This paper is a cross-journal, broad analysis of the state of replication in economics based on our Finance and Economics Discussion Series working paper, Chang and Li (2015a), and our subsequent work forthcoming in the Critical Finance Review. We attempted to replicate research using author-provided data and code files from 67 macroeconomic papers published in 13 well-regarded journals by following a pre-analysis plan.

We defined a successful replication as when we could use author-provided files to qualitatively produce their key results. Under this definition, we were able to replicate 22 of 67 papers (33 percent) independently of the authors.

Excluding six papers that relied on confidential data for all of their results and two papers that provided code written for software we did not possess, we successfully replicated 29 of 59 papers (49 percent) with help from the authors, shown in Figure 1.

The most common reason why we were unable to replicate the remaining 30 papers was that the authors did not provide us data and code replication files. We found that some authors did not provide data and code replication files even when their article was published in a journal with a policy that required submission of such files as a condition of publication, indicating that journal editorial offices did not strictly enforce their own policies, although provision of replication files was more common at journals that had such a policy than at journals that did not.

Figure 2 breaks down our replication success rate by whether the journal a paper was published in had a mandatory data and code policy or not. Excluding the papers in our sample that used proprietary data or had replication code in

---

1 Go to https://doi.org/10.1257/aer.p20171034 to visit the article page for additional materials and author disclosure statement(s).
2 This result was in line with recent evidence in replicating psychology studies by Open Science Collaboration (2015), where the authors failed to replicate the qualitative result of the majority of their sample of psychology experiments.
necessary software and the replicated paper’s use of public data.

Note: Successful replication conditioned on us having the necessary software and the replicated paper’s use of public data.

software that we did not possess, we successfully replicated 23 of 34 papers (68 percent) from journals that had a data and code policy. This rate compares to 6 of 25 papers (24 percent) from journals that did not require such files.

Of the failed replications in Figure 2, about two-thirds failed because we were unable to obtain all of the necessary files from the authors. The remaining one-third failed because either the files that we obtained did not work or they produced results that were contrary to what was published.

The presence of a journal’s mandatory data and code policy does not necessarily imply a causal relationship from the policy to successful replication. Authors selected which journals to submit papers to, taking into account idiosyncratic journal policies such as mandatory submission of replication data and code. However, we found that it was significantly easier to replicate published research that came from journals that required authors to submit their data and code.

I. Methodology and Sample

This section presents an abbreviated version of our methodology and sample. For full details, see Chang and Li (2015a, forthcoming).

We defined our methodology prior to executing any analysis. This preanalysis plan had three benefits: (i) we set a uniform standard for analyzing the results of models; (ii) we avoided hindsight bias in model selection and analysis; and (iii) we avoided pretesting our results.


From our sample of journals, we browsed for original research articles published in issues from July 2008 to October 2013. Within these issues, we identified all papers with the following three characteristics: (i) an empirical component; (ii) model estimation with only US data; and (iii) a key empirical result regarding output, measured by US gross domestic product (GDP), published by the Bureau of Economic Analysis (BEA), in an estimated model.

We found 67 papers that fit our criteria. Of these papers, six papers used proprietary data for all of their key results, so we did not include them in our replication exercise.

For the remaining papers that used public data and were published in journals that maintained data and code archives, we downloaded the replication files provided by the authors through journal online archives. Like McCullough, McGeary, and Harrison (2006) and Vlaeminck and Herrmann (2015), we found that journal data and code archives were incomplete. Of the 35 papers that used public data and that were published in journals that required data and code replication files, we obtained data and code files for 28 papers (80 percent) from journal archives.

When we were unable to obtain replication data and code files from journal websites, either because the mandatory files were missing or because the paper was not subject to a data availability policy, we checked the personal websites of each of the authors for replication.

3 We also searched for papers that used gross domestic income (GDI) as a measure of output for use in our related work on data revisions and measurement error, Chang and Li (2015b), but we did not find any papers that used GDI.
files. If we were unable to locate replication files online, then we e-mailed each of the authors individually requesting the replication files. Of the seven papers that used public data, were subject to a data and code policy, and did not have replication files on the journal’s website, e-mails netted us one additional set of replication files. Therefore, we obtained data and code files for 29 of 35 papers (83 percent) that were published in journals that required submission of data and code replication files. For papers published in journals without a data and code availability policy and that used public data, we located the files for 11 of 26 papers (42 percent).

To determine whether a paper was subject to a data availability policy, we checked the implementation dates of the journal data policies and compared them to the publication and submission dates of the published work. If the journal’s website did not allow us to extract this information, then we queried the editorial office as to when their data availability policy became effective. We did not ask the editorial offices whether a particular paper was subject to a data availability policy. Aside from papers with proprietary data, we found that journal data archives did not provide lists of potentially exempt papers. Therefore, we were unable to determine whether a paper was exempt for a reason other than using proprietary data, although we were not aware of reasons why journals would grant a paper a data and code exemption other than for proprietary data. The authors we queried whose papers we believed were subject to a data availability policy, yet whose replication files we were unable to locate, did not volunteer whether their papers were exempt from the policy, and we did not ask the authors for this information.

