Final draft – Published in Studies in History and Philosophy of Science Part A (2016), 60, 55 – 69.

Can the Behavioral Sciences Self-Correct? A Social Epistemic Study

Felipe Romero

Tilburg University

Abstract

Advocates of the self-corrective thesis argue that scientific method will refute false theories and find closer approximations to the truth in the long run. I discuss a contemporary interpretation of this thesis in terms of frequentist statistics in the context of the behavioral sciences. First, I identify experimental replications and systematic aggregation of evidence (meta-analysis) as the self-corrective mechanism. Then, I present a computer simulation study of scientific communities that implement this mechanism to argue that frequentist statistics may converge upon a correct estimate or not depending on the social structure of the community that uses it. Based on this study, I argue that methodological explanations of the "replicability crisis" in psychology are limited and propose an alternative explanation in terms of biases. Finally, I conclude suggesting that scientific self-correction should be understood as an interaction effect between inference methods and social structures.

Keywords: social structure of science; social epistemology; scientific self-correction; replication; frequentist statistics

1. Introduction

In 2011, a team of psychologists published the following experiment. One group of participants was presented with a set of photographs of places, some of which had an American flag (flag prime condition). Another group (control) was presented with photographs that were the same except that the flags had been digitally removed. Both groups were asked to estimate the time of the day the photographs were taken, and then complete a political belief survey intended to measure their endorsement of the worldview associated with the Republican Party in the United States. For example, they had to indicate their agreement on a 7-point scale to statements such as "Laws designed to protect the environment pose too high a cost on businesses that contribute to the economy." As a result of their study, the authors report:

> Participants who received a single exposure to an American flag exhibited a significant increase in Republican voting intentions, voting behavior, political beliefs, and implicit and explicit attitudes, with some effects lasting 8 months after the initial priming episode Carter et al. (2011).

This result, which I will refer to as the "flag-priming effect," is an example of what is known as a social priming effect. (The underlying hypothesis is that exposing people to socially salient stimuli influences behaviors related to the stimuli, over and above their conscious expectancies.) The result was published in one of the most prestigious journals in psychology, *Psychological Science*. It is also most likely false.

Since the 1990s, social psychologists have been invested in the still-thriving industry of social priming research. However, in the late 2000s, psychologists started reporting failed replication attempts of important social priming findings (Doyen et al., 2012; Pashler et al., 2012; Harris et al., 2013; Shanks et al., 2013). This led other researchers to organize replication efforts, and unfortunately the number of failures to replicate increased.[1] This includes the flag-priming effect (Klein et al., 2014). Particular experiments have been the locus of heated discussions (Yong, 2012; Bower, 2012). But the growing consensus is that social psychology is in a state of crisis.[2]

Social psychology is just one example. Replicability controversies also affect other sciences with recent attention drawn to biomedical research.[3] And more general arguments suggest that the number of false positives in the literature may be con-

[*] I am grateful to Carl Craver, John Doris, Frederick Eberhardt, Anya Plutynski, Roy Sorensen, Jan Sprenger, and Julia Staffel for comments on previous drafts. I also wish to thank Susan Fitzpatrick, Bennett Holman, Nazim Keven, Edouard Machery, Cailin O'Connor, Santiago Olivella, and Michael Strevens for helpful discussion. Previous versions of this work were presented at the 2nd Workshop on Inductive Logic and Confirmation in Science, University of Utah, October 2014; the Agent-Based Modeling in Philosophy Conference, Munich Center for Mathematical Philosophy, Germany, December 2014; and the Logic and Philosophy of Science Department, University of California, Irvine, March 2015. This research was partially supported by ERC Grant No. 640638, "Making Scientific Inferences More Objective".

[1] Some examples are the website PsychFileDrawer.org, intended to be a repository for quick reports of experimental replication attempts; the Reproducibility Project (Open Science Collaboration, 2012); and the "Many Labs" replication project (Klein et al., 2014).

[2] See the special sections in *Perspectives on Psychological Science* in 2012 (volume 7, issue 6); and 2013 (volume 8, issue 4) for a collection of articles discussing the crisis from multiple angles.

[3] Pharmaceutical companies have an interest in academic publications of preclinical studies, and try to replicate results that look promising, oftentimes with little success. In 2012, Amgen (a biotech firm) reported that their scientists could replicate just 6 of 53 "landmark" studies in cancer research (Begley & Ellis, 2012). And in 2011, a team at Bayer HealthCare (Germany) reported a similar experience (Prinz et al., 2011).

siderably larger than what can be expected in theory (Ioannidis, 2005). The cases we know about may just be the tip of the iceberg.

Can science correct these mistakes? The assumption that science is self-corrective often underlies the judgment of philosophers and scientists alike that science is epistemically privileged and truth-conducive. The idea is that, even if sometimes scientists pursue erroneous theories, if they persist in following the scientific method carefully and rigorously, they will eventually correct those mistakes, and find closer approximations to the truth. Philosophers have called this widely held belief "the self-corrective thesis." Here is one rough formulation:

SCT: In the long run, the scientific method will refute false theories and find closer approximations to true theories.

If truth is the goal of science, SCT (or something similar in spirit) is essential. It is also prima facie plausible. But, how can we assess its veracity? The generality of the thesis makes the answer difficult: there is no unique scientific method. Traditional answers focus on the theoretical long-run performance of particular inference methods (i.e., they study whether an inference method will eventually converge on an expected valid inference). Here, however, I show that this approach is insufficient to assess SCT. The prevalence of systematic production of scientific errors, such as the aforementioned, shows that SCT needs to be assessed from a wider social epistemological perspective: we need to study the social conditions under which scientific communities can discover and correct their mistakes, and the theoretical and practical feasibility of those conditions.

From such a perspective, I present a computer simulation study to show that SCT is only true in a *scientific utopia*: a scenario with highly idealized conditions governing the social structure of science. I show that, even in best-case scenarios (i.e., when findings are subject to strict and systematic replication attempts), social structures impose constraints on scientists' work that make it impossible for them to correct their mistakes in the way that SCT contends. This study is framed within the rich literature on the social organization of science.[4] But I take some steps to understand a gap that, as Longino (2015) points out, philosophers have not addressed yet: the gap between the self-corrective ideal and the reality. Specifically, I contribute to understanding the influence of social aspects of science on scientific self-correction in the following three ways: (i) I show quantitatively how some social aspects (i.e., the kind and availability of resources scientists have, their preference for particular patterns of results, and publication practices) affect estimated effect sizes in experimental psychology; (ii) I show that some critiques of classical statistical inference (e.g., the "file drawer" problem) apply in some social structures but not others. This leads us to qualify recommendations that focus primarily on changing statistical inference frameworks and disregard the social context in which such frameworks are deployed; and (iii) I show that some explanations of the causes of replicability controversies are limited, and propose an alternative explanation in the context of psychological research.

I have one caveat about the scope of my discussion. My main case of study is social psychological research as it is a pressing case in recent discussions. The conclusions of this paper apply in that context and extend to fields that rely on controlled experiments with randomized samples and frequentist statistical inference. This covers the vast majority of current research in the behavioral and social sciences but does not cover other fields. In particular, in fields such as anthropology, ecology, and medical research the notions of "experiment", "replication" and the inferential procedures may differ, which brings additional complexities. The discussion of SCT in those contexts deserves further and separate treatment.

The paper is organized as follows. In section 2, I frame SCT historically, reconstructing a tradition that goes back to C. S. Peirce, but the primary goal is to refine SCT so that it fits the methodology of frequentist statistics. I call that formulation SCT*. In section 3, I present the computer simulation study of SCT* and main discussion. And in section 4, I conclude arguing that philosophical attention to methodology can take us only so far in our understanding of the extent to which science self-corrects.

2. Self-Correction in Modern Scientific Methodology

2.1. SCT and Quantitative Induction

Explicit discussions of the idea that science is self-correcting go back at least to the seventeenth century (Laudan, 1981) and persist today (Allchin, 2015). In the twentieth century, perhaps the most prominent advocate of SCT is C. S. Peirce. A recurring idea in his writings is that "science is predestined to reach the truth" (Peirce, CP, 7.78) in an ideal limit of inquiry, which he grounds on a firm belief that science uses self-corrective methods. He discusses inductive inference, and more specifically the inference from samples to populations (or quantitative induction): take a random sample from a population, measure a parameter of interest, and then posit that measurement as the value of the parameter for the population. According to Peirce, such an inference is justified because if we repeat the procedure, taking more measurements of the parameter, we will necessarily get closer and closer to the true value of the parameter for the population. In Peirce's words:

> "[Quantitative induction] is a method which, steadily persisted in, must lead to true knowledge in the long run of cases of its application" (Peirce, CP, 7.207).

Reichenbach was also interested in the self-corrective properties of quantitative (or, as he called it, "enumerative") induction.[5] He acknowledges Peirce's insights, but he goes one step

[4] Several issues discussed here resonate with concerns about the division of cognitive labor (Kitcher, 1990; Solomon, 1992; Strevens, 2003), social conceptions of objectivity (Longino, 1990), and the divergence between individual and collective rationality (Solomon, 2001; Mayo-Wilson et al., 2011).

[5] Reichenbach defines the rule of quantitative induction as follows: "If an initial section of n elements of a sequence x_i is given, resulting in the frequency f^n, and if, furthermore, nothing is known about the probability of the second level for the occurrence of a certain limit p, we posit that the frequency $f^i (i > n)$ *will approach a limit* p within $f^n \pm \delta$ when the sequence is continued [italics added]" (Reichenbach, 1949, p.446).

further and contends that quantitative induction is at the core of scientific discovery. For instance, he writes: “the method of scientific inquiry may be considered as a concatenation of [quantitative] inductive inferences” (Reichenbach, 1938, p.364). His reasons, although not strongly defended, come from his belief that other forms of inference are reducible to that form of inference.[6]

Interestingly, Peirce was writing before the development of modern statistics, but the sort of inferential process that he had in mind lives on in frequentist statistics.[7] Some philosophers have pointed out the connection between Peirce’s ideas about self-correction and the Neyman-Pearson approach to hypothesis testing. For instance, Rescher (1978), Levi (1980), and Hacking (1980) contend that contemporary frequentist statistics instantiates Peirce’s ideas about self-correction in science. And more recently, Mayo (1996, 2005) argues that frequentist statistical methods provide mathematical rigor to Peirce’s SCT, and that Peirce’s SCT offers a rationale for such methods as a form of scientific induction.

Despite the prima facie plausibility and desirability of SCT, Peirce’s case for SCT is questionable because he seems to assume that proving the self-correction of quantitative induction suffices to establish SCT.[8] In this paper, however, I focus on quantitative induction. The reason is not that I endorse either the Peircean implicit idea that discussing quantitative induction exhausts the problem, or the Reichenbachian assumption that all other methods are reducible to quantitative induction. Rather, I want to show that even when we put aside the justification of the self-corrective character of other forms of inference, the justification of the self-corrective character of quantitative induction in its strongest version (namely, its instantiation in modern frequentist statistics) still faces problems.

In what follows, I refine a frequentist statistical formulation of SCT, which I call SCT*. I take this formulation to be in line with the philosophical tradition that I just discussed. But more importantly, I think that if we want to assess whether science self-corrects we have to look at the methods that scientists use, and frequentist statistical inference is the standard methodological approach in many disciplines. I first review quickly some basic concepts of frequentist statistics using the flag-priming experiment as an example (Section 2.2), and then proceed to the formulation of SCT* (Section 2.3). Readers familiar with null hypothesis significance testing, parameter estimation, and meta-analysis can skip to Section 2.3.

