More social science studies just failed to replicate. Here's why this is good.

What scientists learn from failed replications: how to do better science.

By Brian Resnick | @B_resnick | brian@vox.com | Aug 27, 2018, 11:00am EDT

Psychologists are still wondering: "What's going on in there?" They're just doing it with greater rigor. | Enis Aksoy/Getty Creative Images

One of the cornerstone principles of science is replication. This is the idea that experiments need to be repeated to find out if the results will be consistent. The fact that an experiment can be replicated is how we know its results contain a nugget of truth. Without replication, we can't be sure.

For the past several years, social scientists have been deeply worried about the replicability of their findings. Incredibly influential, textbook findings in psychology — like the "**ego depletion"** theory of willpower, or the "**marshmallow test**" — have been bending or breaking under rigorous retests. And the scientists have learned that what they used to consider commonplace methodological practices were really just recipes to generate false positives. This period has been called the "**replication crisis**" by some.

And the reckoning is still underway. Recently, a team of social scientists — spanning psychologists and economists — **attempted to replicate** 21 findings published in the most prestigious general science journals: *Nature* and *Science.* Some of the retested studies have been widely influential in science and in pop culture, like a 2011 paper on whether **access to search engines hinders** our memories, or whether reading books **improves** a child's theory of mind (meaning their ability to understand that other people have thoughts and intentions different from their own).

On Monday, they're **publishing** their results in the journal *Nature Human Behavior.* Here's their take-home lesson: Even studies that are published in the top journals should be taken with a grain of salt until they are replicated. They're initial findings, not ironclad truth. And they can be really hard to replicate, for a variety of reasons.

Rigorous retests of social science studies often yield less impressive results

The scientists who ran the 21 replication tests didn't just repeat the original experiments — they made them more rigorous. In some cases, they increased the number of participants by a factor of five, and preregistered their study and analysis designs before a single participant was brought into the lab.

All the original authors (save for one group that couldn't be reached), signed off on the study designs too. Preregistering is like making a promise to not deviate from a plan and inject bias into the results.

Here are the results: 13 of the 21 results replicated. But perhaps just as notable: Even among the studies that did pass, the effect sizes (that is, the difference between the experimental group and the control group in the experiment, or the size of the change the experimental manipulation made) decreased by around half, meaning that the original findings likely overstated the power of the experimental manipulation.

"Overall, our study shows statistically significant scientific findings should be interpreted rather cautiously until they have been replicated, even if they have been published in the

most renowned journals," Felix Holzmeister, an Austrian economist and one of the study co-authors, says.

It's not always clear why a study doesn't replicate. Science is hard.

Many of the papers that were retested contained multiple experiments. Only one experiment from each paper was tested. So these failed replications don't necessarily mean the theory behind the original findings is totally bunk.

For instance, the famous "Google Effects on Memory" paper — which found that we often don't remember things as well when we know we can search for them online — did not replicate in this study. But the experiment chosen was a word-priming task (i.e., does thinking about the internet make it harder to retrieve information), and not the more real-world experiment that involved actually answering trivia statements. And other **research since** has bolstered that paper's general argument that access to the internet is shifting the relationship we have with, and the utility of, our own memories.

There could be a lot of reasons a result doesn't replicate. One is that the experimenters doing the replication messed something up.

Another reason can be that the study stumbled on a false positive.

One of the experiments that didn't replicate was from University of Kentucky psychologist Will Gervais. The experiment tried to see if getting people to think more rationally would make them less willing to report religious belief.

"In hindsight, our study was outright silly," Gervais says. They had people look at a picture of Rodin's ***The Thinker*** or another statue. They thought *The Thinker* would nudge people to think harder.

"When we asked them a single question on whether they believe in God, it was a really tiny sample size, and barely significant … I'd like to think it wouldn't get published today," Gervais says. (And know, this **study** was published in *Science* a top journal.)

In other cases, a study may not replicate because the target — the human subjects — has changed. In 2012, MIT psychologist David Rand published a paper in *Nature* on human cooperation. The experiment involved online participants playing an economics game. He argues that a lot of online study participants have since grown familiar with this game,

which makes it a less useful tool to probe real-life behaviors. His experiment didn't replicate in the new study.

Finding out why a study didn't replicate is hard work. But it's exactly the type of work, and thinking, that scientists need to be engaged in. The point of this replication project, **and others like it**, is not to call out individual researchers. "It's a reminder of our values," says Brian Nosek, a psychologist and the director of the **Center for Open Science**, who collaborated on the new study. Scientists who publish in top journals should know their work may be checked up on. It's also important, he notes, to know that social science's inability to be replicable is in itself a replicable finding.

Often, when studies don't replicate, it's not that the effort totally disproves the underlying hypothesis. And it doesn't mean the original study authors were frauds. But replication results do often significantly change the **story we tell about the experiment**.

For instance, I recently wrote about a **replication** effort of the famous "marshmallow test" studies, which originally showed that the ability to delay gratification early in life is correlated with success later on. A new **paper** found this correlation, but when the authors controlled for factors like family background, the correlation went away.

Here's how the story changed: Delay of gratification is not a unique lever to pull to positively influence other aspects of a person's life. It's a consequence of bigger-picture, harder-to-change components of a person.