For the papers for which we were able to obtain data and code replication files, we attempted to replicate the key results of the papers using only the instructions provided in the author readme files. If the readme files were insufficient or if the replication files were incomplete (or both) and the paper was subject to a replication policy, then we e-mailed the corresponding author (if no corresponding author, then the first author) for either clarification or to request the missing files. If we did not receive a response within a week, then we queried the second author, and so on, until we contacted all authors. If we already contacted the authors to request data or code but were having difficulty executing the code, then we only queried the authors whom we did not yet contact.

We defined a successful replication as when the authors or journal provided data and code files that allowed us to qualitatively reproduce the key results of the paper. For example, if the paper estimated a fiscal multiplier for GDP of 2.0, then any multiplier greater than 1.0 would produce the same qualitative result (i.e., there is a positive multiplier effect and that government spending is not merely a transfer or crowding out private investment). Interested readers can find all of our replication results in the online Appendix. We defined success using this extremely loose definition to get an upper bound on what the replication success rate could be. We allowed for some reworking of the provided files, following the procedure of McCullough, McGeary, and Harrison (2006).

When available in the readme, we attempted to run the software version-operating system combination specified by the authors. When the replication files failed to execute on a given software version-operating system combination, the author readme did not specify a particular software version-operating system combination, and it appeared that the data and code were complete, we e-mailed the authors to find out which combination they used.

II. Conclusion and Recommendations

In this article, we attempted to replicate 67 papers from 13 well-regarded economics journals using author-provided data and code replication files by following a preanalysis plan. We designed a broad sampling frame that spanned different journals and covered a large number of
original research articles. We replicated 22 of 67 papers (33 percent) by using only the authors’ data and code files, and an additional seven papers with assistance from the authors. The most common cause of our inability to replicate findings was that authors did not provide files to the journal replication archives. The second most common cause was the files we obtained either did not work or produced results different than the original paper. Our overall success rate for papers that used public data and had replication code written for software that we owned was 29 of 59 papers (49 percent).

We now turn to some recommendations that we feel would improve the ability for researchers to replicate and extend research.

Mandatory data and code files should be a condition of publication.

Our replication success rate was significantly higher when we attempted to replicate papers from journals that had a mandatory replication data and code policy, as opposed to a data-only policy or no policy.

The journal’s data and code archive should indicate whether a paper without replication files in the archive is exempt from the journal’s replication policy.

This entry would have four virtues: (i) it would be low-cost for the journal; (ii) it would save authors who are exempt from submitting replication files from needing to respond to queries about replication files; (iii) it would save would-be-replicators from searching for replication files for papers that are exempt from the journal’s policy; and (iv) it would identify those authors who are not compliant with the journal’s policy.

Readme files should indicate the operating system-software version combination used in the analysis.

Although it was not a focus of our paper, we noticed minor discrepancies for a selected subset of papers when running programs on different versions of Matlab.

Readme files should contain an expected model estimation time.

Many estimation routines can take a considerable amount of time to execute, even under the best of circumstances. We encountered a few instances where we believed an estimation was executing, only to find out weeks later that the programs were stuck in an infinite loop and were supposed to run in much less time. In addition, frequently programs are not written to optimize computation time and also frequently written without a progress bar, so there is no way to track the expected completion time of estimation. A low-cost alternative to a progress bar is simply writing the expected estimation time in the readme file.

Code that relies on random number generators should set seeds and specify the random number generator.

Optimization algorithms often rely on a set of initial conditions, which are commonly specified through a random number generator. For any research that relies on a random number generator, exact replication requires the same set of numbers that are generated in the published article.

Readme files should clearly delineate which files should be executed in what order to produce desired results.

We now turn to two recommendations that will improve the ability of researchers to extend work, in addition to replicating it.

Authors should provide raw data in addition to transformed series.

While only the transformed data are needed to replicate published results, raw data facilitate extensions of research.

Programs that replicate estimation results should carry out the estimation.

Some of these replication files that we obtained, instead of estimating the models, took the relevant parameters as given to produce results in tables and figures. For verification of published results, and particularly for purposes of extending research, we assert that code that actually estimates the relevant models would be far more useful.

4 We are elated to report that the working paper version of this article (Chang and Li 2015a) caused the Review of Economic Dynamics to change its data and code policy to include several of our recommendations, very soon after the paper was posted to the Board of Governors of the Federal Reserve System Finance and Economics Discussion Series on October 5, 2015. (An added bonus: this change marks one of the few instances we can identify a causal effect without a randomized control trial!) The Review of Economic Studies also changed its data and code policy to include our recommendations sometime between October 5, 2015 and February 21, 2016, but we are unsure whether they changed their policy because of our working paper.
REFERENCES


