Scientific Self-Correction: The Bayesian Way

Felipe Romero* Jan Sprenger†

March 11, 2019

Abstract

The enduring replication crisis in many scientific disciplines casts doubt on the ability of science to self-correct and to produce reliable knowledge. There are different approaches for addressing this challenge. Social reformists hypothesize that the social structure of science such as the credit reward scheme must be changed. Methodological reformists suggest changes to the way data are gathered, analyzed and shared, such as compulsory pre-registration or multi-site experiments with different research teams. Statistical reformists argue more specifically that science would be more reliable and self-corrective if null hypothesis significance tests (NHST) were replaced by a different inference framework, such as Bayesian statistics. On the basis of a simulation study for meta-analytic aggregation of effect sizes, we articulate a middle ground between the different reform proposals: statistical reform alone won't suffice, but moving to Bayesian statistics eliminates important sources of overestimating effect sizes.

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 a representative sample of published papers, and more specifically the main effect a paper reports, scientists have tried to assess the seriousness of the crisis in a systematic way. The outcome of these studies is sobering: the rate of statistically significant findings drops dramatically and the observed effect sizes are often much lower (for the fields of psychology, experimental economics and cancer biology, respectively: Open Science Collaboration

*Faculty of Philosophy, RU Groningen, Oude Boteringestraat 52, 9712 GL Groningen, The Netherlands. Email: f.romero@rug.nl. Website: www.feliperomero.org.

†Center for Logic, Language and Cognition (LLC), Department of Philosophy and Educational Science, Università degli Studi di Torino, Via Sant'Ottavio 20, 10124 Torino, Italy. Email: jan.sprenger@unito.it. Webpage: www.laeuferpaar.de.

2015; Camerer et al. 2016; Nosek and Errington 2017). While the appropriate interpretation of replication failures is debatable (e.g., Gilbert et al. 2016), there is a shared sentiment that science is not as reliable as it is supposed to be and that something needs to change.

There is a wide range of possible reforms addressing the crisis. Some researchers identify as the main causes of low replicability the adverse effects of social and structural factors (Bakker, van Dijk, and Wicherts 2012; Nuijten et al. 2016; Romero 2017) and questionable research practices ("QRPs": Simmons, Nelson, and Simonsohn 2011). This assessment suggests to adopt **social reforms**, such as educating researchers about statistical cognition and methodology (Schmidt 1996; Lakens, Scheel, and Isager 2018) and creating greater incentives for replication work, for example by publishing and co-citing replications alongside original studies (Koole and Lakens 2012) or by establishing a separate career track for confirmatory research (Romero 2018). Other authors suggest **methodological reforms**: pre-registering studies and stating the data analysis plan prior to the actual experiment (Quintana 2015), sharing experimental data for "successful" as well as "failed" studies (Assen et al. 2014; Munafò et al. 2017) and promoting multi-site experiments as an antidote to various forms of bias (Klein et al. 2014). On the purely statistical end, numerous authors identify "classical" statistical inference based on Null Hypothesis Significance Tests (NHST) as the main culprit (Rosenthal 1979; Cohen 1994; Goodman 1999a,a; Ioannidis 2005; Ziliak and McCloskey 2008), which strongly suggests to adopt **statistical reforms**. These include moving toward Bayesian inference (Goodman 1999b; Rouder et al. 2009; Lee and Wagenmakers 2014), likelihood-based inference (Royall 1997), inference based on effect sizes confidence intervals (Fidler 2005; Cumming 2012, 2014) and even purely descriptive data summaries (Trafimow and Marks 2015).

Most likely all these reform proposals have some merit, but statistical reforms in particular are often regarded as necessary interventions that would improve science even if social and other methodological aspects of science remained the same. Hence, in this paper we evaluate the case for statistical reform conducting a systematic computer simulation study. The study is framed with respect to a prominent philosophical thesis: the self-corrective nature of science (e.g., Peirce 1931–1935; Mayo 1996). A strong version of the self-corrective thesis (SCT) asserts that scientific method guarantees convergence to true theories in the long run (Laudan 1981): by staying on the path of scientific method, errors in published research will eventually be discovered, corrected and wed out. In the context of statistical inference and the replication crisis, we consider a more precise and testable version of SCT: sequential replications of an experiment will eventually "reveal the truth" (Romero 2016).

SCT* Given a series of direct 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 direct 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 error 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. Given our focus on statistical reform, we ask specifically: Does replacing NHST with Bayesian statistics make science more self-corrective? Or are such claims sensitive to the conventions and biases that affect experimental design and data reporting? Taking these contextual factors into account, the results of our study suggest that statistical reform alone will not suffice. However, in many conditions that are typical of scientific practice, switching from NHST to Bayesian inference reduces overestimation of effect size substantially and makes experimental research more reliable.

The paper is structured as follows: Section 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, Section 7 discusses the philosophical 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, it 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 negation of the null hypothesis (e.g., $\mu \neq 0$ or $\mu_1 \neq \mu_2$). When we test such hypotheses against each other, we obtain a so-called **two-sided hypothesis test**: an experimental design where strong deviations in either direction from the "null value" count as evidence

against the null hypothesis, and in favor of the alternative.[1]

If we assume that the data in both experimental conditions (e.g., student grades for on-site and remote teaching) are Normally distributed with unknown variance, 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 indicates a strong divergence 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" ($p < .01$). [2]

The contrast model is **Bayesian inference**, where probabilities express subjective degrees of belief in a scientific hypothesis (Bernardo and Smith 1994; Howson and Urbach 2006). $p(H)$ quantifies your prior degree of belief in hypothesis H whereas $p(H|D)$, the conditional probability of H given D, quantifies your posterior degree of belief in H—that is, the degree of belief in H after learning data 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**

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

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 evidence discriminates between the two hypotheses, and which hypothesis explains the data 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: $BF_{01} = 1/BF_{10}$.

[1] One-sided, that is, directional, tests also exist, but they are used much less frequently than two-sided tests. For a discussion, see Wagenmakers, Wetzels, Borsboom, and van der Maas (2011).

[2] The implicit logic of NHST—to "reject" the null hypothesis and to declare a result statistically significant evidence if it deviates strongly from the null value—has been criticized for a long time in philosophy, statistics and beyond (e.g., Edwards, Lindman, and Savage 1963; Hacking 1965; Spielman 1974). In this paper, we shall not enter into this foundational debate (for pros and cons, see for example Romeijn 2014; Sprenger 2016; Mayo 2018; van Dongen et al. forthcoming).

3 Model Description and Simulation Design

The design of our simulation model follows Romero 2016, with the crucial difference that the choice of the statistical framework (i.e., Bayesian vs. frequentist/NHST inference) is added as an exogenous variable. The focus of the research question is different, too: While Romero's 2016 paper analyzes whether SCT* still holds when relaxing ideal, utopian conditions for scientific inquiry, our paper studies how the validity of SCT* is affected by choosing a statistical framework.

