Critical Finance Review, 2019, 8: 1–9

Editorial: Replication in Financial Economics

Campbell R. Harvey*

Duke University and NBER, USA; cam.harvey@duke.edu

ABSTRACT

All of the top general-purpose economics journals have a data and code-sharing policy. As of this writing, the *Journal of Finance* has a code-sharing policy, the *Journal of Financial Economics (JFE)* requires authors to share code if the results are challenged, and the *Review of Financial Studies (RFS)* is currently developing a policy.[1] Replication is essential to any scientific endeavor, so why has the area of financial economics lagged behind?

Keywords: Replication, Research Culture, Hard misconduct, Soft misconduct, Proprietary data, Robustness, P-hacking, Reverse p-hacking, Falsification, Fabrication, Plagiarism

JEL Codes: G00, G01, G10, G11, G12, G13, G14, G15, G17, G18, G19, G21, G22, G23, G24, G28, G29, G31, G32, G33

[*] I have benefitted from conversations with many of my colleagues over the years, particularly the detailed email exchanges in 2010. For this current paper, I appreciate the input of Ivo Welch, David Hirshleifer, Manju Puri, Itay Goldstein, Matt Spiegel, Toni Whited, Anna Cieślak, A. Subrahmanyam, and Jan Höffler. Kay Jaitly provided editorial assistance.

[1] The *American Economic Review* policy (which applies to all AEA journals, originated in 2005 and was revised on July 10, 2019) is available at https://www.aeaweb.org/journals/policies/data-code. It includes a central role for the AEA Data Editor, a data and code repository, a generic checklist for authors, a template for replicating teams, as well as a protocol that allows for reproduction via a third party if the data cannot be posted online. The *Journal of Finance* code-sharing policy is available at https://afajof.org/wp-content/uploads/files/policies-and-guidelines/CodePolicy.pdf. The policy for the *Journal of Financial Economics* is contained in the acceptance letter: "We expect authors to cooperate by providing code if a controversy arises where 'replicating authors' cannot replicate the published work and the results of the replication attempt make the conclusions of the published paper invalid."

ISSN 2164-5744; DOI 10.1561/104.00000080

1 Introduction

My role in advocating for data and code-sharing policies began in 2010 when I was editor of the *Journal of Finance*. Collaborating with the editors of the *JFE* and the *RFS*, I proposed a policy that could potentially be adopted across all of the top three finance journals (and one that many other finance journals could also adopt). The policy had the following elements: (1) nonproprietary data would be posted; (2) code would be posted; (3) the author would not be expected to "support" (answer questions about) the data or code; and (4) the editor would have discretion to allow for a period of exclusivity if the author was using the data for other projects.

The draft policy was circulated to associate editors and prominent members of the finance profession. The initiative failed.[2] It is important to understand some of the reasons it failed.

2 Cost

By far the most important pushback was the cost imposed on authors.[3] Even though authors were not expected to support the code or data, it seemed unrealistic to think they would ignore inquiries such as "There is a mistake in your code"—even if there was no mistake. In addition, the greatest cost would be borne by the most productive and influential authors. The case that this was optimal for the profession was not obvious in that it could reduce the amount of innovation. Indeed, there were powerful testimonials from top academics about their nightmarish experiences in sharing data. Many believed we did not have a problem. Why impose a costly "fix" to something that was not broken?

3 Finance is Different from Economics

In contrast to economics, a field in which much of the research uses freely available macroeconomic data, financial economics largely relies on proprietary data. Though CRSP and Compustat are widely available, they are proprietary. Many other data sources that we use in financial economics are also proprietary. Having a strong data-sharing policy would potentially change the direction of research. Would we prefer to learn something from an interesting but highly proprietary

[2] See Harvey (2014).

[3] Duvendauck *et al.* (2017), reflecting on the progress since the publication of Dewald *et al.* (1986), find that only 28 of 333 economics journals regularly make data available and explore the reasons why. Gertler *et al.* (2018) recently surveyed 203 economics papers and found that fewer than one in seven provided materials needed for replication.

database—even if the data are not available for replication? Most thought the answer was yes.

4 The *QJE* Model

In 2010, all of the top journals in economics had a data and code-sharing policy except for the *Quarterly Journal of Economics (QJE)*. Many believed this contributed to the *QJE*'s competitive advantage. Many believed that the journal received interesting papers that used proprietary data, which could not be submitted to the *American Economic Review*. The competitors to the top finance journals are not just the other top finance journals, but also the economics journals. A more liberal view of data and code sharing could potentially lure excellent papers from the economics journals to the top finance journals. While the *QJE* now has the identical sharing policy as the *AER*,[4] the argument of attracting papers that would usually be headed to a top economics journal to a top finance journal still holds.

