Researchers overturn landmark study on the replicability of psychological science

By Peter Reuell
Harvard Staff Writer

Category: HarvardScience
Subcategory: Culture & Society
KEYWORDS: psychology, psychological science, replication, replicate, reproduce, reproducibility, Center for Open Science, Gilbert, Daniel Gilbert, King, Gary King, Science, Harvard, FAS, Faculty of Arts and Sciences, Reuell, Peter Reuell
Summary: A 2015 study claiming that more than half of all psychology studies cannot be replicated turns out to be wrong. Harvard researchers have discovered that the study contains several statistical and methodological mistakes, and that when these are corrected, the study actually shows that the replication rate in psychology is quite high – indeed, it is statistically indistinguishable from 100%.

RELATED LINKS:

According to two Harvard professors and their collaborators, a 2015 landmark study showing that more than half of all psychology studies cannot be replicated is actually wrong.

In an attempt to determine the replicability of psychological science, a consortium of 270 scientists known as The Open Science Collaboration (OSC) tried to replicate the results of 100 published studies. More than half of them failed, creating sensational headlines worldwide about the "replication crisis" in psychology.

But an in-depth examination of the data by **Daniel Gilbert** (Edgar Pierce Professor of Psychology at Harvard University), **Gary King** (Albert J. Weatherhead III University Professor at Harvard University), **Stephen Pettigrew** (doctoral student in the Department of Government at Harvard University), and **Timothy Wilson** (Sherrell J. Aston Professor of Psychology at the University of Virginia) has revealed that the OSC made some serious mistakes that make this pessimistic conclusion completely unwarranted:

The methods of many of the replication studies turn out to be remarkably different from the originals and, according to Gilbert, King, Pettigrew, and Wilson, these "infidelities" had two important consequences.

First, they introduced statistical error into the data which led the OSC to significantly underestimate how many of their replications should have failed by chance alone. When this error is taken into account, the number of failures in their data is no greater than one would expect if all 100 of the original findings had been true.

Second, Gilbert, King, Pettigrew, and Wilson discovered that the low-fidelity studies were four times more likely to fail than were the high-fidelity studies, suggesting that when replicators strayed from the original methods, they caused their own studies to fail.

Finally, the OSC used a "low powered" design. When Gilbert, King, Pettigrew, and Wilson applied this design to a published data set that was known to have a high replication rate, it too showed a low replication rate, suggesting that the OSC's design was destined from the start to underestimate the replicability of psychological science.

Individually, Gilbert and King said, each of these problems would be enough to cast doubt on the conclusion that most people have drawn from this study, but taken together, they completely repudiate it. The flaws are described in a commentary published March 4 in Science.

Like most scientists who read the OSC's article when it appeared, Gilbert, King, Pettigrew, and Wilson were shocked and chagrined. But when they began to scrutinize the methods and reanalyze the raw data, they immediately noticed problems—problems that started with how the replicators had selected the 100 original studies.

"If you want to estimate a parameter of a population," said King, "then you either have to randomly sample from that population or make statistical corrections for the fact that you didn't. The OSC did neither."

"What they did," added Gilbert, "is create an idiosyncratic, arbitrary list of sampling rules that excluded the majority of psychology's subfields from the sample, that excluded entire classes of studies whose methods are probably among the best in science from the sample, and so on. Then they proceeded to violate all of their own rules. Worse yet, they actually allowed some replicators to have a choice about which studies they would try to replicate. If they had used these same methods to sample people instead of studies, no reputable scientific journal would have published their findings. So the first thing we realized was that no matter what they found—good news or bad news—they never had any chance of estimating the reproducibility of psychological science, which is what the very title of their paper claims they did."

"And that was just the beginning," King said. "If you are going to replicate a hundred studies, some will fail by chance alone. That's basic sampling theory. So you have to use statistics to estimate how many of the studies are expected to fail by chance alone because otherwise the number that actually do fail is meaningless."

According to King, the OSC did this, but they made a critical error.

"When they did their calculations, they failed to consider the fact that their replication studies were not just new samples from the same population. They were often quite different from the originals in many ways, and those differences are a source of statistical error. So we did the calculation the right way and then applied it to their data. And guess what? The number of failures they observed was just about what you should expect to

observe by chance alone—even if all one hundred of the original findings were true. The failure of the replication studies to match the original studies was a failure of the replications, not of the originals."

Gilbert noted that most people assume that a replication is a "replica"' of the original study.

"Readers surely assumed that if a group of scientists did a hundred replications, then they must have used the same methods to study the same populations. In this case, that assumption would be quite wrong. Replications always vary from originals in minor ways of course, but if you read the reports carefully, as we did, you discover that many of the replication studies differed in truly astounding ways—ways that make it hard to understand how they could even be called replications."

