NATURE | NEWS

Over half of psychology studies fail reproducibility test

Largest replication study to date casts doubt on many published positive results.

Monya Baker

27 August 2015

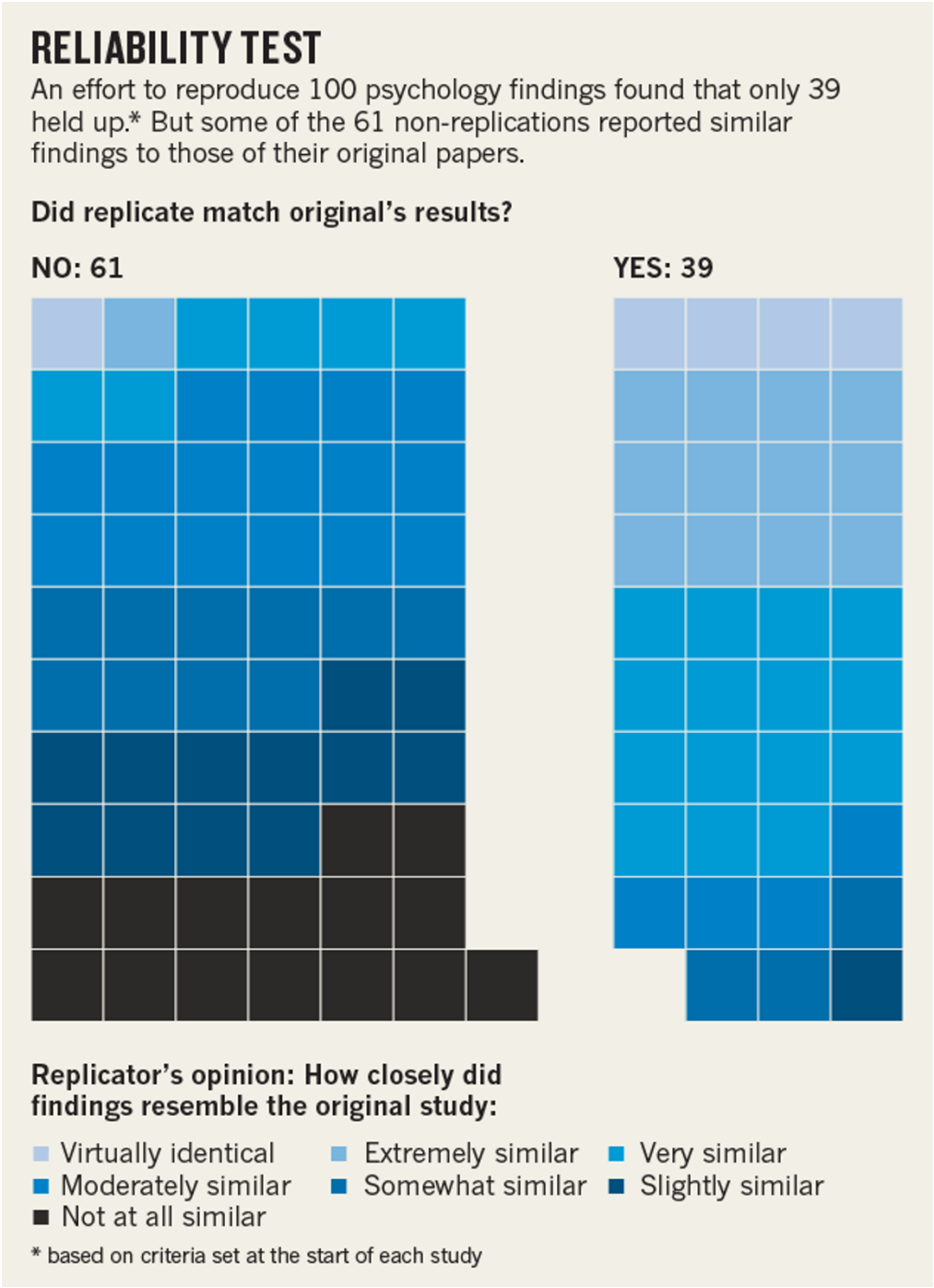
Brian Nosek's team set out to replicate scores of studies.