2.2. Hypothesis Testing, Parameter Estimation, and Meta-analysis

Within frequentist statistics, one dominant approach is Null Hypothesis Significance Testing (NHST). This approach is controversial because it synthesizes two procedures with different philosophical assumptions, Fisher’s test of significance and Neyman-Pearson decision theory (Gigerenzer, 2004). Despite the criticism, NSHT has been dominant in the behavioral sciences. Here I summarize the basics. The flag-priming experiment has two hypotheses. (I refer specifically to experiment 2 in the original article). One, the null (H_0), is the hypothesis that there is no difference between the two groups (i.e., flag-priming condition and control condition) in terms of their mean endorsement of a Republican worldview. That is, if μ_1 and μ_2 are the mean values of Republican endorsement for the flag-priming condition and control condition respectively, then $H_0 : \mu_1 - \mu_2 = 0$. The alternative hypothesis (H_1), is that there is actually a difference, that is $H_1 : \mu_1 - \mu_2 \neq 0$. Now, suppose the null hypothesis is true. If you repeat the experiment a large number of times, then, given random variation (i.e., fluctuations in your measurements due to chance), some of your results will suggest that the difference is positive and some will suggest that it is negative; but overall, most results will show that there is no difference between the two groups. To analyze the data from such an experiment, it is common to perform a *t*-test, that is, a test that examines whether the difference between two groups (control and treatment) could have happened if there were no genuine effect. Figure 1 shows the distribution of possible outcomes if the null hypothesis is true. The x-axis is the possible outcome of your experiment, or *t*-value (for our purposes, it suffices to say that a *t*-value is computed using the means of the two samples, their size, and their pooled standard deviation). The y-axis is the likelihood of getting such an outcome. For instance, if the null hypothesis is true, then it is more likely to get a result closer to zero.

NHST offers a procedure for “rejecting” the null hypothesis. You assume that the null hypothesis is true. You then define a *significance threshold* (e.g., the vertical lines in Figure 1) that demarcates the area for unlikely results. In standard practice, scientists define the thresholds to make the area comprise a small part of the whole distribution (e.g., 5%). You then compute a *p*-value, which is the probability of obtaining your datum or more extreme data, under the assumption that the null hypothesis is true. If your experimental result is unlikely (i.e., $p < 0.05$), you have a statistically significant result, and you reject the null hypothesis. For example, if you observe a positive *t*-value and a *p*-value less than 0.05, then an alternative distribution that is centered more to the right, like the one in Figure 2, makes the datum likely.[9]

Under the assumption that a particular alternative distribution is true (e.g, one in which most results are distributed around a

[6] He says, “All non deductive methods of the calculus of probability reduce to one kind of inference: the inference of induction by enumeration. Since all inductive methods of science, including the theory of indirect evidence and the formation of scientific theories, are interpretable in terms of inferences supplied by the calculus of probability, this result establishes the thesis that all forms of inductive inference are reducible to one form, to the inference of induction by enumeration” (Reichenbach, 1949, p.viii).

[7] Reichenbach, on the other hand, was aware of such developments. He cites the work of R. A. Fisher, J. Neyman, and E. Pearson in *The Theory of Probability* (Reichenbach, 1949).

[8] Peirce famously characterized different forms of scientific inference (e.g., qualitative induction and abduction), but he does not offer separate arguments for those forms being self-corrective, nor does he show that the self-corrective character of quantitative induction generalizes. For this reason, Laudan (1981, p.293) and von Wright (1965, p.226), while accepting Peirce’s conclusion that quantitative induction is self-corrective, conclude that his defense of SCT more generally falls short.

[9] The distribution depends on the sample size n and the effect size d. I explain the notion of effect size below. In this figure, $d = 0.5$ and $n = 32$.

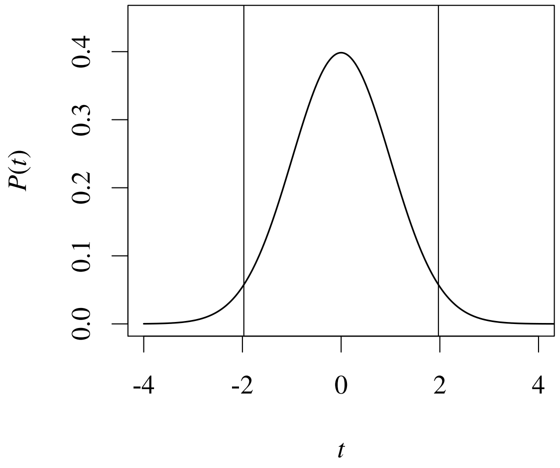

Figure 1: *t*-distribution for the null hypothesis.

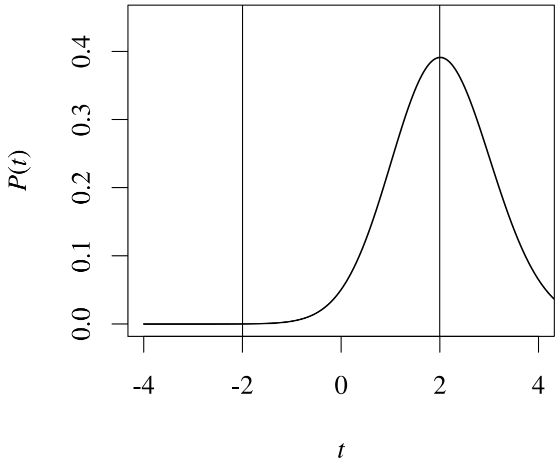

Figure 2: *t*-distribution for an alternative hypothesis.

positive magnitude because participants in the flag-prime condition have a stronger endorsement of the Republican worldview), there is a probability of detecting the result, called *statistical power*, which corresponds to the area of that distribution outside the boundaries of the significance threshold (roughly 50% in Figure 2). All other things being equal, having a larger sample in your experiment gives you a more powerful experiment.

Now, it is possible to reject the null hypothesis when it is true. This error is called a false positive or Type I error. For example, you conclude that there is a difference between the two groups when there actually isn't. One of the virtues of NHST (if not the most important) is that it allows you to control the long-run probability of such errors. The procedure does not say that if you get a significant result, then the null is necessarily false. Neither does it imply that the alternative hypothesis is true. Instead, assuming frequentism is correct, after obtaining a significant result, you are justified in acting as if there is an effect, because in the long run, you would be rejecting a true null hypothesis by mistake only 5% of the time, which is regarded as an acceptable risk.[10]

As I said earlier, the flag-priming experiment is very likely a false positive. How do we know this? In response to the growing skepticism about psychological findings, in 2013 a group of psychologists launched a collaborative replication project (Klein et al., 2014). 36 labs all over the world attempted to replicate a set of 13 studies, one of which was the flag-priming experiment. (In this particular experiment, only the data from replications attempted in the United States were counted, a total of 25 labs.) None of the labs found the flag-priming effect. The original study yields an effect size $d = 0.5$ and $p < 0.05$. But of the twenty-five replication attempts reported by Klein et al. (2014) only one reported $p < 0.05$, with an effect in the opposite direction. (See explanation of d values below.)

NHST, and in particular the practice of rejecting and accepting hypotheses based uniquely on p-values, has been strongly criticized on the basis that NHST practitioners have had a tendency to confuse statistical significance with scientific import (Ziliak & McCloskey, 2008). As a result, some methodologists have suggested that researchers should abandon NHST and estimate more informative parameters, *effect sizes* and *confidence intervals* in particular (Schmidt, 1996; Cumming, 2012):

> *Effect size.* An effect size is a measure of the difference for the value of the parameter between the control and the experimental condition. One common way to report it is Cohen's d: the difference between the means of the distributions for the null and the alternative hypotheses in standard deviation units (e.g., imagine the two distributions in Figure 1 and Figure 2 are in the same plot, and measure a standardized distance between the top of the two). Psychologists have conventions to interpret Cohen's d values: d around 0.2 is considered small, around 0.5 is considered medium, and around 0.8 is considered large. In the original flag-priming experiment $d = 0.5$ (non-negligible).[11]
>
> *Confidence interval.* Confidence intervals give an idea of how precise the estimation is: more precise estimates have narrower confidence intervals. A common practice is to report a 95% confidence interval. More precisely, if we

[10] Critics of frequentism question this interpretation of significant results, particularly for effects with low probabilities. I assume this interpretation here, however, because it is common in practice. For a philosophical discussion about how statistical evidence should be interpreted see Sprenger (2016).

[11] The use of standardized effect size measures such as Cohen's d is questionable, particularly to report treatment effectiveness in clinical trials (Stegenga, 2015). However, I use Cohen's d as it is still the most popular measure of effect sizes in cognitive science research.

> repeat the experiment long enough, 95% of the computed intervals (one for each experiment) will contain the true effect size. The original flag-priming experiment has the 95% confidence interval [.01, .99] (computed by Klein et al., 2014).

In an example, the shift from NHST to the estimation of effect sizes and confidence intervals is a move from the question "Are people subject to the flag-priming effect?" to "How strong is people's endorsement of a Republican world view after being flag-primed?" While the answer to the first question could be "yes", the answer to the second could give more insights about the importance and noticeability of the effect. Now, estimating effect sizes accurately makes apparent the need of two practices, *replication* and *meta-analysis*:

> *Replication*. Replicability is the gold standard of scientific findings. Methodologists in psychology distinguish two types of replications (Schmidt, 2009). On the one hand, *direct replications* are experiments that mirror the original experimental design, allowing variations only in factors that shouldn't be regarded as causally relevant to the original result (e.g., sample size).[12] On the other hand, *conceptual replications* are experiments designed to find an effect that would be expected were the original effect real. While these terms are used mostly in psychology, a similar distinction applies to other fields. Cartwright (1991), for instance, discusses the distinction in the context of economics.
>
> *Meta-analysis*. Meta-analyses aggregate data from multiple experiments that investigate the same phenomenon (direct or conceptual replications) (Schmidt, 1992). The standard models work by computing weighted averages of effect sizes and confidence intervals. For instance, the aggregated effect size of the 25 replication attempts of the flag-priming experiment is $d = 0.03$ (contrast this value with the $d = 0.5$ reported in the original experiment), and not statistically significant ($p = 0.38$) (Klein et al., 2014). Assuming that the replication attempts were properly carried out, this result is thought to be closer to the true value.

2.3. SCT: SCT in Modern Frequentist Statistics*

The concepts of null hypothesis significance testing (NHST) and meta-analysis relate in the following way. For the frequentist, when we measure an effect size we are justified in positing that estimate as the true effect for practical purposes. The estimate could be far from the true value, but replications of the experiment will give us other estimates, and their aggregation by meta-analysis will correct the estimation. The larger the number of experiments we aggregate, the narrower the confidence intervals, and, therefore, the more precise the estimate. Using these concepts, we can state a strong contemporary formulation of the self-corrective thesis, in terms of frequentist statistics, and in line with the philosophy of science tradition mentioned above:

SCT*: Given a series of replications of an experiment, the meta-analytical aggregation of their effect sizes will converge on the true effect size (with a narrow confidence interval) as the length of the series of replications increases.

SCT* offers a rationale for core scientific practices (here illustrated in the context of frequentist statistics): experiments using random samples,[13] estimation of parameters, and aggregation of multiple experimental results.

Now, notice that SCT* (and SCT in general) is completely silent about social aspects of the process it describes. However, we can more realistically think of the process as one in which multiple scientists intervene. In the next section, therefore, I discuss the conditions under which SCT* is true as well as its limitations from a social epistemological perspective.