To examine the self-corrective abilities of Bayesian and frequentist inference, we first need a **measure of effect size**. They come in different categories: measures of correlation for paired data (e.g., Pearson's r), measures of statistical association between categorical variables (e.g., risk difference or risk ratio in medicine), or standardized measures of how a sample means differs across two experimental conditions. Since the latter measures are most common in the behavioral sciences—arguably the disciplines hit most by the replication crisis—, we adopt them in our model, too. Specifically, we study the aggregation of observed effect sizes from experiments with two independent samples and Normally distributed data with unknown population means. There is a wide range of statistical techniques for studying such data and the effect size can be adequately summarized by Cohen's d:

Cohen's d If $\overline{X_1}$ and $\overline{X_2}$ denote the observed sample means in two experimental conditions, then Cohen's d is equal to their standardized difference:

$$d = \frac{\overline{X_1} - \overline{X_2}}{S_P}$$

where S_P denotes the pooled standard deviation of the data.[3]

Conventionally, a d 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 $H_0 : d = 0$, stating equal means in both groups. Frequentists leave the alternative hypothesis $H_1 : d \neq 0$ unspecified whereas Bayesians put a diffuse prior over the various values of d, typically a Cauchy distribution such as $H_1 : d \sim \text{Cauchy}(0, 1/\sqrt{2})$ (Rouder et al. 2009).[4]

[3] 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.

[4] Here, we abuse notation by using d simultaneously as denoting a measure of observed effect size and, in the specification of H_0 and H_1, as an unknown parameter. That parameter is in reality a function of the unknown group means μ_1 and μ_2 and the unknown variance σ^2. (Recall that $X \sim N(\mu_{1,2}, \sigma^2)$.) Speaking strictly, we should therefore write $H_0 : \delta = 0$, with the true effect size δ defined as $\delta = (\mu_1 - \mu_2)/\sigma$ (Rouder et al. 2009).

Using the statistics software *R*, we randomly generate 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 representative of a positive effect, we choose $d = 0.41$, in agreement with meta-studies that consider this value typical of effect sizes in behavioral research (Richard, Bond, and Stokes-Zoota 2003; Fraley and Vazire 2014). The data for the groups is randomly generated. For the first group, the mean is zero (with standard deviation $\sigma = 1$) while for the second, the mean corresponds to the hypothesized effect size (either $d = 0$ or $d = 0.41$). 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 effect size estimates. Figure 1 shows the observed effect sizes from 10 replications and how they are aggregated into an overall meta-analytic estimate.[5]

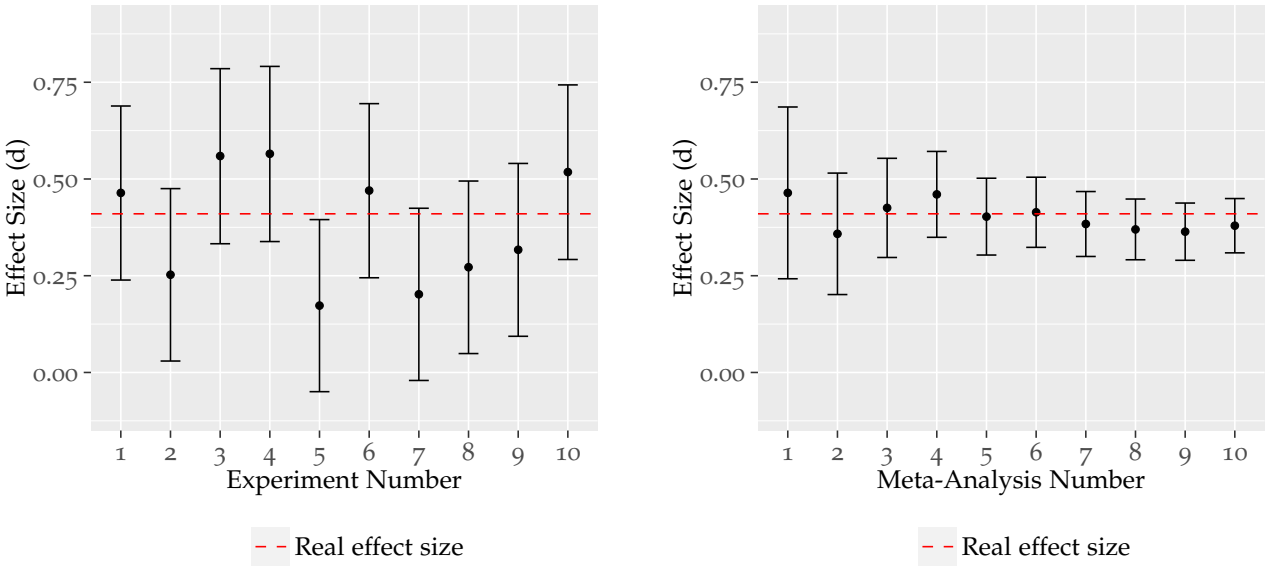

Figure 1: Observed effect sizes with confidence intervals in the replication of an experiment (left figure) and the corresponding aggregated effect size estimate (right figure). Data generated under the assumption $d = 0.41$.

[5]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).

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 into 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.

Variable 1: Sufficient vs. 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), sample size is often unsatisfactory 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 (=1-type II error rate) 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 is bound at the 5% level and power relative to $d = 0.41$ equals 95%. This condition of **sufficient resources** and sufficient sample size (N=156) 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%.

While the limited resources condition (N=36) is invariant across the competing statistical frameworks, the Bayesian needs to conceptualize the sufficient resources condition differently: type I and type II error rates belong conceptually to frequentist inference. Their analogue is to control the **probability of misleading evidence** (Royall 2000; Schönbrodt and Wagenmakers 2018). We should design an experiment such that the Bayes factor will, with high probability, not state evidence for H_1 when H_0 is true, or vice versa. Suppose $K > 0$ represents a conventionally agreed evidence threshold such that $BF_{10} > K$ means (substantial) evidence for H_1. Transferring the ideas of controlling type I error and achieving sufficient power to the Bayesian framework,

