NAOMI A. P. STARK

No reproducibility without preproducibility

Instead of arguing about whether results hold up, let's push to provide enough information for others to repeat the experiments, says **Philip Stark**.

From time to time over the past few years, I've politely refused requests to referee an article on the grounds that it lacks enough information for me to check the work. This can be a hard thing to explain.

Our lack of a precise vocabulary — in particular the fact that we don't have a word for 'you didn't tell me what you did in sufficient detail for me to check it' — contributes to the crisis of scientific reproducibility. In computational science, 'reproducible' often means that enough information is provided to allow a dedicated reader to repeat the calculations in the paper for herself. In biomedical disciplines, 'reproducible' often means that a different lab, starting the experiment from scratch, would get roughly the same experimental result.

In 1992, philosopher Karl Popper wrote: "Science may be described as the art of systematic oversimplification — the art of discerning what we may with advantage omit." What may be omitted depends on the discipline. Results that generalize to all universes (or perhaps do not even require a universe) are part of mathematics. Results that generalize to our Universe belong to physics. Results that generalize to all life on Earth underpin molecular biology. Results that generalize to all mice are murine biology. And results that hold only for a particular mouse in a particular lab in a particular experiment are arguably not science.

SCIENCE SHOULD BE 'SHOW ME', NOT 'TRUST ME'.

Communicating a scientific result requires enumerating, recording and reporting those things that cannot with advantage be omitted. This harks back to the idea of science as a way to build knowledge through careful experimentation. Ushering in the Enlightenment era in the late seventeenth century, chemist Robert Boyle put forth his controversial idea of a vacuum and tasked himself with providing descriptions of his work sufficient "that the person I addressed them to might, without mistake, and with as little trouble as possible, be able to repeat such unusual experiments".

Much modern scientific communication falls short of this standard. Most papers fail to report many aspects of the experiment and analysis that we may not with advantage omit — things that are crucial to understanding the result and its limitations, and to repeating the work. We have no common language to describe this shortcoming. I've been in conferences where scientists argued about whether work was reproducible, replicable, repeatable, generalizable and other '-bles', and clearly meant quite different things by identical terms. Contradictory meanings across disciplines are deeply entrenched.

The lack of standard terminology means that we do not clearly distinguish between situations in which there is not enough information to attempt repetition, and those in which attempts do not yield substantially the same outcome. To reduce confusion, I propose an intuitive, unambiguous neologism: 'preproducibility'. An experiment or analysis is preproducible if it has been described in adequate detail for others to undertake it. Preproducibility is a prerequisite for reproducibility, and the idea makes sense across disciplines.

The distinction between a preproducible scientific report and current common practice is like the difference between a partial list of ingredients and a recipe. To bake a good loaf of bread, it isn't enough to know that it contains flour. It isn't even enough to know that it contains flour, water, salt and yeast. The brand of flour might be omitted from the recipe with advantage, as might the day of the week on which the loaf was baked. But the ratio of ingredients, the operations, their timing and the temperature of the oven cannot.

Given preproducibility — a 'scientific recipe' — we can attempt to make a similar loaf of scientific bread. If we follow the recipe but do not get the same result, either the result is sensitive to small details that cannot be controlled, the result is incorrect or the recipe was not precise enough (things were omitted to disadvantage).

Depending on the discipline, preproducibility might require information about materials (including organisms and their care), instruments and procedures; experimental design; raw data at the instrument level; algorithms used to process the raw data; computational tools used in analyses, including any parameter settings or ad hoc choices; code, processed data and software build environments; or analyses that were tried and abandoned.

Peer review is hamstrung by lack of preproducibility: referees and editors cannot provide serious quality control unless they are given enough information. Preproducibility will bring us closer to the ideals of the Enlightenment, providing crucial evidence about whether a reported result is correct and about how far the result can be generalized.

Science should be 'show me', not 'trust me'; it should be 'help me if you can', not 'catch me if you can'. If I publish an advertisement for my work (that is, a paper long on results but short on methods) and it's wrong, that makes me untrustworthy. If I say: "here's my work" and it's wrong, I might have erred, but at least I am honest. If you and I get different results, preproducibility can help us to identify why — and the answer might be fascinating.

Just as I have pledged not to review papers that are not preproducible, I have also pledged not to submit papers without providing the software I used, and — to the extent permitted by law and ethics — the underlying data. I urge you to do the same. The commitment that Boyle made to the scientific community is even more crucial today. ■

Philip B. Stark *is a professor of statistics who specializes in inference at the University of California, Berkeley.*
e-mail: stark@stat.berkeley.edu