Philosophy of Science (2022), **89**, 970–979
doi:10.1017/psa.2022.50

PSA PHILOSOPHY OF SCIENCE ASSOCIATION

PHILOSOPHY OF SCIENCE

SYMPOSIA PAPER

What FiveThirtyEight Can Teach Us about Solving the Replication Crisis

Bennett Holman

Underwood International College Yonsei University, Seoul, South Korea and Faculty of Humanities, University of Johannesburg, Johannesburg, South Africa
Email: bholman@yonsei.ac.kr

(Received 14 October 2021; revised 15 April 2022; accepted 25 April 2022; first published online 30 May 2022)

Abstract

Numerous philosophers of science have argued that the incentive structure of science is a major contributor to the replication crisis. In this article I review FiveThirtyEight's election forecasting model and show that it confronts similar problems in political polling and successfully mitigates them, not by changing the incentives, but by weighting a pollster's effect on the model by their past reliability. I argue that a similar weighting procedure holds a promising approach to addressing the replication crisis in science.

1. Introduction

In the summer of 2013, Ozge Sigirci reached out to Brian Wansink about spending a summer in his lab as a visiting student. For Sigirci it was a savvy move. Wansink was an academic superstar: The Cornell professor (and former head of the US Department of Agriculture Center for Nutrition Policy) regularly published highly cited work and received national press coverage. But it was also a beneficial relationship for Wansink. Prior to Sigirici's arrival, he sent her a "failed study" and instructed her to comb through the data for statistically significant results. For example, he told her to "see if there are any weird outliers ... [and] if there seems to be a reason they are different, pull them out" and to "think of all of the different ways that you can cut the data and analyze subsets of it to see where the relationship holds." He closed saying, "Work hard, squeeze some blood out of this rock, and we'll see you soon."[1] Following his instructions, Sigirci was able to produce five papers, all coauthored with Wansink.

[1] The quotes from Wansink's email were obtained by Buzzfeed reporter Stephanie Lee (2018). Attempts to reconstruct the analyses lead to the identification of more than 150 inconsistencies in the reported data (van der Zee et al. 2017).

When these practices came to light, Wansink (2016) defended himself saying that,

> hypotheses usually don't "come out" on the first data run. But instead of dropping the study, a person contributes more to science by figuring out when the hypo worked and when it didn't. This is Plan B. Perhaps your hypo worked during lunches but not dinners, or with small groups but not large groups. You don't change your hypothesis, but you figure out where it worked and where it didn't. Cool data contains cool discoveries.

While under certain circumstances post-hoc analyses can be helpful, by publishing the results as if they were the result of a single preplanned analysis, Wansink engaged data dredging—a practice that has long been known to generate spurious results (Selvin and Stuart 1966; Erasmus et al. 2022).

Subsequent investigation showed that such procedures were standard practice in the lab. A graduate student, who left the lab after becoming uncomfortable with Wansink's research practices, revealed that he would intentionally facilitate dredging by creating studies with as many variables as possible (Newburger 2018) and use one-sided hypothesis tests (Brown 2017b). Moreover, he sought out results that would capture media attention (Lee 2018). In short, Wansink was a walking embodiment of the abuse of statistics that have led to the replication crisis.

Yet what often gets overlooked is that his travails began with a blog post intended to help aspiring academics. In his telling, Sigirci's data dredging is a model for an enterprising researcher to emulate. He contrasts her success with a paid postdoc who was now leaving academia and had turned down the same project because it was "low quality." Wansink's message to young academics was publish prolifically even at the cost of sacrificing work-life balance. Yet, what he ultimately displayed in his routine practice is that his prolific publication stemmed from more than just forgoing a few episodes of *Game of Thrones*; in addition to his abuse of statistical methodology, Wansink also was found to have self-plagiarized, republished work without acknowledgment, and published the same work in multiple venues (Brown 2017a).

Aside from his individual stature, the reason why Wansink is worthy of discussing is that he openly described research practices (e.g., fishing, p-hacking) that are presumed to be pervasive. The explosive attention to Wansink came on the heels of a controversy prompted by prominent psychologist Daryl Bem (2011) purporting to show evidence for paranormal abilities. The experiments failed to replicate (Ritchie et al. 2012), but the fact that standard statistical methodology could produce apparent evidence for paranormal abilities drew increased scrutiny to more prominent and central findings in psychology.

