Editorial

Ivo Welch

July 1, 2019

The reexaminations in this CFR special volume succeeded in replicating the original papers but failed on reproducing them. Moreover, they not only failed to reproduce the original results but also failed to reproduce one another. These failures demonstrate that it is usually impossible to reproduce findings without the archived code and data.

My editorial is arguing that it has too often been not in the interest of authors to share code and data. As a profession, we can and need to do more to address our public goods problems, both collectively and individually. Fortunately, our journals have either recently or are now in the process of adopting better policies. They can still do more, but they are at least off to a (late) start.

Some of the questions that I have been struggling with are:

1. Should academic journals ever publish papers for which authors do not provide data and code? This question applies not only to papers using widely available data but also to papers using proprietary data.

2. Is it even science if no skeptical party can reproduce findings, and all we can count on is the “word” of the author (that he did not commit a mistake)?

3. Is it even science if skeptical parties could have but in actuality never have reproduced findings, and all we have to count on so far is the “word” of the author?

4. What kind of corrections and updates are worth publishing?

5. How can we encourage skeptical examination? What roles should the original authors be allowed in approving them (as referees)?

6. When should our own papers still cite publications if the authors do not provide code and data?
Reproductions in Science

Contrary to popular beliefs and occasional bold and misleading claims that papers have been “vetted,” economic journals have been playing almost no role in confirming the evidence in the papers which they are endorsing by publishing them. The referees and editors are paid little and have no obligation to reproduce any findings. (It is not unusual for referees to accept or to reject incredulous findings on face value, both without any evidence!)

Thus, a paper that has “survived” peer review is effectively endorsed by the journal without independent evidence about the papers’ claims, one way or another. It is not a question of journals not having enough resources. Journals could “pay” researchers in the currency of pages dedicated to reproductions and replications only of accepted papers. But they have deliberately chosen not to do so. (This was one of the reasons for the founding of the CFR.)

Taking results on faith may not be ideal individually, but it is inevitable. With finite time and resources, every researcher must stand on the shoulders of others. No one can singularly reproduce and confirm all findings on which their research and knowledge are based. Thus, the decision of our journals not to invest resources for independent skeptical verifications of all papers they publish hurts all our readers. We should all insist on actual reproductions and replications before we accept each others’ evidence, build on other findings, and cite papers as valid precedence.

It is unlikely that trained economists are less aware of incentives to cheat than scientists in other disciplines. Even if researchers do not commit fraud or cherry-pick results, only skeptical reproductions can confirm that there were no honest mistakes, such as data problems or bugs. Curiously, unlike in other fields, there have been almost no retractions in financial economics—voluntary or involuntary. I doubt that this is because financial economists have been much better scientists than those in other disciplines.

The possibility of incorrect results greatly limits the usefulness of research based on proprietary data. In my opinion, other parties must not just be able to reexamine the data, they must actually have reexamined the data. Karl Popper’s famous quote that “non-reproducible single occurrences are of no significance to science” still applies. Findings are only worthy of building on if they have been reexamined by parties with different incentives. Sunshine is not only the best disinfectant; it is the only disinfectant. Again, from the perspective of our profession, any findings that have not been reproduced by at least a few skeptical teams should be considered “speculative” or even “faith-based”—at least until they have been reproduced.

Although most of our interest is in understanding economic behavior, I also have many academic meta-concerns. I am concerned about lack of processes correcting “bad habits drifts.” What if an original researcher obtained an incorrect sample, or forgot to remove missing values, or poorly defined a variable (e.g., forgetting to add dividends when calculating stock returns)? Can these practices now be considered valid precedence? Should this researcher still get full credit for the findings if they are later found to hold up if/when executed

---

1 And, if caught, they are trained to explain why our profession just got what it deserved. If the profession had cared more, it would have instituted better incentives! They just wanted to compete.

