Acad. Quest. (2021) 34.2
DOI 10.51845/34su.2.11

Review Essay

Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth, Stuart Ritchie, 2020, Metropolitan Books, pp. 368, $22.26 hardcover.

Can Science be Saved?

John Staddon

Stuart Ritchie is a cognitive psychologist, a lecturer at the University of London's King's College. A few years ago he had an experience that seems to have been the impetus for this lively and important book.

In 2011 Daryl Bem, a well-known social psychologist at Cornell University, published a series of experiments in a mainstream peer-reviewed journal.[1] Bem claimed to have demonstrated *precognition.* He used a very simple procedure: one hundred subjects had to guess (36 trials each) which of two curtains (on a computer screen) concealed a randomly assigned picture. A third of the pictures were erotic, two-thirds were not. Subjects failed to guess correctly when the pictures were neutral but did better than chance when they were erotic. Since the guesses occurred before the pictures were presented, this counts as precognition. The effects were relatively small, but "statistically significant"; 53.1 percent of choices correctly anticipated the erotic pictures, but only 49.8 percent anticipated the non-erotic ones. Nine similar experiments followed, eight of which showed significant results. This "breakthrough research" created quite a stir.

The experiments were simple and then-graduate-student Ritchie and a couple of collaborators each tried to replicate the first of them. They failed to find any evidence for precognition.[2]

If true, Bem's results would be an astonishing challenge to everything from physics to psychology. A failed replication should therefore be of the greatest interest—like the famous Michelson-Morley experiment which failed to find an expected change in

1 Daryl J. Bem, "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect," *Journal of Personality and Social Psychology* 100, no. 3 (2011), 407– 25.

2 A battle ensued. Ritchie *et al.*'s paper was eventually published. Bem and collaborators mustered considerable support, but overall the consensus is "Not Proven."

John Staddon is James B. Duke Professor Emeritus, Department of Psychology and Neuroscience, Duke University. A recent book is his *Scientific Method: How science works, fails to work or pretends to work.* (Taylor and Francis, 2017). He last appeared in these pages with his article "What's Really Wrong with America," in the Winter 2020 issue.

the velocity of light and eventually led to Einstein's special relativity theory. Sometimes a null result can have huge implications.

But not, apparently, for the editor of the *Journal of Personality and Social Psychology,* who rejected the Ritchie failure-to-replicate paper without review. In other words, standard policy for a prestigious psychology journal in 2011 was "no replications, please, we're scientists"!

In 2005, medical statistician John Ioannides, in a paper provocatively titled "Why most published research findings are false" had already highlighted what is now known as the "replication crisis": the fact that a majority of research findings in biomedicine could not be replicated. The issue has been reviewed at length by the National Association of Scholars and is now a general concern.[3] But still, six years after Ioannides's revelation, replication had apparently failed to catch the attention of the editor at *JPSP*.

Science Fictions covers replication and many other ways that science can fail and is failing. Ritchie's point is that the failures are to a large extent systemic. No one ever accused Bem of faking his data. The problem was partly the methods he used, but it was chiefly the problem of incentive to find something spectacular and the impediments to adequate criticism of the work.

Ritchie begins pessimistically:

> Science, the discipline in which we should find the harshest skepticism, the most pin-sharp rationality and the hardest-headed empiricism, has become home to a dizzying array of incompetence, delusion, lies and self-deception. In the process, the central purpose of science—to find our way ever closer to the truth—is being undermined.

At the end of the book he recounts an admonition just like one I also received from a reviewer of my own *Scientific Method* book: "Isn't it irresponsible to write something like that? Won't you encourage a free-for-all, where people use your arguments to justify their disbelief in evolution, or in the safety of vaccines, or in man-made global warming?" Well, no; as Ritchie points out, science is, or should be, all about criticism: "How do they know?" should always be the

3 *The Irreproducibility Crisis: Causes, Consequence, and the Road to Reform*, National Association of Scholars, April 9, 2018, https://www.nas.org/reports/the-irreproducibility-crisis-of-modern-science/full-report

response to any new claim. Now it is rarely asked.

I thought that I was reasonably familiar with science's problems, but Ritchie's book convinced me that I had greatly underestimated them. Here are just a few examples.

