Ben, The One Simple Step To Eliminate Data Fraud

The recent discussions of irregularities in a dataset created by UCLA PhD student Michael LaCour, which led to a publication in *Science* with Columbia Professor Don Green, have begun to focus on what sorts of policies can be implemented to avoid the fradulent creation of social science data. Some suggestions have included more thorough verification of data and code (as is currently being done by the *American Journal of Political Science*), asking for more thorough peer review, requirements for greater transparency in data and analysis, and instituting stricter oversight of graduate student research.

By and large, none of these policy proposals will be particularly effective in preventing or capturing fraud after it has been committed. Unless one is looking for signs of fraud, it is often difficult to spot. Given the rarity of fraud, that's also a massive investment in time and resources to solve a rare problem. The best way to avoid fraud is to take a preventative approach. In other words, creating institutional structures that prevent those intending to commit scientific fraud from being able to do so. One strategy I particularly like is the notion of open science notebooks, which in addition to preventing fraudulent analyses also provide a check against p-hacking, multiple comparisons, and "the garden of forking paths". But even this level of transparency doesn't stop the kind of fraud alleged to have occurred in this case: that is, the generation of fake raw data before any analysis occurred.

The evidence raised by David Broockman, Josh Kalla, and Peter Aronow about LaCour's data suggests that it was likely produced through the addition of simple random noise to an existing public opinion dataset. No measure of posthoc review or transparency is going to be able to fully prevent an individual from committing fraud at this stage of the process. But, a surprisingly simple policy can prevent that and its name is Ben.

When I was a PhD student at Northwestern University, I spent a lot of time collecting data in the Political Science Research Laboratory with a variety of research subjects

(students, university staff, and the general public) for my own research and for others. Ben, an undergraduate research assistant at the time, was a pivotal figure in these data collection efforts. My advisor's policy was that all data collection had to go through Ben. This meant that while I could design a survey-experimental questionnaire and program it into SurveyMonkey (the tool we used at the time), Ben had to extract the data from SM and send it to both my advisor and me in raw format.

For every study, Ben would reset our SM account with a private password that only he had access to. When the study was over, he would extract the data, send it to us via email to create an electronic record trail, and then reset the password to a shared key so that we could move on to the next study or match the received data against the original. We could therefore trace the data from its digital origins into our analysis workflow. We could only work with data that came through Ben and we had no other way of obtaining those data because we intentionally locked ourselves out of our own data until the study was over. Ben had no incentive to fabricate data because he had no stake in the results and typically was not informed about the nature of the research itself so it would have been impossible for him to fabricate, even if he wanted to.

While this is a fairly low tech solution to an important problem that might otherwise be solved through complicated digital record-keeping, it was one that worked incredibly well. Ben was a simple, affordable institution we created to hold ourselves accountable. I would highly recommend all research labs get one.

Published: 2015-05-27

[Feed]

← Older post

Newer post →