2 Replicability is not enough. Only replication is. Moreover, one replicating team is not enough, especially because subsequent teams are aware of the original study and thus already start with a prior. Such priors increase the incentive to stop digging when the results look similar compared to when they look different.
correctly? Does it matter whether the miscoded missing values were 1% of the sample or 99%? What if all the code was wrong and the coefficient and signs just happened to have been correct? What if the T-statistic was overreported as 2.5 instead of 2.0? What if a loose data definition or careless technique, which could have but luckily made little difference to the inference of the original paper, was then adopted in dozens of follow-up papers, where eventually more and more follow-up results depend critically thereon? (The next paper may need the extra and incorrect 0.5 for its significance, and now it can point to good precedence!) Is there any utility to and should (top) journals still publish correct versions, given that a poorly executed paper had a lucky guess on the correct inference? Do the correcting authors deserve no, some, or all of the intellectual credit? How do we judge the marginal contributions of two papers with similar findings, one original paper executed earlier but poorly (the correct findings could plausibly have been coincidence), and one follow-up paper executed later but well (the findings can be considered reasonable and valid)? When do the means justify the end? How important does our profession consider care and incremental improvement?

Mistakes in research can be insidious, in that they may not matter greatly in the beginning but can compound when they become widely adopted. Eventually, they can become accepted practice (often making it easier to find “results”) and many researchers and referees have stakes in them (Holderness (2016b)).

A Hypothetical Case Study

I want to illustrate our (collective) professional dilemma with a specific hypothetical case. A researcher (she) wants to study whether the X-Y results another researcher (he) published are related to variable Z. Another decade has passed since X-Y was published. She spends one year recollecting and updating the X-Y data, because he has not made it available. When she attempts to reproduce his results, she finds that they do not have the reported T-statistic of 2.2 (a p-value of 3%), but only a T-value of 1.5 (20%), both in the original and the extended sample.

How should she proceed?

• Without the ability to check the original code and data, can she say that she is right and he was wrong?

• Her ultimate goal was not a reproduction but a replication. If she cannot find X-Y in her new data, how can she autopsy the problem if she cannot reproduce the reported baseline? Is her problem the different technique, the new data sample, an error by the original authors or her, or something else? (Should the literature continue to believe X-Y, and is Z still worth considering?)

• Does she have an ethical obligation to correct the record by spending her next year(s) writing up and trying to publish the insignificant X-Y results (especially if it dismays the original author)?

• If other papers—written, published, or refereed by him—later claim similar results in other contexts, is the 1.5 finding now obsolete? Can this author (as referee) reject her paper on the premise that X-Y evidence was found in other contexts or is each paper its own valid target?[2] This situation is not just a hypothetical question, but a situation that I had to contemplate. The target was a junior researchers, whose career depended on his paper critiquing an influential paper by authors in very
influential positions. They and some of their friends actively lobbied me and others for support for their position.]

How should our journals and profession proceed?

- Why do popular magazines employ fact-checkers, while even our most profitable scientific journals have not invested any resources into fact checking? Why do they not employ an assistant whose job it is to reproduce and replicate every study before publication?

- When journals endorse a paper by publishing it, do they have a responsibility to report when good-faith attempts to reproduce the results fail?

- If not at the original journal, where do we expect the peer review process to correct failures? Do other journals share the responsibility?

- The original paper presumably had a finding that was important enough to warrant publication. Is this kind of T=1.5 vs. T=2.2 discrepancy enough to warrant publication? At what decline in significance does a correction warrant publication?

- Is it important to update our literature to newer data? What if the old data had stayed at T=2.2, but after including more data, the cumulative T dropped to 1.5? At what decline in significance does it warrant correcting the record of evidence?

- Should the original author (with his T=2.2 findings) ever be allowed to referee this paper?

- Will the journal ask the original author to referee the submission anonymously, effectively allowing him to block her publication?

Sidenote: Does she deserve an objective referee reports? I am appalled that our journals believe that the answer is no. He (or another member of his research clique) would not be disqualified as an anonymous referee, despite the obvious conflict of interest. With only 1 in 10 papers accepted, it is most likely that he can block publication.

Some of my friends who are editors have argued that they are aware of and accept bias, because it gives them the most informed referees. However, this is a specious argument. First, the evidence strongly suggests that editors of the most competitive journals tend to reject papers for which they have any negative referees. Second, the editors could still consult him, but disclose his identity to the authors. His opinions should not influence the decision and it should not be presented to her as an objective evaluation. (This is the policy at the CFR.)

