Critical Finance Review, 2019, 8: 11–13

Editorial: Replication? Do You Even Have Access to the Data?

Matthew Spiegel

Yale School of Management; matthew.spiegel@yale.edu

You see an article and wonder: Are the results due to a "carefully selected" data filter? Would a small change negate or overturn the article's primary point? Perhaps an innocent coding error explains everything? Maybe you have reason to suspect that the data was simply fabricated? In all of these cases, what is the very least that you need to determine if your suspicions are right? The data. Having it available (even on a very limited scale) is the most basic criteria any article claiming scientific validity needs to meet. Results must be verifiable. Journals should insist on it.

At a minimum, one can claim a paper is verifiable if somebody in the academic community (beyond the paper's authors and close colleagues) has access to the original data. Right now, our profession does not mandate even this limited standard. To that end, I would suggest journals only consider papers that use data accessible by some third party academic, somewhere, under conditions similar to what the paper's authors went through to get it (If need be, with a limited embargo period after publication). When I say "some academic," I do mean some. One would be fine in my view. When I say "somewhere, under conditions similar to the paper's authors," that should be taken literally. If the authors paid $250 million for the data, well then the condition would be to pay $250 million. If the authors had to travel to a faraway government office and run the code on that government's computers, that is the condition to meet. A rule like this would make it possible for at least one academic, without a vested interest in the article's success, to check its results. Nothing more. Even this weak policy would impose some minimal level of verifiability on published papers.

Objections to data availability rules typically revolve around the fact that few databases are universally available. Nothing proposed here would add any hurdles to the publication of papers based on data many cannot access. Sorry, that is life. Even a widely used database like CRSP costs a significant amount of money. If a researcher's school does not subscribe and the funds are not otherwise available, then that person does not have access. No university or academic has access to every database in common use and that seems unlikely to change any time soon. Fortunately, verifiability does not require universal access to a paper's data, just access by some third party somewhere.

ISSN 2164-5744; DOI 10.1561/104.00000079

The policy outlined here is, I believe, the minimal policy an academic science needs to ensure disinterested parties can confirm its output. Does that mean there are better, more stringent rules one might consider? Sure. While access to a database is the very least one needs to check a paper's results, having a copy of the computer code would be enormously helpful. One issue is how to go from the current policy in which replicability may be impossible to one where it is. In my experience, even the data availability proposal outlined above will generate protests from a large fraction of the academic finance profession. If we are going to make an initial call for potential replicability, then I would suggest we initially implement the weakest policy that accomplishes that goal, in the hope that also minimizes opposition.

Beyond data availability, there are also increasing calls to have the code behind a paper made public. I suspect that a code availability policy will prove to be a bridge too far and we will remain with the status quo. In part, people object to publically releasing their code because it will impose yet another publication requirement on top of an already oppressive process. Today, it can take years for a paper to go from submission to publication. During that time, referees demand extensive rewrites and additions that authors have to comply with. Add to that the tendency of referees to reject articles that failed to test for some implausible alternative and today's articles easily hit 100 pages in total when you include the various appendices. Those 20 tables, most of which nobody will ever look at, likely required well over 20 programs to produce. Telling authors to provide the code is going to induce a lot of teeth gnashing. Let us be honest, nobody is going to release the code they initially used to produce a table. It would be embarrassing! Most academics write code to use once and never again. It likely lacks the necessary commentary to explain what each section or line does. It probably contains a lot of debugging code as well. Any self-respecting author is going to want to add the former and clean out the latter—an enormous job! This is added to the time and effort it already took to produce the final 100 plus page paper along with numerous rewrites forced upon the authors during the editorial process.

If people really want to move to an equilibrium where authors happily make their code available, we need to reduce other (far less useful) demands on authors. Start with letting the authors write their papers rather than the referees. If a referee produces a 5-page report with 30 to do items, tell the authors they can choose which to follow, after which the paper will be published. That alone would likely reduce paper lengths considerably. Ideally, we should get papers back down to 20 or 30 pages. With just three to five tables, authors would have only a small number of programs to clean up post-acceptance. Having spent far less time producing the published paper, perhaps the additional burden of providing the code will not seem so daunting. Indeed, I suspect most of the profession would jump at a trade like this. For their part, journals agree to limit referee reviews and

keep papers to 30 manuscript pages. In exchange, authors have to explain how others can obtain their data and code.

Alas, I think that the journals are unlikely to crack down on referees any time soon. That means authors will continue to produce 100 plus page papers in self-defense. In the meantime then, how about we just insist on making it possible for someone, somewhere to independently access the paper's data so that they can see if the results hold up. It seems like the least we should require if we are going to insure the veracity of published results.