A Worrying Trend for Psychology’s 'Simple Little Tricks'

In yet another setback for the field, researchers have failed to replicate two studies showing that basic techniques can reduce racial achievement gaps and improve voter turnout.

Over the past decade, social psychologists have dazzled us with studies showing that huge social problems can seemingly be rectified through simple tricks. A small
grammatical tweak in a survey delivered to people the day before an election greatly increases voter turnout. A 15-minute writing exercise narrows the achievement gap between black and white students—and the benefits last for years.

“Each statement may sound outlandish—more science fiction than science,” wrote Gregory Walton from Stanford University in 2014. But they reflect the science of what he calls “wise interventions”—strategies that work because they’re savvy to subtle psychology behind our everyday lives. In many ways, such strategies represent the ultimate test for psychology—a chance to show that all the academic theorizing and small-scale lab experiments can actually be used to influence people’s minds in the messy, real, complicated world.

They seem to work, if the stream of papers in high-profile scientific journals is to be believed. But as with many branches of psychology, wise interventions are taking a battering. A new wave of studies that attempted to replicate the promising experiments have found discouraging results. At worst, they suggest that the original successes were mirages. At best, they reveal that pulling off these tricks at a large scale will be more challenging than commonly believed.

“It highlights the perils of treating a complex psychological study as fortune cookie wisdom.”

Consider a recent study by Christopher Bryan (then at Stanford, now at University of Chicago), along with Walton and others. During the 2008 U.S. presidential election, they sent a survey to 133 Californian voters. Some were asked: “How important is it to you to vote in the upcoming election?” Others received the same question but with a slight tweak: “How important is it to you to be a voter in the upcoming election?”

Once the ballots were cast, the team checked the official state records. They found that 96 percent of those who read the “be a voter” question showed up to vote, compared to just 82 percent of those who read the “to vote” version. A tiny linguistic tweak led to a huge 14 percentage point increase in turnout. The team
repeated their experiment with 214 New Jersey voters in the 2009 gubernatorial elections, and found the same large effect: changing “vote” to “be a voter” raised turnout levels from 79 percent to 90 percent.

Why? The team explained that, as seen in earlier studies, nouns (“voter”) create a much stronger sense of self-identity than verbs (“vote”) because they define who we are rather than what we do. “People may be more likely to vote when voting is represented as an expression of self—as symbolic of a person’s fundamental character—rather than as simply a behavior,” they wrote. That’s a classic wise intervention—a simple thing that draws on earlier psychological work to change people’s behavior in subtle but profound ways.

When Alan Gerber heard about the results, he was surprised. As a political scientist at Yale University, he knew that previous experiments involving thousands of people had never mobilized voters to that degree. Mail-outs, for example, typically increase turnout by 0.5 percentage points, or 2.3 if especially persuasive. And yet changing a few words apparently did so by 11 to 14 percentage points. Whoa, if true.

Gerber remained open-minded. “When something has an outsized effect, skepticism is understandable but if your attitude is overwhelming skepticism, you’ll reject a lot of very good science,” he says. So he repeated Bryan’s experiment. His team delivered the same survey to 4,400 voters in days leading up to the 2014 primary elections in Michigan, Missouri, and Tennessee. And they found that using the noun version instead of the verb one had no effect on voter turnout. None. Their much larger study, with 20 to 33 times the participants of Bryan’s two experiments, completely failed to replicate the original effects.

Melissa Michelson, a political scientist at Menlo College, isn’t surprised. She was never quite convinced about how robust Bryan’s results were, or how useful they would be. “I’ve conducted hundreds of get-out-the-vote experiments and most of the time you’re having live conversations with targeted voters that are meant to hit certain points, but aren’t scripted word-for-word,” she says. “The idea that you’d have to train your canvassers to use nouns instead of verbs just didn’t sound
realistic. Many of us were waiting to see more data with larger samples in with different populations, and that’s exactly what Gerber has provided.”

Jan Leighley from American University agrees. The small sample size of the original study “would have tanked the paper from consideration in a serious political science journal,” she says.