For example, Roy Baumister's theory of self-control posited that volitional actions depend on an exhaustible inner resource and that if, as in the original study, you have just spent five minutes resisting the urge to eat chocolate chip cookies, you will be less able to persist in attempting to solve an unsolvable puzzle (Baumeister et al. 1998). A meta-analysis of 83 studies of ego-depletion found that the effect was robust across a wide range of experimental paradigms (Hagger et al. 2010). Yet when a large multinational attempt was made to replicate the finding with a preregistered plan of analysis, the effect size could not be distinguished from zero (Hagger et al. 2016).

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

Similar results have been obtained for several of John Bargh's classic experiments on social priming (e.g., Bargh, Chen, and Burrows 1996; cf., Chivers 2019). Likewise, the much-vaunted finding that striking a "power pose" can cause people to behave more assertively (Carney et al. 2010) failed to stand up to scrutiny (e.g., Simmons and Simonsohn 2017). Indeed, in a massive collaboration which replicated 100 prominent psychology experiments, only 40 percent successfully replicated (Open Science Collaboration 2015).

In diagnosing this issue, a number of philosophers of science have faulted the reward structure of science (Bright 2017; Romero 2017; Zollman 2019; Heesen 2021). Indeed, Remco Heesen (2018) has gone so far as to claim that the reward structure of science makes reproducibility problems inevitable. While agreeing that the reward structure of science incentivizes unreliable work, I argue that reliable judgments can nonetheless be derived.

In section 2, I review these concerns and begin my argument that they are overly pessimistic. In particular, I argue that they tacitly assume that our best estimate of the truth will be determined by individual incentives and/or that our best judgment relies on a blind amalgamation of all the available evidence. However, if there are means of discriminating reliable studies, it undercuts the inevitability the reproducibility crisis. Indeed, such work is already now underway by the Open Science Collaboration (Alipourfard et al. 2021). However, because such work is inchoate, I first turn in section 3 to a proof of principle by reviewing the election forecasting model assembled by FiveThirtyEight. Here I show that despite misaligned incentive structures in polling, the FiveThirtyEight election model can generate remarkably accurate projections. In section 4, I use this case study to illustrate how the concerns raised by philosophers are addressed in election forecasting and anticipate some objections to my proposal.

2. Incentive structures and the division of cognitive labor

Recent concerns about the incentive structure in science stand in stark contrast to the views offered by earlier philosophers of science. Most notably, both Phillip Kitcher (1990) and Michael Strevens (2003) argued that credit incentives within science permits a community of selfish fame-seeking scientists to distribute themselves in a way that would promote truth. These "invisible hand" type arguments begin from the premise that there are some puzzles the scientific community is attempting to solve and that there are numerous approaches that might be successful in providing a solution. If scientists were only concerned with the truth, then every scientist would pursue the project most likely to solve the puzzle. In contrast, in a community where scientists are selfish credit seekers, when highly promising projects become saturated it will be rational to attempt a less promising project. Accordingly, Kitcher and Strevens argue, the scientific community will discover more truths if individual scientists selfishly try to maximize credit.

Recent reflections have cast doubt on this idealization of scientific inquiry. As Felipe Romero (2017) argues, these arguments depend on assumptions that often do not hold. First, invisible-hand arguments assume that scientific races for priority have a "winner-contributes-all" character such that once a discovery is made there is no additional epistemic benefit for a researcher to demonstrate the same effect.

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

Romero points out that this depends on the assumption that discoveries are genuine and replicable. However, given that a substantial portion of studies are spurious and would fail to replicate if tested, the winner-takes-all priority rule fails to adequately value replication work. Accordingly, a group of selfish credit-seeking scientists will focus on the discovery of novel effects to the exclusion of the replication work that needs to be done to ensure that previous discoveries are genuine.

Moreover, it may well be that the reward structure incentivizes researchers to "rush to print" and publish specious results (Heesen 2018). Specifically, Heesen argues that as long as (1) there is a trade-off between speed and reproducibility; (2) that scientists get rewarded for publishing; and (3) that peer review has to assess long-term merit of the paper, then (4) the reward structure of science inevitably incentivizes scientists to be less rigorous than would be optimal from a community perspective. Furthermore, if scientists can turn published work into more resources (e.g., better jobs, grants) before the scientific community can show that their work is unreliable, then scientists are doubly incentivized to crank out unreliable research (Heesen 2021). In such cases, researchers can run a "scientific Ponzi scheme" in which apparent successes in the short term are leveraged to attract enough additional resources and prestige that the researcher can absorb the eventual reputation hit when some of their old research fails to replicate because they have produced so many new novel findings (Zollman 2019). This can come crashing down, as it did for Brian Wansink, but it need not.

