University of Groningen

Self-correction in science

Bruner, Justin P.; Holman, Bennett

Published in:
Studies in History and Philosophy of Science. Part A

DOI:
[10.1016/j.shpsa.2019.02.001](https://doi.org/10.1016/j.shpsa.2019.02.001)

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2019

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Bruner, J. P., & Holman, B. (2019). Self-correction in science: Meta-analysis, bias and social structure. *Studies in History and Philosophy of Science. Part A*, 78, 93-97.
<https://doi.org/10.1016/j.shpsa.2019.02.001>

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Self-correction in science: Meta-analysis, bias and social structure

Justin P. Bruner^a, Bennett Holman^{b,*}

^a Department of Theoretical Philosophy and Centre for Philosophy, Politics and Economics, University of Groningen, Groningen, the Netherlands

^b Underwood International College, Yonsei University, Seoul, South Korea

HIGHLIGHTS

- Clarifies what it means for science to self-correct.
- Meta-analysis as a corrective to rampant publication bias.
- Compares different meta-analytic techniques in the presence of publication bias.

1. Introduction

Frustrated with the high failure rate of their cancer drug development program, researchers at the biopharmaceutical company Amgen began making it standard practice to confirm published findings in-house prior to investing in a line of research (Begley & Ellis, 2012). Many of these landmark studies had spawned subfields or led to the development of drugs that reached the clinical trial stage. In 2011, Amgen published the results of nearly a decade of adopting this practice: 89% of trials failed to replicate. Amongst the reasons for the failures, former Amgen researcher Glenn Begley noted that academics were under significant pressure to produce statistically significant results and had few incentives to publish negative data. Similar failures in other fields (Klein et al., 2014; Open Science Collaboration, 2012) have prompted what is generally referred to as “the replication crisis”. Such spectacular failures threaten not just the individual studies they cast doubt on, but point to something wrong with the larger social structure of inquiry, particularly in publication practices.

Especially in cases where the effect of some intervention is being estimated from a sample, reproduction is crucial to ensure that the original finding is in fact a real effect and not a statistical artifact, i.e., part of normal variation. However, if there is some kind of systematic bias regarding which studies are published, then we end up with a skewed picture of the world. For example, suppose we live in a world—perhaps not too far from our own—in which only positive studies with statistically significant results are published. Now consider what the evidence looks like for a drug that, in reality, has no effect. Most of the research conducted will find no effect and be shelved rather than published, but because of the nature of statistical evidence, some studies will find a significant result and be submitted to a journal. As a result, when the published literature is consulted on this topic, it uniformly reports that the drug works. Indeed, if we conducted a meta-

analysis on this data we would come to the same conclusion, even though there is in fact no effect.

This much is common sense. If we amalgamate evidence to generate an all-things-considered judgment, but we feed into our amalgamation process a biased set of data, we will not arrive at the right answer. Garbage in, Garbage out. Indeed, as many have argued, whether or not science is epistemically reliable turns on aspects of its social structure (Biddle, 2007; Longino, 2002; Fernandez-Pinto, 2014; Holman, 2018) and norms that govern publication practices (Jukola, 2015; Romero, 2016; Sismondo, 2009). In the fictional scenario described above, common sense holds science is not self-correcting. We think there is reason to question this common-sense understanding of science.

In this paper we begin with Felipe Romero's (2016) rigorous formal argument that while science might be self-correcting in ideal circumstances, it is not self-correcting in the face of the many social biases encountered in actual research. Though it may seem counter-intuitive, we show that just because the underlying data are skewed it does not necessarily mean that our all-things considered judgements will be inaccurate. Overall, we conclude with a qualified, but nonetheless positive, assessment of the self-correction thesis: Self-correction is possible in non-ideal cases, but requires a commitment to learn from our mistakes. Ultimately, we contend that if science is self-correcting, it is not because the methods it uses unerringly lead to truth, but because of a social commitment to eliminate error.

Our paper proceeds as follows. In section 2, we summarize Romero's argument that science does not self-correct and discuss his simulations that show that meta-analyses of biased data yield systematically inaccurate estimates. In section 3 we describe a statistical procedure which has recently been developed for generating unbiased estimates from biased data and show that applying it to Romero's simulations corrects for previously observed inaccuracies. We conclude in section 4 and agree with Romero that the social practices of science are crucial to

* Corresponding author. Underwood International College, Yonsei University, 85 Songdogwahak-ro, Yeonsu-gu, Incheon, 21983, South Korea.