5 Other Fields

Our colleagues in the physical and biological sciences would be perplexed by the history of the replication debate in finance. If an author refuses to share the original data in their fields, it is highly likely that their paper would be retracted. Retracted papers are much less likely to be cited, with a few unfortunate exceptions, such as the Wakefield study published in *The Lancet*. Many followed the demise of (former) Cornell food science professor Brian Wansink. So far, he has had 40 papers retracted.[5] Last May, the prestigious *Journal of the American Medical Association* summarily retracted all six of his publications because "we do not have access to the original data." Retractionwatch.com lists more than 10,000 retractions from peer reviewed journals. A large number of retractions are due to scientific misconduct—not honest mistakes.[6]

There have been no retractions of papers printed in any of the top three finance journals.

6 Hard and Soft Misconduct

On December 6, 2000, the Office of Science and Technology Policy in the White House published the *Federal Research Misconduct Policy*[7] which required all federal

[4] https://academic.oup.com/qje/pages/Data_Policy

[5] http://retractiondatabase.org/RetractionSearch.aspx#?auth%3dBrian%2bWansink

[6] See Fang *et al.* (2013).

[7] https://ori.hhs.gov/index.php/federal-research-misconduct-policy

agencies or departments that were supporting research (both within the departments or agencies and outside, such as universities) to implement misconduct policies within a year. One prominent example is the US Department of Health and Human Services' Office of Research Integrity,[8] which is charged with monitoring the scientific quality of federally funded research in public health. These federal misconduct regulations have been promulgated in individual universities' policies and procedures.[9] Most of the focus, however, is on "hard" misconduct: fabrication, falsification, and plagiarism—often referred to as FFP.[10]

I believe we (in financial economics and other fields) have a much bigger problem with so-called soft misconduct (e.g., strategic data choices and p-hacking; Harvey, 2017).[11] Indeed, it is insightful to revisit the Wansink case in which no accusations of fabrication, falsification, or plagiarism were made. The concern was about soft research misconduct. While my example is drawn from food science, other fields such as social psychology openly talk of a "replication crisis" because many of their key results have failed to hold up.[12] For example, Open Science Collaboration (2015) revisited 100 studies published in top psychology journals of which 97 reported "significant" results. The replication found only 36 significant effects.

7 Elmo Stickers

Consider the research of Wansink *et al.* (2012). They tested the hypothesis: If apples have an Elmo sticker on them, children aged 8 to 11 are more likely to eat an apple than a cookie. Preliminary research was conducted in 2008. Coauthor Payne emails Wansink in September 2008:[13]

> I have attached some initial results of the kid study to this message for your report. Do not despair. It looks like stickers on fruit may work (with a bit more wizardry).

[8] https://ori.hhs.gov/research-misconduct

[9] See, for example, Duke University's version at https://provost.duke.edu/sites/all/files/FHB_App_P.pdf (Section R on page 34).

[10] The federal policy defines these terms as follows: "Fabrication is making up data or results and recording or reporting them." "Falsification is manipulating research materials, equipment, or processes, or changing or omitting data or results such that the research is not accurately represented in the research record." "Plagiarism is the appropriation of another person's ideas, processes, results, or words without giving appropriate credit." Also important is the following statement: "Research misconduct does not include honest error or differences of opinion."

[11] A broader definition of misconduct is detailed in National Academies of Science (2017, p. 65). They detail "other serious deviations" or non-FFP misconduct. Resnick *et al.* (2015) provide misconduct definitions in 22 different countries. The UK policy is much broader than the current US policy.

[12] See, for example, Andrew Gelman at https://statmodeling.stat.columbia.edu/2018/05/07/replication-crisis-centered-social-psychology/

[13] See https://www.buzzfeednews.com/article/stephaniemlee/brian-wansink-cornell-p-hacking

The research continued. Wansink emailed on January 7, 2012.

> One sticking point is that although the stickers increase apple selection by 71%, for some reason this is a p-value of 0.06. It seems to me that it should be lower. Do you want to take a look at it and see what you think? If you can get the data, and it needs some tweaking, it would be good to get that one value below 0.05.

The paper was published in a top journal, *JAMA Pediatrics*, with the original p-value. It was self-retracted and replaced with a new version[14] on September 17, 2017, with a new p-value of 0.02.

The paper was fully retracted in December 2017 after many readers questioned Wansink's research practices.[15]

Consider the emailed advice that Wansink offered to a PhD student:[16]

> P-hacking shouldn't be confused with deep data dives—with figuring out why our results don't look as perfect as we want.

