First results from psychology's largest reproducibility test

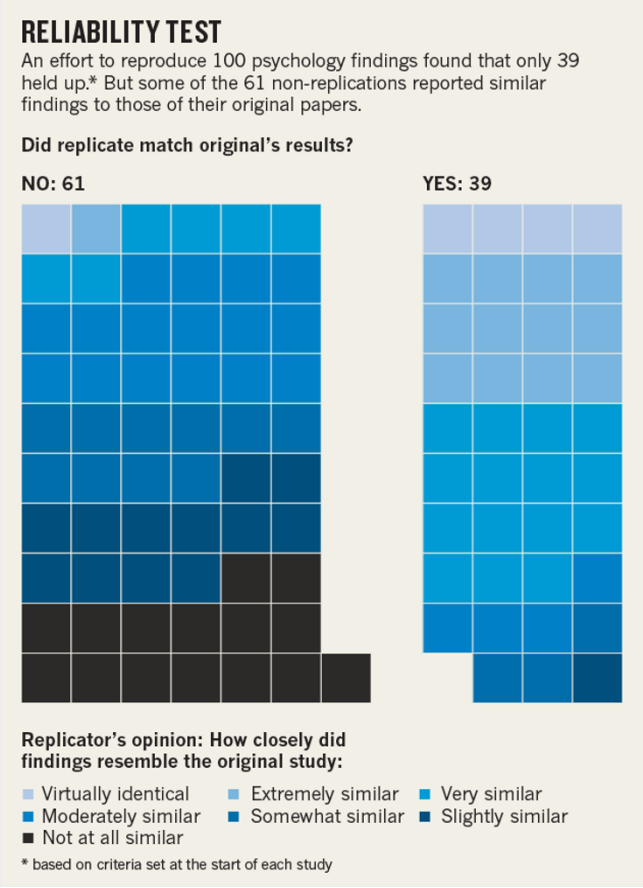
Crowd-sourced effort raises nuanced questions about what counts as replication.

Monya Baker

30 April 2015

An ambitious effort to replicate 100 research findings in psychology ended last week — and the data look worrying. Results posted online on 24 April, which have not yet been peer-reviewed, suggest that key findings from only 39 of the published studies could be reproduced.

But the situation is more nuanced than the top-line numbers suggest (See graphic, 'Reliability test'). Of the 61 non-replicated studies, scientists classed 24 as producing findings at least "moderately similar" to those of the original experiments, even though they did not meet pre-established criteria, such as statistical significance, that would count as a successful replication.



The results should convince everyone that psychology has a replicability problem, says Hal Pashler, a cognitive psychologist at the University of California, San Diego, and an author of one of the papers whose findings were successfully repeated. "A lot of working scientists assume that if it's published, it's right," he says. "This makes it hard to dismiss that there are still a lot of false positives in the literature."

But Daniele Fanelli, who studies bias and scientific misconduct at Stanford University in California, says the results suggest that the reproducibility of findings in psychology does not necessarily lag behind that in other sciences. There is plenty of room for improvement, he adds, but earlier studies have suggested that reproducibility rates in cancer biology and drug discovery could be even lower[1,2]. "From my expectations, these are not bad at all," Fanelli says. "Though I have spoken to psychologists who are quite disappointed."

Collaborative checking

The project, known as the "Reproducibility Project: Psychology", is the largest of a wave of collaborative attempts to replicate previously published work, following reports of fraud and faulty statistical analysis as well as heated arguments about whether classic psychology studies were robust. One such effort, the 'Many Labs' project, successfully reproduced the findings of 10 of 13 well-known studies[3].

In the latest attempt, launched in 2011, more than 250 scientists around the world tried to reproduce key findings from a sample of articles published in 2008 in three leading psychology journals. Fanelli says that no other project has attempted such a systematic assessment. "It's the best data you can have on reproducibility in psychology."

The replication teams followed a highly structured protocol in which experiments, statistical tests, study sizes and replication criteria were set before the study began[4]. The project was led by the Center for Open Science in Charlottesville, Virginia, which made all protocols openly available online, along with the results of each study as they came in. On 24 April, it posted an aggregated spreadsheet of the final conclusions.

When *Nature*'s news team contacted Brian Nosek, a psychologist at the Center for Open Science who leads the project, he declined to discuss the work because the results had been submitted to *Science,* and, in a subsequent email, added that data files required some

clean-up. The centre subsequently took access to the spreadsheet and some analysis files away from public view. It states they are being made “temporarily private while the manuscript is under review at *Science*”.

Not black and white

Nosek has written that there is no simple, standard way to answer whether or not a study replicates[4], and the spreadsheet shows that researchers trying to replicate studies were asked in a number of ways whether their experiment’s results matched the original results. On the basis of criteria set at the start of each study, there were 39 replications and 61 non-replications. But replicators were also asked how similar — in their opinion — their overall findings, and the key effect measured in their replication, were to the original study. This nuanced approach revealed that 24 formal non-replicates had at least moderately similar findings.

Rigid, all-or-nothing categories are not useful in such situations, says Greg Hajcak, a clinical psychologist at Stony Brook University in New York. Hajcak authored one study that could not be reproduced, but for which replicators said they found “extremely similar” results that did not reach statistical significance. (The study looked at how making mistakes affected people’s physiological responses.) Many sets of experiments showed markedly similar relationships in the original and the replication study, Hajcak says. “What does it mean to say it’s not a replication, if you have the exact same pattern of results?”

Teams did not always replicate experiments exactly. One non-replicate was an Israeli study about factors that promote personal reconciliation. The original study posed a scenario involving a colleague who had to miss work because of maternity leave or military duty. But vignettes prepared for a replication study in the United States involved a wedding and a honeymoon. Another non-replicate investigating response times to fear-related stimuli in children used an iPad instead of a larger touch-screen monitor to display images, and found that children sometimes interacted less precisely with the smaller screen.

The online reports - and *Nature*'s interviews - suggest that for the most part, original authors and replicators had very collegial interactions. In some cases, the original authors commended their replicators for streamlining experimental techniques, or suggested future, collaborative work.

Some psychologists contacted by *Nature* felt that greater awareness of reproducibility problems was making the field more robust. “I’m already in the practice of building in replications — it has become necessary to get published,” says Jamie Campbell, a psychologist at Saskatchewan University in Saskatoon, Canada.

But more work must be done, says Pashler. “We need to have some changes in the field so that replication becomes more widespread and more represented in journals.”

Nature | doi:10.1038/nature.2015.17433

References

1. Prinz, F., Schlange, T. & Asadullah, K. *Nature Rev. Drug Discov.* **10**, 712 (2011).
2. Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012).
3. Klein, R. A. *et al. Social Psychol.* **45**, 142–152 (2014).
4. Open Science Collaboration. *Perspect. Psychol. Sci.* **7,** 657–660 (2012).