STATE OF THE WORLD'S SCIENCE 2018

MAKE RESEARCH REPRODUCIBLE

Better incentives could reduce the alarming number of studies that turn out to be wrong when repeated

By Shannon Palus

Illustrations by Neil Webb

KATIE CORKER WONDERED what temperature the coffee was supposed to be. She was doing a psychology experiment—well, redoing an experiment. The original findings, suggesting that holding something warm can make a person behave warmly, had been published in 2008 in the prestigious journal *Science* to a flurry of media coverage. Yet as Corker tried to retrace each step in the study, there were so many unknowns: the temperature of the hot coffee distributed to subjects, how quickly the mug cooled in their hands.

Corker, a psychologist at Grand Valley State University, was trying what few scientists attempt: to carefully replicate research and publish the results. The goal, in her case, was to find out whether she, working in another laboratory with a different group of subjects, would find the same effect as the *Science* study, which had been conducted by just one research group with only 94 participants clutching coffee or therapeutic pads of varying temperatures. In theory, this is how science is supposed to work: as a self-correcting process in which researchers build on the findings of others.

For decades it has been something of an open secret that a chunk of the literature in some fields is plain wrong. In biomedicine, the truth became clear in 2012. At the time, C. Glenn Begley was a vice president and global head of hematology and oncology research at the pharmaceutical company Amgen, overseeing the development of cancer drugs based, in part, on promising breakthroughs from academia. After a decade in the gig, he wanted to know why some projects looking into promising targets for drugs were being halted. He turned to the company's files and found that, incredibly, often the problem lay with the preclinical research, something that his teams double-checked before pouring money and resources into basing a treatment on it. "To my horror, I discovered that 90 percent of the time, we were unable to reproduce what was published," says Begley, who is now CEO of the Australian firm BioCurate. A study would later find that failures to replicate preclinical work in the field of biomedicine eat up $28.2 billion every year in the U.S. Begley even sent Amgen scientists to some labs to watch them try to replicate their own results. They failed, too.

Meanwhile the crisis was becoming apparent in psychology. Nearly 300 scientists were volunteering their time to repeat experiments in 100 papers in the field as part of University of Virginia psychologist Brian Nosek's Reproducibility Project: Psychology. In 2015 they declared that just 36 percent of the repeated experiments showed significant results in line with the original findings.

Although the landmark reproducibility studies have been in biomedicine and psychology, the issue is not confined to those fields. Lorena A. Barba, an engineer at George Washington University, who works in computational fluid dynamics, spent a full three years collaborating with a student to reconstruct a complex simulation from her own lab on how flying snakes, which leap off tree branches to glide through the air, wiggle as they soar. The new results were consistent, but she learned that sifting through other people's code to piece together what they did can be a nightmare. She essentially encountered the same problem that Corker did with the hot cups of coffee. Scientists are focused on publishing results, not necessarily on every mundane step of how they arrived at them. "There's just not a lot written down," Corker says. She got lucky, though: the original first author of the coffee study was "very willing to work with us." She also collaborated with a chemist to standardize how quickly the test apparatus changed temperature. "I found it more challenging than some of the original research I've done," she says.

Long-ingrained scientific habits such as an aversion to sharing techniques for fear of being scooped often work counter to the goal of reproducibility. Barba's own field was born in a veil of secrecy in Los Alamos, N.M., during the Manhattan Project, as researchers designing the first nuclear weapons used early computers to calculate how blasts of air and energy would ripple off exploding bombs. The Manhattan Project, of course, provided fuel to large swaths of the hard sciences. Scientists at the time actively tried to prevent outsiders from replicating their work.

Furthermore, journals and tenure committees often prize new, flashy results instead of piecemeal advances that carefully build on the existing literature. "My training was about trying to find the unexpected effect," says Charlotte Tate, a social and personality psychologist at San Francisco State University. She jokes that members of her field "run around with this model that we have to get on the *Daily Show*." This attitude is not just vanity: flashy results are often how you secure a job. Those quietly fact-checking the work of others or spending extra hours toiling to ensure that their code is easy for another researcher to understand do not earn a name in lights—or even at the top of a stack of resumes.