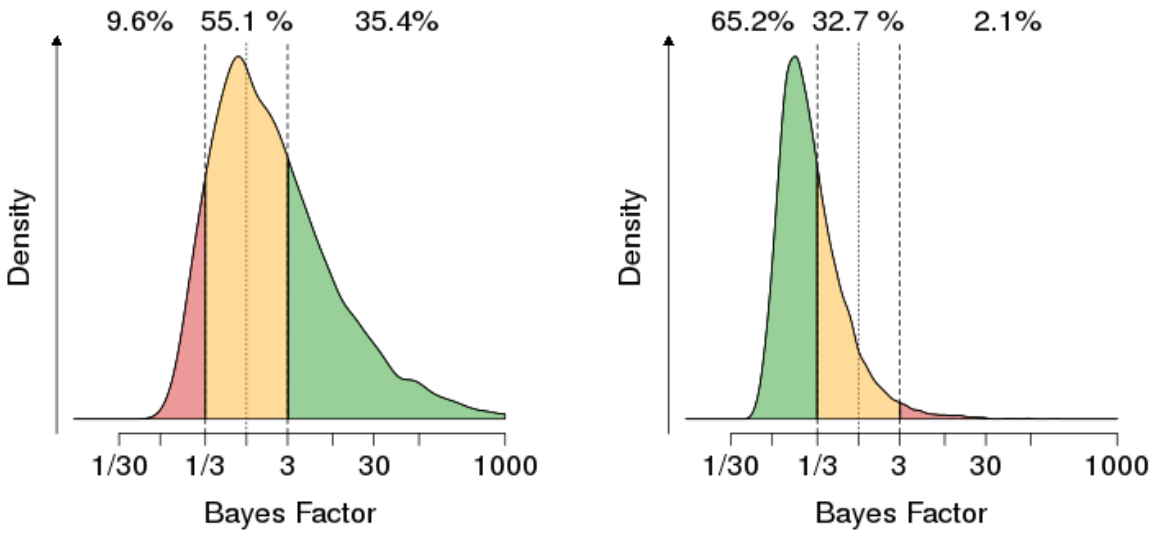

Figure 2: Graphical illustration of the distribution of Bayes factors and probability of misleading evidence in the Bayesian Framework for N=36 and hypothesized effects of $d = 0$ and $d = 0.4$. Figure produced with the BFDA app https://shinyapps.org/apps/BFDA/ developed by Felix Schönbrodt.

sample size N should be chosen such that

$$p_{H_0}(BF_{10}(X_1, \ldots, X_N) > K) \leq 0.05 \qquad p_{H_1}(BF_{10}(X_1, \ldots, X_N) > K) \geq 0.95.$$

In other words, the probability of finding misleading evidence against the null hypothesis should be smaller then 5%, and the probability of finding evidence for the alternative hypothesis when it is true should be at least 95%. See Figure 2 for a graphical illustration of the distribution of the Bayes factor. The appropriate choice of K depends on the strength of the evidence that is relevant for the particular application (e.g., $K = 3, 6, 10 \ldots$). We use $K = 3$, the conventional borderline between weak and moderate evidence. This choice will be motivated in more detail below. For our representative effect sizes of $d = 0$ and $d = 0.41$, the appropriate sample size in the Bayesian sufficient resources condition is equal to N=190.

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.,

polygamous 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).

Variable 3: Suppressing Inconclusive Evidence

In an ideal world, results supporting the null hypotheses would be as likely to be reported as results supporting the alternative. After all, any systematic suppression of evidence can be expected to lead to a distorted picture of effect size. However, in practice, non-significant outcomes of NHST are often filtered out and end up in the proverbial file drawer, diminishing the reliability of published research (Rosenthal 1979; Ioannidis 2005; Fanelli 2010). We distinguish between two conditions: an ideal condition where **all results are published**, also non-significant ones (i.e., results with a p-value exceeding .05), and a non-ideal condition where **only results significant at the 5% level are published**, and enter the meta-analytic effect size aggregation.

To date, there has not yet been a systematic study of evidence filtering in Bayesian statistics. Nonetheless, it is fair to assume that similar psychological mechanisms may operate. The role of the p-value or significance level is then played by the Bayes factor. 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. Therefore it is hard to extract a clear "story" from the study, making the results more difficult to sell and eventually, to publish. In agreement with this rationale—and in order to draw a meaningful comparison between NHST and Bayesian statistics—we distinguish two conditions in Bayesian inference. In the ideal condition, **all observed Bayes factors are reported**, 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.

We consider two Bayesian explications of "weak evidence": the range $1/6 <$

$BF_{10} < 6$ that Schönbrodt and Wagenmakers (2018) use for the Bayes factor design analysis, and the more narrow range $1/3 < BF_{10} < 3$. The second option strikes us as more adequate. First, the proposal $K = 6$ does not correspond to an interpretation of Bayes factors anchored in existing conventions. Most researchers use a quasi-logarithmic scale where the interval 1–3 corresponds to anecdotal 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, and more importantly, we need to compare the Bayesian and frequentist meta-analysis on equal grounds. This means that the intuitive meaning of the $p < .05$ significance threshold needs to be mapped to an appropriate Bayesian threshold. Since frequentists consider p-values between .05 and .10 as weak or anecdotal evidence, barely worth a mention (as witnessed by formulations such as "marginally significant" and "trend"), the border between weak and moderate evidence in the Bayesian framework—that is, $K = 3$—strikes us as a natural "translation" of the frequentist threshold $p = .05$. Third, $BF_{10} \approx 3$ is typically obtained in Bayesian re-analyses of frequentist experiments with an observed significance level of $p \approx .05$ (Benjamin et al. 2018). Hence, our choice of $K = 3$ ensures a level playing field between Bayesian and frequentist inference.

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

We begin with the ideal case where resources are sufficient and all evidence is

	$d = 0.41$				$d = 0$				$d = 0.41$				$d = 0$			
	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	✓	✓	✓	✓	✓	✓	✓	✓								

Table 1: The 16 possible simulation scenarios.

published (scenario S1). Figure 3 shows the progression of the meta-analysis over time, for a total of 25 replications. We observe that both the Bayesian and the frequentist framework validate SCT*: the aggregated effect size accurately estimates the actual effect size and convergence is fast, too.