E-mail addresses: j.p.bruner@rug.nl (J.P. Bruner), bholman@yonsei.ac.kr (B. Holman).

understanding science, but disagree about where we should look to determine whether science is self-correcting.

2. Truth and the end of science

The self-correcting thesis figures most notably in the work of Peirce who defined truth as the beliefs that we have at the end of inquiry. Lying behind this idea is that if we are estimating some population value by repeatedly taking samples, our estimate of the population parameter approaches the true value as the number of samples increase. Though Peirce was writing prior to the development of modern statistics, frequentist statistical approaches adopt Peirce's notion of quantitative induction and rigorously demonstrate the truth of his conjectures (Mayo, 1996).

Of course, whether all science can be understood as on par with quantitative induction is questionable. Strong cases have been made that the conceptual understanding embedded in scientific theories exhibits “catastrophic” changes over the course of the history of science and we should expect such changes to continue (Stanford, 2015). Yet setting aside such concerns, Romero has recently argued that even the kind of quantitative induction championed by Peirce and Mayo is not self-correcting.

Two virtues of Romero's argument are its clarity and its precision. His position pertains to a version of the self-correcting thesis he calls SCT*, which holds that: “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” (p. 58). Romero notes that SCT* is silent about the social conditions of the production of evidence and it is his contention that whether SCT* holds depends crucially on such factors.

To examine this rigorously, Romero conducts a simulation of a scientific community all investigating some question of interest. Depending on the simulation, there is either no effect or a medium size effect, but agents in the simulation are blind to this fact and must learn about the world by conducting experiments and reading the research of their colleagues. Each member of the community conducts an experiment and then chooses whether to publish it depending on the norms of their community. Specifically, Romero considers cases where there is a social norm to only publish statistically significant results (a publication bias) and the related case where researchers publish only if the experimental group outperforms the control (a direction bias). Finally, he also explores the effect of having adequate funding by comparing research that relies on large and small sample sizes. Romero finds that SCT* holds in a scientific utopia where studies are well-funded and researchers publish their results without fear or favor.

In non-utopian communities, however, Romero finds SCT* frequently fails. To see why, consider the case in which the actual effect size is 0.41, the null hypothesis is that the effect size is zero and there is publication bias. By chance, a study can identify the effect as less than it actually is (say, 0.21). In this case, the measured effect size is non-zero, but may nonetheless not be large enough to allow the scientist to reject the null hypothesis. As a result, the study is not published. On the other hand, results exaggerating the effect size *are* published. In a case where a study found an effect size of 0.61, the null hypothesis would be handily rejected and thus these results will be published. Taken together, a publication bias of this kind results in the systematic publication of studies overestimating the effect size.

On the basis of simulations which instantiate this type of problem, Romero claims that “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” (p. 66). The challenge that Romero proposes to the defender of SCT* is to show that science can be self-correcting as we move from the ideal to the rough and tumble world where human biases are an unavoidable part of inquiry. We believe this challenge can be met.

We first describe a newly developed statistical procedure that can be

used to generate accurate estimates of an effect size from the kind of data Romero considers. We then replicate Romero's simulations with this more sophisticated meta-analytic technique. We show that in applying this procedure, agents in our simulation produce reliable estimates of the effect outside of a scientific utopia.

3. Generating reliable estimates from unreliable data

When a meta-analysis is conducted, the trials themselves functionally serve as the data. As with analysis of primary data, blindly applying statistical procedures can go horribly astray. All statistical procedures contain assumptions and proper statistical analysis involves ensuring that these assumptions are met. An assumption of meta-analysis is that the published literature is not systematically biased and this assumption should be assessed (Borenstein, Hedges, Higgins, & Rothstein, 2009).

Such an assessment is possible and depends on some basic features of statistical distributions. Broadly speaking, most studies produce estimates that are approximately equal to the true value. Of course, just by chance, some studies will provide an overestimate, while others an underestimate. Most importantly, it is a consequence of the central limit theorem that irrespective of the distribution of the variable, the distribution of the sample means of that variable will be approximately normal.¹ Consequently, we can generate precise expectations for what to anticipate when we look at an array of sample means. If they are not normally distributed, then we will be alerted to the fact that an assumption of our meta-analysis may be violated. When this occurs it is inappropriate to proceed with a standard meta-analytic aggregation procedure. However, when there is evidence that the normality assumption is violated due to publication bias, violations of this assumption *can be corrected for*. In simple terms, when the underlying data is unbiased, standard meta-analyses give unbiased estimates; when the underlying data is biased because non-significant results are not published, it is possible to still generate an unbiased estimate.