The wide use of conflicted referees can shape a literature just as conflicted papers can. I consider the existing conflicts of interest by referees to be worse than the conflicts of interests by authors. Unlike author conflict of interest, referee conflict of interest is never disclosed and remains completely opaque to the authors and to the profession. I know that The practice of consulting critiqued authors as anonymous referees has also been widespread. It has resulted in hundreds of conflicted decisions and created deep skepticism about the scientific merit of the refereeing process.
• What is the discipline-wide incentive effect on writing papers when authors know that previously published authors become referees on the same subject? Is it a problem when one particular perspective becomes (relatively) easier to publish as more and more papers with the same perspective are being published?

• How will the profession ever learn whether X-Y holds or not?

• What if there have been twenty other attempts, which came up with T-statistics of between 1.0 and 2.0? Will all or any of these results be published or even be written up? Should journals publish many of them, so readers can eventually learn the truth by adopting a published majority view?

• Do we owe it to other (and especially junior) researchers to try to protect their research careers from suffering the same one-year data collection fate on X-Y later on?

• Do we grant absolution to papers with p-values of 20% (instead of 3% that they claimed)?

If the original author found that the p-value was 20% but reported 3% (perhaps even intentionally), it means that the hypothesis is true 4 in 5 times. More powerful tests and years of more data collection will eventually bring it down to 0. Does he really deserve the intellectual credit (in 4 out of 5 cases) for sloppiness and misconduct if the results later turn out correct? Only reproductions (not replicability) can ever discourage sloppiness and misconduct.

• What will be the outcome if our processes keep rewarding sloppiness and make it difficult to report corrected results?

**Incentives Matter**

Without a tradition of disclosure and active policing, the probability of misconduct discovery is lower in financial economics than it is in our sister disciplines that have such traditions. It is an irony that we may well have the worst incentives. After all, we have been trained to understand and exploit incentives. If authors do not make code and data readily available, they know that the hurdle to induce other skeptical researchers to check the accuracy and robustness of their published results is usually too high.

Moreover, if it ever were to happen that they were found out to have been careless (or worse, dishonest), we all know how they can contend that it was not their fault. They just responded appropriately: *If financial economics had cared about uncovering mistakes and fraud, and/or if they had wanted to incentivize care, they would have bothered with it. By revealed preference, they did not care.*

To make matters worse, academic researchers do not live in a system of absolute but relative standards. Our publication and peer evaluation system is competitive. It is more difficult for careful and honest research and researchers to survive if unscrupulous peers can publish faster and offer more amazing results.

And worse again, being alone in providing code and data paints a target on one’s back. Reverse p-hacking out of context (i.e., relative to other papers) may make it appear that these papers are worse than those that do not. When sharing is not mandatory, voluntary submission of code and data is a sign of great integrity,
regardless of whether their results ultimately held up or not. I thank all researchers who have done so in the past.

I suspect that, over time, journals and referees have even come to expect “sprucing up” of reported results. Research is expected to “put its best foot forward.” Over time, referees would have inferred that a more objective version of the results would likely be less remarkable. (And, probably, they usually are!) Over time, many researchers turned into advocates rather than scientists in order to survive. All of us have struggled with the gray area of which data presentation is and which is not ethical and customary. Over time, submissions to our best journals, which only accept one in ten submissions, were printing ever-more remarkable results. More boring results were no longer worthy. Their authors do not get tenure and disappear.

**Proprietary Data and Darwinian Selection**

Many researchers would argue that we need papers based on proprietary data and thus grant exceptions to the code and a good data provision requirement for them. I consider this argument to be too dangerous.

There are great benefits to exempting proprietary-data studies, especially if others can, in principle, reproduce the study by obtaining access to the same proprietary data on similar terms. (Nevertheless, I still fear that the incentives to do so are insufficient, for all of the reasons given above.) Which proprietary data studies have actually been independently reproduced?

However, the costs of exempting proprietary-data studies are even greater. In a Darwinian world of natural selection, even if all authors examining proprietary data set were without exception honest, over time, the papers most likely to be published would be those with the most amazing results due to honest data or coding errors or simply sampling variation. Given the journal process, with its assignment of referees of future papers to these published authors, their unreexaminable findings may become gospel. That is, other authors using other proprietary (and non-proprietary) data can expect to draw them as referees. They are unlikely to succeed in publishing contravening and especially “boring” contravening findings, for all the reasons above. Our peer review system has then effectively blessed a new sect in our religion.