3. Scientific Utopia and SCT* In Silico

Imagine an ideal scenario for scientists to make discoveries, to confirm their theories up to a more than merely acceptable degree, and to disconfirm false ones. I will call this scenario *scientific utopia*.[14] In this section I present a computer simulation study of SCT* in the scientific utopia and in less utopian scenarios.[15]

Suppose a community of frequentist scientists is interested in one purported effect. The epistemic goal of the community is to discover that effect. To do so, they perform experimental replications and cumulative meta-analyses (Cumming, 2012; de Winter & Happee, 2013; van Assen et al., 2014). The basic set-up is as follows. Suppose scientists A and B are part of the community. First, A performs the original experiment. The background assumptions for frequentist statistics to work are met: A succeeds in randomizing her samples and applying the treatment selectively. Then, B attempts a replication of A's experiment, and performs a meta-analysis aggregating her results and A's results. A third scientist C, then, runs another replication, and aggregates her results with A's and B's results, and so on. The long-run performance of the community is assessed in terms of the reliability (i.e., how disperse the estimates are) and

[12] Direct replicability does not entail scientific importance, although any scientifically important finding is in principle directly replicable. Also, direct replicability does not rule out methodological flaws in the experiment's design.

[13] See Worrall (2010) for a critical discussion about the epistemic weight of randomized control trials.

[14] The scientific utopia that I describe here is an echo of Sir Francis Bacon's utopian society in *New Atlantis* (1627). In his novel, Bacon portrays an institution called Solomon's House, which is structured with the ultimate end of expanding human knowledge. Division of labor plays an important role at Solomon's House: some members design experiments, others execute them, and others compile results and generalize findings. For a recent use of the term "scientific utopia" see Nosek & Bar-Anan (2012) and Nosek et al. (2012), who used it in discussions about bad practices in psychological research.

[15] One could label this computer simulation study as *systems-oriented* social epistemology: a study of an epistemic system in terms of its practices and protocols and how they affect the epistemic outcomes of its members (Goldman, 2009). In this case, the epistemic system is the institution of science (or more precisely, a scientific community of frequentist scientists) and the epistemic outcome is the estimation of a parameter of interest.

validity (i.e., how close to the truth the aggregate estimate is) of the parameter given by the meta-analyses as more evidence is gathered. In a computer simulation study it is possible to assess these properties since the value for the real effect is part of the simulation assumptions.

I characterize the utopia in terms of social conditions required for such a procedure to succeed. The conditions are the following:

1. Sufficient resources: In the utopia, scientists have enough time, participants, assistants, and funds to run all the experiments and direct replications they want. In what sense can these resources be sufficient? These resources determine how large the experiment's sample size n is, so they determine the experiment's statistical power (i.e., the probability of detecting an effect of a given magnitude under the assumption that the effect is real).[16] In other words, what I mean by resources being sufficient to find an effect d is that n is large enough to yield a statistical power for detecting d or a larger effect.
2. No direction bias: In the utopia, scientists report the results of their experiments regardless of whether they are consistent or not with previous theoretical expectations. In particular, they report their findings regardless of whether the direction (sign) of the effect d is expected. A direction bias could be a mistake of an individual scientist, and in this sense they wouldn't be aspects of the social structure of science. However, direction bias also results from institutional pressures, and it is in this sense that they constitute a social structural problem.
3. Negative results are published: In the utopia, the editorial system publishes results related to the effect regardless of their statistical significance. That is, publication is not based on whether the test's significance level (p-value) is below some threshold, so negative (i.e., non statistically significant) results are published.[17]

Perhaps unsurprisingly, SCT* is true in a scenario in which conditions 1–3 are satisfied. To illustrate this, in the next section I show a simulation example in which the conditions are met.

3.1. SCT* in Utopia

Suppose that the community is trying to find an effect. Let's assume that if the effect is true, then it is the typical effect reported in social psychological research. This corresponds to $d = 0.41$ (Richard et al., 2003; Fraley & Vazire, 2014). Here I give the intuition of how this process works in the utopia using a simulation run. For the reader interested in technical details, the footnotes in this section contain the mathematical functions used in the simulations.

Consider the sufficient resources condition. The standard recommendation (Cohen, 1992) is that the sample size n should be sufficient for a statistical power of 0.8 (i.e., n should be such that 80% of the time scientists would detect the effect). Strictly speaking for any n an experiment with $n + 1$ subjects yields higher statistical power, but experiments cannot have infinite sample sizes, so n has to be a specific value. I will assume that in utopia scientists have enough resources to go beyond the desirable 0.8 power requirement, and can run their experiments to obtain 0.95 power, assuming that they aim to detect the typical effect size.[18] In this example, that means that if scientists run their experiments with $n = 156$ and the effect is $d = 0.41$, then 95% of the time they will reject the null hypothesis.[19] Figure 3 shows the theoretical t-distribution for the assumed real effect size and sample size.[20]

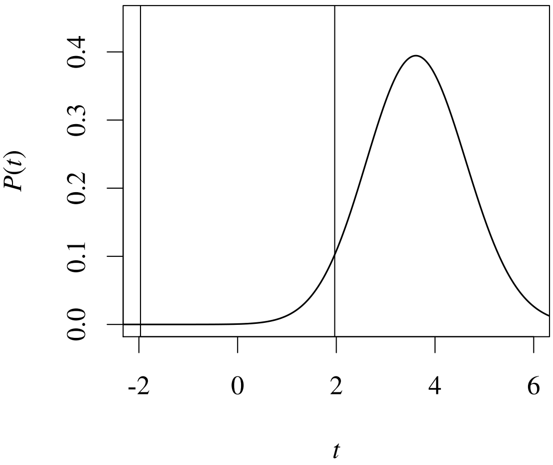

Figure 3: t-distribution for $d = 0.41$ and $n = 156$.

Suppose A runs the first experiment. I simulate this process as a function that produces a random deviate from the theoretical t-distribution and gives the result (i.e., an observed effect

[16] In frequentist statistics there is a relation between four variables: sample size, effect size, statistical power, and the false positive rate (Cohen, 1992). Standard practices fix the false positive rate. If we fix the false positive rate and the effect size, the two remaining variables covary: as the sample size of an experiment increases, its power also increases. In other words, if the effect is real, then the larger (or smaller) the sample, the more (or less) likely one will be to detect the effect (or a larger effect). In theory, this means that when scientists have limited resources, understood as limited access to samples, they will be more likely to run low-power experiments.

[17] Negative results should not be confused with an effect of negative magnitude, since non-significant results can have either positive or negative magnitudes.

[18] Indeed, it is difficult to justify the 0.8 value since it is asymmetrical with the significance level. 0.95 would be symmetrical (Machery, 2012, fn.4).

[19] To compute this sample size, one has to produce the right theoretical distribution associated with the effect that would have the desired percentage of its area falling outside the significance threshold(s). This is usually done by numerical approximation. I used G*power, a tool to compute power analysis for different tests (Faul et al., 2007).

[20] The distribution is a non-central t-distribution, which represents the possible outcomes of a two-tailed between-subjects t-test when there is an effect. In the distribution for the null hypothesis, t is distributed around 0. When there is a real effect, t is distributed around a different value. To generate the distribution, one requires the *degrees of freedom* df, given by df $= 2n - 2$; and a *non-centrality parameter* ncp, given by ncp $= d/\sqrt(2/n)$.

size) in Cohen's *d* units as output.[21] Suppose A finds an effect size of 0.46. Given condition 1, A publishes the result. Figure 4 shows the published studies. A's finding is the first data point in the figure, with the 95% confidence interval.[22] The dotted horizontal line shows the real effect size ($d = 0.41$).

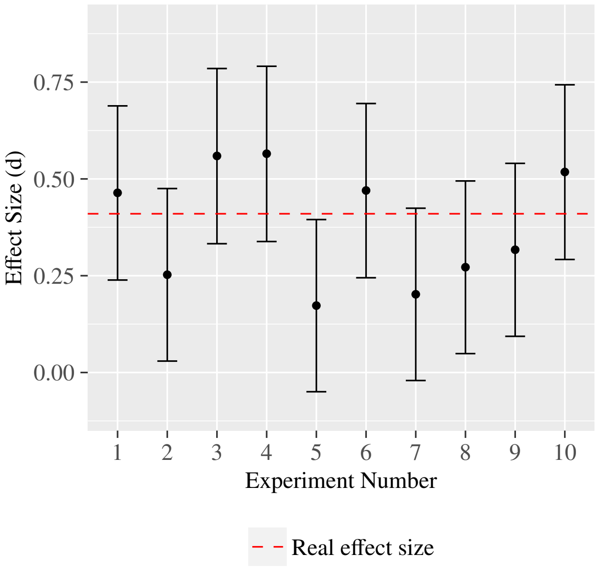

Figure 4: Publications of experiment replications.

Now, B thinks A's experiment is well designed. But they are in a scientific utopia, so B does not need to trust A. Given condition 2, B can *directly* replicate the experiment herself, because that is not going to slow her down in her other projects, and it is not going to drain her funding sources. So B calls A and A kindly sends B a detailed description of the protocols and materials. B draws 156 subjects from her infinite pool to ensure high statistical power, and runs a direct replication (i.e., we take another random deviate from the same theoretical *t*-distribution). Suppose B finds an effect size of 0.25, which corresponds to the second data point in Figure 4.

Now, B does not make an assessment of the effect based only on her result. She aggregates the effect size she found and the one A reported in a meta-analysis to get a more precise estimate of the real effect size.[23] Figure 5 shows the progression of meta-analyses in time (or cumulative meta-analysis). Because A is the first person running the experiment, A's meta-analysis simply yields the effect size of her experiment. B's meta-analytical result is the second data point in the figure. Now, the process repeats for the rest of the community. Each scientist runs her own replication and aggregates her finding with all the previous findings. And as expected, the estimated effect size gets closer to the true effect, and the corresponding confidence intervals get more precise.

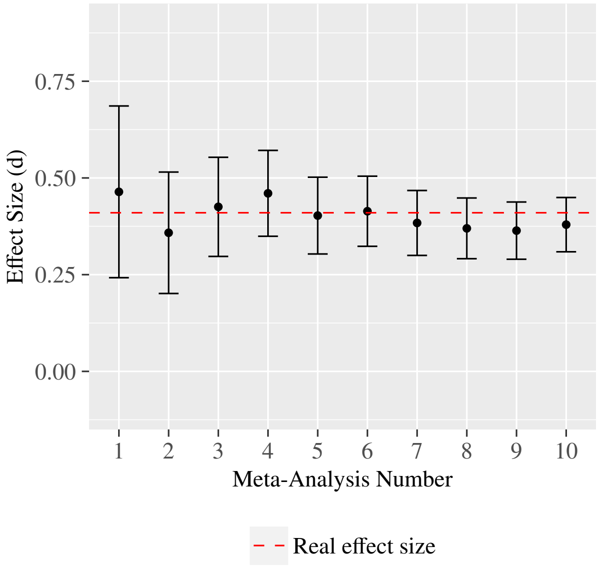

Figure 5: Meta-analyses of previous publications.

This scenario constitutes a baseline. It captures the intuitions that motivate SCT*, and illustrates how self-correction occurs in the long run, within a frequentist statistical framework, as the tradition that I traced back to Peirce contends.

But what happens to SCT* when the social organization of science is less than utopian? I address this question in the next two sections.