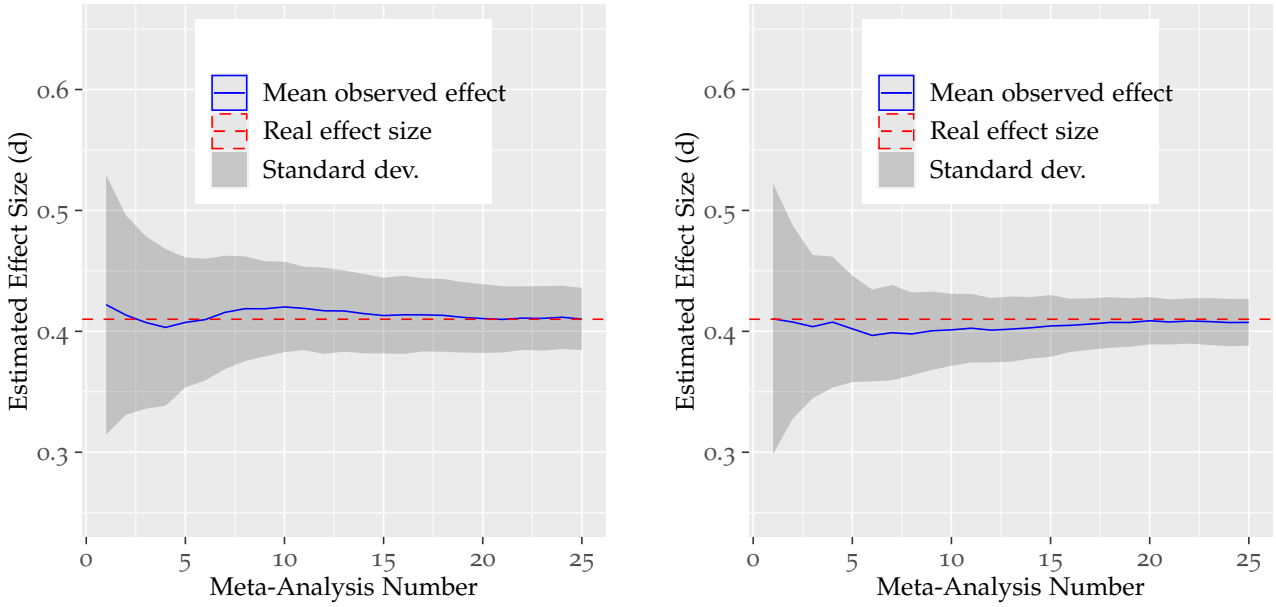

Figure 3: Meta-analytic results for Bayesian and frequentist inference in Scenario 1 (=the ideal condition with $d = 0.41$), as a function of the number of replications. The shaded area represents the data that are within one standard deviation of the observed mean.

Turning to less ideal and more realistic scenarios, Figure 4 reveals that there is no notable difference between Bayesian and frequentist inference as long as "negative results" (i.e, results with inconclusive evidence) are published. When the alternative hypothesis is true, meta-analytic estimates are accurate in both frameworks (scenarios S1–S4); when the null hypothesis is true, both frameworks are vulnerable to direction bias (scenarios S7–S8). This sensitivity to the value of d can be explained easily: when the alternative hypothesis is true, few experiments will yield estimates with a negative magnitude. As a consequence, the presence or absence of direction bias does not affect the meta-analytic aggregation substantively.

All in all, the evaluation of scenarios S1–S8 does not make the case for statistical reform: whatever the merits of Bayesian inference from a foundational point of view, in a classical estimation problem like the aggregation of observed effect sizes, it is not more reliable than frequentist inference with NHST.

Figure 5 shows the results of scenarios S9–S16 where a file drawer effect is op-

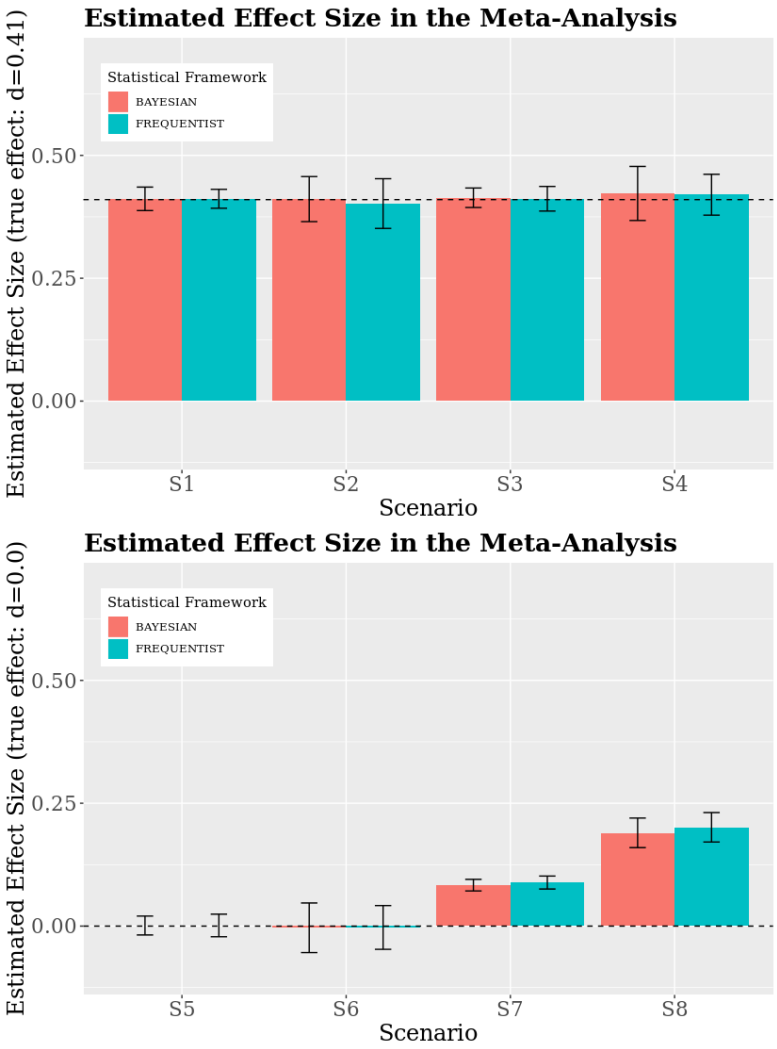

Figure 4: Meta-analytic results for Bayesian inference (red bars) and frequentist inference (blue bars) in conditions S1–S8 after 25 replications. Upper graph: scenarios S1–S4 where $d = 0.41$, lower graph: scenarios S5—S8 where $d = 0$. All inconclusive evidence is published. The dashed line represents the true effect size, the error bars show one standard deviation.

erating 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 $1/3 < BF_{10} < 3$ 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 largely overestimates the actual effect size (e.g., $d \approx 0.3$ in S15 and $d \approx 0.5$ in S16 while in reality, $d = 0$). 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

