S1–S4); when the null hypothesis is true, both frameworks are vulnerable to direction bias (scenarios S7–S8). Indeed, when the alternative hypothesis is true, few experiments will yield estimates with a negative magnitude and the presence of direction bias will not compromise the meta-analytic aggregation substantively.

Figure 3 shows the results of scenarios S9–S16 where a file drawer effect is operating and inconclusive, “non-significant” evidence is suppressed. To recall, this means that data from experiments with $p \geq .05$ or with a Bayes factor in the range

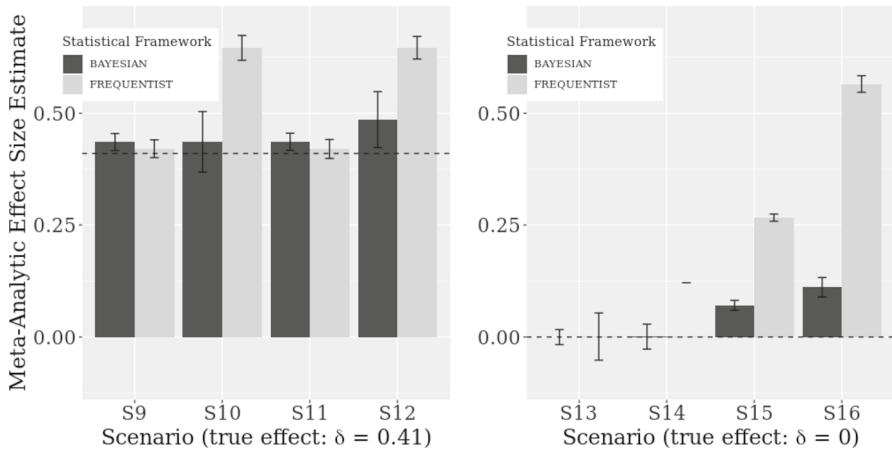


Fig. 3 Meta-analytic effect size estimates for Bayesian inference (dark bars) and frequentist inference (light bars) in conditions S9–S16 after 25 reported experiments. Left graph: scenarios S9–S12 where $\delta = 0.41$, right graph: scenarios S13–S16 where $\delta = 0$. All inconclusive evidence is suppressed. The dashed line represents the true effect size, the error bars show one standard deviation

$\frac{1}{3} < BF_{10} < 3$ do not enter the meta-analysis. In some of these scenarios, especially when the null hypothesis is true and direction bias is present, the frequentist excessively overestimates the actual effect size (e.g., $d \approx 0.25$ in S15 and $d \approx 0.55$ in S16 while in reality, $\delta = 0$). The reason is that the frequentist conception of “significant evidence” filters out evidence for the null hypothesis and acts as an amplifier of direction bias: only statistically significant effect sizes with positive magnitude enter the meta-analysis (e.g., $d \geq 0.47$ in S16). By contrast, the Bayesian also reports evidence that speaks strongly for the null hypothesis (i.e., $d \approx 0$) and obtains just a weak positive meta-analytic effect.

A similar diagnosis applies when the alternative hypothesis is true, regardless of direction bias. Consider scenarios S10 and S12. Due to the limited resources and the implied small sample size, only large effects meet the frequentist threshold $p < .05$, leading to a substantial overestimation of the actual effect ($d \approx 0.65$ in both scenarios, instead of the true $\delta = 0.41$). The overestimation in the Bayesian case, by contrast, is negligible for S10 and moderate for S12 ($d \approx 0.5$).

Thus, Bayesian inference performs considerably better when inconclusive evidence is not published, as it often happens in empirical research. There is thus a (partial) case for statistical reform: Bayesian analysis of experiments leads to more accurate meta-analytic effect size estimates when the experimental conditions are non-ideal and inconclusive evidence is suppressed.⁶

⁶ Note that these conclusions are sensitive to choice of the threshold $\frac{1}{K} < BF_{10} < K$ in the exclusion of inconclusive evidence. If the threshold is made more stringent, e.g., $K = 6$ instead of our $K = 3$, there are also some scenarios where the frequentist analysis performs better. However, we have argued in Sect. 3 that such a comparison would not be appropriate since the evidence thresholds of both frameworks should match each other, and $K = 6$ should be compared to a more severe frequentist conception of evi-

The next two sections present two extensions that model other practically relevant situations.