3.2. *Non-utopian Scenarios: Setup*

As many authors have pointed out, scientists work with limited resources (Kitcher, 1990, 1993), they are subject to biases (Ioannidis et al., 2014; Anderson, 2015, §5), and negative results are typically not published (Rosenthal, 1979). In the next two sections, I show quantitatively in the model how this reality theoretically affects SCT*, and could practically affect self-correction of social psychological research. But first I discuss some philosophical work and empirical evidence regarding how conditions 1–3 are often violated, along with an explanation of how I capture them in the model. The particular values for the

[21] Suppose *t* is the random deviate from the *t*-distribution. The standard formula to obtain *d* is $d = t\sqrt{(2/n)}$. This formula, however, produces a biased *d* when *n* is small. The formula $d_{\text{unb}} = (1-3/(4\text{df}-1))\times d$ is an unbiased estimate. Here what I refer to as *d* is really d_{unb}. See Cumming (2014, pp.294–295) for discussion of this correction.

[22] The confidence intervals depend on the variance of the observed effect size. Given an observed effect size *d*, a good approximation of *d*'s variance is given by $v = \frac{2}{n} + \frac{d^2}{2n^2}$ (Borenstein et al., 2009, p.27). The variance not as sensitive to *d* as it is to *n* (at least for the ranges of both values that I consider here). In all simulations, replications in the same scenario have the same sample size. Hence it is expected in Figure 4 to see very similar confidence intervals (even though they are slightly different).

[23] I use the *fixed-effect model* of meta-analysis. This model assumes that all experiments under consideration are estimating the same true effect size. It produces a weighted mean of the experiments' effect sizes, in which the weighting factors depend on the variance v_i (see previous footnote) of each experiment. More precisely, suppose we have a series of experiments. Then, the mean effect size M is given by $M = \frac{\Sigma w_i d_i}{\Sigma w_i}$, for all i in the series, where $w_i = 1/v_i$ (i.e., the inverse of the variance). The variance of M is given by $v_M = \frac{1}{\Sigma w_i}$. See Cumming (2012, pp.210–213) and Borenstein et al. (2009, pp.63–67).

	$d = 0.41$				$d = 0$				$d = 0.41$				$d = 0$			
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16
Sufficient Resources	✓		✓		✓		✓		✓		✓		✓		✓	
No direction bias	✓	✓			✓	✓			✓	✓			✓	✓		
Negative Results are Published	✓	✓	✓	✓	✓	✓	✓	✓								

Table 1: Simulation Scenarios

parameters intend to capture the context of social and personality psychology. However, importantly, the qualitative conclusions and philosophical implications that I will derive are not tied to particular parameter values so they apply to the behavioral sciences more generally.

1. Limited resources: The economic structure of society has played an important role in shaping scientific research in the past century (Stephan, 2012). These days, science is a professional activity, and scientists have more pressures to find and secure funding sources, which could impact the quality of their scientific output.[24] In most research programs, scarce resources of different kinds, such as money to pay participants, staff to collect data, and time lead experimenters to work with small samples (e.g., collecting data from babies in developmental psychology experiments requires a great deal of patience and time), These resources deficiencies and extreme pressure to publish forces researchers to run low-power experiments. Now, in practice we know that statistical power in published research in psychology has been historically low —slightly below 0.5 for medium effects.[25] Shockingly, this means that researchers have been roughly as good as a fair coin to detect effects, assuming that the effects are real. This fact may not be fully explained by lack of resources.[26] Nonetheless, keeping everything else constant, in cases where scientists do have limited access to participants, power tends to be lower.

   In the model: I relax the "sufficient resources" condition by reducing the sample size n from 156 to 36, which, for a real effect size $d = 0.41$, reduces statistical power to 0.4 —a value that is slightly below 0.5, as common in practice in psychology.

2. Direction bias: Feminist philosophers of science have identified multiple sources of bias and their negative (and sometimes positive) impact in scientific objectivity in specific cases (Longino, 1990; Solomon, 1992; Anderson, 2015). The form of bias that concerns me here arises when scientists have theoretical commitments. Such commitments help them to envisage experiments and also expect particular patterns of results. This kind of expectation may not be a problem in early stages of a research program, but it is in later stages (Douglas, 2009, p.102). In particular, this expectation could lead to motivated reasoning and create an "egocentric bias towards one's own data" (Solomon, 2001, p.57) in which the researcher filters the available data to only a subset that gets processed and published. For example, the firm believer in social priming theory is less likely to predict an effect in the opposite direction for the flag-priming effect (i.e., that participants in the flag prime condition would exhibit a decrease in Republican voting intentions). This could lead her to disregard evidence for that prediction. In the best case, researchers are unaware of their direction biases. In the worst case, direction biases are a conscious practice. As a consequence of the professionalization of contemporary science, researchers have substantially less freedom and incentives to pursue research programs that their colleagues think are misconceived (Stanford, 2015, §1.1). These pressures influence the publication system. For example, after a journal has published a series of articles supporting a trend, its threshold for accepting articles with contradictory evidence is higher. Evidence of biases at the editorial level (Lee, 2013) and peer review level (Lee et al., 2013) make this story plausible. Also, direction biases explain why results that researchers think are striking and have positive magnitude have higher odds of being published (Hopewell et al., 2009).

   In the model: I relax the "no direction bias" condition by filtering effects depending on their sign. Specifically, scientists have a bias for effects of positive magnitude ($d > 0$). This means that scientists do not produce results of negative magnitude ($d < 0$).

3. Publication of positive results only: In practice, there is a tacit rule in many disciplines that says that only statistically significant results (also referred to in the literature as "positive results") are to be published. This is intuitively

[24] This is arguably a recent challenge. At the origins of modern science and up to a transition period that started around the 19th century, scientists were supported largely by private wealthy patrons, and their enterprise was fairly protected from a market economy.

[25] Cohen (1962) reviewed 70 articles in the *Journal of Abnormal Psychology* and found a that statistical power was on average 0.48. Twenty-four years later, Sedlmeier & Gigerenzer (1989) repeated the same study in the same journal, and they found power to be (median) 0.44. Maxwell (2004) reported little improvement, as did Fraley & Vazire (2014) more recently for social and personality research.

[26] Another possible explanation is that, for a good while, most psychologists didn't really care about statistical power, in part because it was not central in the statistics curricula, nor a requirement for journal submissions.

expected given the widespread practice of null hypothesis significance testing (NHST). Fanelli (2010) reports shocking metrics of this phenomenon. He analyzed 2434 papers from all disciplines that report hypothesis testing, and shows that as we go down in "the hierarchy of sciences" (i.e., a hypothesized hierarchy that Fanelli traces back to Comte, in which physics is at the top and the behavioral sciences at the bottom—of course, we need not agree that there is such a hierarchy), published results tend to be more significant than non-significant. In particular, psychology and psychiatry have the highest percentage of significant results (91.5%). And the rate of published non-significant results has also diminished in the last decades (Fanelli, 2012).

In the model: I relax the "negative results are published" condition by making statistical significance the threshold for publication. As explained before, an original experiment (or a replication) is a random deviate from a t-distribution. Publishing a result only if it is statistically significant means that the result will be published and included in the meta-analysis only if the deviate produces a p-value below 0.05. There is an interdependence between direction bias and publication of positive results only. When scientists publish only statistically significant results and there is an effect, having a direction bias that is consistent with the effect or not makes no difference.

Another parameter in the model that is worth examining explicitly is the assumed real effect size. A good heuristic to design experiments is to try to find a theoretically expected effect (e.g., such as the typical $d = 0.41$). In reality, of course, the effect could be larger or non-existent. In the former case the effect would be easier to detect. But what happens when there is not a real effect? To model such a possibility, I consider scenarios in which scientists assume that the effect they are trying to find is the typical effect, but in reality $d = 0$ (or close enough to zero to be of any scientific import). Being mistaken in such a case renders a false positive.

Relaxing conditions 1–3 in the way just discussed, and considering that effects might be non-existent produces 16 possible scenarios (i.e., 4 parameter changes, each of which has 2 possible states). Table 1 lists all these scenarios. I discuss all of them in the next three sections. In section 3.3 I focus on the effects of relaxing only conditions 1 and 2 (scenarios S1–S8) —the scenarios in which non-significant results are published. Condition 3 is studied in section 3.4. In section 3.5 I discuss some implications for replicability controversies.

3.3. Non-utopian Scenarios: Relaxing Conditions 1 and 2.

Scenarios S1–S8 in Table 1 are the scenarios in which negative results are published. These will be sufficient to establish my main social epistemological conclusion. The check mark (✓) states whether the utopian assumption is in place. S1, for example, is the utopia (all conditions are met); and S6 is a scenario in which negative results are published, scientists don't have direction biases, they work with limited resources to detect a typical effect, but they are pursuing a phenomenon that does not exist.

I simulate and aggregate the results of 500 communities of 50 scientists. In reality it does not happen that 500 teams pursue the same effect. However, I use this value to obtain an idea of the *spread* of possible estimations that a single community can make. On the other hand, I chose 50 scientists in each community to have an idea of an upper value that may occur in practice (i.e., 50 independent teams replicating a single experiment). This value is optimistic. As mentioned before, Klein et al. (2014) reported 36 independent attempts to replicate the same experiments, but most experiments in psychology are not replicated that many times.

Figure 6 and Figure 7 show the simulation results. The horizontal dotted line shows the real effect ($d = 0.41$ or $d = 0$). The shaded area shows one standard deviation above and below the mean effect size. That is, most communities fell within the shaded area.

Here are some qualitative observations about these simulation scenarios. The first observation is

- In all scenarios, at the beginning, a community's assessment of an effect size is very imprecise. This can be seen in the gap between the standard deviation lines. However, by the time the community performs the 50th experiment, the 50th meta-analysis converges on some effect size with a reduced margin of error.

- Reducing samples size (and therefore lowering statistical power) reduces the *speed* of convergence. This can be seen clearly by comparing the limited resources scenarios S2 and S6 with the sufficient resources scenarios S1 and S5 respectively: the spread in the standard deviation lines is wider in the former than in the latter.

- Direction biases might be intuitively expected to have an important impact on SCT*. This is partially confirmed. When there is not a real effect, direction bias leads communities to systematically report an effect when there isn't really one, as shown in S7 and S8. Less intuitive perhaps is that direction bias has no noticeable impact when there is a real effect *and* resources are sufficient. That is, directional bias in reporting results does not entail a directional bias in the long run estimation: compare S3 (direction bias) with S1.

- Limited resources *and* direction bias jointly inflate effect size estimates, as shown in S4 and S8.