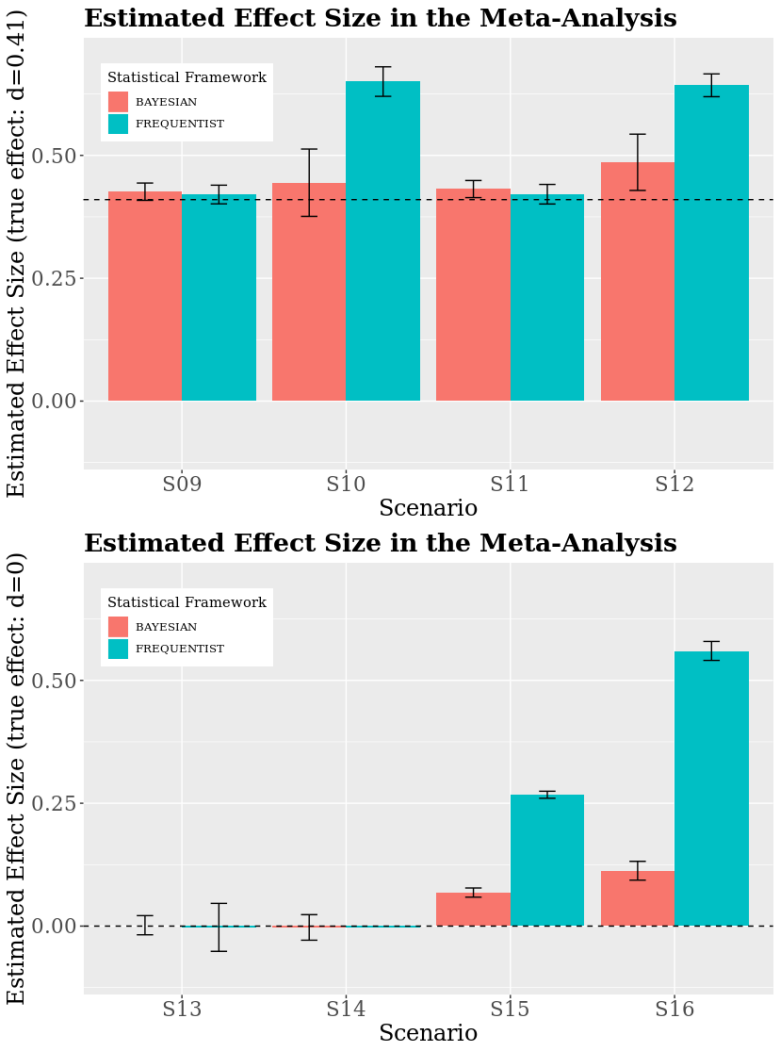

Figure 5: Meta-analytic results for Bayesian inference (red bars) and frequentist inference (blue bars) in conditions S9–S16 after 25 replications. Upper graph: scenarios S9–S12 where $d = 0.41$, lower graph: scenarios S13—S16 where $d = 0$. All inconclusive evidence is suppressed. The dashed line represents the true effect size, the error bars show one standard deviation.

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 notes just a very weak positive 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 $d = 0.41$). The overestimation in the Bayesian case, by contrast, is

negligible for S10 and moderate for S12 ($d \approx 0.5$).

Thus, while there is little difference between the reliability of frequentist and Bayesian inference in the absence of a file drawer effect, Bayesian inference performs considerably better when inconclusive evidence is not published. Such a publication bias is, unfortunately, often found in empirical research. Thus, we conclude that the statistical reformist is partially right: Bayesian design and 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]

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, van Dijk and Wicherts (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 the 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 which is accepted or rejected on the basis of 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

[6]Note that these conclusions are sensitive to choice of the threshold $K = 3$ in the Bayes factor design analysis and the exclusion of inconclusive evidence. If the threshold is made more stringent, e.g., $K = 6$, there are also some scenarios where the frequentist analysis performs better. However, we have argued in Section 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 evidence, e.g., $p < .01$.

[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*.

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 proverbial file drawer. We extend our original model by investigating how the performance of frequentist and Bayesian inference depends on the proportion of inconclusive evidence that is actually published.

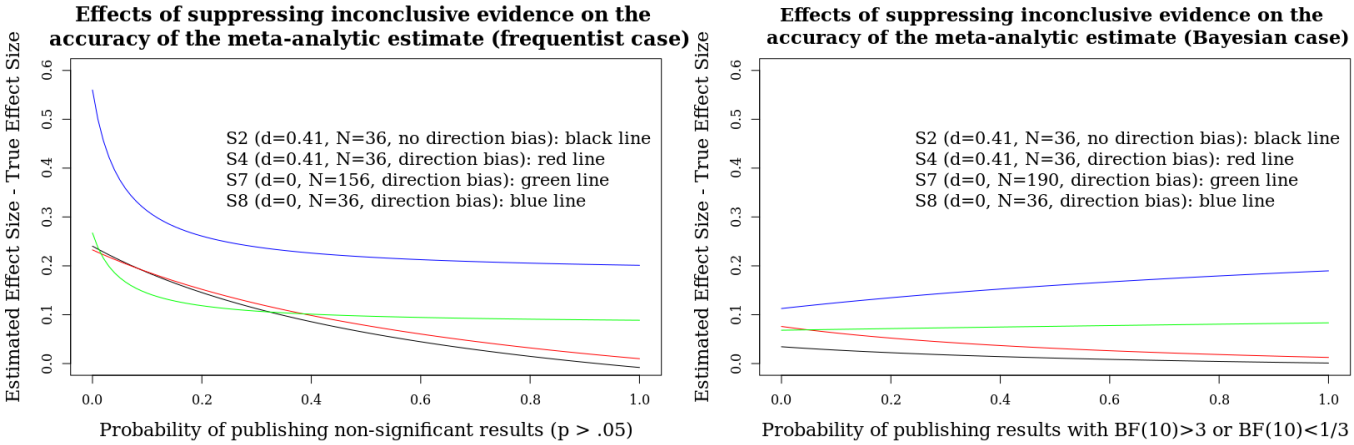

Figure 6: Overestimation of parameter value as a function of the probability of suppressing inconclusive evidence. Left graph = frequentist analysis, right graph = Bayesian analysis.

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 ($1/3 < BF_{10} < 3$). Figure 6 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 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 $d = 0.41$ may be a good long-term average for the effect size of true alternative hypotheses in behavioral research, effect sizes will typically spread over a wide range, ranging from small and barely observable effects ($d \approx 0.1$) to very large and striking effects (e.g., $d \approx 1$). This depends also 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. Figure 7 and 8 show for both frameworks how the difference between estimated and true effect size varies as a function of the true effect size.

