Researcher discusses the the science replication crisis

November 21 2018, by Emily Velasco

Credit: Caltech

If there's a central tenet that unites all of the sciences, it's probably that scientists should approach discovery without bias and with a healthy dose of skepticism. The idea is that the best way to reach the truth is to allow the facts to lead where they will, even if it's not where you intended to go.

But that can be easier said than done. Humans have unconscious biases that are hard to shake, and most people don't like to be wrong. In the past several years, scientists have discovered troubling evidence that those biases may be affecting the integrity of the research process in many fields.

The evidence also suggests that even when scientists operate with the

best intentions, serious errors are more common than expected because even subtle differences in the way an experimental procedure is conducted can throw off the findings.

When biases and errors leak into research, other scientists attempting the same experiment may find that they can't replicate the findings of the original researcher. This has given the broader issue its name: the replication crisis.

Colin Camerer, Caltech's Robert Kirby Professor of Behavioral Economics and the T&C Chen Center for Social and Decision Neuroscience Leadership Chair, executive officer for the Social Sciences and director of the T&C Chen Center for Social and Decision Neuroscience, has been at the forefront of research into the replication crisis. He has penned a number of studies on the topic and is an ardent advocate for reform. We talked with Camerer about how bad the problem is and what can be done to correct it; and the "open science" movement, which encourages the sharing of data, information, and materials among researchers.

What exactly is the replication crisis?

What instigated all of this is the discovery that many findings—originally in medicine but later in areas of psychology, in economics, and probably in every field—just don't replicate or reproduce as well as we would hope. By reproduce, I mean taking data someone collected for a study and doing the same analysis just to see if you get the same results. People can get substantial differences, for example, if they use newer statistics than were available to the original researchers.

The earliest studies into reproducibility also found that sometimes it's hard to even get people to share their data in a timely and clear way.

There was a norm that data sharing is sort of a bonus, but isn't absolutely a necessary part of the job of being a scientist.

How big of a problem is this?

I would say it's big enough to be very concerning. I'll give an example from social psychology, which has been one of the most problematic areas. In social psychology, there's an idea called priming, which means if I make you think about one thing subconsciously, those thoughts may activate related associations and change your behavior in some surprising way.

Many studies on priming were done by John Bargh, who is a well-known psychologist at Yale. Bargh and his colleagues got young people to think about being old and then had them sit at a table and do a test. But the test was just a filler, because the researchers weren't interested in the results of the test. They were interested in how thinking about being old affected the behavior of the young people. When the young people were done with the filler test, the research team timed how long it took them to get up from the table and walk to an elevator. They found that the people who were primed to think about being old walked slower than the control group that had not received that priming.

They were trying to get a dramatic result showing that mental associations about old people affect physical behavior. The problem was that when others tried to replicate the study, the original findings didn't replicate very well. In one replication, something even worse happened. Some of the assistants in that experiment were told the priming would make the young subjects walk slower, and others were told the priming would make them walk more quickly—this is what we call a reactance or boomerang effect. And what the assistants were told to expect influenced their measurements of how fast the subjects walked, even though they were timing with stopwatches. The assistants' stopwatch

measures were biased compared to an automated timer. I mention this example because it's the kind of study we think of as too cute to be true. When the failure to replicate came out, there was a big uproar about how much skill an experimenter needs to do a proper replication.

You recently explored this issue in a pair of papers. What did you find?

In our first paper, we looked at experimental economics, which is something that was pioneered here at Caltech. We took 18 papers from multiple institutions that were published in two of the leading economics journals. These are the papers you would hope would replicate the best. What we found was that 14 out of 18 replicated fairly well, but four of them didn't.

It's important to note that in two of those four cases, we made slight deviations in how the experiment was done. That's a reminder that small changes can make a big difference in replication. For example, if you're studying political psychology and partisanship and you replicate a paper from 2010, the results today might be very different because the political climate has changed. It's not that the authors of the original paper made a mistake, it's that the phenomenon in their study changed.

In our second paper, we looked at social science papers published between 2010 and 2015 in *Science* and *Nature*, which are the flagship general science journals. We were interested in them because these were highly cited papers and were seen as very influential.

We picked out the ones that wouldn't be overly laborious to replicate, and we ended up with 21 papers. What we found was that only about 60 percent replicated, and the ones that didn't replicate tended to focus on things like priming, which I mentioned before. Priming has turned out to

be the least replicable phenomenon. It's a shame because the underlying concept—that thinking about one thing elevates associations to related things—is undoubtedly true.

How does something like that happen?

One cause of findings not replicating is what we call "p-hacking." P-value is a measure of the statistical likelihood that your hypothesis is true. If the p-value is low, an effect is highly unlikely to be a fluke due to chance. In social science and medicine, for example, you are usually testing whether changing the conditions of the experiment changes behavior. You really want to get a low p-value because it means that the condition you changed did have an effect. P-hacking is when you keep trying different analyses with your data until you get the p-value to be low.

A good example of p-hacking is deleting data points that don't fit your hypothesis—outliers—from your data set. There are statistical methods to deal with outliers, but sometimes people expect to see a correlation and don't find much of one, for example. So then they think of a plausible reason to discard a few outlier points, because by doing that they can get the correlation to be bigger. That practice can be abused, but at the same time, there sometimes are outliers that should be discarded. For example, if subjects blink too much when you are trying to measure visual perception, it is reasonable to edit out the blinks or not use some subjects.

Another explanation is that sometimes scientists are simply helped along by luck. When someone else tries to replicate that original experiment but doesn't get the same good luck, they won't get the same results.

