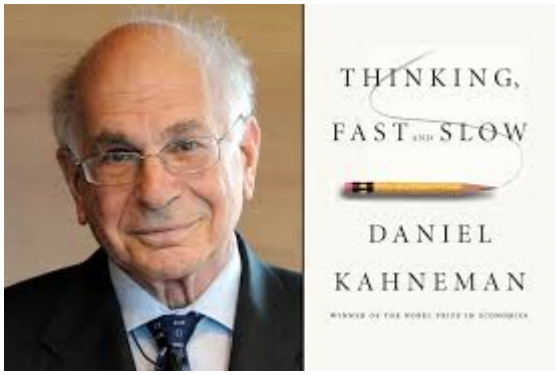

Daniel Kahneman: “train wreck”

You might think that’s a funny question, because it is all the rage now in academia. Last month’s Harvard Law Review featured an article by Ryan Bubb & Richard Pildes arguing that behavioral economics does not go far enough. But an article in this month’s Pacific Standard by Jerry Adler reveals a growing problem with its experimental data. (Pacific Standard is an outstanding magazine, by the way: you should read it). The problem has become so rampant in behavioral psychology that the field’s founder, Nobel prizewinner Daniel Kahneman, warned in an e-mail to colleagues that the situation threatens to turn the whole discipline into a “train wreck.”

Anyone who does field research can engage in outright fakery, but according to Adler, a particular kind of deception appears to have become rampant in behavioral psychology: “P-hacking.” “P” represents a number measuring the probability that a particular result occurred by chance — that “nature is throwing you a curve”, in the article’s parlance — and scholarly convention holds that your “p-number” must be less than 5%, i.e. there is a less than 5% probability that your results came about by chance:

> How does *p*-hacking work? One common approach is called “data peeking,” a technique that involves taking advantage of the real-time chance fluctuations that nature throws your way as you’re conducting an experiment. For instance, rather than determining in advance how many subjects you will test, you might pause after every five or 10 or 20 to analyze your data up to that point, and you stop when you’ve gotten the results you want. Or maybe the effect you’re looking for shows up in women but not in men, so you only report women’s results. Or you correct for subjects’ height—maybe tall people are less affected by standing on a stage, since they’re used to looking down on everyone else anyway. Or maybe the hot sauce wasn’t hot enough, so you start over with a new batch and declare it a new experiment. In all cases, the old results go in a cabinet that gets

> emptied into a dumpster when you move offices, illustrating what scientists call "publication bias"—the selective reporting of positive reports—or, more colloquially, the "file-drawer effect."

P-hacking allowed a psychologist Joseph P. Simmons and a couple of colleagues to "prove" a result that was downright ridiculous: if undergraduates listened to the Beatles' "When I'm Sixty-Four," this would make them younger:

> To get the results they wanted, they divided their sample of 34 undergraduates into three groups and played for them either the Beatles track, an instrumental called "Kalimba," or "Hot Potato," a children's song by The Wiggles. Thus they could compare results in four different ways: each song matched against one other or all three together. Then the scientists looked at all the answers they collected from their questionnaire. With hundreds of possible comparisons, it was, Simmons says, "highly likely" that they would find at least one pairing that showed just the sort of statistically significant correlation they were fishing for. This turned out to be the group that heard "When I'm Sixty-Four" versus the "Kalimba" group, "adjusted" for the ages of the subjects' fathers.

Adler describes a series of major papers that have been withdrawn because reviewers discovered the p-hacking. **Note to self**: when you read a paper where the p-value is just under 5%, be careful. If that author has a lot of them, be *very* careful.

But here is the good news: a group of scientists has developed a sustained campaign called the Reproducibility Project that seeks to, well, reproduce the results of famous papers. I remember as a kid my science teachers would have on their desks a satirical scholarly publication entitled the *Journal of Irreproducible Results*, but Adler reports that actually reproducing results run counter to academic trends: you get no points for originality of topic or inquiry, and if you wind up *confirming* results you have not made any new findings. So the Reproducibility Project is crowdsourced and involves the collaboration of laboratories that usually compete. Earlier attempts to engage in reproducibility work on a smaller scale found that not only could classic behavioralist studies be fully replicated, if anything they *under*estimated their findings. That is very important news.

Science is ***supposed*** to work by being reproducible, but the professional and institutional pressures mean that there are huge disincentives to do so. That actually sounds like the sort

of finding you would might get from behavioral economics. And unlike neoclassical economics, particularly of the freshwater variety, the behavioralists are taking embarrassing failures to heart. If in half a century, we find that psychology has managed to supplant economics as the master science, the success of the Reproducibility Project might form a large part of the answer.