5 Extension 1: the probabilistic file drawer effect

The preceding simulations have modeled the file drawer effect as the exclusion of *all* non-significant p -values. In practice, it will depend a lot on the context whether inconclusive evidence is published or not. Bakker et al. (2012) report studies according to which the percentage of unpublished research in psychology may be greater than 50%. Especially in conceptual replications and other follow-up studies it is plausible that evidence contradicting the original result may be discarded (e.g., by finding fault with oneself and repeating the experiment with a slightly different design or test population). Then, disciplines with an influential private sector such as medicine may be especially susceptible to bias in favor of significant evidence: as indicated by the effect size gap between industry-funded and publicly funded studies, sponsors are often disinterested in publishing research on an apparently ineffective drug (Wilholt 2009; Lexchin 2012).

By contrast, there is an increasing number of prestigious journals that accepts submissions according to the “registered reports” model: before starting to collect the data, the researcher submits a study proposal that is accepted or rejected based on the study’s theoretical interest and the experimental design.⁷ This means that the paper will be published regardless of whether the results are statistically significant or not. Moreover, in large-scale replication projects such as Open Science Collaboration (2015) or Camerer et al. (2016) that examine the reproducibility of previous research, the evidence is published regardless of direction or size of the effect.

Taking all this together, we can expect that *some* proportion of statistically inconclusive studies will make it into print, or be made publicly available, while a substantial part of them will remain in the file drawer. We extend the results of the model analytically to investigate how the performance of frequentist and Bayesian inference depends on the proportion of inconclusive evidence that is actually published.

Like in the baseline condition, we model the suppression of inconclusive evidence as not reporting non-significant results, i.e., $p > .05$ and Bayes factors with weak, anecdotal evidence ($\frac{1}{3} < BF_{10} < 3$). Figure 4 plots effect size overestimation in both frameworks as a function of the probability of publishing studies with inconclusive evidence.

For the frequentist, estimates get more accurate when more statistically non-significant studies are published. Notably, when direction bias is present, publishing

Footnote 6 (continued)

dence. Moreover, the dependence of performance on the scenario implies that we cannot give a general answer to the question of which Bayesian cutoff criterion performs as well as $p = .05$.

⁷ Some of the better known journals who offer this publication model are *Nature Human Behaviour*, *Cortex*, *European Journal of Personality* and the *British Medical Journal Open Science*.

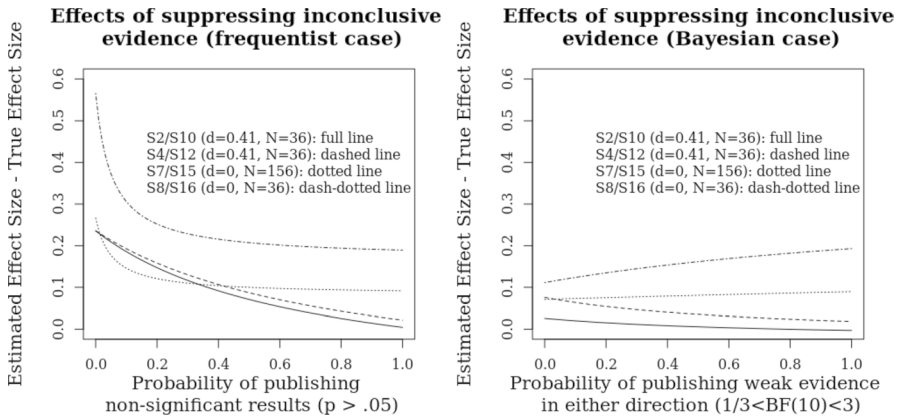


Fig. 4 Difference between estimated and true effect size as a function of the probability of suppressing inconclusive evidence, that is, the prevalence of the file drawer effect. Left graph = frequentist analysis, right graph = Bayesian analysis

just a small proportion of those studies is already an efficient antidote to large overestimation. This is actually logical: when direction bias is present and statistically non-significant results are suppressed, only studies with extreme effects are published and including *some* non-significant results will already be a huge step toward more realistic estimates.