Many emphasize the role that better training—on how to write a bullet-proof "methods" section of a paper or carefully document code so that it is legible to others—can play in helping

the crisis. Barba is in this camp, noting that people who use code in their work would do well to take a software etiquette class so that they can present well-documented code alongside their results. She also uses a technology known as version control, which records any changes made to a file, to make the evolution of her team's code as legible as possible. The tool is standard in software development but, bafflingly to Barba, not yet in science. "There's this fundamental tension between doing an experiment and documenting the experiment," says Charles Fracchia, who is trying to increase the detail and depth of experiment logs in biomedicine through his company BioBright. One of his tools, DarwinSync, records data from every instrument possible, including seemingly unimportant things such as whether a computer was plugged in or running on a battery or the amount of ambient light in a room, in case those details are later revealing. In the case of Corker's replication attempt, if the original study had better assessed the mugs' temperatures, that would have set her up with more information to rerun the trial later.

But time-intensive solutions and expensive equipment are not enough. "There's no reward for doing things right," Barba says. The trick, Nosek says, is to rework the incentives to ensure "what's good for a scientist is what's good for science." For instance, agencies that fund research could choose to finance only projects that include a plan for making their work transparent. In 2016 the National Institutes of Health rolled out new application instructions and review questions to encourage scientists seeking grant money to improve the reproducibility of their work. The NIH now asks for more information about how the study builds on previous work and a list of variables that could impact the investigation, such as the sex of rat subjects (a previously overlooked factor that led many studies to describe phenomena found in male rats as universal).

And all the questions that a funder can ask up front could also be asked by journals and reviewers. For Nosek, a promising solution lies in what is known as registered reports, a preregistration of studies in which scientists submit research analysis and design plans for publication before they actually do it. Peer reviewers then evaluate the methodology—if it is sound, if it builds on past findings—and the journal promises to print the results no matter what they are. The reward of a paper comes for carefully thought-out experiments, not flashy results. Some wonder if such a change would simply produce boring science. Nosek contends that is not the case. He is currently completing a pair of investigations to examine the impact and quality of the early registered reports that have been published; preliminary results suggest that they are cited just as often as traditional papers. Still, he notes that relying too heavily on preregistered studies could encourage safer research, potentially overcorrecting the problem. He sees the model operating in tandem with the traditional results-focused model, one that is friendly to haphazard discoveries, the "accidental arrival of things," he says.

A harder problem to solve is the pressure for researchers to produce breakthroughs to make a living. A larger cultural shift would need to take place, Nosek notes. Right now it is not necessarily enough to carefully trod down intriguing paths that turn out to be empty, expanding the map of knowledge by illuminating the dead ends. We do not live in a world where fact-checkers become famous.

Yet the reproducibility problem does not necessarily mean that science is fundamentally broken. "Progress is dependent on failures," says Richard M. Shiffrin, a psychologist at Indiana University Bloomington, who is skeptical of the attention being paid to the "crisis." He argues that focus on irreproducibility stands to overshadow the advances that science has brought us. Those who do see the crisis as real do not always disagree with his assessment. Begley notes that the problem has real consequences: so many findings fail under scrutiny that drugs are arriving slower and at higher costs than they would under a cleaner system. "We spend a lot of time chasing red herrings," he says.

The effects in the coffee study turned out to be one of them. Corker's work, which she completed with hot and cold pads, ultimately showed there was no evidence that holding something warm could make you act warmer. Although the original work appeared in a topflight journal, the replication effort can be found in a comparatively smaller one. It was a breakthrough of a different kind, one met with less pizzazz.

Shannon Palus is a freelance journalist and staff reporter at Wirecutter, which is part of the New York Times Company. Her work has appeared in Slate, *Popular Science*, the *Atlantic*, *Discover*, *Audubon*, Quartz, *Smithsonian* and Retraction Watch.