- Looking only at scenarios S1–S4, someone might think that SCT* is not seriously compromised by social conditions, because most communities are is still finding the effects, just more slowly (S2) and with slight overestimations (S4). Such an assessment is too optimistic. Scenarios S1–S4 are favorable for SCT* because they assume that there is always an effect to be discovered. Importantly, the more worrisome cases, however, are S7 and S8: scientists

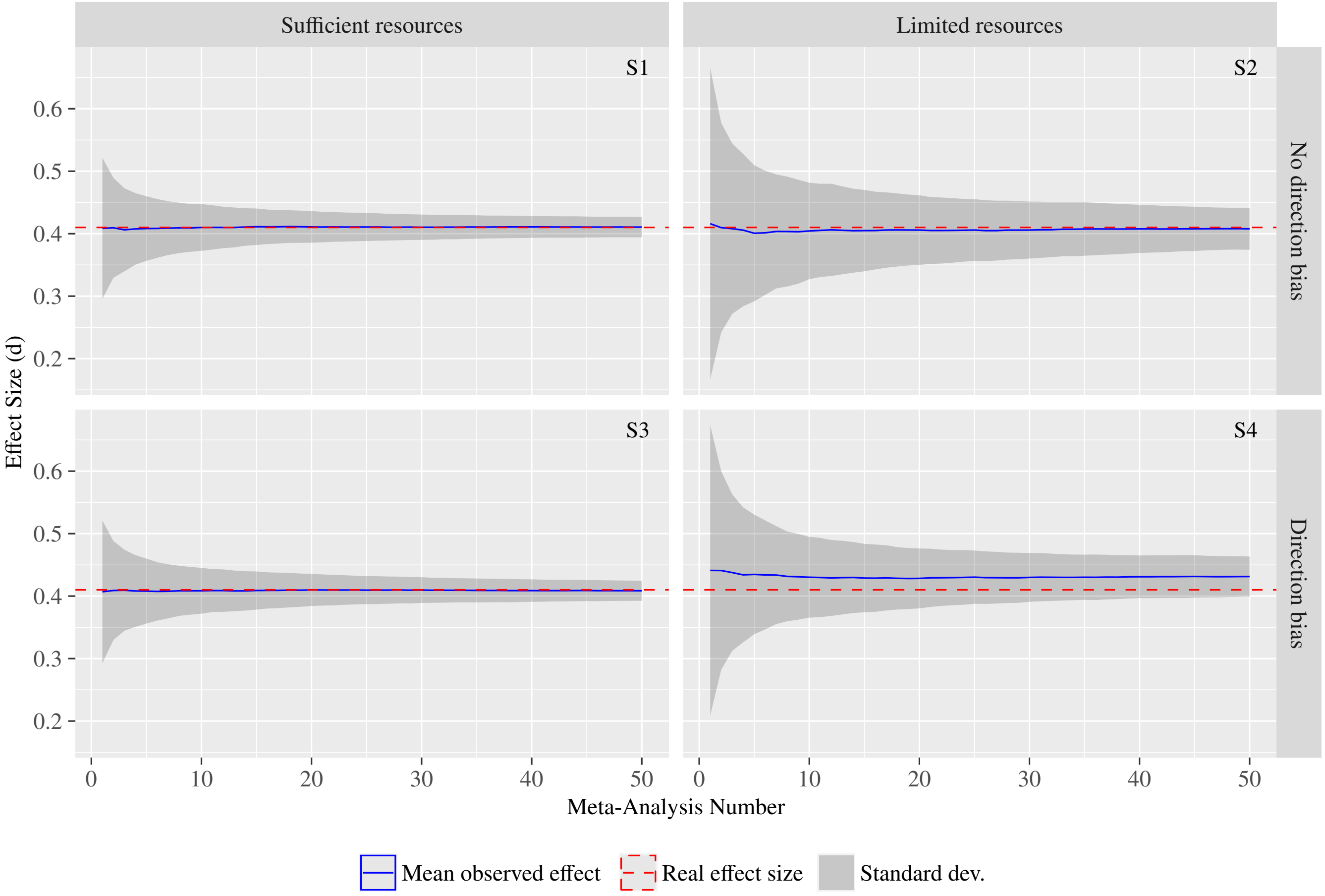

Figure 6: Simulation Results for Scenarios S1–S4: Real effect size $d = 0.41$

systematically report that effects exists when they don't exist at all.

Scenarios S1–S8 show how the social structure of science affects SCT*. In order for the process described by SCT* (i.e., experimental replications and aggregation of evidence by meta-analysis) to be effectively self-corrective, certain specific highly idealized social conditions have to be in place. That is, SCT* is true in a scientific utopia (S1), but its truth is compromised in non-utopian scenarios that reflect the social organization of contemporary practice. Of course, scenarios S1–S8 model the context of psychological research, so claims about magnitudes do not generalize, but the same social conditions affect research programs in the behavioral science more broadly.

Now, I will turn to the second set of scenarios: those in which we relax the condition that negative results are published.

3.4. *Non-utopian scenarios: Critique of NHST and Critique of the Critique*

As mentioned earlier, NHST is widely used and therefore utopian "negative results are published" condition is systematically not met in practice. This leads to the "file drawer problem" (Rosenthal, 1979)—the problem that negative results might stay in file drawers because scientists lack incentives to publish them, which introduces biases for significant results in the literature. The file drawer problem constitutes a strong criticism of NHST. In this section I show what this critique consists in using the model, and then proceed to some important qualifications. Figure 8 and Figure 9 show the scenarios in which condition 3 is not met.

Why did the practice of publishing only statistical significant results gain traction and momentum in the life sciences? Primarily perhaps because of its simplicity. Also because of difficulties inherent to non-significant results. Traditionally, non-significant results are regarded as inconclusive: a result could be non-significant because there is actually no effect to find, but also due to other complications that prevent experimenters from detecting a real effect. Additionally, the null understood as the hypothesis that an effect size is *exactly* zero, is almost certainly always rejected: an experiment with a sufficiently large sample size, almost certainly detects some difference (Meehl, 1967).[27] For these reasons, the rules for accepting and publishing null hypotheses (and whether we should do it at all) are controversial.

[27] However, scientists aware of this problem do not try to reject the hypothesis that an effect size is exactly zero, but that the effect is trivially small (Friston, 2012). See also Machery (2014, pp.271–273).

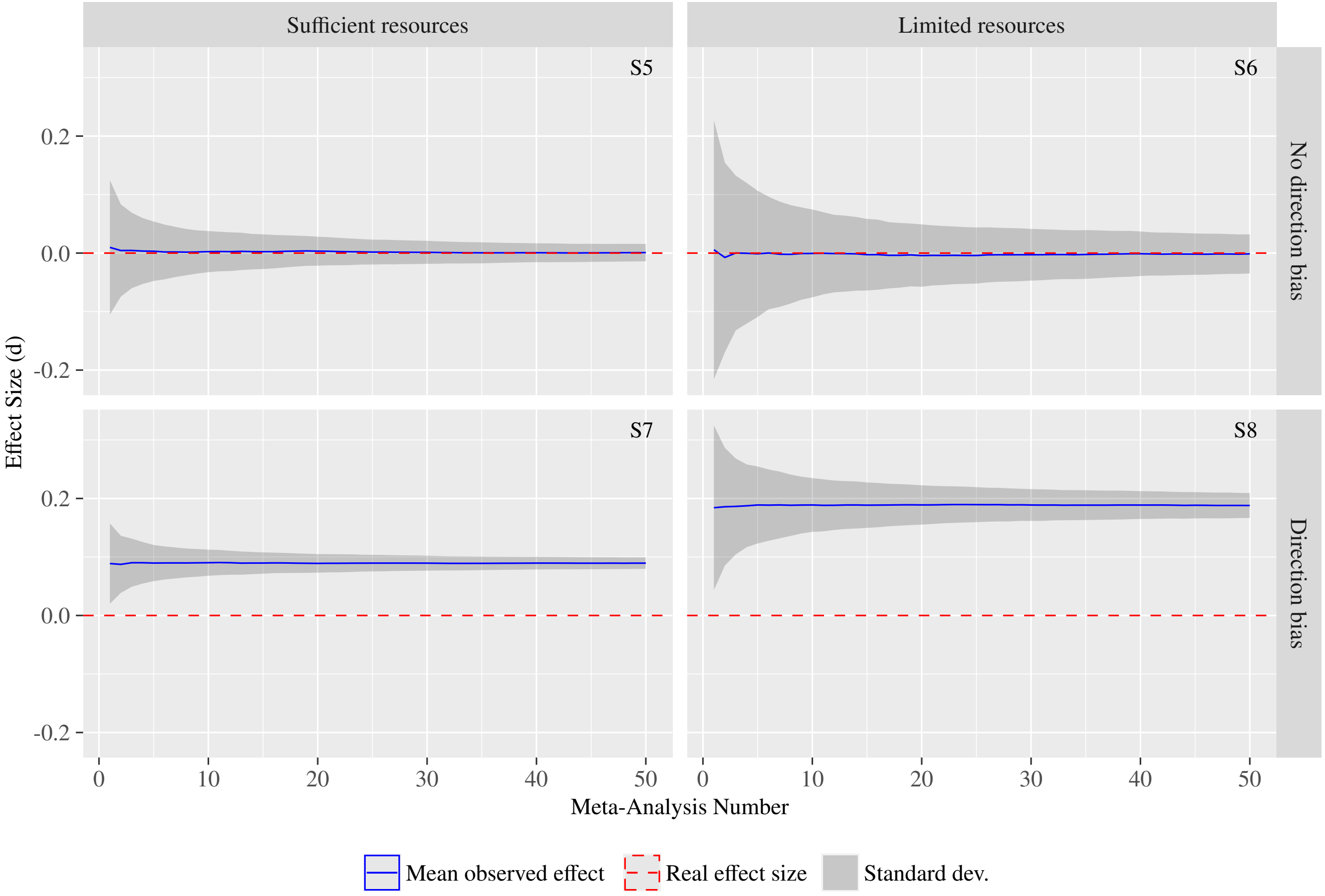

Figure 7: Simulation Results for Scenarios S5–S8: Real effect size $d = 0$

Considering these difficulties, one can be tempted to reason incorrectly as follows: the rule of not publishing non-significant results filters noise out of the system; hence, a system in which this rule is in place is in a better position to converge on the true value of the effect in the long run. S9 shows, however, that this reasoning is incorrect. Given the sufficient resources condition, 95% of all experiments would be statistically significant. In S9, communities overestimate the effect when they ignore the non-significant results (5%).[28]

Now, in addition to the observations about the relation between sample size and speed of convergence, and direction bias and overestimation discussed in the previous section, here are some observations about these scenarios:

- S10 shows that if the community publishes only significant results and resources are not sufficient to detect the effect, there is a large overestimation of the effect size in the long run. This effect has been previously discussed by Ioannidis (2008) and Button et al. (2013).[29] To grasp intuitively why such large overestimation occurs, consider Figures 2 and 3 again. In both distributions, the effect size is the same ($d = 0.41$). But Figure 2 is a low-power context (0.5), whereas Figure 3 is a high-power context (0.95). Now, notice that while increasing statistical power shifts the distribution for *the same* effect more to the right, the significance threshold that determines what gets published is roughly the same (it is not exactly the same, but the difference is negligible) in both situations. Hence, in the low-power context more results that suggest a small effect will not be published in comparison to the high-power context, which leads to a larger overestimation of the effect size in the former.
- The overestimation in S9 and S11 is small (at least for what is considered a difference in psychology). This suggests that, at least in the long run, not publishing non-significant results is not dramatic if resources are sufficient.

[28] For simplicity, in this scenario non-significant results don't get published at all, which is not the case in practice. Nonetheless, this does not threaten the qualitative conclusion: so long as the majority of significant results get published and some of the non-significant results don't get published, there will be overestimations.

[29] Ioannidis (2008) offers a theoretical argument and a survey of empirical evidence that early discovery effects are inflated. He calls this problem the "Winner's curse": the scientist who finds an effect with a low-power study is more likely to find an overestimated effect.

