NATURE | NEWS

Muddled meanings hamper efforts to fix reproducibility crisis

Researchers tease out different definitions of a crucial scientific term.

Monya Baker

14 June 2016

A semantic confusion is clouding one of the most talked-about issues in research. Scientists agree that there is a crisis in reproducibility, but they can't agree on what 'reproducibility' means.

The muddle is hampering communication about the problem and efforts to address it, a meeting last week on improving the reproducibility of preclinical research was told.

Most scientists — at least, those in biomedical research — have the idea that reproducible findings are those that give generally consistent results across slight variations in experimental set-up, says Ferric Fang, a microbiologist at the University of Washington in Seattle. "Reproduction is taking the idea of a scientific project and showing that it is robust enough to survive various sorts of analysis," he says. That is, that it supports an expectation, for example that 'reproducible' preclinical results are those worth taking forward to clinical trials.

"I don't think we can define reproducibility across the board"

But Fang adds that other definitions of the term are in common use. A second one is much narrower: a finding is reproducible if another researcher gets the same results when doing exactly the same experiment. On this interpretation, a fragile experiment that works under certain conditions in the laboratory, but not in other contexts, is still 'reproducible'. And a third definition holds that a reproducible experiment is merely one that has been published with a sufficiently complete description — such as detailed methods — for another scientist to repeat it.

All these definitions point to the various problems that plague research. Scientists don't want experiments that are poorly documented or unreliable, or that don't give similar findings when the methods are slightly tweaked. But all of these issues have at times been framed as issues of reproducibility. "Reproducibility is shorthand for a lot of problems," Jon Lorsch, head of the National Institute of General Medical Sciences, told attendees at the meeting, which was held in Bethesda, Maryland.

Without a shared understanding of the term, it can be unclear how scientists should respond when told that someone is "unable to reproduce" results in their paper, adds Ulrich Dirnagl, a stroke researcher at the Charité Medical University in Berlin. Challenges to research should be more clearly explained, he says.

Expanded terms

Instead of advocating for a common definition, several scientific leaders are calling for an expanded set of terms. Earlier this month, researchers at the Meta-Research Innovation Center at Stanford in California proposed three[1]: methods reproducibility, results reproducibility and inferential reproducibility, mapping roughly onto the three concepts described by Fang.

But the term can be split according to other kinds of distinctions. Victoria Stodden, a data scientist at the University of Illinois at Urbana-Champaign, makes the distinction between 'empirical' reproducibility (supplying all the details necessary for someone to physically repeat and verify an experiment) and 'computational' and 'statistical' reproducibility, which refer to the resources needed to redo computational and analytical findings. Achieving each type of reproducibility calls for different remedies, she says.

Last year, a white paper by the American Society for Cell Biology in Bethesda dismissed reproducibility as a catch-all term, and introduced a four-tier definition instead. According to this paper, "analytic replication" refers to attempts to reproduce results by reanalysing original data; "direct replication" refers to efforts to use the same conditions, materials and methods as an original experiment; "systematic replication" describes efforts to produce the same findings using different experimental conditions (such as trying an experiment in a different cell line or mouse strain), and "conceptual replication", which refers to attempts to demonstrate the general validity of a concept, perhaps even using different organisms.

Agreeing on ways to assess reproducibility can be fraught with complications — even within a research field. Last year, the

Reproducibility Project: Psychology, which conducted 100 replication studies to assess psychology publications, used five indicators of whether a study had been successfully replicated[2]. Earlier this year, a paper that aimed to distil best practices for neuroimaging outlined 10 levels of reproducibility in such experiments across three categories, called 'measurement stability', 'analytical stability' and 'generalizability'. They differed according to whether, for example, an attempted repeat experiment used the same scanners, analysis methods, subject population, stimulus type and so on across 11 variables.

Ultimately, each discipline may need to come to its own definition, says Lee Ellis, a surgical oncologist at the University of Texas MD Anderson Cancer Center in Houston. "I don't think we can define reproducibility across the board," he says.

Arguments over such distinctions are not just fruitless wordplay, says Dirnagl. An appreciation of the nuances of reproducibility could help researchers to communicate when they can't reach common ground on apparently differing findings. Some of science's most enlightening results arise when an effect is partially reproducible — seen under some conditions but not others. "Science advances through differential reproduction," he says.

Nature | doi:10.1038/nature.2016.20076

References

1. Goodman, S. N. *et al Sci. Trans. Med.* **8** 341 (2016).
2. Open Science Collaboration. Science http://dx.doi.org/10.1126/science.aac4716 (2015).