

In Psychology And Other Social Sciences, Many Studies Fail The Reproducibility Test

[Richard Harris](#) August 27, 2018 11:09 AM ET



A researcher showed people a picture of The Thinker in an effort to study the link between analytical thinking and religious disbelief. In hindsight, the researcher called his study design "silly". The study could not be reproduced.

Peter Barritt/Getty Images

The world of social science got a rude awakening a few years ago, when researchers concluded that many studies in this area appeared to be deeply flawed. Two-thirds could not be replicated in other labs.

Some of those same researchers now report those problems still frequently

crop up, even in the most prestigious scientific journals.

But their [study, published Monday](#) in *Nature Human Behaviour*, also finds that social scientists can actually sniff out the dubious results with remarkable skill.

First, the findings. [Brian Nosek](#), a psychology researcher at the University of Virginia and the executive director of the Center for Open Science, decided to focus on social science studies published in the most prominent journals, *Science* and *Nature*.

"Some people have hypothesized that, because they're the most prominent outlets they'd have the highest rigor," Nosek says. "Others have hypothesized that the most prestigious outlets are also the ones that are most likely to select for very 'sexy' findings, and so may be actually less reproducible."

To find out, he worked with scientists around the world to see if they could reproduce the results of key experiments from 21 studies in *Science* and *Nature*, typically psychology experiments involving students as subjects. The new studies on average recruited five times as many volunteers, in order to come up with results that were less likely due to chance.

The results were better than the average of [a previous review of the psychology literature](#), but still far from perfect. Of the 21 studies, the experimenters were able to reproduce 13. And the effects they saw were on average only about half as strong as had been trumpeted in the original studies.

The remaining eight were not reproduced.

"A substantial portion of the literature is reproducible," Nosek concludes. "We are getting evidence that someone can independently replicate [these findings]. And there is a surprising number [of studies] that fail to replicate."

One of the eight studies that failed this test came from the lab of [Will Gervais](#),

when he was getting his PhD at the University of British Columbia. He and a colleague had run a series of experiments to see whether people who are more analytical are less likely to hold religious beliefs. In one test, undergraduates looked at pictures of statues.

"Half of our participants looked at a picture of the sculpture, 'The Thinker,' where here's this guy engaged in deep reflective thought," Gervais says. "And in our control condition, they'd look at the famous stature of a guy throwing a discus."

People who saw *The Thinker*, a sculpture by August Rodin, expressed more religious disbelief, [Gervais reported in *Science*](#). And given all the evidence from his lab and others, he says there's still reasonable evidence that underlying conclusion is true. But he recognizes the sculpture experiment was really quite weak.

"Our study, in hindsight, was outright silly," says Gervais, who is now an assistant professor at the University of Kentucky.

[A previous study](#) also failed to replicate his experimental findings, so the new analysis is hardly a surprise.

But what interests him the most in the new reproducibility study is that scientists had predicted that his study – along with the seven others that failed to replicate – were unlikely to stand up to the challenge.

As part of the reproducibility study, about 200 social scientists were surveyed and asked to predict which results would stand up to the re-test and which would not. Scientists filled out a survey in which they predicted the winners and losers. They also took part in a "prediction market," where they could buy or sell tokens that represented their views.

"They're taking bets with each other, against us," says [Anna Dreber](#), an economics professor at the Stockholm School of Economics, and coauthor of

the new study.

It turns out, "these researchers were very good at predicting which studies would replicate," she says. "I think that's great news for science."

These forecasts could help accelerate the process of science. If you can get panels of experts to weigh in on exciting new results, the field might be able to spend less time chasing errant results known as false positives.

"A false positive result can make other researchers, and the original researcher, spend lots of time and energy and money on results that turn out not to hold," she says. "And that's kind of wasteful for resources and inefficient, so the sooner we find out that a result doesn't hold, the better."

But if social scientists were really good at identifying flawed studies, why did the editors and peer reviewers at *Science* and *Nature* let these eight questionable studies through their review process?

"The likelihood that a finding will replicate or not is one part of what a reviewer would consider," says Nosek. "But other things might influence the decision to publish. It may be that this finding isn't likely to be true, but if it is true, it is super important, so we do want to publish it because we want to get it into the conversation."

Nosek recognizes that, even though the new studies were more rigorous than the ones they attempted to replicate, that doesn't guarantee that the old studies are wrong and the new studies are right. No single scientific study gives a definitive answer.

Forecasting could be a powerful tool in accelerating that quest for the truth.

That may not work, however, in one area where the stakes are very high: medical research, where answers can have life-or-death consequences.

Jonathan Kimmelman at McGill University, who was not involved in the new

study, says when he's asked medical researchers to make predictions about studies, the forecasts have generally flopped.

"That's probably not a skill that's widespread in medicine," he says. It's possible that the social scientists selected to make the forecasts in the latest study have deep skills in analyzing data and statistics, and their knowledge of the psychological subject matter is less important.

And forecasting is just one tool that could be used to improve the rigor of social science.

"The social-behavioral sciences are in the midst of a reformation," says Nosek. Scientists are increasingly taking steps to increase transparency, so that potential problems surface quickly. Scientists are increasingly announcing in advance the hypothesis they are testing; they are making their data and computer code available so their peers can evaluate and check their results.

Perhaps most important, some scientists are coming to realize that they are better off doing fewer studies, but with more experimental subjects, to reduce the possibility of a chance finding.

"The way to get ahead and get a job and get tenure is to publish lots and lots of papers," says Gervais. "And it's hard to do that if you are able run fewer studies, but in the end I think that's the way to go — to slow down our science and be more rigorous up front."

Gervais says when he started his first faculty job, at the University of Kentucky, he sat down with his department chair and said he was going to follow this path of publishing fewer, but higher quality studies. He says he got the nod to do that. He sees it as part of a broader cultural change in social science that's aiming to make the field more robust.

You can reach Richard Harris at rharris@npr.org.