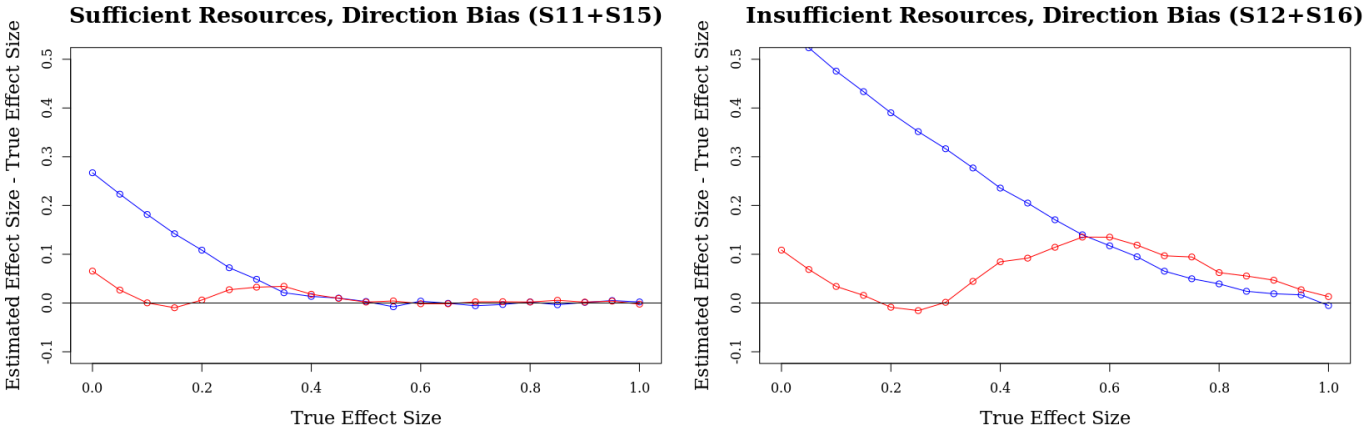

Figure 7: Difference between estimated and true effect size as a function of the true effect size (measured by standardized means difference), for scenarios with direction bias and suppression of inconclusive evidence. Blue line = frequentist analysis, red line = Bayesian analysis.

When direction bias is present (Figure 7), 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 less and less results will be suppressed. In these cases, the

file drawer effect does not compromise the accuracy of the meta-analytic estimation procedure.

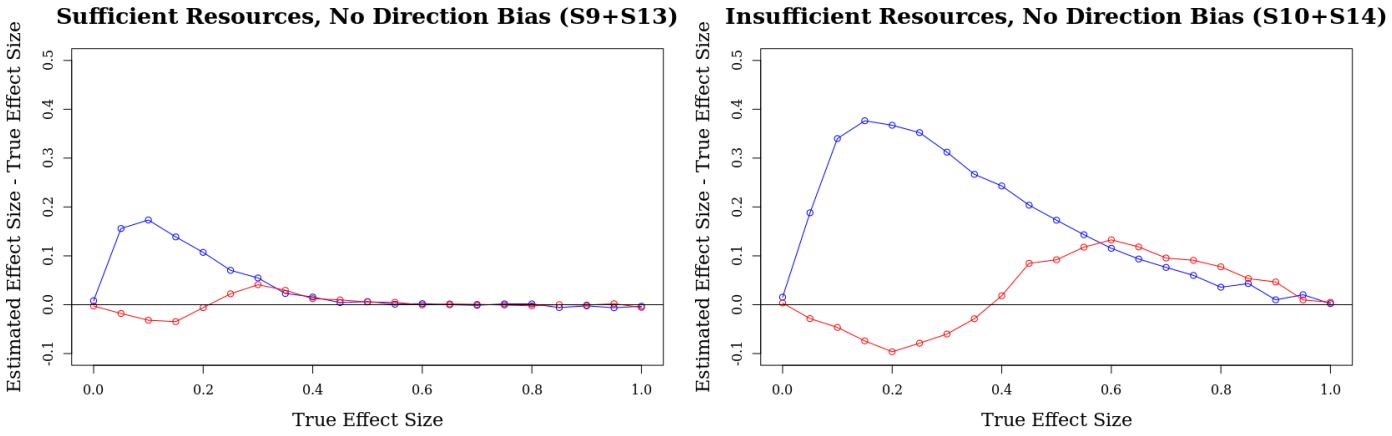

Figure 8: Difference between estimated and true effect size as a function of the true effect size (measured by standardized means difference), for scenarios with suppression of inconclusive evidence and *no* direction bias. Blue line = frequentist analysis, red line = Bayesian analysis.

Turning to the case of no direction bias, shown in Figure 8, two observations are striking. First, the frequentist graph ceases to be monotonically decreasing: small 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. 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 Figure 9 visualizes these explanations by plotting the probability density function of d, and the range of censured observations.

Second, the Bayesian slightly *underestimates* 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 Figure 9. For small positive effects such as $d = 0.1$ or $d = 0.2$, the Bayesian is more likely to obtain results that favor the null hypothesis with $BF_{01} > 3$, than results that favor the alternative with $BF_{10} > 3$. However, the amount of underestimation does not affect the qualitative interpretation of the effect size in

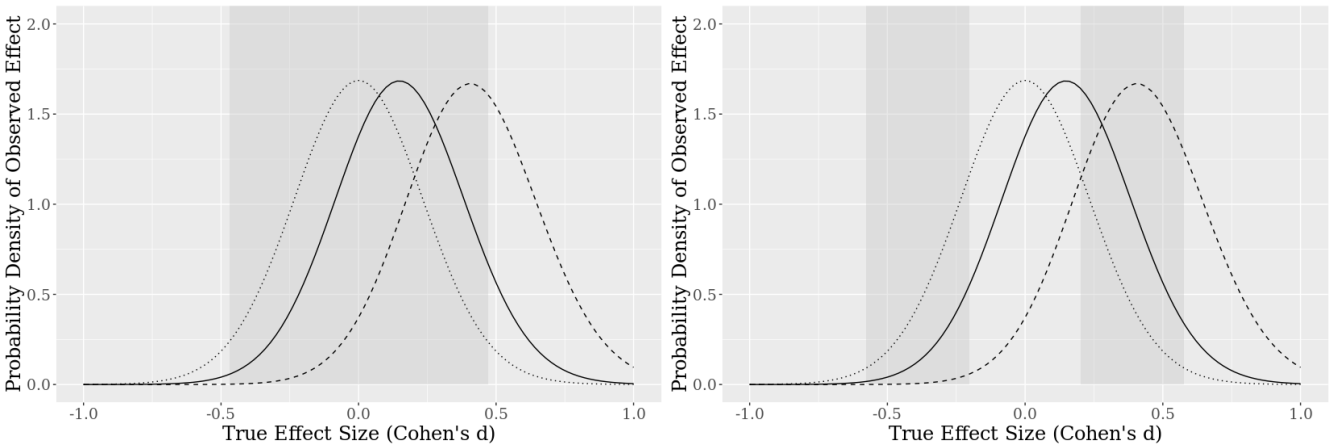

Figure 9: Probability density functions for the expected value of the standardized sample mean in a single experiment for $N = 36$ and different values of the real effect size. Full line: $d = 0.15$, dashed line: $d = 0.41$, dotted line: $d = 0$. The censured regions (i.e., observations that do not enter the meta-analysis because $p > .05$ or $1/3 < BF_{10} < 3$) are shaded in grey. Left graph: frequentist case, right graph: Bayesian case.

question.