The accuracy of the Bayesian estimates, however, does not depend much on the probability of publishing inconclusive studies—the overestimation is more or less invariant under the strength of the file drawer effect. Indeed, the Bayesian estimates are already accurate when all inconclusive evidence is suppressed. Using Bayesian inference instead of NHST may act as a safeguard against effect size overestimation in conditions where the extent of publication bias is unclear and potentially large. As soon as 20–30% of statistically non-significant results are published, however, frequentist estimates become similarly accurate.

6 Extension 2: a wider range of effect sizes

While $\delta = 0.41$ may be a good long-term *average* for the effect size of true alternative hypotheses in behavioral research, true effect sizes will typically spread over a wide range, ranging from small and barely observable effects ($\delta \approx 0.1$) to very large and striking effects (e.g., $\delta \approx 1$). This also depends on the specific scientific discipline and the available means for filtering noise and controlling for confounders. To increase the generality of our findings, we examine a wider range of true effect sizes. We focus on those conditions where Bayesians and frequentists reach different conclusions—that is, scenarios S9–S16 where inconclusive evidence is suppressed. Figures 5 and 6 show for both frameworks how the difference between estimated and true effect size varies as a function of the true effect size.

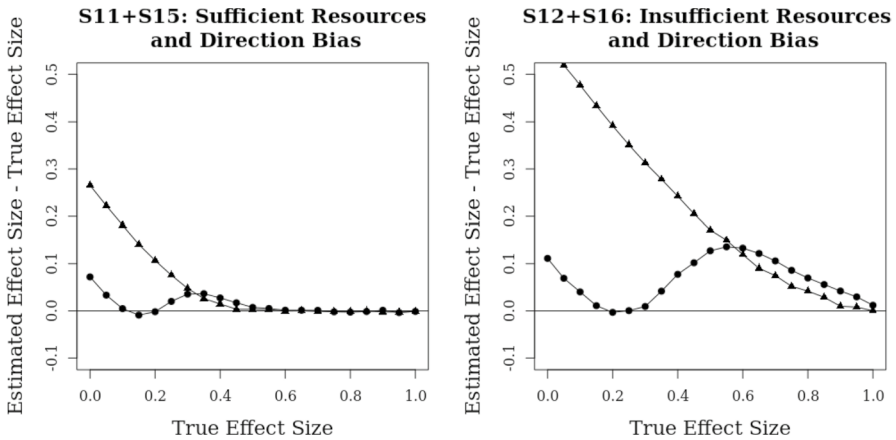


Fig. 5 Difference between estimated and true effect size as a function of the true effect size (measured by standardized means difference), for scenarios with direction bias and suppression of inconclusive evidence. Triangles = frequentist case, circles = Bayesian case, with linear interpolation

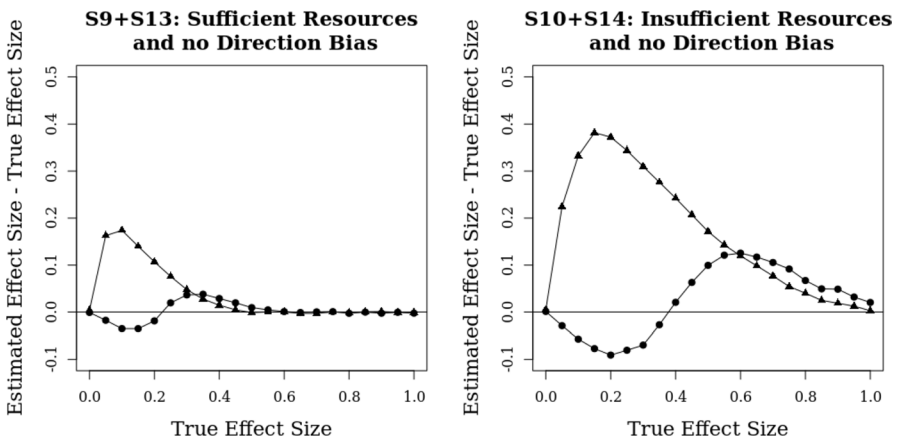


Fig. 6 Difference between estimated and true effect size as a function of the true effect size (measured by standardized means difference), for scenarios with suppression of inconclusive evidence and *no* direction bias. Triangles = frequentist case, circles = Bayesian case, with linear interpolation