In science, too often, the first demonstration of an idea becomes the lasting one. Replications are a reminder that in science, this isn't supposed to be the case. Science ought to embrace and learn from failure.

The "replication crisis" in psychology has been going on for years now. And scientists are reforming their ways.

The "replication crisis" in psychology, as it is often called, started **around 2010**, when a paper using completely accepted experimental methods was published purporting to find evidence that people were capable of perceiving the future, which is impossible. This prompted a **reckoning**: Common practices like drawing on small samples of college students were found to be insufficient to find true experimental effects.

Scientists thought if you could find an effect in a small number of people, that effect must be robust. But often, significant results from small samples turn out to be statistical flukes.

(For more on this, read our **explainer** on p-values.)

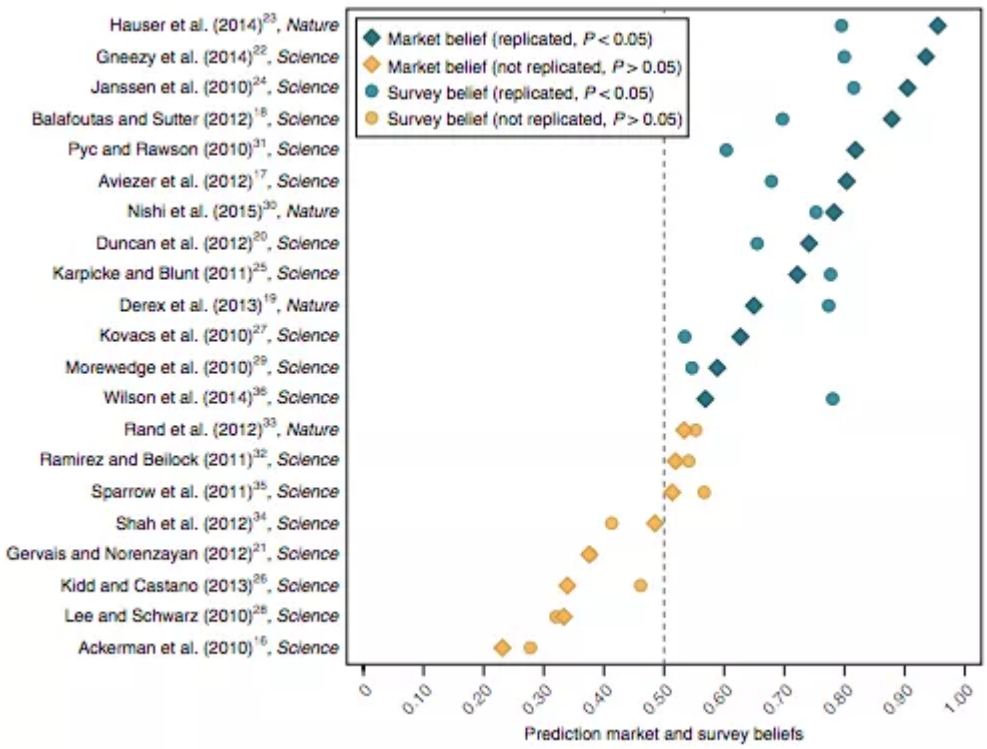
The crisis intensified in 2015 when a group of psychologists, which included Nosek, published a **report** in *Science* with evidence of an overarching problem: When 270 psychologists tried to replicate 100 experiments published in top journals, only around 40 percent of the studies held up. The remainder either failed or yielded inconclusive data. And again, the replications that did work showed weaker effects than the original papers. The studies that tended to replicate had more highly significant results compared to the ones that just barely crossed the threshold of significance.

Another important reason to do replications, Nosek says, is to get better at understanding what types of studies are most likely to replicate, and to sharpen scientists' intuitions about what hypotheses are worthy of testing and which are not.

As part of the new study, Nosek and his colleagues added a prediction component. A group of scientists took bets on which studies they thought would replicate and which they thought wouldn't. The bets largely tracked with the final results.

As you can see in the chart below, the yellow dots are the studies that did not replicate, and they were all unfavorably ranked by the prediction market survey.

"These results suggest [there's] something systematic about papers that fail to replicate," Anna Dreber, a Stockholm-based economist and one of the study co-authors, says.



Nature Human Behavior

One thing that stands out: Many of the papers that failed to replicate sound a little too good to be true. Take this 2010 paper that finds simply washing hands negates a **common human hindsight bias**. When we make a tough choice, we often look back on the choice we passed on unfavorably and are biased to find reasons to justify our decision. Washing hands in an experiment "seems to more generally remove past concerns, resulting in a metaphorical 'clean slate' effect," the study's abstract **stated**.

It all sounds a little too easy, too simple — and it didn't replicate.

All that said, there are some promising signs that social science is getting better. More and more scientists are **preregistering their study designs**. This prevents them from cherry-picking results and analyses that are more favorable to their favored conclusions. Journals are getting better at demanding larger subject pools in experiments and are increasingly **insisting that scientists share all the underlying data** of their experiments for others to assess.

"The lesson out of this project," Nosek says, "is a very positive message of reformation. Science is going to get better."