As an example, Gilbert described an original study that involved showing White students at Stanford University a video of four other Stanford students discussing admissions policies at their university. Three of the discussants were White and one was Black. During the discussion, one of the White students made offensive comments about affirmative action, and the researchers found that the observers looked significantly longer at the Black student when they believed he could hear the others' comments than when he could not.

"So how did they do the replication? With students at the University of Amsterdam!" Gilbert said. "They had Dutch students watch a video of Stanford students, speaking in English, about affirmative action policies at a university more than 5000 miles away."

In other words, unlike the participants in the original study, participants in the replication study watched students at a foreign university speaking in a foreign language about an issue of no relevance to them.

But according to Gilbert, that was not the most troubling part.

"If you dive deep into the data, you discover something else," Gilbert said. "The replicators realized that doing this study in the Netherlands might have been a problem, so they wisely decided to run another version of it in the US. And when they did, they basically replicated the original result. And yet, when the OSC estimated the reproducibility of psychological science, they excluded the successful replication and included only the one from the University of Amsterdam that failed. So the public hears that 'Yet another psychology study doesn't replicate' instead of 'Yet another psychology study replicates just fine if you do it right and not if you do it wrong' which isn't a very exciting headline. Some of the replications were quite faithful to the originals, but anyone who carefully reads all the replication reports will find many more examples like this one."

"These infidelities were a problem for another reason," King added, "namely, that they introduce additional error into the data set. That error can be calculated, and when we do,

it turns out that the number of replication studies that actually failed is about what we should expect if every single one of the original findings had been true. Now, one could argue about how best to make this calculation, but the fact is that OSC didn't make it at all. They simply ignored this potent source of error, and that caused them to draw the wrong conclusions from their data. That doesn't mean that all one hundred studies were true, of course, but it does mean that this article provides no evidence to the contrary."

"So we now know that the infidelities created statistical noise," said Gilbert, "but was that all they did? Or were the infidelities of a certain kind? In other words, did they just tend to change the original result, or did they tend to change it in a particular way?"

"To find out," said King, "we needed a measure of how faithful each of the hundred replications was. Luckily, the OSC supplied it."

Before each replication began, the OSC asked the original authors to examine the planned replication study and say whether they would endorse it as a faithful replication of their work, and about 70 percent did so.

"We used this as a rough index of fidelity, and when we did, we discovered something important: The low-fidelity replications were an astonishing four times more likely to fail," King said. "What that suggests is that the infidelities did not just create random statistical noise—they actually biased the studies toward failure."

In their Technical Comment, Gilbert, King, Pettigrew, and Wilson also note that the OSC used a "low powered" design: They replicated each of the 100 studies once, using roughly the number of subjects that were used in the original studies. But according to King, this method artificially depresses the replication rate.

"To show how this happens, we took another published article that had examined the replicability of a group of classic psychology studies," said King. "The authors of that paper had used a very high-powered design—they replicated each study with more than thirty times the original number of participants—and that high-powered design produced a very high replication rate. So we asked a simple question: What would have happened if these authors had used the low-powered design that was used by the OSC? The answer is that the replication rate would have been even lower than the replication rate found by the OSC."

Despite uncovering serious problems with the landmark study, Gilbert and King emphasized that their critique does not suggest any wrongdoing and is simply part of the normal process of scientific inquiry.

"Let's be clear, Gilbert said. "No one involved in this study was trying to deceive anyone. They just made mistakes, as scientists sometimes do. Many of the OSC members are our friends, and the corresponding author, Brian Nosek, is actually a good friend who was both forthcoming and helpful to us as we wrote our critique," Gilbert said. "In fact, Brian is the one who suggested one of the methods we used for correcting the OSC's error

calculations. So this is not a personal attack, this is a scientific critique. We all care about the same things: Doing science well and finding out what's true. We were glad to see that in their response to our comment, the OSC quibbled about a number of minor issues but conceded the major one, which is that their paper does *not* provide evidence for the pessimistic conclusions that most people have drawn from it."

"I think the big take-away point here is that meta-science must obey the rules of science," King said. "All the rules about sampling and calculating error and keeping experimenters blind to the hypothesis—all of those rules must apply whether you are studying people or studying the replicability of a science. Meta-science does not get a pass. It is not exempt. And those doing meta-science are not above the fray. They are part of the scientific process. If you violate the basic rules of science, you get the wrong answer, and that's what happened here."

"This paper has had extraordinary impact," Gilbert said. "It was Science magazine's number three 'Breakthrough of the Year' across all fields of science. It led to changes in policy at many scientific journals, changes in priorities at funding agencies, and it seriously undermined public perceptions of psychology. So it is not enough now, in the sober light of retrospect, to say that mistakes were made. These mistakes had very serious repercussions. We hope the OSC will now work as hard to correct the public misperceptions of their findings as they did to produce the findings themselves."

NOTE: The OSC's reply to Gilbert et al's Technical Comment and Gilbert et al's response to that reply can be found here.