

We're All 'P-Hacking' Now

An insiders' term for scientific malpractice has worked its way into pop culture. Is that a good thing?



PHOTOGRAPH: YULIA REZNIKOV/GETTY IMAGES

IT'S GOT AN entry in the *Urban Dictionary*, been discussed on *Last Week Tonight with John Oliver*, scored a wink from *Cards Against Humanity*, and now it's been featured in a clue on the TV game show *Jeopardy*. Metascience nerds rejoice! The term *p-hacking* has gone mainstream.

Results from a study can be analyzed in a variety of ways, and p-hacking refers to a practice where researchers select the analysis that yields a pleasing result. The *p* refers to the p-value, a ridiculously complicated statistical entity that's essentially a measure of how surprising the results of a study would be if the effect you're looking for wasn't there.

Suppose you're testing a pill for high blood pressure, and you find that blood pressures did indeed drop among people who took the medicine. The p-value is the probability that you'd find blood pressure reductions at least as big as the ones you measured, even if the drug was a dud and didn't work. A p-value of 0.05 means there's only a 5 percent chance of that scenario. By convention, a p-value of less than 0.05 gives the researcher license to say that the drug produced "statistically significant" reductions in blood pressure.

Journals generally prefer to publish statistically significant results, so scientists have incentives to select ways of parsing and analyzing their data that produce a p-value under 0.05. That's p-hacking.

"It's a great name—short, sweet, memorable, and just a little funny," says Regina Nuzzo, a freelance science writer and senior advisor for statistics communication at the American Statistical Association.



COURTESY OF CARDS AGAINST HUMANITY

P-hacking as a term came into use as psychology and some other fields of science were experiencing a kind of existential crisis. Seminal findings were failing to replicate. Absurd results (ESP is real!) were passing peer review at well-respected academic journals. Efforts were underway to test the literature for false positives and the results weren't looking good. Researchers began to realize that the problem might be woven into some long-standing and basic research practices.

Psychologists Uri Simonsohn, Joseph Simmons, and Leif Nelson elegantly demonstrated the problem in what is now a classic paper. “False-Positive Psychology,” published in 2011, used well-accepted methods in the field to show that the act of listening to the Beatles song “When I’m Sixty-Four” could take a year and a half off someone’s age. It all started over dinner at a conference where a group of researchers was discussing some findings they found difficult to believe. Afterward, Simonsohn, Simmons, and Nelson decided to see how easy it would be to reverse-engineer an impossible result with a p-value of less than 0.05. “We started brainstorming—if we wanted to show an effect that isn’t true, how would you run a study to get that result without faking anything?” Simonsohn told me.

They produced their absurd conclusion by exploiting what they called “researcher degrees of freedom”: the little decisions that scientists make as they’re designing a study and collecting and analyzing data. These choices include things like which observations to measure, which variables to compare, which factors to combine, and which ones to control for. Unless researchers have committed to a methodology and analysis plan in advance by preregistering a study, they are, in practice, free to make (or even change) these calls as they go.

The problem, as the Beatles song experiment showed, is that this kind of fiddling around allows researchers to manipulate their study conditions until they get the answer that they want—the grownup equivalent of kids at a slumber party applying pressure on the Ouija board planchette until it spells out the words they’re looking for.

A year later, the team went public with its new and better name for this phenomenon. At a psychology conference in 2012, Simonsohn gave a talk in which he used the term p-hacking for the first time.

“We needed a shorter word to describe [this set of behaviors], and *p-dash-something* seemed to make sense,” Simmons says. “P-hacking was definitely a better term than ‘researcher degrees of freedom’ because you could use it as a noun or an adjective.”

The phrase made its formal debut in a paper the team published in 2014, where they wrote “p-hacking can allow researchers to get most studies to reveal significant relationships between truly unrelated variables.”

They weren’t the first to identify what can go wrong when scientists exploit researcher degrees of freedom, but by coining the term p-hacking, Simonsohn, Simmons, and Nelson had given researchers a language to talk about it. “Our primary goal was to make it easier for us to present our work. The ambitious goal was that it would make it easier for other people to talk to each other about the topic,” Nelson says. “The popular acceptance of the term has outstripped our original ambitions.”

“It is brilliant marketing,” says Brian Nosek, cofounder of the Center for Open Science. The term p-hacking brings together a constellation of behaviors that methodologists have long recognized as undesirable, assigns them a name, and identifies their consequence, he adds. Nosek credits the term with helping researchers “organize and think about how their behaviors impact the quality of their evidence.”

As a wider conversation about reproducibility spread through the field of psychology, rival ways of describing p-hacking and related issues gained attention too. Columbia University statistician Andrew Gelman had used the term “the garden of forking paths” to describe the array of choices that researchers can select from when they’re embarking on a study analysis. Data mining, fishing expeditions and data dredging are other descriptors that had been applied to the act of p-hacking.

IT COMES BEFORE "-VALUE" TO INDICATE THE STRENGTH OF EVIDENCE IN SOME EXPERIMENTS, OR BEFORE "-HACKING" IF DATA ARE MANIPULATED

PHOTOGRAPH: JEOPARDY PRODUCTIONS, INC.

Gelman and his colleague Eric Loken didn't care for these alternatives. In 2013, they wrote that they "regret the spread of the terms 'fishing' and 'p-hacking' (and even 'researcher degrees of freedom')," because they create the "misleading implication that researchers were consciously trying out many different analyses on a single data set." The "garden of forking paths," on the other hand, more aptly describes how researchers can get lost in all the decisions that go into data analysis, and not even realize that they've gone astray.

"People say p-hacking and it sounds like someone's cheating," Gelman says. "The flip side is that people know they didn't cheat, so they don't think they did anything wrong. But even if you don't cheat, it's still a moral error to misanalyze data on a problem of consequence."

Simmons is sympathetic to this criticism. "We probably didn't think enough about the connotations of the word 'hacking,' which implies intentions," he says. "It sounds worse than we wanted it to." He and his colleagues have been very explicit that p-hacking isn't necessarily a nefarious endeavor, but rather a human one, and one that they themselves had been guilty of. At its core, p-hacking is really about confirmation bias—the human tendency to seek and preferentially find evidence that confirms what we'd like to believe, while turning a blind eye to things that might contradict our preferred truths.

The "hacking" part makes it sound like some sort of immoral behavior, and that's not helpful, Simmons says. "People in power don't understand the inevitability of p-hacking in the absence of safeguards against it. They think p-hacking is something that evil people do. And since we're not evil, we don't have to worry about it." But Simmons says that p-hacking is a human default: "It's something that every single person will do, that I continue to do when I don't preregister my studies." Without safeguards in place, he notes, it's almost impossible to avoid.

Still, there's something indisputably appealing about the term p-hacking. "You can't say that someone got their data and garden-of-forking-pathed it," Nelson adds. "We wanted to make it into a single action term."