While there are a few recently developed statistical procedures that perform this feat (e.g., Simonsohn, Nelson, & Simmons, 2014), we concern ourselves here with the p-uniform technique developed by Marcel van Assen, Robbie van Aert and Jelte Wicherts (2015).² To understand the basic idea behind this procedure, consider what would happen if we conducted 1000 trials testing a hypothesis that there was a difference between two groups, when in fact the truth was that there was no difference. A “p-value” tells you the probability that one would obtain a difference at least as large as the one observed if there were actually no difference. So for example, suppose we were testing the claim that some drug prevents heart attacks and we found 7 fewer heart attacks in the treatment group that were administered the drug out of a sample of 500 people. We don't yet know if we can reject the hypothesis that there is no difference between the groups (and conclude the drug works). But if we conduct a statistical test and calculate a p-value of .02, what this tells us is that, if we were in a world where the drug actually did no good, we would see a difference this large (or larger) between the groups in only two 2 trials in 100. This much is basic statistics.³

To understand the p-uniform procedure, let's now return to our imagined scenario where in fact the drug does not work and we repeat the trials 1000 times. Let's further assume a standard cut-off for statistical significance ($\alpha = 0.05$) and that all and only the significant trials are published. Roughly speaking we would expect to find 50

¹ For example, wealth in America is not normally distributed, but if we repeatedly took the average wealth of 1000 random Americans, the distribution of the sample means (i.e. the average wealth of the 1000 people in each sample) would be approximately normal. The larger the sample and the less distorted the initial distribution, the closer the distribution of the sample means comes to being normally distributed.

² For a forerunner of these procedures see Hedges (1984).

³ Note that Romero directs his original argument towards the behavioral and social sciences, but because his argument turns on a critique of meta-analysis, it inherently applies in disciplines where meta-analysis is relied upon. In personal communication, Romero agreed with this sentiment.

published experiments, half that found the drug was better than the placebo and half that found the placebo was better than the drug.⁴

The p-uniform technique confines itself to looking at the portion of the data which should not be affected by publication bias (i.e. statistically significant results). By virtue of the way the p-value is constructed, we expect ten out of 1000 tests to have a p-value between .05 and .04 and another ten trials to have a p-value between .04 and .03. If the null hypothesis is true, we expect the number of trials in any equivalent p-value range to be the same. If the null hypothesis is true, we expect the distribution of p-values to be *uniform*.

On the other hand, if the null hypothesis is false (i.e., if there is a difference between the groups), then we wouldn't expect the distribution of p-values to be uniform. For example, if there were a difference and we ran well-powered trials, most of the obtained results would have p-values less than .05 even if we published all of the studies. This would lead to a non-uniform distribution of p-values providing additional evidence that the original null hypothesis—that there is no difference in populations—is false. However, null hypotheses are not confined to the hypothesis that there is no difference between the groups. Crucially, we can test the hypothesis that there is a small or large difference between the groups. Conditional on us selecting the true effect size, the distribution of p-values is expected to be uniform (for just the same reason it was above when we correctly believed there was no difference). This means our best estimate of the true value of the effect is the value that makes the distribution of p-values (most) uniform. Moreover, note that we do not need to know whether there is a publication bias to compute this statistic. In the next section, we show that this approach to meta-analysis can correct for the concerns raised by Romero (2016).

3.1. Simulating self-correction

We explore via computer simulation a community of scientists attempting to measure some effect. Consider the example provided above involving a medical drug aimed at reducing the likelihood of cardiac arrest. A scientist intent on measuring the effectiveness of the drug runs an experiment involving two groups – one group (the treatment) is exposed to the drug, while the other group (the control) is not. Based on the health outcomes of members of both groups at the end of the study, the scientist can then provide an estimate of the treatment effect. A simple statistical test can determine whether the treatment effect is statistically significant – i.e., whether the null hypothesis that the likelihood of cardiac arrest is the same for both groups can be rejected.

Scientists then publish their study, which in turn is read and consumed by others. Publication occurs, however, with one important caveat. We assume scientists only publish *significant* results which show the drug works. In other words, *only those studies where the scientist is able to reject the null hypothesis* and which support drug efficacy are presented to the community for public consumption. If all scientists behave in this fashion, then the community as a whole will form incorrect beliefs about the actual effect of the drug. Romero nicely demonstrates this through the use of a simple meta-analytic technique.⁵ Aggregating across all published studies, Romero shows that in many cases the community severely overestimates the actual effect of the drug.