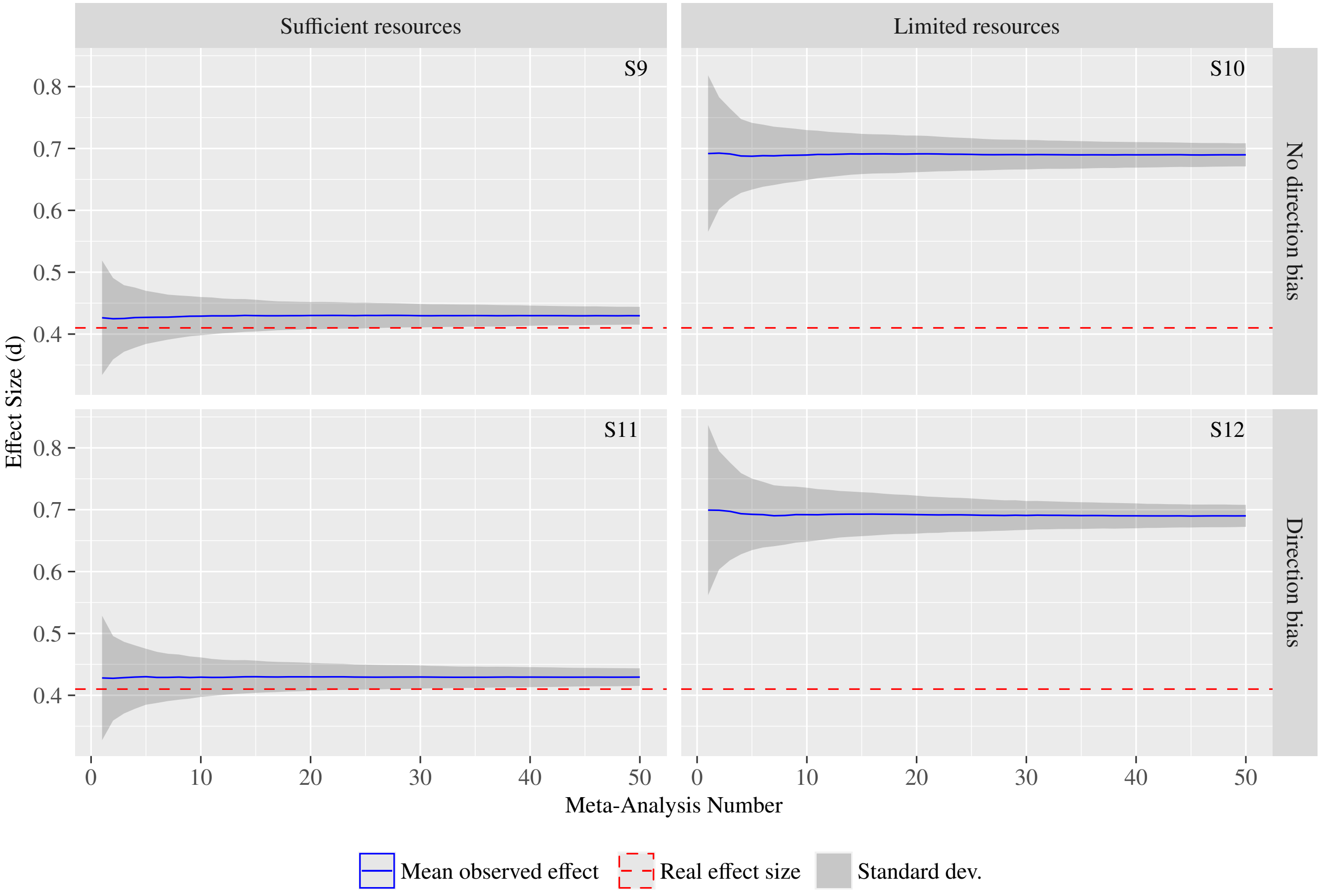

Figure 8: Simulation Results for Scenarios S9–S12: Real effect size $d = 0.41$

- Looking at the simulation scenarios before the 10th replication attempt can give us an idea of how (in)accurate estimations are in practice since in fact few experiments get replicated that many times. Recall (Section 2.2) the qualitative categories to interpret effect sizes: $d = 0.2$ is a small effect, $d = 0.5$ medium, and $d = 0.8$ large. Now, notice that in several scenarios many communities would disagree in this respect (i.e., one slice of the shaded area covers different qualitative categories). For instance, in S2 before the 10th replication some communities find a small effect while others find that it is medium. Also, in S14 before 10 replications, some communities communities could think that an effect is small or even medium when in reality there is no effect.

At this point one may wonder whether some technique could correct the overestimations shown in S9–S12. Some techniques help assess the file drawer problem (and publication bias more generally), but there is not a general correction technique for it. The two most popular techniques are Funnel Plots (a detection technique) and the Trim and Fill algorithm (a detection/adjustment technique based on funnel plots).[30] These techniques, however, do not work in this simulation study. They can detect/adjust publication bias *only* when there is a large amount of dispersion in the sample sizes of the experiments in a meta-analysis (Borenstein et al., 2009, p.290).[31] Here all the replications in a single meta-analysis have the same sample size. This limitation illustrates one reason why there is not yet a reliable procedure to adjust meta-analyses. More generally, trim and fill gives adjusted estimates that are quantitatively different from the unbiased estimate (Duval, 2005, p.134). Hence, methodologists recommend this and other techniques only to assess how

[30] Funnel Plots show asymmetries in distributions of effect sizes that are supposed to be symmetrical if there is no publication bias in a set of studies. The technique does not correct the overestimation, but at least it is useful to raise flags about its possibility. See Palmer (2000) for an application of Funnel Plots in biology. The Trim and Fill algorithm computes adjusted effect sizes by eliminating the asymmetry in a funnel plot.

[31] Techniques to assess publication bias (including Funnel Plots and Trim and Fill) exploit one assumption: small studies have more variance in their effect sizes than large studies. If there is publication bias due to publishing only statistically significant results, then published small studies will not report small effect sizes (because small studies will only obtain statistical significance with large effect sizes). Hence, the sensitivity of the techniques depends on having multiple studies with different sample sizes. Another caveat is that these techniques depend on strong assumptions about the missing studies. Borenstein et al. (2009, pp.277–291) discuss these and other techniques and their limitations in detail.

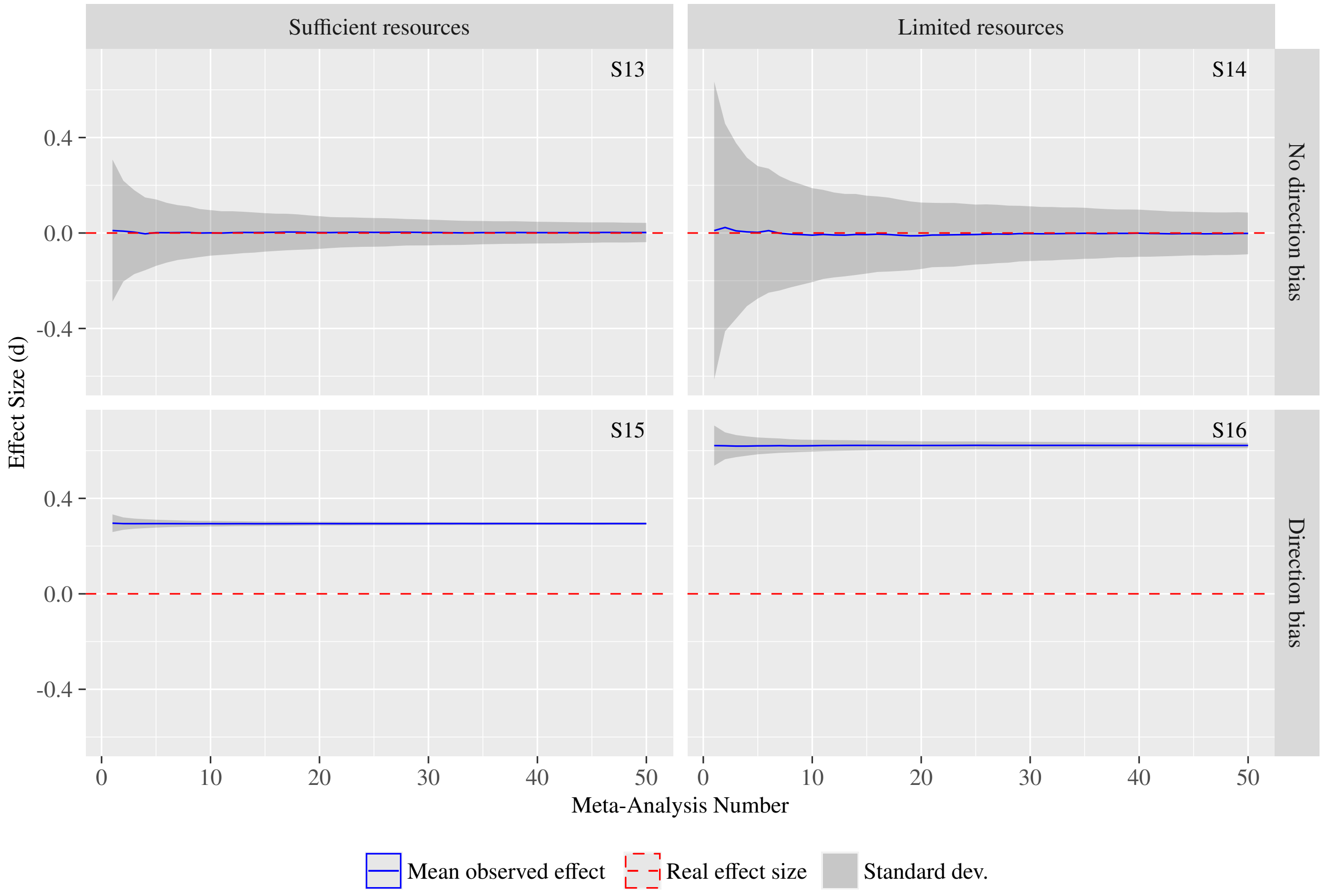

Figure 9: Simulation Results for Scenarios S13–S16: Real effect size $d = 0$

sensitive a meta-analysis is to possibly missing studies, not to correct biased estimations (Sutton, 2009, p.448).

Now, I will turn to an important observation about these scenarios that causes us to re-evaluate the impact of the file drawer problem. In his early work about the file drawer problem, Rosenthal worries about the extreme possibility "that journals are filled with the 5% of the studies that show Type I errors, while the file drawers are filled with the 95% of the studies that show non-significant results" (Rosenthal, 1979, p.638). The underlying question is, what happens in scenarios where there is not a real effect and only statistically significant results get published?

Consider again scenarios S9 and S10. In these scenarios, we observe that communities vastly overestimate effects due to negative (non-significant) results not being published (contrast these scenarios with scenarios S1 and S2, in which negative results are published), and the overestimations are larger when power is low. Scenarios S13 and S14 model the same arrangement of conditions but in a context in which the real effect is zero. An intuitive prediction for these scenarios is that, in the same way as in scenarios S9 and S10, effects would be overestimated (that was, indeed, my prediction when I designed the scenarios). The simulation results for S13 and S14, however, contradict that prediction. Communities in S13 and S14 converge on the true non-existent effect. There is no bias. To understand why, consider that when there is not a real effect, the replications in the simulations are produced by sampling from a null distribution (e.g., Figure 1), and there are significant results at both of its tails. For example, Figure 10 shows the publications of one single community. In such a case, scientists find very extreme opposite results. In the long run, however, such results are balanced in number and strength, which makes them cancel out in the meta-analysis.

In other words, scenarios S13 and S14 show that if there is not an effect to find, most communities, even when publishing only statistically significant results and underpowered studies, find that there is not an effect in the long run. That is, in the context in which the file drawer problem would intuitively seem to be more troublesome it is not. This is a positive theoretical lesson about NHST.