There are many reasons why researchers might be unable to replicate the results of an earlier study. It could be that the original experiment was flawed, and its results were a random fluke. Also, psychologists often tamper with details of their studies in ways that produce positive and publishable results, but also illusory and irreproducible ones.

Gerber doesn’t think any of that is necessarily happening here. Instead, he notes that there are many differences between his experiment and Bryan’s. They involved different people, elections, and years. Bryan used an online survey, while Gerber delivered his over the phone. The second study doesn’t necessarily mean that the verb-noun effect isn’t real, just that it might only show up in some situations and not others. “Failure to generalize might be a better phrase than failure to replicate,” Gerber says.

Bryan agrees. He thinks that Gerber’s study, though much larger, had some fatal flaws. First, the timing was off. Gerber’s team deployed their surveys a few days before their respective elections, whereas Bryan’s team did so the day before or the morning of the polls. “This kind of psychology is ephemeral,” he says. “If you do it days before, you might think, ‘Yeah I really should vote’, and then move on to something else. You have to do it at the point where people are making a decision, and going: Okay, where’s my polling station?”

Second, the stakes were lower. “We ran our study in two major elections that got a lot of media attention,” he says. “The elections that Gerber used... most of them didn’t matter, and nearly half were uncontested.” In this context, emphasizing one’s identity as a voter shouldn’t really matter. (Gerber counters that previous
studies have shown that voter mobilization should be *more* effective, not less, in lower-stakes elections.)

“In that context, I don’t think the theory would have predicted a strong effect. They had little or no chance of getting useful results,” Bryan says. “I’ve heard of a number of political candidates who tried to apply the same idea, but most of the time, it was a significant enough deviation from how we did it that I wasn’t at all confident it would be effective. It highlights the perils of treating a complex psychological study as fortune cookie wisdom.”

He takes some responsibility for that. “The space in high-profile journals is limited, but we can all do a better job of more thoroughly articulating the theory behind our ideas,” he says. And that gulf between theory and practice can mean the difference between a wise intervention and a foolish one.

“*It looked like a way of addressing these very large and troubling racial achievement gaps.*”

While Bryan was trying to get people to polls, Geoff Cohen (also from Stanford) was working to improve the fates of African American children. In 2003 and 2006, his team worked with 158 black seventh-graders from a Northeastern school. Half of them were randomly chosen to write about something that was important to them, from having friends to being musically adept. The other half wrote about something they deemed unimportant.

The exercise lasted just 15 minutes, but it worked wonders. Those who wrote about their values had added 0.3 points to their grade point average by the end of the term, closing the academic gap between them and their white peers by 40 percent. *After two years* (and a few ‘booster’ repetitions of the same exercise), their GPAs were still higher by a quarter of a point.

The exercise worked, Cohen said, because it breaks a vicious and self-fulfilling psychological cycle. Black students have to worry about the negative stereotype that
they underperform at school, and that worry causes so much stress that they actually do underperform—an insidious effect known as stereotype threat. By asking the children to write about their values, Cohen mentally vaccinated them by bolstering their sense of self-worth. According to this theory, only students who are subject to negative stereotypes should benefit, and the poorest performers should benefit most. And that’s exactly what the team found.

Cohen has since replicated his results in other schools. He and others, like Walton, have also tested similar exercises with other groups who suffer from negative stereotypes like women in college physics classes. Time and again, they found that these short, simple tasks could have dramatic, lasting benefits.

At first, so did Paul Hanselman from the University of California, Irvine. Like Gerber, he came from a place of open-minded interest. “Cohen’s original study looked exciting and promising,” he says. “It looked like a way of addressing these very large and troubling racial achievement gaps.”

In 2011, Hanselman and his colleagues repeated the study with 374 minority seventh-graders from 11 schools in a single Midwestern district, with the same materials that Cohen had used. This time, the black students gained just 0.065 GPA points—a much weaker effect than in the original study, but a positive one nonetheless, and one that also lasted for years. But Hanselman, wanting to make the most of his relationships with the school district, repeated the study a second time. “Life would be so much easier if we hadn’t,” he says.