A natural response to the preceding analysis is to recommend aligning the researcher's incentive structure with the production of reliable science. However, two additional results warrant caution. First, merely aligning individual incentives with truth doesn't necessarily lead to individuals who act in truth-promoting ways. On some occasions, it is possible that truth-motivated scientists will be inspired to tell a "noble lie". That is, they will commit scientific fraud because, from their perspective, truthfully reporting their results would be misleading (Bright 2017). Second, on the basis of his analysis, Heesen (2018) argues that "no 'nearby' reward structure fully solves this problem" (664). In particular, Heesen argues that the solutions to prevent publication bias (e.g., publication of negative results) and p-hacking (e.g., preplanned analysis) would still fail to prevent the incentive for individual scientists to rush to publish. In section 4, I argue that reliability weighting in meta-analyses provides just such a solution, but before I do so, it will be helpful to examine a working example of the procedure: the election modeling published by FiveThirtyEight.

3. The signal and the noise

Polling as a means to predict an election has a great deal in common with many of the problems tackled in the sciences.[2] Polls are surveys of a relatively small sample that attempt to estimate a population value (i.e., the margin of victory in an upcoming election). Just as in science, polling agencies must make a series of methodological decisions to derive their estimate. For example, a polling agency must first determine how to generate their sample. One might employ a call center staffed with people,

[2] Dissimilarities, such as a definitive outcome to compare the polls against to assess accuracy, will be addressed at the end of section 4. Readers who are particularly concerned with this apparent dissimilarity may wish to skip ahead before reading this section.

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

utilize robopolls (i.e., automated voice polling), or survey people online. There are also issues with survey design. For example, some pollsters recommend asking the participant to predict the outcome of the election rather than ask for whom they plan to vote (Graefe 2014; cf. Gelman and Azari 2017). Beyond sampling techniques and constructing survey questions, polling agencies must also make choices about how to weight their samples, which might have oversampled some groups and undersampled others. To compensate for this, agencies must employ a "turn out model" and then weight their sample accordingly.

Importantly, as in science, the incentives for polling agencies are often not truth conducive. At the extreme, some polls are used as a means to exploit betting markets (Yeargain 2020). Similarly, "push polls," in which pollsters divulge beneficial or critical information before soliciting the respondent's opinion, can be used to manipulate public opinion (Streb and Pinkus 2004). Less egregiously, campaigns conduct internal polls but generally only release results that reflect favorably on their candidate (Freiss 2015). Even professional firms can be subject to misaligned incentives. As Gelman et al. (2020) note, polling firms that have a house effect that consistently favor one political party may be more likely to be hired by partisan actors and receive coverage on partisan media outlets. Similarly, close to the election, pollsters tend to "herd." This occurs when an agency produces a poll that is out of step with other results and decides to alter their assumptions post-hoc to produce a poll that is better in line with other polls.

All these factors impede accurate forecasts of the election; however, over the past two decades FiveThirtyEight's forecast has been remarkably accurate in predicting American elections. Most recently, while not forecasting the election for Donald Trump in 2016, they gave Trump a 29 percent chance of winning the electoral college, in comparison to other models that gave Trump 15 percent, 8 percent, 2 percent, or less than 1 percent chance of winning (Silver 2016). Moreover, the reasons why FiveThirtyEight's model gave Trump a higher chance of winning were borne out by the election—namely polling errors were correlated.[3] Similarly, though the 2020 election forecast gave Biden a 90 percent chance of winning, Silver cautioned that with a standard-sized polling error in Trump's favor:

> Biden would probably hold on, but he'd . . . narrowly win some states that she [Clinton] narrowly lost . . . Biden would probably be reliant on Pennsylvania in this scenario—a state that is expected to take some time to count its vote—the election might take longer to call . . . . So while Biden *isn't* a normal-sized polling error away from *losing*, he is a normal-sized polling error away from having a messy win that might not come with control of Congress. (Silver 2020b, emphasis in original)

After a normal-sized polling error in Trump's favor, the election outcome remained uncertain for four days and was only resolved when a Biden victory in Pennsylvania became definitive. Democrats narrowly retained the Senate by virtue of two

[3] Polling errors are correlated when the polling error in one state (e.g., Wisconsin) predicts polling error in other states (e.g., the entire Midwest). In 2016, late undecided voters broke consistently for Trump.