When direction bias is present (Fig. 5), the Bayesian estimate comes closer to the true effect. Frequentists largely overestimate small effects due to the combination of direction bias and suppressing inconclusive evidence, but they estimate large effects accurately. This is to be expected since with increasing effect size, almost everything will be significant and fewer results will be suppressed. In these cases, the file drawer effect does not compromise the accuracy of the meta-analytic estimation procedure.

Turning to the case of no direction bias, shown in Fig. 6, two observations are striking. First, the frequentist graph ceases to be monotonically decreasing: small

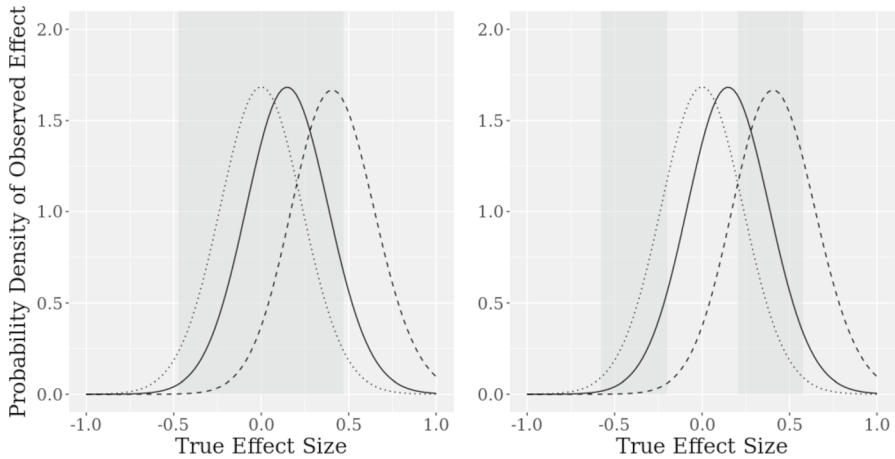


Fig. 7 Probability density functions for the standardized sample mean in a single experiment for $N = 36$ and different values of the real effect size. Full line: $\delta = 0.15$, dashed line: $\delta = 0.41$, dotted line: $\delta = 0$. The suppressed regions (i.e., observations that do not enter the meta-analysis because $p > .05$ or $\frac{1}{3} < BF_{10} < 3$) are shaded in dark. Left graph: frequentist case, right graph: Bayesian case

effects are substantially overestimated while null effects are estimated accurately. This is because, in the case of $N = 36$, all results inside the range $d \in [-0.47; 0.47]$ yield a p -value higher than $.05$ and do not enter the meta-analysis. For a true small positive effect, we will therefore observe many more (large) positive than negative effects and obtain a heavily biased meta-analytic estimate. For a true null effect, however, positive and negative magnitude effects are equally likely to be published and the aggregated estimate will be accurate.⁸ Similarly, when effects are big enough, few results will remain unpublished and the meta-analytic estimate will converge to the true effect size. The left graph in Fig. 7 visualizes these explanations by plotting the probability density function of d , and the range of suppressed observations.

Second, the Bayesian *underestimates* some small effects. This phenomenon is due to a superposition of two effects. Unlike the frequentist, the Bayesian publishes large effects in both directions *and* observed effects close to the null value $d \approx 0$. Intermediate effect size estimates from single studies are not published and left out of the meta-analysis—see Fig. 7. For small positive effects such as $\delta = 0.1$ or $\delta = 0.2$, the Bayesian is more likely to obtain results that favor the null hypothesis with $BF_{10} < \frac{1}{3}$, than results that favor the alternative with $BF_{10} > 3$. However, for these scenarios, the underestimation does not affect the qualitative interpretation of the effect size in question.