This time, they went bigger, recruiting 449 minority children. And this time, they found that the writing exercise had no effect at all.

The critical thing here is that Hanselman has replicated both Cohen’s original experiment and his own successful replication—a rarity in psychology. This means the usual criticism—that the replicating team missed key aspects of the original experiment, as Bryan claims of Gerber—doesn’t quite apply. “We were the same team in the same schools with many of the same teachers and administrators, and there were a lot of subtleties that we controlled in our two trials,” says Hanselman.
“I was mostly impressed by the high quality of their methods,” says Linda Skitka from the University of Illinois at Chicago. “They have a very large sample size, they examined a range of possible contingencies for why the effect might be observed with some students but not others, and they conferred with the original authors and used their exact materials.” But despite those efforts, Hanselman is no closer to explaining why his two replications differed in their results. “A lot of the most obvious things don’t seem to explain the difference, which leaves us with a puzzle,” he says.

It’s possible that the benefits from the first two experiments were flukes, while the third and largest one produced a more statistically reliable (albeit disappointing) result.* Alternatively, in the year of the second study, there was political unrest in Wisconsin during which teachers went on strike; perhaps that affected the school environment. Perhaps the teachers got tired of administering the exercises, or the students took it less seriously. Perhaps, most simply, the difference between the studies is itself a fluke—the result of random chance.

Cohen, as you might imagine, sees things differently. “Bigger is not necessarily better,” he and his colleagues have argued in a written rebuttal. They say that the affirming exercises must be administered delicately, and in scaling up, Hanselman’s team sacrificed attention to detail.

For example, the students can’t know that it’s part of an outsider’s study; they have to see it as something their teachers assigned, because that tells them that their values matter in the classroom. They shouldn’t be told that the exercise would benefit them, either. “The message that this is good for you can be stigmatizing by insinuating to students that they need of help, which undermines the affirmation,” says Cohen. And in both Hanselman’s studies, the teachers broke these rules for many of the students. “This suggests a significant lack of oversight and quality control.”

Hanselman counters that the teachers were more likely to stick to the rules in the second study than the first. “I thought we implemented the activities better the second time around,” he says. “Our interactions with teachers, our ability to recruit
students, even the logistics of preparing thousands and thousands of activities, all seemed to go smoother in the second study.”

---

**It seems that they are like sensitive and delicate flowers, only able to bloom if the conditions are just right.**

This debate mirrors that between Bryan and Gerber. In both cases, there’s a team of independent researchers trying to replicate their peers’ work—one of the cornerstones of science—and to see if the benefits they saw can generalize to new contexts and larger scales. In both cases, those replication attempts, carried out in good faith, have been disappointing. And in both cases, the original experimenters have argued that some crucial detail was missing.

Affirmation activities “are not like the power pellets in the old Pac Man video game, which abruptly give the player extra powers,” says Cohen. “This is why our approach to research has been not to do ‘mass vaccinations.’ Instead, like research on drug therapy, we try to identify the time, place, and persons for which affirmation and other psychological interventions work best.”

It seems, then, that wise interventions are like sensitive and delicate flowers, only able to bloom if the conditions are just right. Walton, Cohen, and their peers have always argued as much. But that’s in itself a problem. If it is so hard for teams of experienced and competent social scientists to get these techniques to work, what hope is there for them to be used more broadly?

Cohen is optimistic, suggesting training liaisons to ensure that the interventions are used correctly. Hanselman is less bullish, noting that if the effects are so variable, it will take very large studies to work out when and where the interventions work, if they do at all. And no matter who is right, it is clear that these wise interventions are not the simple tricks they’re made out to be.
Update: The article originally reported that sample size in Hanselman’s second replication was much larger than his first; that is not the case.

ABOUT THE AUTHOR

ED YONG is a staff writer at The Atlantic, where he covers science.

Twitter  Email