Scientists themselves believe there's a crisis: Nature, one of the two leading general-science journals, found that 52 percent of 1,500 researchers who filled out a 2016 website survey thought there was a "significant crisis" of replicability;

Obstacles to replication: an attempt to replicate fifty-one important preclinical cancer studies was foiled when, "In every single one of the original papers, for every single one of the experiments reported, there wasn't enough information provided for researchers to enable them to repeat the experiment." A later study of 268 biomedical papers again found only one that adequately reported what had been done.

Medical failures to replicate: For years it was believed that Caesarean section was safer for childbirth. But a big 2013 randomized trial found no difference. Guidelines for at-risk babies said to avoid peanuts. But a 2015 randomized trial found the opposite was best. After a heart attack, it was thought that cooling a patient could help. But a 2014 study found the opposite. After a stroke it was believed moving the patient was the best therapy. Nope. A 2016 study found that rest is better. The vacillations of dietary diktats are now so frequent that even the woman-in-the-street is aware of them.[4]

Outright fraud is also a problem. A surprising number of "scientific" results are simply faked. The champ in this respect seems to be Italian surgeon Paulo Macchiarini, who claimed in many published papers to have developed a treatment that allowed him to successfully transplant artificial human tracheas. After painful patient deaths in several countries, Macchiarini was revealed as a data faker and a liar.

Biomedical fraud is disturbingly widespread. Yoshihiro Sato in Japan "had fabricated data for dozens of clinical trials published in international journals."[5] Two instances of fraud at Duke University Medical Center have led to disciplinary action from government agencies and fines in the hundreds of millions of dollars. In 2018, Harvard Medical School and Brigham and Women's Hospital in 2018 reported on thirty-one publications with "falsified and/or fabricated

4 John Staddon, "Diet Reporting—The Real Fake News," *Quillette*, September 18, 2019.
5 Kai Kupferschmidt, "Tide of Lies," *Science*, August 17, 2018.

data." Wikipedia has an entry that lists scientific frauds in many areas and in many countries; the list for biomedicine is much the longest.

Yes, science is in trouble, mostly for the reason of perverse incentives. But first, a technical note.

The NHST Method

Stuart Ritchie is a psychologist and is therefore most familiar with the Null Hypothesis Statistical Test method, which is favored in that area and in most social and biomedical science research. The method is now dominant but was not always so.

The NHST method was invented when the *single-subject* method[6] used in earlier experimental sciences such as physics and chemistry proved impractical. The single-subject method is still used in much of experimental psychology (as opposed to social, personality, clinical, and even cognitive psychology). For example, as a graduate student, I measured visual acuity under two conditions: white on black letters or the reverse, to see which gave the better acuity. I needed only a handful of subjects, as the two conditions could be repeated and compared indefinitely within each subject. Neither averaging nor statistics was necessary.

The single-subject method runs in to difficulties when the experimental treatment itself has an irreversible effect on the subject. If you want to compare two methods of teaching kids to read, for example, you can't teach using method A and have the kid unlearn so you can try method B. But you can compare two randomly selected (that bit is important!) groups of kids, one of which learns under A, the other under B.

This is the NHST method: compare the average scores of two (or more) groups to see which treatment is best. But what if the scores overlap: A on average is better, but some B kids do better than some A kids? Is A really better, or could this degree of difference come about just by chance?[7]

Given a suitable statistical model it is possible to use the variability of the results in the two groups to be compared and decide how likely it is that the mean difference obtained could have happened by chance. R. A. Fisher, who pioneered the method, proposed that if the chance that the observed mean difference could have occurred even if the groups were the same, the *p-value*, is less than 5

6 John Staddon, "Psychology's Other (Non-replication) Problem," *Academic Questions* 32, no. 2 (Summer 2019): 246-256.

7 John Staddon, *Scientific Method* Chapter 3, for a relatively simple account.

percent, it is reasonable to reject the *null hypothesis* (that the two groups are from the same population) in favor of the hypothesis that they are in fact different. If p < .05, the experiment worked: you can assume that one treatment is really better than the other.