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

improbable and close runoff elections in Georgia. As illustrated here, a good forecast does not merely generate a prediction, but articulates the range of uncertainty, a task which FiveThirtyEight has been remarkably successful at.

Though the FiveThirtyEight model has several features that contribute to its success, the most pertinent for our purposes is how it handles polls. For reasons outlined in the preceding text, FiveThirtyEight does not treat all polls equally, but instead factors in a polling agency's reliability score before letting it influence the model. The first step in calculating the reliability score is collecting every poll an agency has conducted that meets certain inclusion parameters (e.g., conducted within three weeks of the election). Next, these polls are compared against the election result to calculate the polling agency's average error. Then, a regression analysis is run to control for known factors of polling error (e.g., sample size, race type, days from the election), and a score is derived from the difference between the polling firms observed error and its predicted error. Finally, the score is adjusted by the pollster's relative performance in the field (i.e., how well did they do compared to other agencies that polled the same race). Further details are available (Silver 2014, 2020a), but what is relevant for our purposes is that when generating predictions of future elections, a pollster's past track record is taken into consideration. Agencies with better track records are given more influence in subsequent predictions.

4. How to mitigate the effects of misaligned incentives

To recap the argument thus far, I first surveyed arguments that the incentive structure of science produces epistemically deleterious outcomes. In the previous section, I reviewed how FiveThirtyEight is able to generate a high-quality forecast despite a series of underlying incentives that distort the underlying polls. In this section, I will show that such "reliability weighting" is particularly adept at confronting the very issues that have been at the root of the pessimism expressed by philosophers of science. Again, how such a procedure would be translated into various areas of scientific inquiry is a nontrivial challenge. While such efforts are underway (Alipourfard et al. 2021), the particulars of how a system would be implemented do not matter for the argument here, at least to refute arguments that the incentive structure of science makes these problems inevitable. The fact that such problems have proved tractable in political polling undercuts the inevitability of such issues and suggests promising avenues for future work. Thus, in this section I will consider the specific concerns expressed in the context of science and show that they have been mitigated in the field of political polling.

Recall that Romero's (2017) primary concern was that the priority rule in science leads to an undervaluation of replication work. As has frequently occurred, the first study of a phenomena finds a strong effect and garners a significant amount of attention, but replications of the study find that the true effect is substantially smaller (Schooler 2011). Indeed, this is almost inevitable if studies are conducted with small samples and only publish statistically significant results (Gelman and Carlan 2014). Even when subsequent studies show the effect to be significantly smaller than previously described, the original authors still earn most of the credit for the discovery. In contrast, note that this is not the case for FiveThirtyEight's credibility rating. Though the first poll in some contest might garner outsized media attention, what matters is

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

accuracy not priority. If a similar reliability weighting practice were adopted in science, early movers who produced inflated estimates of effects would be judged as less reliable than careful attempts to replicate that produced a more accurate estimate of the true effect. Moreover, in the FiveThirtyEight reliability ratings, polling firms are incentivized to seek out places where there are a small number of inflated polls because a firm's weighting is partially determined by relative performance (Silver 2014).

This same feature also addresses Heesen's (2018, 2021) concerns, though the broad strokes of how such a system is implemented matter here. If it were the case that promotion and granting agencies began using reliability ratings in making their decisions, then accuracy rather than publication quantity drives the incentive structures and the preconditions for Heesen's proofs (authors are rewarded for publication) no longer obtain. Yet a system in which scientists are rewarded for accuracy is a trivial case in which the incentive structure no longer produces unreliable science. Far more interesting are cases in which the incentive structures for personal advancement remain in place. Note that, as discussed in section 3, numerous pressures to perform unreliable polls exist and do still in fact lead to the production of unreliable information. What occurs is not a curbing of these incentive structures, but an iterative process of separating the signal from the noise.