There is a caveat, however. It would be too optimistic to think that in practice science could recover from false positives in this way. Recall that the tails of the distribution after the significance threshold correspond to 5% of it, so this procedure is slow. Also, the amplitude of estimations (i.e., extreme results in both directions), which can be seen in the dotted lines in S14,

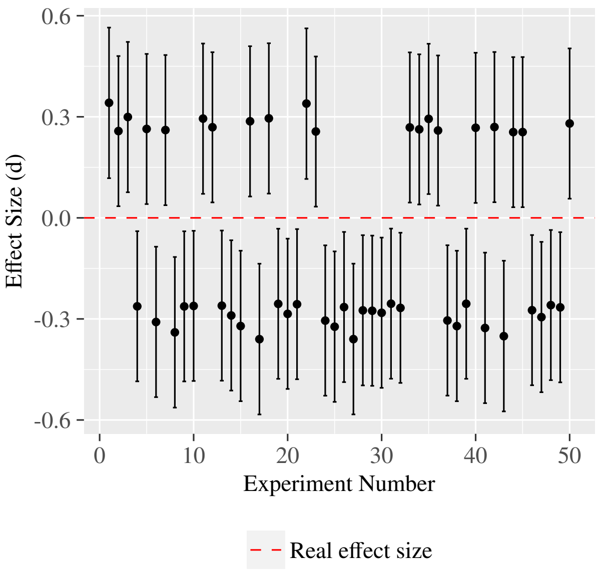

Figure 10: Extreme opposite results - No real effect.

show that the estimation is less reliable, relative to the other scenarios in the set.

The file drawer problem is commonly used in attacks to NHST (Schmidt, 1996; Gill, 1999; Gelman, 2015). The argument is that NHST introduces biases in the literature, such as the overestimations that scenarios S10 and S12 illustrate. And it is followed by the prescriptive conclusion that NHST should be banned. Such a prescription should be qualified, because it assumes that NHST plays a negative role in all circumstances, and I have shown that it does not. Furthermore, the critique is misleading because it overemphasizes the role of inference methods in scientific self-correction. Here I have framed NHST as one of the many conditions that could affect SCT*. In this way, we can see that NHST has flaws in some contexts but not in others. Other factors interact with NHST, and these are by themselves problematic, and even more in some contexts than NHST.

I will now turn to a general discussion, considering implications for replicability controversies, and proposals for future work.

4. General Discussion

4.1. Scientific Self-Correction as an Interaction Effect

Laudan (1981) complains that Peirce and philosophers after him have trivialized SCT by focusing on technical investigations of quantitative induction, and neglecting the study of the self-corrective character of scientific inference in general. My study is consistent with his conclusion, but I take it one step further. It is indeed important to study the long-run properties of inference methods in the abstract. But such studies disregard social conditions that ultimately matter for SCT. Based on such utopian studies it would be a mistake to draw conclusions about whether science (either actual or nomologically possible) is self-corrective or not.

The set of utopian conditions 1–3 is by no means exhaustive. The reader might envisage other ways in which the social structure of science is not utopian, which may affect SCT*. But conditions 1–3 are sufficient to illustrate my main conclusion: while SCT* is true in a scientific utopia, self-correction is a fragile property: once we move away from the utopia and consider less utopian scenarios, the procedure of aggregating experimental evidence by meta-analysis can easily lead communities of frequentist scientists astray.

An advocate of SCT* could try to relativize my conclusions about fragility by saying that it is all just pragmatics: every inference method has conditions under which it works, and of course, if the conditions are violated, then the method will not work as expected. In particular, if frequentist scientists fail to satisfy conditions 1-3, then this would be a failure in their conduct, and not a failure of SCT* itself. Nonetheless, I think this response misses the problem. For, what would it take for frequentist scientists to do things correctly? The answer cannot be that they should move to a utopia: that they should publish every single study, significant or not, in equally prestigious venues, or that every study should be run with massive sample sizes, or that they should purge themselves from all possible biases, and so on. Failing to meet the utopian conditions is not malpractice, but an unavoidable reality. The "correct" way of doing things is infeasible because it is too demanding.

I suggest viewing scientific self-correction appealing informally to the notion of *interaction effect* (i.e., given two variables, the effect on one of them depends on the value of the other). That is, *scientific self-correction depends on two variables: inference methods and social structures, and these two variables interact.* The long-run performance of an inference method (i.e., it's reliability and validity) depends on the particular social structure in which that method is used. I would propose that in the same way as a car would not work if it uses the wrong kind of fuel, inference methods malfunction if they are deployed in an inappropriate social structure. Of course, the present study concerns the context of frequentist statistics, so further work is necessary to understand better this interaction outside of this context. Consider two issues:

- *How do other forms of inference interact with their social context?* For instance, Bayesian inference would not be subject to problems I discussed here (e.g. Bayesian inference does not use the statistical significance as a threshold in a decision procedure). However, before drawing normative recommendations from that fact, we have to consider that Bayesian convergence might be fragile with respect to other social structural conditions. A different simulation study would be required to explore this issue in detail.

- *What are the implications of the fragility of particular inference methods to the more general self-corrective thesis (SCT)?* If my arguments are sound, we face a dilemma: either the research programs that use frequentist statistics do not self-correct, or there is an alternative to SCT* that explains how they do. We have a reason to reject the first

> alternative: the history of superseded scientific theories suggests that error gets corrected.[32] However, it is not obvious how to substantiate the second alternative. Notice that a story according to which error gets corrected by fortunate accidents is inconsistent with SCT (e.g., scientists discover accidentally that a theory is false when testing some unrelated hypothesis but this form of error correction is not guaranteed to work in the long run). We need analytic work characterizing how error correction is (and can be) a result of a "scientific method" as SCT contends.

Now, I will turn to some implications of this study for the replicability crisis in psychology.

4.2. Replicability Controversies Revisited

One central aspect to the replicability crisis in social psychology is that the psychological community has systematically produced false positives. How do we explain this fact? Multiple causes may contribute, and I don't intend to give a full account of them. But I want to stress four implications of this simulation study.

First, as shown in Section 3.4, based on scenarios S13 and S14 we can see that *NHST and low power do not produce long-run overestimations in the context in which the effects in question are truly non-existent*. To be clear, individual scientists do err in this context. But the self-corrective procedure works for the community after a series of replications. This implies that these two factors should play a smaller role as an explanation of the crisis than some methodologists and psychologists have suggested.

Second, direction biases could be an important contributing factor in the replicability crisis. As illustrated in scenarios S15 and S16, direction biases can explain overestimations when there are no real effects better than the other violations of the conditions. This explanation is not a full actual explanation. In particular, I have no empirical evidence of the extent to which direction bias is present in e.g. social priming research. Nonetheless, consider that given the high attractiveness of social priming hypotheses, researchers could be subject to motivated reasoning (that would lead to direction biases) more than in other research programs. If this story is correct, then S15 and S16 show sufficient conditions under which communities of frequentist scientists studying these hypotheses could systematically produce *large* overestimations of non-existent effects.

Third, this study shows the limits of meta-analyses to settle replicability disputes.[33] Both psychologists and philosophers (Doris, 2015, p.49) hope that meta-analysis will eventually help us with these issues. However, we have non-skeptical reasons to think that the experiments that could constitute the input to retrospective meta-analyses for social priming research have been produced in a context that is far from the utopia, in particular with regard to utopian conditions 1 and 3. If this is right, then the study that I have presented here gives us reasons to qualify that hope.

Now, recall (section 1) the distinction between direct replications (i.e., those that mirror the original experimental design) and conceptual replications (i.e., those that intend to test the robustness of underlying hypotheses). In psychology most replications are conceptual and not direct. Makel et al. (2012) systematically analyzed articles in the top 100 psychology journals since 1900, and they found that only 1.07% of them reported replication attempts. And even worse, only 14% of all replications were direct replications. For Pashler & Harris (2012) this is a problem because the preference for conceptual replications over direct replications plus publication bias opens the door for producing false positives, and they strongly advocate for more direct replications. In response to this, my fourth point is that in *all* scenarios in the present simulation study scientists run long series of direct replications, and many don't self-correct. Hence, regardless of whether the assessment that conceptual replications have played a role in the crisis, my study calls into question the efficacy of the normative recommendation of encouraging more direct replications tout court. In particular, normative recommendations should consider the possibility that replicators could be systematically biased.

Now, the fact that we are not in the utopia does not mean that all non-utopian scenarios stand on the same ground. In particular, institutional interventions can address in part the problems of the three utopian conditions.

> *Institutionalize sample size/power requirements to address limited resources problems.* One way of alleviating the consequences of limited resources is to place editorial policies that discourage authors from producing small sample studies. For instance, the journal *Social Psychological and Personality Science* amongst its recent editorial policies states that "authors will be asked to disclose how sample size was determined" and "manuscripts that present underpowered studies without adequate justification will have a greater chance of being rejected without review" (Vazire, 2015, p.2).
>
> *Open Science to address direction biases.* The "Open Science" movement encourages open access to all information associated with a research project (e.g., data, methods, materials, tools, code etc.) This initiative could counteract direction bias under the assumption that it be practically enforced. See Ioannidis et al. (2014) for other recommendations to reduce reporting biases.
>
> *Pre-registration of studies to address problems of publication of only positive results.* Pre-registration of studies has been increasingly implemented in clinical research. The International Committee of Medical Journal Editors (ICMJE) "requires, and recommends that all medical journal editors require, registration of clinical trials in a pub-

[32] Although, as Ioannidis points out, "The fact that some practical progress is made does not mean that scientific progress is happening in an efficient way or that we cannot become even more efficient" (Ioannidis, 2012, p.648).

[33] Recent literature has been concerned with the evidential import of meta-analysis. Stegenga (2011) casts doubts on meta-analysis as a high standard of evidence on the grounds that designing a meta-analysis involves numerous decisions that compromise the reliability of its outcomes. In response, Jukola (2015) argues that the reliability of meta-analysis requires taking into account the social context in which meta-analyses are designed and used.

lic trials registry at or before the time of first patient enrollment as a condition of consideration for publication" (ICMJE, 2015, p.12). This practice has shown its benefits. For instance, Kaplan & Irvin (2015) show that preregistration was strongly associated with the trend toward null findings. This practice could alleviate the problems of publication bias without requiring the publication of every non-significant result.[34]

We need work on assessing the extent to which these normative interventions can make the behavioral sciences more self-corrective. Another set of proposals are what I call "self-corrective labor schemes", i.e. ways of organizing replication work to increase replicability. These proposals include performing multi-site replication projects (Klein et al., 2014), educational replication (Frank & Saxe, 2012; Standing et al., 2014), and adversarial collaboration (Rakow et al., 2015). Additional analyses are required to assess and compare their efficiency.

5. Conclusion

Is science an enterprise that corrects its mistakes? The assumption that it is lies at the core of the justification of science. I have argued by example, however, that when we approach the question as a social epistemological one, we have strong reasons to think that self-correction is very fragile: the social structure of the community in which frequentist inference is deployed greatly affects its long-run performance. I have shown that the kinds of resources scientists have, their theoretical commitments, and the rules for publishing results impact the estimation of parameters in communities of frequentist scientists. The general lesson is not necessarily pessimism about scientific self-correction. Rather, it is that philosophical attention to inference methods in isolation from social context can only give us a partial understanding of the mechanisms of self-correction. In addition to studying whether particular forms of inference self-correct, it is necessary to ask how they interact with their social and institutional context.

References

Allchin, D. (2015). Correcting the "Self-Correcting" Mythos of Science. *Filosofia e Historia da Biologia, Sao Paulo*, *10(1)*, 19–35.

Anderson, E. (2015). Feminist Epistemology and Philosophy of Science. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. (Fall 2015 ed.).

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*, e84896. doi:`10.1371/journal.pone.0084896`.

Begley, C. G., & Ellis, L. M. (2012). Drug Development: Raise Standards for Preclinical Cancer Research. *Nature*, *483*, 531–533.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd. doi:`10.1002/9780470743386.ch30`.