⁴ If there is no effect and we conduct 1000 trials at $\alpha = 0.05$, then 5% of the trials will be significant (by chance). Here we are assuming “two-tailed” significance tests were conducted so that there are cut-off values for both cases where the drug significantly outperforms the placebo and cases where the placebo significantly outperforms the drug. Assuming that there is no effect, both of these are equally likely and so by chance we expect 25 published trials where the drug outperforms the placebo and 25 where the placebo outperforms the drug.

⁵ In particular, we use a fixed-effect model of meta-analysis which calculates a weighted average of the experiment's effect size (in particular, effect size is weighted by the variance of each experiment). This term is then divided by the sum of the inverse of the variance of each experiment. See Borenstein et al. (2009) for more a more detailed presentation.

However, Romero's agents do not assess whether the assumptions of the meta-analysis they perform are met. As mentioned in section 3, the p-uniform method is a sophisticated meta-analytic technique designed to estimate an effect size in the presence of publication bias.⁶ We run simulations similar to those constructed by Romero and determine whether a p-uniform estimate outperforms the standard meta-analytic model considered by Romero (we refer to this as the baseline model). As discussed below, not only is the p-uniform method more reliable than the baseline model, the p-uniform method is able to arrive at highly accurate estimates of an effect size in a variety of different circumstances.

Fig. 1 provides an example of a simulation run. In each round, a series of studies are generated and the first experiment to reject the null hypothesis is published. We then estimate an effect size through the use of the baseline method as well as the p-uniform method, calculating an estimate for every ten rounds, beginning with round ten. As displayed in Fig. 1, the baseline approach systematically overestimates the actual effect size (0.41). The p-uniform method, however, rather rapidly finds its way to the correct effect size. Recall Romero's contention that SCT* holds if: 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. Fig. 2 speaks to this directly. For the same parameter values explored in Fig. 1, we can see that at the 10 round mark, the effect size estimated by the p-uniform method is on average very close to the actual effect size. At round 100, the p-uniform has zeroed in on the actual effect while the baseline method still provides a severe overestimate.

Table 1 reports results for other parameter values explored by Romero and we provide p-uniform estimates (based on 50 simulation runs) of the effect size.⁷ We determine the “error” of the p-uniform estimate by calculating the absolute difference between the estimate and the actual effect size. We then average error terms across our simulations. This is also done for the baseline model. Finally, we compare baseline to p-uniform estimates by considering the ratio of the average error associated with the baseline model to the average error associated with the p-uniform model. As is clear, not only does the p-uniform spectacularly outperform the baseline method, the average error by the 50th round can be extraordinarily small (and is many times smaller than the average error associated with the baseline model). This is especially so when the sample size of the underlying studies is modest ($N = 156$), yet even when studies are underpowered ($N = 36$), the p-uniform still performs quite well.

4. The self-correcting nature of non-self-correcting science

Though our simulations are highly idealized, they nevertheless allow us to think through various possible solutions to the problems revealed by the replication crisis. For example, one popular proposal that has been floated recently is the adoption of the more stringent p-value threshold of 0.005 (Benjamin et al., 2018). While this policy significantly decreases the likelihood the null hypothesis is incorrectly rejected and was in part originally suggested as a means of combating p-hacking, it should be noted that in the context of our simulations, the adoption of a 0.005 threshold makes self-correction *more* difficult. With a 0.005 threshold, only those papers that drastically overestimate the true effect size will be published. The baseline meta-analytic method used by Romero will thus *severely* overestimate the effect. The adoption of more stringent standards for statistical significance does not come without cost.

The p-uniform method, on the other hand, is able to provide a satisfactory estimate of the true effect size in this case (assuming the p-uniform is not underpowered). The most straightforward implication of the simulations in the previous section is that the deleterious effects of

⁶ For more on the p-uniform method as well as simulation code see <https://github.com/RobbievanAert/puniform>.

⁷ We consider just those cases identified by Romero as resulting in incorrect estimates of the effect size.

Fig. 1. Details of one simulation comparing estimates of baseline model and p-uniform method. Meta-analysis was conducted every ten rounds in the presence of a publication bias and a direction bias (real effect size is 0.41).

Fig. 2. Estimate of effect size (with 95% confidence intervals) at round 10 and round 100 in the presence of a publication bias and a direction bias. Actual effect size is 0.41 and N = 36. Data based on 50 simulation runs.