The final concern is that incentive structures can reward fraud and that this can occur even when scientists are truth-seeking (Bright 2017). Cases of malicious fraud do occur in polling. For example, in July 2017, Delphi Analytic released a poll indicating that Kid Rock was leading the race for a Michigan Senate seat over incumbent Debbie Stabenow. The poll quickly ricocheted through the political media, but the team at FiveThirtyEight was able to determine the poll was fraudulent (Enten 2017). Firms that have been found to conduct such polls or are suspected of engaging in fraud are excluded from FiveThirtyEight's election forecast (Silver 2020a). More interesting is the type of fraud engaged in by truth-seeking pollsters. Bright's concern is that a truth-motivated actor may tell a "noble lie" when they believe that their results are out of line with the truth. Though Bright raises the case as a hypothetical, this is exactly what happens when pollsters "herd" as the election draws near. As Silver (2014) notes, "[P]aradoxically while herding may make an individual polling firm's results more accurate, it can make polling averages worse." This is because the individual pollster is using additional information to inform their poll (which is good for the agency), but does so at the cost of sacrificing the independence of each observation (which is bad for the group [and the forecaster]). Accordingly, FiveThirtyEight includes a "herding penalty" when pollsters fail to produce results with the expected amount of variability (Silver 2019).

In short, the very issues with incentive structures that philosophers of science claim make the reproducibility crisis inevitable in science have been successfully mitigated in polling. Of course, whether such procedures could be translated into other disciplines is a matter of conjecture. While I do not have space to address all the possible complications, perhaps the most salient dissimilarity appears to be that at the end of a polling period there is a definitive result that can be used as a benchmark, and there doesn't seem to be any analogue in most areas of science. In response, I briefly consider two possible scenarios that might serve to allay these concerns.

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

In an optimistic scenario, the culmination of scientific work is sufficient to produce a reliable estimate of the relevant effect size. As a matter of empirical fact, scientific communities frequently come to consensus and I have previously argued that there is reason to believe that meta-analysis combined with more advanced tools should converge on the truth even in the face of distorted incentives (Bruner and Holman 2019; Holman 2019). If the long-term results of inquiry are sufficiently definitive and reliable, then they can serve as a benchmark for reliability scores (perhaps iteratively as new data emerges). However, suppose a less happy situation in which consensus is unreliable. In this situation it may yet be possible to create benchmarks by conducting large, rigorous multinational experiments for important findings (e.g., Hagger et al. 2016). Indeed, it might well be in the interest of governmental funding bodies to set aside grants for just such projects. Obviously, this is far from a worked-out proposal for how such benchmarks would be created, but hopefully it serves to illustrate that the creation of benchmarks is not foreclosed *a priori*. That, in turn, is enough to undercut our confidence in the pessimistic pronouncements of philosophers of science.

5. Conclusion

It is quite right to raise concerns that the incentive structures of science may be problematic. Contemporary work in philosophy of science is a step in the right direction from the optimistic invisible-hand-type models that sought to reassure philosophers that departures from rationality need not necessarily lead to postmodern relativism (Kitcher 1990; Strevens 2003). Yet, though the work Romero, Heesen, Zollman, and Bright represent an improvement over previous models, I contend that they have done so at the expense of the primary insight of older work: a principle that Mayo-Wilson, Zollman, and Danks (2011) call "the independence thesis." Specifically, the contention that the epistemic merits of a group can come apart from the epistemic merits of the individuals that comprise it and "that methodological prescriptions for scientific communities and those for individual scientists are logically independent" (653).

In this article I have shown that some of the very same issues that philosophers claim are an inevitable consequence of scientists' current incentive structure have been mitigated in political polling. While this falls short of a proposal for precisely how to achieve the same result within science, I contend that FiveThirtyEight's forecast model is deserving of further study by those interested in taking the next step forward in addressing the replication crisis. With such an approach, individual studies would not necessarily be more replicable, but it would cease to be a crisis if the community could tell—with sufficient time—which findings were reliable. Even though such an approach may leave distorted incentive structures in place, it may nevertheless embody the primary insight of social epistemology: Science does not need fully reliable individuals to produce reliable knowledge.

Acknowledgments. I am grateful to Dan Singer, Patrick Grim, Remco Heesen, and Felipe Romero for their comments on an earlier draft of this paper.