Of course, methods that overstate results can also thrive when papers use non-proprietary data. The most important example may well be Holderness (2016b), which points out that the methods in La Porta et al. (1998) are flawed. Inappropriate aggregation greatly inflated their T-statistics. The subsequent adoption of their method inflated T-statistics and thus spawned thousands of subsequent flawed papers. However, at least Holderness (2016b) could point out the problem and study its effects.

How many published articles using proprietary data in financial economics have later been shown to have been wrong? Zero! Do we truly believe that the authors of such studies are much better and more careful than the authors of studies elsewhere, because they realize that their results cannot later be corrected?

---

3 The CFR would welcome submissions that can credibly defend the methods in La Porta et al. (1998) and explains why Holderness (2016b) (and Holderness (2016a)) is incorrect.
An exemption for proprietary data would push researchers towards exempted areas. Yes, without proprietary data studies, our profession loses some correct insights. Yet it is even more dangerous when false and unrepudiable insights can contaminate our chain of evidence.

There is a place for such studies, just as there is a place for anecdotes. It is on unrefereed websites, like SSRN, which accept all comers without any pretense of peer review, journal endorsement, or scientific replicability. Non-reproducible papers should be viewed only as suggestive. These “papers” can even gather citations. Alternatively, if possible, authors and journals could collaborate on incentivizing a skeptical reproduction by a truly independent team with strong incentives to disagree and permission to examine the same data.

The treatment of non-reproducible papers is a choice that the top journals, themselves hungry for citations, have to coordinate on.

**Pooling or Separating Equilibria**

Financial economics has been living in a pooling equilibrium. We have been hoping that the incidence of bad apples in our pooling equilibrium was small—without evidence one way or another. We must now collectively switch from a pooling to a separating equilibrium.

A separating equilibrium will make two differences.

1. It will reduce the incidence of sloppy and dishonest research. It is true that sending the signal—making data and code available in a user-friendly manner—imposes an individual cost on researchers. Yet this cost is higher for less scrupulous researchers—and only so if all researchers have to share code and data without exception. In a separating equilibrium, we should expect lower individual publication productivity and fewer incredible results, but more publications that are likely to survive skeptical replication.

2. Researchers will be able to use another’s code and data and not constantly have to reinvent the wheel. Socially, it will reduce extensive needless duplications. We will finally be able to stand on the shoulders of others.

No researcher alone benefits from deviating from pooling (by being in the minority of providing their own code and data). We can only benefit if our journals and personnel processes force disclosure on all of us to do so collectively.

**Conclusion**

The original three papers on liquidity, examined in this volume, were written in the early 2000s. It was a different era. The importance of skeptical reproduction was neither as well understood nor as easy as it is today. Computers and storage media were not very reliable. I am no exception. I am not holier than thou. I no longer possess the code or data for my own empirical papers from those years, either. I am as guilty as most of my peers.
However, it is now two decades later. What was appropriate in 2000 should no longer be acceptable in 2020. The effective cost of reliable data and code sharing has been coming down. We can archive code and data now. M-Disc provides media designed not to degrade for centuries. The cloud is offering cheap data storage. The journals can begin to play a role in the archival and provision of code and data.

Let me return to and conclude with my questions from the introduction.

1. I call on our journals not to publish papers if their data and code are not made available to independent skeptical researchers.
2. I call on our journals to commission and publish at least one skeptical reproduction and replication for every empirical publication.
3. I call on our profession not to consider any research finding valid until at least one independent skeptical replication confirms the results.
4. I call on our profession to adopt a culture of progressive improvements (correcting definitions and methods) before they become insidious, and to regularly update empirical results.
5. I call on our journals to develop a culture of retractions and corrections.
6. I call on all of us individually not to name articles in their reference sections (bibliographies), or not to cite them at all, if the authors are not sharing data and code and reproduction is difficult or comes up with different results. Such articles no longer deserve intellectual credit. If anything, their presence is poisoning the well. It makes it more difficult for other scholars to publish similar research, which our profession can then vet, trust, and build on.⁴

References


⁴ In the future, the CFR may solicit papers that quasi-replicate published studies and to do so without citing the “original.” The CFR studies will share their code and data and have credible findings.