The following is even more alarming:

> With field studies, hypotheses usually don't "come out" on the first data run. But instead of dropping the study, a person contributes more to science by figuring out when the hypo worked and when it didn't. This is Plan B. Perhaps your hypo worked during lunches but not dinners, or with small groups but not large groups. You don't change your hypothesis, but you figure out where it worked and where it didn't. Cool data contains cool discoveries.[17]

In September 2018, Wansink resigned from Cornell.[18]

8 Increasing the Cost of Misconduct

Even if we do not have a hard misconduct problem in financial economics (now), having a data and code-sharing requirement makes it much more difficult to fabricate and falsify results.

[14] https://jamanetwork.com/journals/jamapediatrics/fullarticle/2654849

[15] https://jamanetwork.com/journals/jamapediatrics/fullarticle/2659568

[16] https://web.archive.org/web/20170312041524/http:/www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no

[17] Examples of soft misconduct where a researcher is trying to find a "significant" effect are detailed in Harvey (2017) and include: sample selection (start date and end date), sample collection (terminate data collection as soon as a significant effect is found), winsorization choices, outlier exclusion, scaling and transformation of variables, choice of control variables, instrument selection, the choice of estimation method, and the selective reporting of results.

[18] https://www.washingtonpost.com/health/2018/09/20/this-ivy-league-food-scientist-was-media-darling-now-his-studies-are-being-retracted/?utm_term=.17c9c9964d2d

I do believe, however, we have a soft misconduct problem. Having data and code readily available allows researchers to easily check both the accuracy and the robustness of results and increases the cost of misconduct. It is easy for a third party to see, for example, data selection rules that might be driving a particular result. Accordingly, a researcher will think twice about torturing the data until they yield the desired outcome.

9 Three Reasons for Data and Code Sharing

The primary reason for having a data and code-sharing culture[19] is to increase the cost of scientific misconduct. The pressure to publish is overwhelming, and journals much prefer to publish papers with "positive" results (supporting the hypothesis being tested). Papers with positive results gather many more citations. Some authors respond by engaging in p-hacking (Harvey, 2017).

But other reasons exist for data and code sharing. Consider a researcher who has independently written a computer program that extends the analysis in a published paper. It is useful for the researcher to calibrate the base case with the published paper. This reduces the chance of an error in the new program.

The third reason is the so-called reinventing-the-wheel problem. Some mechanical tasks (e.g., forming factor portfolios) require a lot of effort. Why should researchers have to independently repeat the same process over and over? Doing so is a waste of human capital, which could be deployed to more innovative projects. Other fields, such as computer science, have moved to a much more collaborative model. If a researcher is trying to solve a problem in computer science, their first stop is Github.com, where it is likely someone will share code that solves the problem or a similar problem. It makes research much more efficient.[20]

10 Reverse p-Hacking

Harvey (2017) warns of "reverse" p-hacking, a situation in which data and code are made available by Researcher A. Researcher B sets out to invalidate the results of A. B tries hundreds of robustness checks and finds one that weakens or eliminates the results of A. Researcher B writes a paper featuring this singled-out robustness check. Obviously, when trying hundreds of checks the result will be weakened—even, if purely by chance.

[19] It is not sufficient to simply have a policy. Höffler (2017) argues that many of the top journals have stated policies that are not enforced. It is important to develop the norms within the profession that are consistent with the policy.

[20] There is a fourth reason—but it is temporary. There is evidence that papers that have code and data available are more likely to be cited (see Höffler, 2017). Presumably, this advantage will disappear in the future when all papers have code and data sharing.

Hence, journal editors should take this situation into account. In instances of data and code sharing, the hurdle for publishing a paper that refutes the original study needs to increase due to a multiple testing problem.

11 The Ingredients for a Successful Policy

1. Any successful policy should be common across all top finance journals. A researcher that is unwilling to share their nonproprietary data and code will target the journal with the weakest policy.
2. As a condition for final acceptance of a paper, authors should be required to post all nonproprietary data along with the code that generated the data and the results in the paper. Example 1. A researcher uses data from the Federal Reserve Bank of St. Louis in a fixed-income paper. The researcher would post the original data as well as the transformed data along with the computer programs. Example 2. A researcher uses proprietary data from CRSP and Compustat to form factors. The researcher does not post the proprietary data but posts the factor data as well as the code that generated the factors and results.
3. Journal editors have the discretion to delay the general posting of the data and code if the case can be made that the authors are working on related projects and that posting the data and code would impact their respective competitive advantage. This should be the exception rather than the rule and should not extend past 3 years. Further, if the integrity of the research is challenged in the embargo period, the editor should make the data available on an exclusive basis to the challenging researcher.
4. Any successful policy must make it as easy as possible for authors over the longer term. This means that code and data should be documented. If the data are widely available but proprietary, such as CRSP or Compustat, the author(s) should provide some example data (which could be generated arbitrarily) to show that the programs run. This would save many questions from replicators. For non-standard proprietary data, it might be necessary to engage a *third party replicator* (someone that has access to the proprietary data but no conflicts of interest).[21]
5. Journals should institutionalize the replications by posting a record of replications on their websites (e.g., *AEA Data and Code Repository*) or a common website (e.g., a potential AFA Data and Code Repository that would serve all leading finance journals or an interdisciplinary initiative such as *ReplicationWiki*) and assign DOIs to each replication. Indeed, all top PhD programs