⁸ Note that such a canceling-out effect may not be realistic to obtain in practice since most replications will be suppressed. The meta-analytic effect will be unbiased, but with a very large variance and therefore typically be inaccurate.

All in all, omitting weak evidence in favor of either hypothesis leads to more accurate meta-analytic estimates than omitting statistically non-significant results. These observations are especially salient for small effects. SCT*—the thesis about the self-corrective nature of science in sequential replications of an experiment—therefore holds for a wider range of possible effect sizes when replacing NHST with Bayesian inference.

Our findings also agree with the distribution of effect sizes in the OSC replication project for behavioral research (Open Science Collaboration 2015): replications of experiments with large observed effects usually confirm the original diagnosis, while moderate effects often turn out to be small or inexistent in the replication.⁹ While a more detailed and substantive analysis would require assumptions about the prevalence of direction bias and suppressing inconclusive evidence in empirical research, our findings are, at first sight, consistent with patterns observed in recent replication research.

7 Discussion

Numerous areas of science are struck by a replication crisis—a failure to reproduce past landmark results. Such failures diminish the reliability of experimental work in the affected disciplines and the epistemic authority of the scientists that work in them. There is a plethora of complementary reform proposals to leave this state of crisis behind. Three principled strategies can be distinguished. The first strategy, called *statistical reform*, blames statistical procedures, in particular in the continued use of null hypothesis significance testing (NHST). Were NHST to be abandoned and to be replaced by Bayesian inference, scientific findings would be more replicable. The opposed strategy, called *social reform*, contends that the current social structure of science, in particular career incentives which reward novel and spectacular findings, has been the main culprit in bringing about the replication crisis. Between these extremes is a wide range of proposals for *methodological reform* that combines elements of social interaction and statistical method techniques (multi-site experiments, data-sharing, compulsory preregistration, etc.).

In this paper, we have explored the scope of statistical reform proposals by contrasting Bayesian and frequentist inference with respect to a specific thesis about the self-corrective nature of science, SCT*: convergence to the true effect in a sequential replication of experiments. Validating SCT* is arguably a minimal adequacy condition for any statistical reform proposal that addresses the replication crisis. Our model focuses on a common experimental design—two independent samples with normally distributed data—and compares NHST and Bayesian inference in different conditions: an ideal scenario where resources are sufficient and all results are published, as well as less ideal (and more realistic) conditions, where experiments are underpowered and/or various biases affect the publication of a research finding.

⁹ This observation has to be taken with a grain of salt since the OSC replication uses standardized correlation coefficients instead of standardized mean differences.

Our results support a partially favorable verdict on the efficacy of statistical reform. When a substantial proportion of studies with inconclusive evidence are published, both Bayesian inference and frequentist inference with NHST lead to quite accurate meta-analytic estimates and validate SCT*. However, when inconclusive evidence is not published, but strong evidence for a null effect is, Bayesian inference leads to more accurate estimates. In these conditions, which are unfortunately characteristic of scientific practice, statistical reform in favor of Bayesian inference will improve the reproducibility of published studies, validate SCT* and make experimental research more reliable.

The advantage of Bayesian statistics is particularly evident for small effect sizes ($\delta \approx 0.2$), which the frequentist often misidentifies as moderate or relatively large effects. This finding is in line with observations that small effects are at particular risk of being overestimated systematically (Ioannidis 2008). This holds for experimental research (e.g., Open Science Collaboration 2015), but perhaps even more so for observational research. Especially in the context of regression analysis, slight biases due to non-inclusion of relevant variables are almost inevitable and they inflate effect size estimates and observed significance substantially (Bruns and Ioannidis 2016; Ioannidis et al. 2017).

Finally, we turn to the limitations of our study. First, our results do not prove that moving to Bayesian statistics is the best statistical reform: alternative frameworks within the frequentist paradigm (e.g., Cumming 2012; Lakens et al. 2018b; Mayo 2018) could improve matters, too. Assessing and comparing such proposals is beyond the scope of this paper.