References

Alipourfard, Nazanin, Beatrix Arendt, Daniel Jacob Benjamin, Noam Benkler, Michael Bishop, Mark Burstein, Martin Bush et al. 2021. "Systematizing Confidence in Open Research and Evidence (SCORE)." https://doi.org/10.31235/osf.io/46mnb

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

Bargh, John A., Mark Chen, and Lara Burrows. 1996. "Automaticity of Social Behavior: Direct Effects of Trait Construct and Stereotype Activation on Action." *Journal of Personality and Social Psychology* 71(2):230–44.

Baumeister, Roy F., Ellen Bratslavsky, Mark Muraven, and Dianne M. Tice. 1998. "Ego Depletion: Is the Active Self a Limited Resource?" *Journal of Personality and Social Psychology* 74(5):1252–65.

Bem, Daryl J. 2011. "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect." *Journal of Personality and Social Psychology* 100(3):407–25.

Bright, Liam K. 2017. "On Fraud." *Philosophical Studies* 174(2):291–310.

Brown, Nick. 2017a. "A Different Set of Problems in an Article from the Cornell Food and Brand Lab." http://steamtraen.blogspot.com/2017/02/a-different-set-of-problems-in-article.html. Accessed July 14, 2021.

Brown, Nick. 2017b. "Some Instances of Apparent Duplicate Publication from the Cornell Food and Brand Lab." http://steamtraen.blogspot.com/2017/03/some-instances-of-apparent-duplicate.html. Accessed July 14, 2021.

Bruner, Justin P., and Bennett Holman. 2019. "Self-correction in Science: Meta-analysis, Bias and Social Structure." *Studies in History and Philosophy of Science Part A* 78:93–97.

Carney, Dana R., Amy J. C. Cuddy, and Andy J. Yap. 2010. "Power Posing: Brief Nonverbal Displays Affect Neuroendocrine Levels and Risk Tolerance." *Psychological Science* 21(10):1363–68.

Chivers, Tom. 2019. "What's Next for Psychology's Embattled Field of Social Priming." *Nature* 576(7786):200–3.

Enten, Harry. 2017. "Fake Polls Are a Real Problem." *FiveThirtyEight*, August 22, 2017. https://fivethirtyeight.com/features/fake-polls-are-a-real-problem/.

Erasmus, Adrian, Bennett Holman, and John P. A. Ioannidis. 2022. "Data-Dredging Bias." *BMJ Evidence-Based Medicine* 27(4):209–11.

Freiss, Steve. 2015. "Why Political Journalists Shouldn't Report on Internal Polling." *Columbia Journal Review*, August 10, 2015. https://www.cjr.org/analysis/internal_polling_election_data.php. Accessed August 10, 2021.

Gelman, Andrew, and Julia Azari. 2017. "19 Things We Learned from the 2016 Election." *Statistics and Public Policy* 4(1):1–10.

Gelman, Andrew, Jessica Hullman, Christopher Wlezien, and George Elliott Morris. 2020. "Information, Incentives, and Goals in Election Forecasts." *Judgment and Decision Making* 15(5):863–80.

Gelman, Andrew, and John Carlin. 2014. "Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors." *Perspectives on Psychological Science* 9(6):641–51.

Graefe, Andreas. 2014. "Accuracy of Vote Expectation Surveys in Forecasting Elections." *Public Opinion Quarterly* 78(S1):204–32.

Hagger, Martin S., Chantelle Wood, Chris Stiff, and Nikos L. D. Chatzisarantis. 2010. "Ego Depletion and the Strength Model of Self-control: A Meta-analysis." *Psychological Bulletin* 136(4):495–525.

Hagger, Martin S., Nikos L. D. Chatzisarantis, Hugo Alberts, Calvin O. Anggono, Cedric Batailler, Angela R. Birt, Ralf Brand et al. 2016. "A Multilab Preregistered Replication of the Ego-Depletion Effect." *Perspectives on Psychological Science* 11(4):546–73.

Heesen, Remco. 2018. "Why the Reward Structure of Science Makes Reproducibility Problems Inevitable." *The Journal of Philosophy* 115(12):661–74.

Heesen, Remco. 2021. "Cumulative Advantage and the Incentive to Commit Fraud in Science." *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/716235

Holman, Bennett. 2019. "In Defense of Meta-analysis." *Synthese* 196(8):3189–211.

Kitcher, Phillip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* 87 (1):5–22.

Lee, Stephanie. 2018. "Here's How Cornell Scientist Brian Wansink Turned Shoddy Data into Viral Studies About How We Eat." https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking. Accessed July 14, 2021.