Bower, B. (2012). The Hot and Cold of Priming: Psychologists are Divided on Whether Unnoticed Cues Can Influence Behavior. *Science News*, *181*, 26–29. doi:`10.1002/scin.5591811025`.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power Failure: Why Small Sample Size Undermines the Reliability of Neuroscience. *Nature Reviews Neuroscience*, *14*, 365–376.

Carter, T. J., Ferguson, M. J., & Hassin, R. R. (2011). A Single Exposure to the American Flag Shifts Support Toward Republicanism up to 8 Months Later. *Psychological Science*, *22*, 1011–1018.

Cartwright, N. (1991). Replicability, Reproducibility, and Robustness: Comments on Harry Collins. *History of Political Economy*, *23*, 143–155.

Cohen, J. (1962). The Statistical Power of Abnormal-Social Psychological Research: A Review. *The Journal of Abnormal and Social Psychology*, *65*, 145–153.

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*, 155–159.

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

Cumming, G. (2014). The New Statistics: Why and How. *Psychological Science*, *25*, 7–29. doi:`10.1177/0956797613504966`.

Doris, J. (2015). *Talking To Our Selves: Reflection, Ignorance, and Agency*. Oxford: Oxford University Press.

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

Doyen, S., Klein, O., Pichon, C.-L., & Cleeremans, A. (2012). Behavioral Priming: It's All in the Mind, but Whose Mind? *PLoS ONE*, *7*, e29081.

Duval, S. (2005). The Trim and Fill Method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. John Wiley & Sons Ltd.

Fanelli, D. (2010). Positive Results Increase Down the Hierarchy of the Sciences. *PLoS ONE*, *5*, e10068. doi:`10.1371/journal.pone.0010068`.

Fanelli, D. (2012). Negative Results are Disappearing from Most Disciplines and Countries. *Scientometrics*, *90*, 891–904. doi:`10.1007/s11192-011-0494-7`.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:`10.3758/BF03193146`.

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the Quality of Empirical Journals with Respect to Sample Size and Statistical Power. *PLoS ONE*, *9*, e109019. doi:`10.1371/journal.pone.0109019`.

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives on Psychological Science*, *7*, 600–604.

Friston, K. (2012). Ten Ironic Rules for Non-Statistical Reviewers. *NeuroImage*, *61*, 1300–1310. doi:`10.1016/j.neuroimage.2012.04.018`.

Gelman, A. (2015). The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management*, *41*, 632–643.

Gigerenzer, G. (2004). Mindless Statistics. *The Journal of Socio-Economics*, *33*, 587–606.

Gill, J. (1999). The Insignificance of Null Hypothesis Significance Testing. *Political Research Quaterly*, *52*, 647–674.

Goldman, A. I. (2009). Systems-Oriented Social Epistemology. *Oxford Studies in Epistemology*, *3*, 189–214.

Hacking, I. (1980). The Theory of Probable Inference: Neyman, Peirce, and Braithwaite. In D. H. Mellor (Ed.), *Science, Belief, and Behaviour : Essays in Honour of R. B. Braithwaite*. Cambridge: Cambridge University Press.

Harris, C. R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two Failures to Replicate High-Performance-Goal Priming Effects. *PLoS ONE*, *8*, e72467. doi:`10.1371/journal.pone.0072467`.

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.

ICMJE (2015). *Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals*. Technical Report International Committee of Medical Journal Editors.

Ioannidis, J. P., Munafo, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*, 235–241.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Med*, *2*, e124.

Ioannidis, J. P. A. (2008). Why Most Discovered True Associations are Inflated.

[34] I thank an anonymous reviewer for pointing this out.

Epidemiology, *19*, 640–648.

Ioannidis, J. P. A. (2012). Why Science Is Not Necessarily Self-Correcting. *Perspectives on Psychological Science*, *7*, 645–654.

Jukola, S. (2015). Meta-Analysis, Ideals of Objectivity, and the Reliability of Medical Knowledge. *Science & Technology Studies*, *28(3)*, 101–120.

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLoS ONE*, *10*, 1–12. URL: `http://dx.doi.org/10.1371%2Fjournal.pone.0132382`. doi:`10.1371/journal.pone.0132382`.

Kitcher, P. (1990). The Division of Cognitive Labor. *Journal of Philosophy*, *87*, 5–22.

Kitcher, P. (1993). *The Advancement of Science: Science Without Legend, Objectivity Without Illusions*. Oxford University Press.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. J., Bahnik, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., Hasselman, F., Hicks, J. A., Hovermale, J. F., Hunt, S. J., Huntsinger, J. R., IJzerman, H., John, M.-S., Joy-Gaba, J. A., Barry Kappes, H., Krueger, L. E., Kurtz, J., Levitan, C. A., Mallett, R. K., Morris, W. L., Nelson, A. J., Nier, J. A., Packard, G., Pilati, R., Rutchick, A. M., Schmidt, K., Skorinko, J. L., Smith, R., Steiner, T. G., Storbeck, J., Van Swol, L. M., Thompson, D., & van 't Veer, A. E. (2014). Investigating Variation In Replicability: A 'Many Labs' Replication Project. *Social Psychology*, *45*, 142–152.

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

Lee, C. J. (2013). The Limited Effectiveness of Prestige as an Intervention on the Health of Medical Journal Publications. *Episteme*, *10*, 387–402.

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*, 2–17. doi:`10.1002/asi.22784`.

Levi, I. (1980). Induction as Self-Correcting According to Peirce. In D. H. Mellor (Ed.), *Science, Belief, and Behaviour : Essays in Honour of R. B. Braithwaite*. Cambridge: Cambridge University Press.

Longino, H. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press.

Longino, H. (2015). The Social Dimensions of Scientific Knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. (Spring 2015 ed.).

Machery, E. (2012). Power and Negative Results. *Philosophy of Science*, *79*, 808–820.

Machery, E. (2014). Significance Testing in Neuroimagery. In M. Sprevak, & J. Kallestrup (Eds.), *New Waves in Philosophy of Mind*. Palgrave Macmillan.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*, 537–542.

Maxwell, S. E. (2004). The Persistence of Underpowered Studies in Psychological Research: Causes, Consequences, and Remedies. *Psychological Methods*, *9*, 147–163.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. Science and Its Conceptual Foundations series. University of Chicago Press.

Mayo, D. (2005). Peircean Induction and the Error-Correcting Thesis. *Transactions of the Charles S. Peirce Society*, *41*, 299–319.

Mayo-Wilson, C., Zollman, K. J. S., & Danks, D. (2011). The Independence Thesis: When Individual and Social Epistemology Diverge. *Philosophy of Science*, *78*, 653–677.

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*, 103–115.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*, *23*, 217–243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Open Science Collaboration (2012). An Open, Large-Scale, Collaborative Effort to Estimate the Reproducibility of Psychological Science. *Perspectives on Psychological Science*, *7*, 657–660.

Palmer, A. R. (2000). Quasi-replication and the Contract of Error: Lessons from Sex Ratios, Heritabilities and Fluctuating Asymmetry. *Annual Review of Ecology and Systematics*, *31*, 441–480.

Pashler, H., Coburn, N., & Harris, C. R. (2012). Priming of Social Distance? Failure to Replicate Effects on Social and Food Judgments. *PLoS ONE*, *7*, e42510. doi:`10.1371/journal.pone.0042510`.

Pashler, H., & Harris, C. R. (2012). Is the Replicability Crisis Overblown? Three Arguments Examined. *Perspectives on Psychological Science*, *7*, 531–536.

Peirce, C. S. (CP). *The Collected Papers of Charles Sanders Peirce*. Vols 1-6, ed. Charles Hartshorne and Paul Weiss (1931-1935), Vols. 7-8 ed. Arthur W. Burks (1958). Cambridge, MA: Belknap Press of Harvard University Press.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe It or Not: How Much Can We Rely on Published Data on Potential Drug Targets? *Nature Reviews Drug Discovery*, *10*, 712.

Rakow, T., Thompson, V., Ball, L., & Markovits, H. (2015). Rationale and guidelines for empirical adversarial collaboration: A Thinking & Reasoning initiative. *Thinking & Reasoning*, *21*, 167–175. doi:`10.1080/13546783.2015.975405`.

Reichenbach, H. (1938). *Experience and Prediction*. The University of Chicago Press.

Reichenbach, H. (1949). *The Theory of Probability*. Berkeley, University of California Press.

Rescher, N. (1978). *Peirce's Philosophy of Science : Critical Studies in his Theory of Induction and Scientific Method*. Notre Dame: University of Notre Dame Press.

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*, 331–363.

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

Schmidt, F. L. (1992). What Do Data Really Mean? Research Findings, Meta-Analysis, and Cumulative Knowledge in Psychology. *American Psychologist*, *47*, 1173–1181.

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

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*, 90–100.

Sedlmeier, P., & Gigerenzer, G. (1989). Do Studies of Statistical Power Have An Effect on the Power of Studies? *Psychological Bulletin*, *105*, 309–316.

Shanks, D. R., Newell, B. R., Lee, E. H., Balakrishnan, D., Ekelund, L., Cenac, Z., Kavvadia, F., & Moore, C. (2013). Priming Intelligent Behavior: An Elusive Phenomenon. *PLoS ONE*, *8*, e56515. doi:`10.1371/journal.pone.0056515`.

Solomon, M. (1992). Scientific Rationality and Human Reasoning. *Philosophy of Science*, *59*, 439–455.

Solomon, M. (2001). *Social Empiricism*. Bradford Bks. MIT Press. URL: `https://books.google.nl/books?id=5ptFPwAACAAJ`.

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

Standing, L. G., Grenier, M., Lane, E. A., Roberts, M. S., & Sykes, S. J. (2014). Using Replication Projects in Teaching Research Methods. *Psychology Teaching Review*, *20*, 96–104.

Stanford, P. K. (2015). Unconceived alternatives and conservatism in science: the impact of professionalization, peer-review, and Big Science. *Synthese*, (pp. 1–18). doi:`10.1007/s11229-015-0856-4`.

Stegenga, J. (2011). Is Meta-Analysis the Platinum Standard of Evidence? *Studies in History and Philosophy of Science Part C*, *42*, 497–507.

Stegenga, J. (2015). Measuring Effectiveness. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *54*, 62–71.

Stephan, P. E. (2012). *How Economics Shapes Science*. Harvard University Press.

Strevens, M. (2003). The Role of the Priority Rule in Science. *Journal of Philosophy*, *100*, 55–79.

Sutton, A. J. (2009). Publication Bias. In H. Cooper, L. V. Edges, & J. C. Valentine (Eds.), *Handbook of Research Synthesis and Meta-Analysis, 2nd Edition*. The Russell Sage Foundation.

Vazire, S. (2015). Editorial. *Social Psychological and Personality Science*, . doi:`10.1177/1948550615603955`.

de Winter, J., & Happee, R. (2013). Why Selective Publication of Statistically Significant Results Can Be Effective. *PLoS ONE*, *8*, e66463. doi:`10.1371/journal.pone.0066463`.

Worrall, J. (2010). Do We Need Some Large, Simple Randomized Trials in Medicine? In M. Suárez, M. Dorato, & M. Rédei (Eds.), *EPSA Philosophical Issues in the Sciences* (pp. 289–301). Dordrecht: Springer Netherlands. doi:10.1007/978-90-481-3252-2_27.

von Wright, G. H. (1965). *The Logical Problem of Induction*. New York: Barnes and Noble.

Yong, E. (2012). A Failed Replication Attempt Draws a Scathing Personal Attack from a Psychology Professor. *Discover Magazine Blog*, .

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