But what *is* the correct p-value? Five percent is completely arbitrary, after all, yet it has become the standard.[8]

Fisher's method has been applied with little thought to much of social and biomedical science. You compare drug X against placebo Y and if the result is insignificant, you know that X is probably ineffective; but you do not know what *is* effective. In experiments like Bem's ESP study or the very many social psychology experiments looking at hypothesized effects such as "implicit bias" or social attitudes, the cost of finding an effects when there is none seems to be negligible. But the cost is in fact very high, for reasons that Ritchie points out. Subsequent studies will be based on the error and will propagate it. The result is often a cross-cited body of credible error which, if it happens to agree with existing prejudices, proliferates and corrupts the body of science. In response to this, one group of scientists has argued that the statistical significance should never be used as a deciding factor;[9] while others, more generous, simply propose setting a much higher standard, say p < .001 vs. p < .05.

The NHST method is prone to other errors, including "p-hacking." Suppose, for example, that an epidemiologist is interested in the causes of depression. He thinks that, say, "screen time" on mobile devices is the problem: too much screen time causes depression. He can't do an experiment; he can't hire a bunch of people and force them to spend X or Y amount of time on their phones. But he can measure (however inadequately) the amount of time people spend on their phones and correlate that with the amount of depression they report, while controlling for other factors.

Suppose this researcher finds that in fact depression is only weakly correlated with screen time: should he give up? Not if he is like Brian Wansink, a discredited Cornell University food researcher.[10] Confronted with a similar situation, a study that failed to confirm his hypothesis, Wansink

8 p <= .05 is popular possibly because it allows a sufficient number of accidental positives, so that with enough effort a publishable result can be achieved, even if every experiment is really null.

9 Comment, "Scientists Rise Up Against Statistical Significance," *Nature*, March 20, 2019.

10 See also *Scientific Method* and John Staddon, "Peer Review: the Publication Game and 'the Natural Selection of Bad Science'," James G. Martin Center for Academic Renewal, February 2, 2018.

famously commented to a graduate student: "There's got to be something here we can salvage." Salvaging, for our epidemiologist, takes the form of looking for correlations other than the one with which he began. Suppose he finds that although screen time doesn't work, income is significantly correlated with depression. What he should do, if he believes the (significant) correlation is not accidental, is use the correlation not as an excuse to publish, but as a new hypothesis to be independently tested. But in fact what Wansink, and many others like him, do is to write up the study as if the income-depression link *was* the hypothesis with which they began—"p-hacking." If he is willing to do that, our duplicitous epidemiologist will also likely slip in the observation—to journalists if not journal editors—that low income causes depression, a completely unwarranted conclusion when he has done no experiment but only measured correlations.

What should be done to correct these malpractices? Here I differ slightly with author Ritchie, who suggests, among other things, that scientists should abandon "small studies," which are likely by chance to show spuriously large effects, in favor of large ones, with many subjects, simply because effects found in a large sample are more likely to be replicable. There are problems with this solution, however. True, a large sample has more "power" and hence allows for a more stringent p-criterion: .01 instead of .05, say. But in fact, researchers trying to publish large-sample studies are very happy with the 5 percent criterion when a more stringent criterion might make the work unpublishable.

A large sample and a lax p value has a bad feature: it allows weak effects to achieve significance and be published—and possibly approved by the FDA. (So we may wind up with an approved drug that improves patient outcomes by 10 percent, rather than one that is at 80 percent, say.) In this way drug companies may get to market a drug that is little more effective than the one it supersedes but is a lot more costly for the patient.

So, don't publish a "p-hacked" result and do worry about effect size as well as statistical significance. What then to do about a failed study? Ritchie points out that *null results matter too, so* why not publish them? Publication was certainly warranted in the case of the failed Bem replications, but is it always? Science is an evolutionary, trial-and-error process. There are many more ways to be wrong than right. I suspect that there are simply too many failed studies to publish. There are alternatives to the standard hard-copy publication

route, which in any case has its own problems, many of which Ritchie discusses.[11] But in general the function of a failed study is to guide the researcher's future work—at least that's how it has worked in the past. Scientists from Humphry Davy and Marie Curie to B. F. Skinner tried out many ideas and techniques before coming up with their breakthroughs. None of their early missteps were separately published. The proper course after a failed study or even a "significant" study which nevertheless has exceptions is to keep trying until you find the reasons for essentially all the exceptions.

