World view

By Stuart Buck

Beware performative reproducibility

Well-meant changes to improve science could become empty gestures unless underlying values change.

Almost a decade ago, at Arnold Ventures — a US$2-billion philanthropic organization in Houston, Texas — we realized that using evidence to direct our giving required having more confidence in the evidence itself. As vice-president of research, I found myself deep in efforts to improve science, dispersing more than $60 million in grants to make sure researchers can build on others' results. I was part of the discussions that led to widely adopted guidelines promoting transparency and openness, the clinical-trial repository Vivli and the launch of the Center for Open Science, a non-profit organization in Charlottesville, Virginia.

I've seen much positive change since then. But sometimes I worry that we might end up with the worst of all worlds: the pretence of reproducibility without the reality.

In 2012, very few people had even heard of preregistration, the anti-bias practice of specifying, in writing, intended analyses and hypotheses at the start of an experiment. Doing so was a requirement for our grantees.

These days, it seems all scientists know what preregistration is. Most agree that it can help to reduce publication bias and *P*-hacking — when data are tweaked to produce significant *P* values. Major professional societies now endorse the practice: the American Economic Association's registry lists more than 4,700 studies, and the American Psychological Association has created a set of 'Preregistration Standards for Quantitative Research in Psychology'. Indeed, there are some 75,000 registered research projects on the Center for Open Science's Open Science Framework repository.

A similar story can be told about data-sharing through Zenodo from CERN, Europe's particle-physics laboratory near Geneva, Switzerland; Figshare from London-based analytics firm Digital Science; many National Institutes of Health repositories; and more. Although still far from routine in many disciplines, the rate at which scholarly articles share their underlying data is growing: one study put it at increasing from around 0% in 2000 to almost 20% in 2018 (S. Serghiou *et al. PLoS Biol.* **19**, e3001107; 2021).

But robust, sustainable change depends on whether underlying cultural values have altered, not just surface signals. If they haven't, then open-science practices can become just another hoop to jump through, a form of virtue signalling or a smokescreen.

I've seen it happen. At a conference a few months before the pandemic, a scholar told me how, in his department, everyone wrote lengthy pre-analysis plans that would, in theory, constrain *P*-hacking. In practice, he admitted, researchers could give cherry picking free rein, counting on the fact that no one has the time or patience to read a 100-page pre-analysis plan and compare it with the later publication.

> "Open-science practices can become just another hoop to jump through, a form of virtue signalling."

More-systematic evidence comes from the COMPare Project led by Ben Goldacre at the University of Oxford, UK, an effort my department funded. That team reviewed publications from 67 clinical trials in top medical journals, and compared them against original descriptions. Only 9 matched. Of the others, 354 preregistered outcomes went unreported; another 357 outcomes were "silently added" (B. Goldacre *et al. Trials* **20**, 118; 2019).

Meanwhile, many preregistrations are too vague. In one study, reviewers were asked to count the number of hypotheses in 106 preregistrations. They agreed only 14% of the time (M. Bakker *et al. PLoS Biol.* **18**, e3000937; 2020).

What about data sharing? The FAIR principles stipulate that shared data should be 'findable, accessible, interoperable and reusable'. A 2020 analysis across 15 psychology journals concluded that the majority of data sets "were neither complete nor re-usable" (J. N. Towse *et al. Behav. Res.* https://doi.org/gkzk; 2020).

I worry that, by adopting the trappings of reproducibility, poor-quality work can look as if it has engaged in best practices. The problem is that sloppy work is driven by a scientific culture that overemphasizes exciting findings. When funders and journals reward showy claims at the expense of rigorous methods and reproducible results, reforms to change practice could become self-defeating. Helpful new practices, rules and policies are transformed into meaningless formalities on the way to continuing to grab headlines at any cost.

That said, I do see values shifting. In the first few years that Arnold Ventures began supporting these efforts, some researchers reacted with open hostility, using phrases such as "replication police". Now such criticism is rare (at least in public). And in some communities, researchers now prioritize work that others can build on. For example, organizations such as the Society for Improving Psychological Science embody a groundswell of energy and idealism from mostly younger researchers.

Still, what really matters is whether scientists feel empowered and rewarded for doing robust work, publishing null results and following the data. Idealism from early-career scientists must be matched by strong signals from senior leaders and institutions that it is possible to be hired and get tenure while engaging in best practices. A hopeful sign is that some university job advertisements now ask about an applicant's commitment to open-science practices.

That sort of cultural change is where the real challenge lies.

Stuart Buck was vice-president of research at Arnold Ventures in Houston, Texas, from 2012 to 2021. He is now a consultant on rigorous research practices.
stuartbuck@gmail.com