

How to rein in the horsemen of irreproducibility

The changes I see in how science is done

Dorothy Bishop describes how four threats to reproducibility, recognized but unaddressed for decades, might finally be brought under control

More than four decades in to my scientific career, I find myself a bit of an outlier among my academic contemporaries: I'm strongly identified with the movement to make the practice of science more robust. It's not that my colleagues are unconcerned about doing science well; it's just that many of them don't recognise that there are serious problems with current practices. I, in contrast, think that within two decades we will look back on the past 60 years, particularly in biomedical science, and marvel at how much time and money has been wasted on flawed research.

How can that be? We know how to formulate and test hypotheses with rigorously controlled experiments; we account for unwanted variation with appropriate statistical techniques. We appreciate the need to document and replicate observations. Yet the literature is strewn with papers that do not follow these agreed-on precepts of scientific practice.

Many researchers persist in doing research in a way virtually guaranteed not to deliver meaningful results. They ride, oblivious as zombies, with what I refer to as the four horsemen of the reproducibility Apocalypse: publication bias, low statistical power, p-hacking and HARKing (hypothesising after results are known). My generation and the one before did little to rein them in.

Back in 1975, psychologist Tony Greenwald noted that science is prejudiced against null hypotheses: We even refer to sound work supporting a null hypothesis as "failed experiments." That prejudice results in publication bias: researchers are less likely to write up studies showing no effect; editors less likely to accept them, and so no one can learn from them. The term "file drawer problem" referring to work that goes unpublished, thereby encouraging redundant studies, dates back to my own graduate school days, though people did not take it seriously. That began to change for two reasons. First, clinicians realized that publication bias harms patients: if there are twenty studies of a drug and only the one that shows a benefit is published, we see a distorted view of drug efficacy. Second, the growing use of meta-analyses, which combine results across studies, made clear that these give misleading results if unpublished work is left out.

Low power followed a similar trajectory. Although I had statistics courses as an undergraduate in the early 1970s, I learned nothing about statistical power. That left me and my contemporaries prey to a common practice. Quite simply, if a study has a small sample size, and the effect of an experimental manipulation is small, then odds are you won't detect the effect, even if it is there. It is wasteful to conduct studies that are underpowered, but researchers have stubbornly persisted in pursuing them, treating statisticians who point it out as killjoys. Jacob Cohen published an entire book on the topic in 1977. Ten years later, biostatistician Robert Newcombe wrote that "Small studies continue to be carried out with little more than a blind hope of showing the desired effect." In fields such as clinical trials and

genetics, funders have forced great improvements in working practices by insisting that studies be adequately powered, but other fields have yet to catch up.

I stumbled upon p-hacking before knowing the term. Back in the 1980s I reviewed the literature on brain lateralization and developmental disorders, and I noticed that, while many studies claimed links between handedness and dyslexia, they tended to change the definition of atypical handedness from one study to the next – even among different studies from the same research group. I published a sarcastic note, including a simulation showing how easy it was to find an effect if you explored the data after collecting results. Subsequently I noticed similar phenomena in other fields: the practice of trying many analyses and only reporting the 'significant' results. This practice, or "p-hacking" had become endemic in most branches of science that tested null hypotheses, and few researchers realised how seriously it could distort findings. That started to change only in 2011 with an elegant, comic paper showing how to craft analyses to prove listening to the Beatles made undergraduates younger. "Undisclosed flexibility," the authors explained, "allows presenting anything as significant."

The term HARKing (for hypothesizing after the results are known) was coined in 1998. Like p-hacking, it is so widely done that researchers assume it is good practice. Researchers look at the data, pluck out a finding that looks exciting, and write a paper to 'tell a story' around this result. Of course, researchers need to be free to explore their data for unexpected findings, but p-values are meaningless when taken out of context. To use a gambling analogy, if you pray for number 8 to come up in roulette and it does, that's unusual enough to encourage belief in a benevolent deity. However, if have to wait for 36 spins of the wheel before it comes up, you may become more sceptical.

This overview of these problems, older than most junior faculty, might seem gloomy, but new forces are reining in these horsemen.

First the field of meta-science has been blossoming, and with it documentation and awareness of the problems. We can no longer dismiss concerns as purely theoretical. Second, social media enables criticisms to be raised and explored soon after publication. Third, more journals are adopting the Registered Report format, where editors evaluate the experimental question and study design before results are collected – a strategy that thwarts publication bias, p-hacking and Harking. Finally, and most importantly, those who fund research have become more concerned, and more strict: with requirements to make data and scripts open and describe methods fully.

I anticipate we are approaching a tipping point where the four horsemen might be persuaded to dismount.

Dorothy Bishop is an experimental psychologist at the University of Oxford, UK

dorothy.bishop@psy.ox.ac.uk