Table 1

Average estimate of effect size and average error for the p-uniform method. Table also displays the ratio of average error for baseline method and average error for p-uniform method. Simulations assume both direction bias and publication bias.

		Rounds	
		50	100
Effect = 0; N = 156	Average effect	0.007	0.009
	Average error	0.045	0.037
	Error ratio	8.5	10.2
Effect = 0.41; N = 156	Average effect	0.405	0.416
	Average error	0.024	0.019
	Error ratio	3.0	3.9
Effect = 0; N = 36	Average effect	0.011	0.019
	Average error	0.113	0.084
	Error ratio	7.1	9.5
Effect = 0.41; N = 36	Average effect	0.394	0.417
	Average error	0.086	0.061
	Error ratio	5.3	7.4

publication bias can often be eradicated when we attend to the way in which we aggregate data. Meta-analytic techniques can be *adapted* and *modified* to better deal with the various ways science as practiced departs from the ideal. Such adaptations allow us to construct procedures that meet Romero's challenge and demonstrate that SCT* still holds in cases that are far from utopian.

Yet this is not to say that any single aggregation procedure is a best response to *all* forms of bias. The p-uniform method, for instance, is not as

attractive of a corrective tool when dealing with direction bias alone. When scientists publish all studies (whether they be statistically significant or not) that register an effect in a particular direction, the p-uniform method provides an estimate that (while more accurate than the baseline method) fails to precisely track the underlying effect in the short-run.⁸ Failures of this kind should be expected and it would be presumptuous to think there is a truly magic aggregation scheme capable of dealing with all forms of bias.⁹ Yet the lack of a panacea does not entail that science is incapable of self-correction. While publication bias can be counteracted with the p-uniform method, other departures from the ideal (such as direction bias) can be dealt with in various ways. For instance, the 'trim and fill procedure' reliably estimates an effect size by correcting for asymmetries in the aggregate meta-data (Duval & Tweedie, 2000). Where one aggregation procedure fails, another may succeed.

Collectively the addition of such procedures add to the scientific community's arsenal of techniques to probe for error. As Mayo (1996, p. 5) noted, "the history of mistakes made in a type of inquiry gives rise to a list of mistakes that we would either work to avoid (before trial planning) or check if committed (after-trial checking)." We contend that the accumulation of such techniques reveals a norm of error elimination. In light of demonstrations that experimental techniques are flawed, scientific communities often respond by attempting to avoid the conditions that produce the error (e.g., instituting blinding procedures to avoid investigator bias) or developing the means to correct the error (e.g., the p-uniform technique).¹⁰

While Mayo (1996) focuses on errors at the level of the individual experiment, we wish to draw attention to the fact that these very same practices also apply to systematic effects that produce errors at the community (or meta-analytic) level. Thus far the philosophical discourse on the replication crisis has focused on ways to prevent errors by changing the social structure of science which gives rise to the problems in the first place (Romero, 2017, 2018; Heesen, 2018). However, applying Mayo's work on learning from error reveals numerous other strategies for how to respond to the replication crisis. In some cases, like the p-uniform technique, the changes leave problematic practices in place and work to find corrections for them. As such the p-uniform method is a concrete illustration of the ability to "to 'subtract out'" (Mayo, 1996, p. 5) the influence of deleterious error (publication bias).¹¹

Nevertheless, we concede that there may be cases where no currently available amalgamation scheme on offer is able to provide an accurate estimate. This reveals an important sense in which SCT* does not fully

⁸ This is due to the fact that the p-uniform method only considers papers that can reject the null hypothesis of no effect. As a result, this means the p-uniform method can be underpowered as it only draws on a fraction of the number of published papers. Yet even in this case SCT* holds; in the long run the p-uniform technique will converge on the correct effect size.

⁹ A recent paper (Aert, Wicherts, & van Assen, 2016) by the developers of the p-uniform method provides a careful outline of when and where the p-uniform method can be reliably applied.

¹⁰ Which is not to say that eliminating error is the only norm that govern scientific inquiry. As Elliott (2018) points out, there are many cases where other values take precedent. For example, in the case of chemical regulation, the state of California reasonably adopted less accurate methods of determining chemical safety even though they were less accurate, because they radically increased the number of chemicals that could be screened and thus, on the whole, increased the government's ability to protect the public.