Don’t trust everything you read in the psychology literature. In fact, two thirds of it should probably be distrusted.

In the biggest project of its kind, Brian Nosek, a social psychologist and head of the Center for Open Science in Charlottesville, Virginia, and 269 co-authors repeated work reported in 98 original papers from three psychology journals, to see if they independently came up with the same results.

The studies they took on ranged from whether expressing insecurities perpetuates them to differences in how children and adults respond to fear stimuli, to effective ways to teach arithmetic.

According to the replicators' qualitative assessments, as previously reported by *Nature*, only 39 of the 100 replication attempts were successful. (There were 100 completed replication attempts on the 98 papers, as in two cases replication efforts were duplicated by separate teams.) But whether a replication attempt is considered successful is not straightforward. Today in *Science,* the team report the multiple different measures they used to answer this question[1].



The 39% figure derives from the team's subjective assessments of success or failure (see graphic, 'Reliability test'). Another method assessed whether a statistically significant effect could be found, and produced an even bleaker result. Whereas 97% of the original studies found a significant effect, only 36% of replication studies found significant results. The team also found that the average size of the effects found in the replicated studies was only half that reported in the original studies.

There is no way of knowing whether any individual paper is true or false from this work, says Nosek. Either the original or the replication work could be flawed, or crucial differences between the two might be unappreciated. Overall, however, the project points to widespread publication of work that does not stand up to scrutiny.

Although Nosek is quick to say that most resources should be funnelled towards new research, he suggests that a mere 3% of scientific funding devoted to replication could make a big difference. The current amount, he says, is near-zero.

Replication failure

The work is part of the Reproducibility Project, launched in 2011 amid high-profile reports of fraud and faulty statistical analysis that led to an identity crisis in psychology.

John Ioannidis, an epidemiologist at Stanford University in California, says that the true replication-failure rate could exceed 80%, even higher than Nosek's study suggests. This is because the Reproducibility Project targeted work in highly respected journals, the original scientists worked closely with the replicators, and replicating teams generally opted for papers employing relatively easy methods — all things that should have made replication easier.

But, he adds, "We can really use it to improve the situation rather than just lament the situation. The mere fact that that collaboration happened at such a large scale suggests that scientists are willing to move in the direction of improving."

The work published in *Science* is different from previous papers on replication because the team actually replicated such a large swathe of experiments, says Andrew Gelman, a statistician at Columbia University in New York. In the past, some researchers dismissed indications of widespread problems because they involved small replication efforts or were based on statistical simulations.

But they will have a harder time shrugging off the latest study, says Gelman. “This is empirical evidence, not a theoretical argument. The value of this project is that hopefully people will be less confident about their claims.”

Publication bias

The point, says Nosek, is not to critique individual papers but to gauge just how much bias drives publication in psychology. For instance, boring but accurate studies may never get published, or researchers may achieve intriguing results less by documenting true effects than by hitting the statistical jackpot; finding a significant result by sheer luck or trying various analytical methods until something pans out.

Nosek believes that other scientific fields are likely to have much in common with psychology. One analysis found that only 6 of 53 high-profile papers in cancer biology could be reproduced[2] and a related reproducibility project in cancer biology is currently under way. The incentives to find results worthy of high-profile publications are very strong in all fields, and can spur people to lose objectivity. “If this occurs on a broad scale, then the published literature may be more beautiful than reality," says Nosek.

The results published today should spark a broader debate about optimal scientific practice and publishing, says Betsy Levy Paluck, a social psychologist at Princeton University in New Jersey. “It says we don't know the balance between innovation and replication.”

The fact that the study was published in a prestigious journal will encourage further scholarship, she says, and shows that now “replication is being promoted as a responsible and interesting line of enquiry”.

Nature | doi:10.1038/nature.2015.18248

References

1. Open *Science* Collaboration. *Science* http://dx.doi.org/10.1126/science.aac4716 (2015).
2. Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012)