Mayo-Wilson, Connor, Kevin Zollman, and David Danks. 2011. "The Independence Thesis: When Individual and Social Epistemology Diverge." *Philosophy of Science* 78(4):653–77.

Newburger, Emma. 2018. "Students Who Worked in Cornell Food Lab Say Director's Retracted Studies Stain Reputations." *The Cornell Daily Sun*, February 8, 2018. https://cornellsun.com/2018/02/08/cornell-professors-continuous-retractions-stained-lab-reputation-students-say/. Accessed July 14, 2021.

https://doi.org/10.1017/psa.2022.50 Published online by Cambridge University Press

Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716.
Ritchie, Stuart J., Richard Wiseman, and Christopher C. French. 2012. "Failing the Future: Three Unsuccessful Attempts to Replicate Bem's 'Retroactive Facilitation of Recall' Effect." *PloS One* 7(3): e33423.
Romero, Felipe. 2017. "Novelty Versus Replicability: Virtues and Vices in the Reward System of Science." *Philosophy of Science* 84(5):1031–43.
Schooler, Jonathan. 2011. "Unpublished Results Hide the Decline Effect." *Nature* 470(7335):437.
Selvin, Hanan C., and Alan Stuart. 1966. "Data-Dredging Procedures in Survey Analysis." *The American Statistician* 20(3):20–23.
Silver, Nate. 2014. "How FiveThirtyEight Calculates Pollster Ratings." *FiveThirtyEight*, September 25, 2014. https://fivethirtyeight.com/features/how-fivethirtyeight-calculates-pollster-ratings/. Accessed August 9, 2021.
Silver, Nate. 2016. "Why FiveThirtyEight Gave Trump a Better Chance of Winning Than Almost Everyone Else." *FiveThirtyEight*, November 11, 2016. https://fivethirtyeight.com/features/why-fivethirtyeight-gave-trump-a-better-chance-than-almost-anyone-else/. Accessed August 9, 2021.
Silver, Nate. 2019. "The State of the Polls 2019." *FiveThirtyEight*. https://fivethirtyeight.com/features/the-state-of-the-polls-2019/. Accessed August 9, 2021.
Silver, Nate. 2020a. "Our New Polling Averages Show Biden Leads Trump by 9 Points Nationally." *FiveThirtyEight*, June 18, 2020. https://fivethirtyeight.com/features/our-new-polling-averages-show-biden-leads-trump-by-9-points-nationally/. Accessed August 9, 2021.
Silver, Nate. 2020b. "Biden's Favored in Our Final Presidential Forecast, But It's a Fine Line between a Landslide and a Nail-Biter." *FiveThirtyEight*, November 3, 2020. https://fivethirtyeight.com/features/final-2020-presidential-election-forecast/. Accessed August 9, 2021.
Simmons, Joseph P., and Uri Simonsohn. 2017. "Power Posing: P-curving the Evidence." *Psychological Science* 28(5):687–93.
Streb, Matthew J., and Susan H. Pinkus. 2004. "When Push Comes to Shove: Push Polling and the Manipulation of Public Opinion." In *Polls and Politics: The Dilemmas of Democracy*, edited by Michael Genovese and Mathew Streb, 95–115. Albany: State University of New York Press.
Strevens, Michael. 2003. "The Role of the Priority Rule in Science." *The Journal of Philosophy* 100 (2):55–79.
van der Zee, Tim, Jordan Anaya, and Nicholas J. L. Brown. 2017. "Statistical Heartburn: An Attempt to Digest Four Pizza Publications from the Cornell Food and Brand Lab." *BMC Nutrition* 3(1):1–15.
Wansink, Brian. 2016. "The Grad Student Who Never Said No." https://web.archive.org/web/20170312041524/http:/www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no. Accessed July 14, 2021.
Yeargain, Tyler. 2020. "Fake Polls, Real Consequences: The Rise of Fake Polls and the Case for Criminal Liability." *Missouri Law Review* 85(1):129–90.
Zollman, Kevin. 2019. "The Scientific Ponzi Scheme." Unpublished manuscript, July 27, 2019. https://philsci-archive.pitt.edu/16264/

Cite this article: Holman, Bennett. 2022. "What FiveThirtyEight Can Teach Us about Solving the Replication Crisis." *Philosophy of Science* 89 (5):970–979. https://doi.org/10.1017/psa.2022.50