All in all, it is notable that "two wrongs seem to make a right": omitting weak evidence in favor of the alternative *and* in favor of the null hypothesis leads to more accurate meta-analytic estimates than omitting statistically non-significant results only, as the frequentist does. These observations are especially salient for small effects. SCT*—the thesis about the self-corrective nature of science in sequential replications of an experiment—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.[8] 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.

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

7 Conclusion

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 suggestions what science should do in order 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 tests (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.

Our results support a partially positive verdict on the efficacy of statistical reform. When studies with inconclusive evidence—or at least a substantial proportion of them—are published, both Bayesian inference and frequentist inference with NHST lead to quite accurate estimates and validate SCT*. However, when inconclusive evidence is not published, as it often happens in scientific practice, Bayesian inference leads to more accurate effect size estimates. Specifically, the Bayesian framework avoids the large overestimations that frequently occur in NHST. Thus, SCT* holds for a wider range of scenarios using Bayesian inference and in these conditions, statistical reform will indeed improve the reproducibility of published studies and the reliability of experimental research.

Building on the extensions of our model where different effect size ranges are studied, we can qualify this conclusion further. The advantage of Bayesian statistics is particularly salient for small effect sizes, which the frequentist often misidentifies

as moderate or relatively large effects. This finding is in line with observations from replication research that small effects are at particular risk of being overestimated systematically.

Our study has evident limitations, too. First, our results do not prove that moving to Bayesian statistics is the best way to statistical reform: alternative frameworks within the frequentist paradigm (e.g., Lakens, Scheel and Isager 2018; Mayo 2018) could improve matters equally, or even more so. Assessing such proposals is beyond the scope of this paper. Second, 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 the credit reward scheme, funding replication work, etc.) 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. All other things being equal, adopting Bayesian principles for designing and analyzing experiments leads to more accurate effect size estimates, without incurring substantial drawbacks. Ultimately, this thesis should also be tested against a track record of published research: Are effect size estimates from experiments using Bayesian statistics really more robust in replication than their frequentist counterparts? Or does our model neglect important factors which offset the advantages of Bayesian inference? We are curious whether the recent surge of Bayesian methods in experimental research (e.g., Rouder et al. 2009; Lee and Wagenmakers 2014) will provide us with data that allow us to answer these questions.

References

Assen, Marcel A. L. M. van, Robbie C. M. van Aert, Michele B. Nuijten, and Jelte M. Wicherts (2014). Why Publishing Everything Is More Effective than Selective Publishing of Statistically Significant Results. *PLoS ONE 9*, e84896.

Bakker, Marjan, Annette van Dijk, and Jelte M. Wicherts (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science 7*, 543–554.

Benjamin, Daniel J., James O. Berger, Magnus Johannesson, Brian A. Nosek, E. J. Wagenmakers, Richard Berk, Kenneth A. Bollen, Björn Brembs, Lawrence Brown, Colin Camerer, David

Cesarini, Christopher D. Chambers, Merlise Clyde, Thomas D. Cook, Paul De Boeck, Zoltan Dienes, Anna Dreber, Kenny Easwaran, Charles Efferson, Ernst Fehr, Fiona Fidler, Andy P. Field, Malcolm Forster, Edward I. George, Richard Gonzalez, Steven Goodman, Edwin Green, Donald P. Green, Anthony G. Greenwald, Jarrod D. Hadfield, Larry V. Hedges, Leonhard Held, Teck Hua Ho, Herbert Hoijtink, Daniel Hruschka, Kosuke Imai, Guido Imbens, John P.A. Ioannidis, Minjeong Jeon, James Holland Jones, Michael Kirchler, David Laibson, John List, Roderick Little, Arthur Lupia, Edouard Machery, Scott E. Maxwell, Michael McCarthy, Don A. Moore, Stephen L. Morgan, Marcus Munafó, Shinichi Nakagawa, Brendan Nyhan, Timothy H. Parker, Luis Pericchi, Marco Perugini, Jeff Rouder, Judith Rousseau, Victoria Savalei, Felix D. Schönbrodt, Thomas Sellke, Betsy Sinclair, Dustin Tingley, Trisha Van Zandt, Simine Vazire, Duncan J. Watts, Christopher Winship, Robert L. Wolpert, Yu Xie, Cristobal Young, Jonathan Zinman, and Valen E. Johnson (2018). Redefine statistical significance. *Nature Human Behaviour 2*, 6–10.

Bernardo, José M. and Adrian F. M. Smith (1994). *Bayesian Theory*. New York: Wiley.

Borenstein, Michael, Larry V. Hedges, Julian P. T. Higgins, and Hannah R. Rothstein (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd.

Camerer, Colin F., Anna Dreber, Eskil Forsell, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Johan Almenberg, Adam Altmejd, Taizan Chan, Emma Heikensten, Felix Holzmeister, Taisuke Imai, Siri Isaksson, Gideon Nave, Thomas Pfeiffer, Michael Razen, and Hang Wu (2016). Evaluating Replicability of Laboratory Experiments in Economics. *Science*. https://doi.org/10.1126/science.aaf0918.

Cohen, Jacob (1994). The Earth is Round ($p < .05$). *Psychological Review 49*, 997–1001.

Cumming, Geoff (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-analysis*. Multivariate applications book series. Routledge.

Cumming, Geoff (2014). The New Statistics: Why and How. *Psychological Science 25*, 7–29.

Douglas, Heather E. (2009). *Science, Policy, and the Value-Free Ideal*. Pittsburgh: University of Pittsburgh Press.

Edwards, Ward, Harold Lindman, and Leonard J. Savage (1963). Bayesian statistical inference for psychological research. *Psychological Review 70*, 193–242.

Fanelli, Daniele (2010). Positive Results Increase Down the Hierarchy of the Sciences. *PLoS ONE 5*, e10068.

Fidler, Fiona (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. Chris and Simine Vazire (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS ONE 9*, e109019.

Gilbert, Daniel T., Gary King, Stephen Pettigrew, and Timothy D. Wilson (2016). Comment on "Estimating the reproducibility of psychological science". *Science 351*, 1037–1037.

Goodman, Stephen N. (1999a). Toward Evidence-Based Medical Statistics 1: The *P* value Fallacy. *Annals of Internal Medicine 130*, 995–1004.

Goodman, Stephen N. (1999b). Toward Evidence-Based Medical Statistics 2: The Bayes Factor. *Annals of Internal Medicine 130*, 1005–1013.

Hacking, Ian (1965). *Logic of Statistical Inference*. Cambridge: Cambridge University Press.

Hopewell, Sally, Kirsty Loudon, Mike J Clarke, Andrew D Oxman, and Kay Dickersin (2009). Publication Bias in Clinical Trials Due to Statistical Significance or Direction of Trial Results. *Cochrane Database of Systematic Reviews 1*, MRoooo06.

Howson, Colin and Peter Urbach (2006). *Scientific Reasoning: The Bayesian Approach* (3rd ed.). La Salle, IL: Open Court.