I end this section on a note of profundity: there is no "science algorithm," no gold standard scientific method.

Incentives, the Real Problem

A solution to the p-hacking problem that has become quite popular also shows the problem at the heart of bad science: "From 2005, the International Committee of Medical Journal Editors, recognizing the massive problem of publication bias,[12] ruled that all human clinical trials should be publicly registered before they take place—otherwise they wouldn't be allowed to be published in most top medical journals." The idea is that the hypothesis to be tested should be made public before the study is begun, so it can be compared with the hypothesis tested in the final manuscript—to see that no p-hacking has occurred.

This sounds like a good solution and registration has been widely adopted.[13] But it seems to presume that scientists must be legally enjoined to do their math correctly and describe their procedures accurately. External regulation of the details of scientific practice is a potential death blow to the spontaneous creativity that is the essence of great science. That regulation is now thought necessary points to a rot at its heart.

The rot has to do with why people become scientists and the conditions under which they work. In past times, science was often a vocation, done for love of the subject, not to make a living. Not that ambition played no part: Isaac Newton fought bitterly for priority with rivals like Leibniz and Hooke. Even cautious and retiring Charles Darwin was devastated at the thought that Alfred Russel

11 See John Staddon, "How Is Science Judged? How Useful Is Peer Review?," James G. Martin Center for Academic Renewal, January 31, 2018.

12 The reference here is to the so-called "file-drawer problem," where failed experiments are never reported, giving published positives unwarranted credibility.

13 Chris Chambers, *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice* (Princeton, N.J.: Princeton University Press, 2017).

Wallace's paper on natural selection might scoop him. But now science has become a career for most scientists. The ways they are rewarded and their research is supported have promoted much of the bad behavior that Stuart Ritchie so ably documents.

Science has become a career rather than a vocation, and remuneration must be based on measurable proxies. But proxies for research excellence have numerous problems. Citations are an obvious example: a paper that makes an error may gain more citations than the paper that corrects it. An excellent paper in a small field will usually get fewer citations than a mediocre paper in a large one. And, as Ritchie points out, proxies can be "gamed." If number of publications is important, career-driven scientists will turn to the "LPU strategy,"[14] splitting their product into the largest number of separately publishable packages. If shared authorship counts, the number of multi-author papers will increase. Ritchie describes the travails of famous psychologist Robert Sternberg, onetime president of the American Psychological Association, who had to step down from a prestigious editorship after being criticized for self-citation, self-plagiarism, and other practices aimed at increasing his publication and citation counts.

Finally there is the number of scientists. Science as a profession has grown exponentially over the past one-hundred years or so; ninety percent of all the scientists that have ever lived are alive today. The question is: are there enough solvable scientific problems available to keep them all usefully occupied? This problem is rarely mentioned, but does science have an "endless frontier?" Believers will point out that at the end of the nineteenth century physicist Lord Kelvin is reported to have said (though there is some dispute) "[t]here is nothing new to be discovered in physics now. All that remains is more and more precise measurement." He was wrong, of course: quantum mechanics and relativity followed a few years after.

But will Kelvin always be wrong, especially about the social sciences: just how many solvable questions are available at any time? The "softer" social sciences lack a firm theoretical structure, so that a bright researcher can always come up with a new term and purport to test for it. Current examples are things like self-esteem, white fragility, implicit bias, stereotype threat. The point is that as the

14 "Least Publishable Unit," by faux-science analogy with the "British Thermal Unit."

proper motivation for science—curiosity and the desire to understand nature—is overshadowed by ideology and the careerist incentives Ritchie describes, there is little to stop the steady decay of the sciences into tools of activism.

Stuart Ritchie has written a thoughtful, well-researched and surprisingly readable book on a difficult but genuinely important topic. Science, and the freedom of inquiry on which it depends, is at the heart of Western civilization. It is perhaps no coincidence that the very words "Western civilization" are now taboo on many campuses. *Science Fictions* can help us understand how corrupt science has become. To cure this corruption we must also understand the social forces that have brought it about. Perhaps that will be a topic for Ritchie's next book.