¹¹ Though the dynamics of inquiry are different, the same considerations for addressing error apply to industry-funded science as well. For example, in their recent survey of the topic, Holman and Elliott (2018) identify numerous concrete means to address funding-bias, but nearly every corrective aims to prevent the error by some form of "pretrial (meta-) planning" in the form of reshaping the scientific community (e.g., public funding, firewalls, trial registries, etc.). Nothing here is to suggest that these solutions should not be considered. Nevertheless, extrapolating Mayo's view on error to the level of the community structure suggests that philosophers may find new and more promising ways to combat industry bias if they move beyond an exclusive focus on the prevention of errors and begin to consider how to amplify and subtract out sources of bias.

capture the notion of self-correction. Science is a dynamic process where knowledge is not only updated, but the methods of knowledge acquisition themselves evolve. To show that science does not self-correct, one must show both: (1) scientific methods are unreliable in practice and (2) where methods are unreliable, scientists fail to correct them.

One might object that this conceptualization sets an impossible standard. How does the critic of scientific self-correction show that not only are scientists using flawed methods, but that scientists are incapable of correcting the flaws in the methods they use? It seems that the former requires knowledge of the truth and the latter foreknowledge of the future. If so, scientific self-correction is in danger of being reduced to an article of faith, impossible to support one way or the other.

However, we think the replication crisis indicates this problem is more tractable than it first appears. The p-uniform technique was not the only procedure developed in the wake of the crisis (Simonsohn et al., 2014). Indeed, there was a community-wide recognition of publication bias and its negative consequences, which, along with a commitment to eliminating error and an incentive structure which rewards discovery and the creation of novel inferential tools, soon led to the development of a number of effective aggregation techniques. More than just an odd quirk of history, we think that this reveals something profound and hopeful about science and its ability to self-correct. The very salience of the issues that gave rise to Romero's (2016) argument that science was systematically biased, also stimulated scientists to improve their methods (van Assen, van Aert and Wicherts, 2015). In this case, at least, it was the initial failure to self-correct that led to self-correcting science.

References

- Aert, R., Wicherts, J., & van Assen, M. (2016). Conducting meta-analysis based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, 11, 713–729.
- Assen, M., Aert, R., & Wicherts, J. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological Methods*, 20, 293–309.
- Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, 483(7391), 531.
- Benjamin, et al. (2018). Redefine statistical significance. *Nature Human Behavior*. <https://doi.org/10.17605/OSF.IO/MKY9J>.
- Biddle, J. (2007). Lessons from the viox debate: What the privatization of science can teach us about social epistemology. *Social Epistemology*, 21, 21–39.
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester: John Wiley and Sons.
- Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics: Journal of the International Biometric Society*, 56, 455–463.
- Elliott, K. (2018). *Tapestry of values*. Oxford: Oxford University Press.
- Fernandez-Pinto, M. (2014). Philosophy of science for globalized privatization: Uncovering some limitations of critical contextual empiricism. *Studies In History and Philosophy of Science Part A*, 47, 10–17.
- Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational and Behavioral Statistics*, 9, 61–85.
- Heesen, R. (2018). Why the reward structure of science makes reproducibility problems inevitable. *Journal of Philosophy*, 115(12), 661–674.
- Holman, B. (2018). Philosophers on drugs. *Synthese*. <https://doi.org/10.1007/s11229-017-1642-2>.
- Holman, B., & Elliott, K. (2018). The promise and perils of industry-funded science. *Philosophy Compass*, 13, e12544.
- Jukola, S. (2015). Longino's theory of objectivity and Commercialized research. In S. Wagenknecht, N. Nersessian, & H. Andersen (Vol. Eds.), *Empirical philosophy of science. Studies in applied philosophy, epistemology and rational ethics: Vol. 21*. Cham: Springer.
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., ... Cemalcilar, Z. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45(3), 142.
- Longino, H. (2002). Science and the common good: Thoughts on Philip Kitcher's science, truth and democracy. *Philosophy of Science*, 69, 573–577.
- Mayo, D. (1996). *Error and the growth of experimental knowledge*. University of Chicago Press.
- Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, 7(6), 657–660.
- Romero, F. (2016). Can the behavioural sciences self-correct? A social epistemic study. *Studies In History and Philosophy of Science Part A*, 60, 55–69.
- Romero, F. (2017). Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, 84(5), 1031–1043 December 2017.
- Romero, F. (2018). Who should do replication labor? *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/2515245918803619>.
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science*, 9, 666–681.
- Sismondo, S. (2009). Ghosts in the machine: Publication planning in the medical sciences. *Social Studies of Science*, 39, 171–198.
- Stanford, K. (2015). Catastrophism, uniformitarianism, and a scientific realism debate that makes a difference. *Philosophy of Science*, 82, 867–878.