Second, the claim in favor of Bayesian statistics depends crucially on the assumption that researchers would publish evidence for the null hypothesis when the statistical framework supports such a conclusion (compare Sect. 3). One could object to this assumption by saying that such studies would just count as “failed” and that the evidence would nonetheless be suppressed (e.g., think of a clinical trial showing that a particular medical drug does *not* cure the target disease). Such situations certainly occur, but on the other hand, the null hypothesis *does* often play a major role in scientific inference and hypothesis testing: it is simple, has higher predictive value and can express important theoretical relations such as additivity of factors, chance effects and absence of a causal connection (e.g., Gallistel 2009; Morey and Rouder 2011; Sprenger and Hartmann 2019, ch. 9). In such circumstances, evidence for the null is of major theoretical interest. Moreover, evidence for a point null hypothesis is often the target of medical research that assesses the equivalence of two treatments, i.e., those aiming at establishing “theoretical equipoise” (Freedman 1987). Such research is greatly facilitated by a statistical framework that allows for a straightforward quantification of evidence for the null hypothesis. We therefore conjecture that statistical frameworks where evidence for the null can be expressed on the same scale as evidence for the alternative would lead to more “null” results being reported. Being able to state strong evidence *against* the targeted alternative hypothesis (e.g., that a specific intervention works) will also make the allocation of future resources easier compared to just stating “failure to reject the null”.

Third, statistical reform does not cure all the problems of scientific inference. We have not discussed here which concrete steps for social reform (e.g., changing incentive

structures and funding allocation schemes) would be most effective in complementing statistical reforms. The interplay of reform proposals on different levels is a fascinating topic for future research in the social epistemology of science. At this point, we can just observe that the file drawer effect seems to be particularly detrimental to reliable effect size aggregation, and that proposals for social and methodological reform should try to combat it. Compulsory pre-registration of experiments is a natural approach, but studying the efficacy of that strategy has to be left to future work.

Increasing the reliability of published research remains a complex and challenging task, involving reform of the scientific enterprise on various levels. What we have shown in this paper is that the choice of the statistical framework plays an important role in this process. Under the imperfect conditions where experimental research operates, adopting Bayesian principles for designing and analyzing experiments leads to more accurate effect size estimates compared to NHST, without incurring major drawbacks. Regardless of whether or not one likes Bayesian inference, it would be desirable to evaluate the model empirically—for example, by imposing the use of Bayesian statistics on an entire subdiscipline and then measuring how publication bias and replicability rates change. Such a project would not be easy to implement, but yield valuable insights about the mechanisms underlying the replication crisis.

Acknowledgements The authors would like to thank Mattia Andreoletti, Noah van Dongen, Daniël Lakens, Jürgen Landes, Barbara Osimani, Jan-Willem Romeijn, two anonymous reviewers of this journal, as well as audiences at the EPSA17 in Exeter, SILFS 2017 in Bologna, ECAP17 in Munich, the Reasoning Club conference in Turin, the “Perspectives on Scientific Error” workshop in Tilburg, and the weekly seminar of the philosophy department of the University of Sydney for their valuable comments and feedback.

Author Contributions The authors contributed equally to the scientific ideas, the design of the study, the data analysis and the drafting of the manuscript. Simulation code: FR. Visualization: JS.

Funding This research was financially supported by the European Research Council (ERC) through Starting Investigator Grant No. 640638.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

Availability of data and materials Simulations and figures were generated in *R*. The simulation data and the figures are openly available at the archive of the Open Science Foundation (OSF) at <https://osf.io/yejqg/>.

Code availability The *R* code used by the authors to generate data and figures is also openly available at the OSF archive at the URL <https://osf.io/yejqg/>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission

directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554.
- Benjamin, D., Berger, J., Johannesson, M., Nosek, B., Wagenmakers, E., Berk, R., et al. (2018). Redefine statistical significance. *Nature Human Behavior*, 2(1), 6–10.
- Bernardo, J. M., & Smith, A. F. M. (1994). *Bayesian Theory*. New York, NY: Wiley.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.
- Bruns, S. B., & Ioannidis, J. P. A. (2016). p-curve and p-hacking in observational research. *PLoS ONE*, 11(2), e0149144. <https://doi.org/10.1371/journal.pone.0149144>.
- Bruns, S. B., Asanov, I., Bode, R., Dungen, M., Funk, C., Hassan, S. M., et al. (2019). Reporting errors and biases in published empirical findings: Evidence from innovation research. *Research Policy*, 48, 103796.
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., et al. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433–1436. <https://doi.org/10.1126/science.aaf0918>.
- Cohen, J. (1994). The Earth is round ($p < .05$). *Psychological Review*, 49, 997–1001.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Multivariate applications book series. London: Routledge.
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25(1), 7–29.
- De Winter, J., & Dodou, D. (2015). A surge of p-values between 0.041 and 0.049 in recent decades (but negative results are increasing rapidly too). *Peer J*, (3), e733. <https://doi.org/10.7717/peerj.733>.
- Douglas, H. (2009). *Science, policy and the value-free ideal*. Pittsburgh: Pittsburgh University Press.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70, 193–242.
- Fanelli, D. (2010). Positive results increase down the hierarchy of the sciences. *PLoS ONE*, 5(4), e10068. <https://doi.org/10.1371/journal.pone.0010068>.
- Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90(3), 891–904.
- Fidler, F. (2005). From statistical significance to effect estimation: Statistical reform in psychology, medicine and ecology. Ph.D. thesis, University of Melbourne. <https://doi.org/10.1080/13545700701881096>.
- 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(10), e109019. <https://doi.org/10.1371/journal.pone.0109019>.
- Freedman, B. (1987). Equipose and the ethics of clinical research. *New England Journal of Medicine*, 317(3), 141–145.
- Freese, J., & Peterson, D. (2018). The emergence of statistical objectivity: Changing ideas of epistemic virtue and vice in science. *Sociological Theory*, 36(3), 289–313.
- Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, 116, 439–453.
- Goodman, S. N. (1999a). Toward evidence-based medical statistics 1: The *P* value fallacy. *Annals of Internal Medicine*, 130, 995–1004.
- Goodman, S. N. (1999b). Toward evidence-based medical statistics 2: The Bayes factor. *Annals of Internal Medicine*, 130, 1005–1013.
- Hacking, I. (1965). *Logic of statistical inference*. Cambridge: Cambridge University Press.
- Hopewell, S., Loudon, K., Clarke, M. J., Oxman, A. D., & Dickersin, K. (2009). Publication bias in clinical trials due to statistical significance or direction of trial results. *Cochrane Database of Systematic Reviews*, 1, MR000006. <https://doi.org/10.1002/14651858.mr000006.pub3>
- Howson, C., & Urbach, P. (2006). *Scientific reasoning: the Bayesian approach* (3rd ed.). La Salle, IL: Open Court.