Hrdy, Sarah (1986). Empathy, polyandry, and the myth of the coy female. In Ruth Bleier (ed.), *Feminist approaches to scienceFeminist approaches to science*, pp. 119–146. New York: Teachers College Press.

Hubbard, Ruth (1990). *The Politics of Women's Biology*. Rutgers University Press.

Ioannidis, John P. A. (2005). Why most published research findings are false. *PLoS Medicine 2*, e124.

Jeffreys, Harold (1961). *Theory of Probability* (3rd ed.). Oxford: Oxford University Press.

Kass, Robert E. and Adrian E. Raftery (1995). Bayes Factors. *Journal of the American Statistical Association 90*, 773–795.

Klein, Richard A., Kate A. Ratliff, Michelangelo Vianello, Reginald B. Jr Adams, Stephan Bahnik, Michael J. Bernstein, Konrad Bocian, Mark J. Brandt, Beach Brooks, Claudia Chloe Brumbaugh, Zeynep Cemalcilar, Jesse Chandler, Winnee Cheong, William E. Davis, Thierry Devos, Matthew Eisner, Natalia Frankowska, David Furrow, Elisa Maria Galliani, Fred Hasselman, Joshua A. Hicks, James F. Hovermale, S. Jane Hunt, Jeffrey R. Huntsinger, Hans IJzerman, Melissa-Sue John, Jennifer A. Joy-Gaba, Heather Barry Kappes, Lacy E. Krueger, Jaime Kurtz, Carmel A. Levitan, Robyn K. Mallett, Wendy L. Morris, Anthony J. Nelson, Jason A. Nier, Grant Packard, Ronaldo Pilati, Abraham M. Rutchick, Kathleen Schmidt, Jeanine L. Skorinko, Robert Smith, Troy G. Steiner, Justin Storbeck, Lyn M. Van Swol, Donna Thompson, and A. E. van 't Veer (2014). Investigating Variation In Replicability: A 'Many Labs' Replication Project. *Social Psychology 45*, 142–152.

Koole, Sander L. and Daniel Lakens (2012). Rewarding Replications. *Perspectives on Psychological Science 7*, 608–614.

Lakens, Daniël, Anne M. Scheel, and Peder M. Isager (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science 1*, 259–269.

Laudan, Larry (1981). Peirce and the Trivialization of the Self-Corrective Thesis. In *Science and Hypothesis*, Volume 19 of *The University of Western Ontario Series in Philosophy of Science*, pp. 226–251. Springer Netherlands.

Lee, Carole J., Cassidy R. Sugimoto, Guo Zhang, and Blaise Cronin (2013). Bias in Peer Review. *Journal of the American Society for Information Science and Technology 64*, 2–17.

Lee, Michael D. and Eric-Jan Wagenmakers (2014). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge: Cambridge University Press.

Lexchin, Joel (2012). Sponsorship bias in clinical research. *The International Journal of Risk & Safety in Medicine 24*, 233–242.

MacCoun, Robert J. (1998). Biases in the interpretation and use of research results. *Annual Review of Psychology*, 259–287.

Mayo, Deborah (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Science Wars*. Cambridge: Cambridge University Press.

Munafò, Marcus R., Brian Nosek, Dorothy V. M. Bishop, Katherine Button, Christopher D. Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J. Ware, and John P.A. Ioannidis (2017). A manifesto for reproducible science. *Nature Human Behaviour 1*, 0021.

Nosek, Brian A and Timothy M Errington (2017). Reproducibility in cancer biology: Making sense of replications. *eLife 6*, e23383.

Nuijten, Michèle B., Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods 48*, 1205–1226.

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

Peirce, Charles Sanders (1931–1935). *The Collected Papers of Charles Sanders Peirce*, Volume I–VI. Cambridge, Mass.: Harvard University Press.

Quintana, Daniel 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.

Richard, F. D., Charles F. Jr. Bond, and Juli J. Stokes-Zoota (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology 7*, 331–363.

Romeijn, Jan-Willem (2014). Philosophy of Statistics. In E. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Retrieved from https://plato.stanford.edu/archives/sum2018/entries/statistics/.

Romero, Felipe (2016). Can the Behavioral Sciences Self-Correct? A Social Epistemic Study. *Studies in the History and Philosophy of Science*.

Romero, Felipe (2017). Novelty vs. Replicability: Virtues and Vices in the Reward System of Science. *Philosophy of Science 84*, 1031–1043.

Romero, Felipe (2018). Who Should Do Replication Labor? *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.31234/osf.io/ab3y9.

Rosenthal, Robert (1979). The File Drawer Problem and Tolerance for Null Results. *Psychological Bulletin 86*, 638–641.

Rouder, Jeffrey N., Paul L. Speckman, Dongchu Sun, Richard D. Morey, and Geoffrey Iverson (2009). Bayesian *t* Tests for Accepting and Rejecting the Null Hypothesis. *Psychonomic Bulletin & Review 16*, 225–237.

Royall, Richard (1997). *Statistical Evidence: A Likelihood Paradigm*. Chapman & Hall, London.

Royall, Richard (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, 760–768.

Schmidt, Frank L. (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods 1*, 115–129.

Schönbrodt, Felix D. and Eric-Jan Wagenmakers (2018). Bayes factor design analysis: Planning for Compelling Evidence. *Psychonomic Bulletin & Review 25*, 128–142.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science 22*, 1359–1366.

Spielman, Stephen (1974). The Logic of Tests of Significance. *Philosophy of Science 41*, 211–226.

Sprenger, Jan (2016). Bayesianism vs. Frequentism in Statistical Inference. In *The Oxford Handbook of Probability and Philosophy*, pp. 185–209. Oxford University Press UK.

Trafimow, David and Michael Marks (2015). Editorial. *Basic and Applied Social Psychology 37*, 1–2.

van Dongen, Noah N. N., Johnny B. van Doorn, Quentin F. Gronau, Don van Ravenzwaaij, Rink Hoekstra, Matthias N. Haucke, Daniël Lakens, Christian Hennig, Richard D. Morey, Saskia Homer, Andrew Gelman, Jan Sprenger, and Eric-Jan Wagenmakers (forthcoming). Multiple perspectives on inference for two simple statistical scenarios. *The American Statistician*.

Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, and Han L. J. van der Maas (2011). Why Psychologists Must Change the Way They Analyze Their Data: The Case of Psi. *Journal of Personality and Social Psychology 100*, 426–432.

Wilholt, Torsten (2009). Bias and values in scientific research. *Studies in History and Philosophy of Modern Science A*, 92–101.

Ziliak, Stephen T. and Deirdre N. McCloskey (2008). *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Economics, cognition, and society. University of Michigan Press.