NATURE | NEWS

Psychology's reproducibility problem is exaggerated – say psychologists

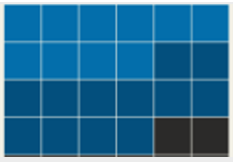
Reanalysis of last year's enormous replication study argues that there is no need to be so pessimistic.

Monya Baker

03 March 2016

Is psychology facing a 'replication crisis'? Last year, a crowdsourced effort that was able to validate fewer than half of 98 published findings[1] rang alarm bells about the reliability of psychology papers. Now a team of psychologists has reassessed the study and say that it provides no evidence for a crisis.

"Our analysis completely invalidates the pessimistic conclusions that many have drawn from this landmark study," says Daniel Gilbert, a psychologist at Harvard University in Cambridge, Massachusetts, and a co-author of the reanalysis, published on 2 March in *Science*[2].



Over half of psychology studies fail reproducibility test

But a response[3] in the same issue of *Science* counters that the reanalysis itself depends on selective assumptions. And others say that psychology still urgently needs to improve its research practices.

Statistical criticism

In August 2015, a team of 270 researchers reported the largest ever single-study audit of the scientific literature. Led by Brian Nosek, executive director of the Center for Open Science in Charlottesville, Virginia, the Reproducibility Project attempted to replicate studies in 100 psychology papers. (It ended up with 100 replication attempts for 98 papers because of problems assigning teams to two papers.) According to one of several measures of reproducibility, just 36% could be confirmed; by another statistical measure, 47% could[1]. Either way, the results looked worryingly feeble.

> **"Both optimistic and pessimistic conclusions about reproducibility are possible, and neither are yet warranted"**

Not so fast, says Gilbert. Because of the way the Reproducibility Project was conducted, its results say little about the overall reliability of the psychology papers it tried to validate, he argues. "The number of studies that actually did fail to replicate is about the number you would expect to fail to replicate by chance alone — even if all the original studies had shown true effects."

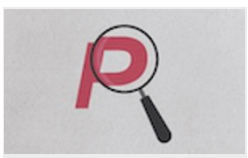
Reproducibility: A tragedy of errors

Gilbert's team tries to calculate an expected replication rate based on the Reproduciblity Project's protocol. Some replication attempts weren't faithful copies of the original paper, for instance — making a replication less likely, Gilbert says. And each replication study was attempted just once, giving the project limited statistical power to confirm the original findings. To make his point, Gilbert refers to an earlier work by Nosek, called the Many Labs Replication Project, in which 36 separate laboratories tried to replicate 13 findings. That project validated 10 findings, but looking at any one lab's replication effort might have suggested failure.

Overall, Gilbert's team concludes, the Reproducibility Project failed to account for important sources of uncertainty in its analysis. "While no study is perfect, it is clear that the reproducibility of psychological studies is not as bad as the original *Science* article characterized," says Kosuke Imai, a statistician at Princeton University in New Jersey.

Neither good nor bad

But the reanalysis does not show that psychology's reproducibility rate is high, says Uri Simonsohn, a social psychologist at the University of Pennsylvania in Philadelphia. He notes that other statistical reassessments of the project (one of which was published in *PLoS ONE* last week[4], and another of which Simonsohn has reported on the blog Data Colada) both suggest that about one-third of the replications are inconclusive, in that they neither strongly confirm nor refute the original results.



In general, replication efforts are likely to be more-reliable guides to the existence (and magnitude) of effects in psychology experiments than are the original studies, says Andrew Gelman, a statistician at Columbia University in New York. That's in part because what is published in the original studies tends to be the statistical 'flukes' that are left standing after the researchers have cast around to find publishable, positive

results. In contrast, for replication projects analysis plans are put in place before a study begins.

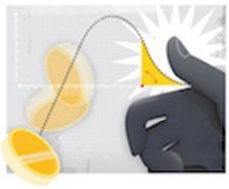
Statistics: *P* values are just the tip of the iceberg

Nosek says that he is glad to see other researchers poring over the data. But he and his co-authors respond in *Science* that Gilbert's "optimistic assessment is limited by statistical misconceptions". Although some of the Reproducibility Project's replications didn't faithfully copy the methods of the papers they were testing, in many cases the original authors endorsed the methods used, Nosek says. The Many Labs project is a misleading comparison, he and his co-authors say. And the Reproducibility Project used five measures of reproducibility that coalesced around similar values, which together suggest that most effects -- even if they do exist -- are likely to be much smaller than the original results suggested.

Nosek tells *Nature* that the Gilbert team has crafted an exploratory hypothesis that is very optimistic but cannot be seen as definitive. "Both optimistic and pessimistic conclusions about reproducibility are possible, and neither are yet warranted," he and his co-authors conclude in their response.

Economics versus psychology

Another paper published today in *Science* reports results from a project to replicate economics studies[5]. It found that at least 11 of 18 studies could be reproduced, increasing to 14 when using different criteria to assess reproducibility. But Nosek, who was not involved in the study, says that it is hard to conclude from this that economics has a higher replication rate than psychology. Not only was the number of studies replicated small, he says, but the studies focused on simple relationships.



Scientific method: Statistical errors

Nonetheless, the size of effects found in the replication and original studies are closer in the economics study than in the psychology one, says study author, Colin Camerer, a behavioural economist at the California Institute of Technology in Pasadena. "It is like a grade of B+ for psychology versus A– for economics," he says." There is room for improvement in both fields."

Overall, knowing the exact replication rate for any discipline is not essential for knowing what needs to happen next, says Steve Lindsay, a psychologist at the University of Victoria in Canada and interim editor of the journal *Psychological Science*. "We have a lot of reasons to believe that a lot of psychologists have for a long time tended to systematically exaggerate the effects of what they publish," he says; the real urgency is to improve bad practices.

Nature | doi:10.1038/nature.2016.19498

References

1. Open *Science* Collaboration *Science* **349**, aac4716 (2015).
2. Gilbert, D. T., King, G., Pettigrew, S. & Wilson, T. D. *Science* **351**, 1037 (2016).
3. Anderson, C. J. et al. *Science* **351**, 1037b (2016).
4. Etz, A. & Vanderkerchkhove, J. *PLoS ONE* http://dx.doi.org/10.1371/journal.pone.0149794 (2016)
5. Camerer, C. F. *et al. Science* http://dx.doi.org/10.1126/science.aaf0918 (2016).