In the sciences, you're supposed be impartial and say,

"Here's my hypothesis, and I'm going to prove it right or wrong." So, why do people tweak the results to get an answer they want?

At the top of the pyramid is outright fraud and, happily, that's pretty rare. Typically, if you do a postmortem or a confessional in the case of fraud, you find a scientist who feels tremendous pressure. Sometimes it's personal—"I just wanted to be respected"—and sometimes it's grant money or being too ashamed to come clean.

In the fraudulent cases, scientists get away with a small amount of deception, and they get very dug in because they're really betting their careers on it. The finding they faked might be what gets them invited to conferences and gets them lots of funding. Then it's too embarrassing to stop and confess what they've been doing all along.

There are also faulty scientific practices less egregious than outright fraud, right?

Sure. It is the scientist who thinks, "I know I'm right, and even though these data didn't prove it, I'm sure I could run a lot more experiments and prove it. So I'm just going to help the process along by creating the best version of the data." It's like cosmetic surgery for data.

And again, there are incentives driving this. Often in Big Science and Big Medicine, you're supporting a lot of people on your grant. If something really goes wrong with your big theory or your pathbreaking method, those people get laid off and their careers are harmed.

Another force that contributes to weak replicability is that, in science, we rely to a very large extent on norms of honor and the idea that people care about the process and want to get to the truth. There's a tremendous

amount of trust involved. If I get a paper to review from a leading journal, I'm not necessarily thinking like a police detective about whether it's fabricated.

A lot of the frauds were only uncovered because there was a pattern across many different papers. One paper was too good to be true, and the next one was too good to be true, and so on. Nobody's good enough to get 10 too-good-to-be-trues in a row.

So, often, it's kind of a fluke. Somebody slips or a person notices and then asks for the data and digs a little further.

What best practices should scientists follow to avoid falling into these traps?

There are many things we can do—I call it the reproducibility upgrade. One is preregistration, which means before you collect your data, you publicly explain and post online exactly what data you're going to collect, why you picked your sample size, and exactly what analysis you are going to run. Then if you do very different analysis and get a good result, people can question why you departed from what you preregistered and whether the unplanned analyses were p-hacked.

The more general rubric is called open science, in which you act like basically everything you do should be available to other people except for certain things like patient privacy. That includes original data, code, instructions, and experimental materials like video recordings—everything.

Meta-analysis is another method I think we're going to see more and more of. That's where you combine the results of studies that are all trying to measure the same general effect. You can use that information

to find evidence of things like publication bias, which is a sort of groupthink. For example, there's strong experimental evidence that giving people smaller plates causes them to eat less. So maybe you're studying small and large plates, and you don't find any effect on portion size. You might think to yourself, "I probably made a mistake. I'm not going to try to publish that." Or you might say, "Wow! That's really interesting. I didn't get a small-plate effect. I'm going to send it to a journal." And the editors or referees say, "You probably made a mistake. We're not going to publish it." Those are publication biases. They can be caused by scientists withholding results or by journals not publishing them because they get an unconventional result.

If a group of scientists comes to believe something is true and the contrary evidence gets ignored or swept under the rug, that means a lot of people are trying to come to some collective conclusion about something that's not true. The big damage is that it's a colossal waste of time, and it can harm public perceptions of how solid science is in general.

Are people receptive to the changes you suggest?

I would say 90 percent of people have been very supportive. One piece of very good news is the Open Science Framework has been supported by the Laura and John Arnold Foundation, which is a big private foundation, and by other donors. The private foundations are in a unique position to spend a lot of money on things like this. Our first grant to do replications in experimental economics came when I met the program officer from the Alfred P. Sloan Foundation. I told him we were piloting a big project replicating economics experiments. He got excited, and it was figuratively like he took a bag of cash out of his briefcase right there. My collaborators in Sweden and Austria later got a particularly big grant for $1.5 million to work on replication. Now that there's some momentum, funding agencies have been reasonably generous, which is

great.

Another thing that's been interesting is that while journals are not keen on publishing a replication of one paper, they really like what we've done, which is a batch of replications. A few months into working on the first replication paper in experimental economics funded by Sloan, I got an email from an editor at *Science* who said, "I heard you're working on this replication thing. Have you thought about where to publish it?" That's a wink-wink, coy way of saying "Please send it to us" without any promise being made. They did eventually publish it.

What challenges do you see going forward?

I think the main challenge is determining where the responsibility lies. Until about 2000, the conventional wisdom was, "Nobody will pay for your replication and nobody will publish your replication. And if it doesn't come out right, you'll just make an enemy. Don't bother to replicate." Students were often told not to do replication because it would be bad for their careers. I think that's false, but it is true that nobody is going to win a big prize for replicating somebody else's work. The best career path in science comes from showing that you can do something original, important, and creative. Replication is exactly the opposite. It is important for somebody to do it, but it's not creative. It's something that most scientists want someone else to do.

What is needed are institutions to generate steady, ongoing replications, rather than relying on scientists who are trying to be creative and make breakthroughs to do it. It could be a few centers that are just dedicated to replicating. They could pick every fifth paper published in a given journal, replicate it, and post their results online. It would be like auditing, or a kind of Consumer Reports for science. I think some institutions like that will emerge. Or perhaps granting agencies, like the National Institutes of Health or the National Science Foundation, should

be responsible for building in safeguards. They could have an audit process that sets aside grant money to do a replication and check your work.

For me this is like a hobby. Now I hope that some other group of careful people who are very passionate and smart will take up the baton and start to do replications very routinely.

Provided by California Institute of Technology