[21] See Pérignon *et al.* (2019) and Hurlin and Pérignon (2019).

should require their students to replicate a study published in a top finance journal, perhaps in the summer after their first year. This would also greatly reduce the cost to authors over the longer term. Future replicating researchers would have not just the original study to guide them, but also a record of a replication, thus reducing the chance that the original author is bothered with questions.

6. The American Finance Association should take the lead and appoint a Data Editor.

12 Conclusions

Let's be clear about the costs of implementing such a policy. First, authors must engage in extra work to prepare data and code for posting. Second, extra work is involved in the inevitable answering of questions. Third, the extra work is disproportionately imposed on the most productive members of the profession—potentially leading to reduced innovation.

I would argue that, if implemented efficiently, the benefits of a data and code-sharing policy outweigh the costs. Such a policy has many benefits. First, the trust in our research is supported by a strong replication culture. Currently, the field of finance's weak or nonexistent policy is inconsistent with the broad trends in economics and other areas of science.[22] If we aspire to develop ideas that influence policy makers, our results will be discounted without a strong replication policy. Second, we increase the cost of the difficult-to-observe soft misconduct. Researchers will think twice before p-hacking a result when their code and data are available. Third, research will become more efficient. Researchers will easily be able to calibrate their own programs to current published work, and it will no longer be necessary to reinvent the wheel for mechanical, but time costly, exercises that have been published before.

References

Dewald, W. G., J. G. Thursby, and R. G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *American Economic Review*. 76: 587–603.

[22] For example, research supported by the National Science Foundation, requires data sharing. See, *Proposal and Award Policies and Procedures Guide*, (February 25, 2019) Chapter XI, D4, "Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections, and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing." https://www.nsf.gov/pubs/policydocs/pappg19_1/pappg_11.jsp#XID4

Duvendauck, M., R. Palmer-Jones, and W. R. Reed. 2017. “What is Meant by “Replication” and Why Does it Encounter Resistance in Economics?” *American Economic Review: Papers and Proceedings*. 107: 46–51.

Fang, F. C., R. G. Steen, and A. Casadevall. 2013. “Misconduct Accounts for the Majority of Retracted Scientific Publications.” *Proceedings of the National Academy of Sciences of the United States of America*. 110: 1138–1143.

Gertler, P., S. Galiani, and M. Romero. 2018. “How to Make Replication the Norm.” *Nature*. 554: 417–420.

Harvey, C. R. 2014. “Reflections on Editing the Journal of Finance.” In: *Secrets of Economics Editors*. Edited by M. Szenberg and L. Ramrattan. Cambridge: MIT Press. 67–82.

Harvey, C. R. 2017. “The Scientific Outlook in Financial Economics.” *Journal of Finance*. 72: 1399–1440.

Höffler, J. H. 2017. “Replication and Economics Journal Policies.” *American Economic Review: Papers and Proceedings*. 107: 52–55.

Hurlin, C. and C. Pérignon. 2019. “Reproducibility Certification in Economics Research.” SSRN *Working Paper*.

National Academies of Science. 2017. *Fostering Integrity in Research*. Washington, DC: National Academies Press.

Open Science Collaboration. 2015. “Estimating the Reproducibility in Psychological Science.” *Science*. 349: 943–951.

Pérignon, C., K. Gadouche, C. Hurlin, R. Silberman, and E. Dobonnel. 2019. “Certifying Reproducibility with Confidential Data.” *Science*. 365: 6449.

Resnick, D. B., L. M. Rasmussen, and G. E. Kissling. 2015. “An International Study of Research Misconduct Policies.” *Accountability in Research*. 22: 249–266.

Wansink, B., D. Just, and C. Payne. 2012. “Can Branding Improve School Lunches?” *Archives of Pediatrics and Adolescent Medicine*. 166: 473–478. Retracted, December 2017.