- Hrdy, S. (1986). Empathy, polyandry, and the myth of the coy female. In R. Bleier (Ed.), *Feminist approaches to science* (pp. 119–146). New York, NY: Teachers College Press.
- Hubbard, R. (1990). *The politics of women's biology*. New Brunswick: Rutgers University Press.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2, e124. <https://doi.org/10.1371/journal.pmed.0020124>.
- Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, 19(5), 640–648.
- Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, 127(605), F236–F265.
- Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. J., Bahnik, S., Bernstein, M. J., et al. (2014). Investigating variation in replicability: a ‘Many Labs’ replication project. *Social Psychology*, 45(3), 142–152.
- Koole, S. L., & Lakens, D. (2012). Rewarding replications. *Perspectives on Psychological Science*, 7, 608–614.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 1(2), 270–280.
- Lakens, D. (2015). On the challenges of drawing conclusions from p -values just below 0.05. *PeerJ*, 3, e1142. <https://doi.org/10.7717/peerj.1142>.
- Lakens, D. (2019). The practical alternative to the p -value is the correctly used p -value. <https://doi.org/10.31234/osf.io/shm8v>, <https://osf.io/shm8v>, deposited on PsyArXiv.
- Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018a). Improving inferences about null effects with Bayes factors and equivalence tests. *The Journals of Gerontology: Series B*, 75, 45–57.
- Lakens, D., Scheel, A. M., & Isager, P. M. (2018b). Equivalence testing for psychological research: a tutorial. *Advances in Methods and Practices in Psychological Science*, 1, 259–269.
- Laudan, L. (1981). Peirce and the trivialization of the self-corrective thesis. *Science and hypothesis* (pp. 226–251). The University of Western Ontario Series in Philosophy of Science, Vol. 19. Dordrecht: Springer Netherlands.
- Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1), 2–17.
- Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: a practical course*. Cambridge: Cambridge University Press.
- Lexchin, J. (2012). Sponsorship bias in clinical research. *The International Journal of Risk & Safety in Medicine*, 24, 233–242.
- MacCoun, R. J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, 49, 259–287.
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *The American Psychologist*, 70, 487–98.
- Mayo, D. (1996). *Error and the growth of experimental knowledge*. Chicago, IL: University of Chicago Press.
- Mayo, D. (2018). *Statistical inference as severe testing: How to get beyond the science wars*. Cambridge: Cambridge University Press.
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16, 406–419.
- Munafò, M. R., Nosek, B., Bishop, D. V. M., Button, K., Chambers, C. D., du Sert, N. P., et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021. <https://doi.org/10.1038/s41562-016-0021>.
- Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: Making sense of replications. *eLife*, 6, e23383. <https://doi.org/10.7554/eLife.23383>.
- Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4), 1205–1226.
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*. <https://doi.org/10.1126/science.aac4716>.
- Peirce, C. S. (1931–1935). *The collected papers of Charles Sanders Peirce*, Vol. I–VI. Cambridge, MA: Harvard University Press.

- Quintana, D. S. (2015). From pre-registration to publication: a non-technical primer for conducting a meta-analysis to synthesize correlational data. *Frontiers in Psychology*, 6, 1549. <https://doi.org/10.3389/fpsyg.2015.01549>.
- Richard, F. D., Bond, C. F. J., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, 7(4), 331–363.
- Romeijn, J. W. (2014). Philosophy of statistics. In E. Zalta (Ed.), *The Stanford encyclopedia of philosophy*, Retrieved April 27, 2020 from <https://plato.stanford.edu/archives/sum2018/entries/statistics/>.
- Romero, F. (2016). Can the behavioral sciences self-correct? A social epistemic study. *Studies in History and Philosophy of Science Part A*, 60, 55–69.
- Romero, F. (2017). Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, 84, 1031–1043.
- Romero, F. (2018). Who should do replication labor? *Advances in Methods and Practices in Psychological Science*, 1(4), 516–537.
- Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, 14, e12633. <https://doi.org/10.1111/phc3.12633>.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638–641.
- 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(2), 225–237.
- Royall, R. (1997). *Statistical evidence: a likelihood paradigm*. London: Chapman & Hall.
- Royall, R. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 95(451), 760–768.
- Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1(2), 115–129.
- Schönbrodt, F. D., & Wagenmakers, E. J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25, 128–142.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.
- Spiegelman, S. (1974). The logic of tests of significance. *Philosophy of Science*, 41(3), 211–226.
- Sprengr, J. (2013). Testing a precise null hypothesis: the case of Lindley’s paradox. *Philosophy of Science*, 80, 733–744.
- Sprengr, J. (2016). Bayesianism versus frequentism in statistical inference. In *The Oxford handbook of probability and philosophy* (pp. 185–209). Oxford: Oxford University Press.
- Sprengr, J., & Hartmann, S. (2019). *Bayesian philosophy of science*. Oxford: Oxford University Press.
- Tendeiro, J., & Kiers, H. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, 24, 774–795.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37, 1–2.
- van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M., (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE*, 9(1), e84896. <https://doi.org/10.1371/journal.pone.0084896>.
- van Dongen, N. N. N., van Doorn, J. B., Gronau, Q. F., van Ravenzwaaij, D., Hoekstra, R., Hauke, M. N., et al. (2019). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*, 73, 328–339.
- Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of Psi. *Journal of Personality and Social Psychology*, 100(3), 426–432.
- Wilholt, T. (2009). Bias and values in scientific research. *Studies in History and Philosophy of Modern Science A*, 40, 92–101.
- Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan Press.