A Replication Recipe: List Your Ingredients before You Start Cooking

Andrew C. Chang*

May 14, 2018

Abstract

I argue that researchers should do replications using preanalysis plans. These plans should specify at least three characteristics that would act as stopping rules for the replicator: (1) how much flowtime the replicator will spend, (2) how much money and effort (working hours) the replicator will spend, and (3) the intended results and the precision of the replication necessary for "success". A researcher's replication will be "successful" according to context-specific criteria in the preanalysis plan. I also argue that the two biggest drawbacks of preanalysis plans—(1) that they discount unexpected but extraordinary findings and (2) that they make it difficult for researchers to prespecify all possible actions in their decision trees—are less relevant for replications compared with new research. I conclude with describing a preanalysis plan for replicating a paper on housing demand and household formation.

JEL Codes: B41; C80; C81; R21

Keywords: Data and Code Files; Household Formation; Housing Demand; Preanalysis; Prespecification; Publication Bias; Replication

*Senior Economist, Division of Research and Statistics, Board of Governors of the Federal Reserve System. 20th St. and Constitution Ave., NW, Washington DC 20551 USA. +1 (657) 464-3286. a.christopher.chang@gmail.com. https://sites.google.com/site/andrewchristopherchang/

†The views and opinions expressed here are mine and are not necessarily those of the Board of Governors of the Federal Reserve System. I prepared this paper for a special issue on definitions of "successful" and approaches to replications in *Economics: The Open-Access, Open-Assessment E-Journal*, with an example of how I would conduct a replication. The original call for papers is at http://www.economics-ejournal.org/special-areas/special-issues/the-practice-of-replication. I thank anonymous referees, Annette Brown, Christopher Karlsten, and the editor W. Robert Reed for helpful comments. Any errors are mine.

1 Replication Recipe

Researchers should do replications using preanalysis plans that set at least three criteria: (1) the flowtime of the intended replication; (2) the budget of the replication, in terms of money and working hours; and (3) the set of estimates and the degree of precision that would define a "successful" replication.

Prespecification of the amount of flowtime and the budget for a replication would provide you, as a replicator, with explicit stopping rules. These stopping rules would be useful because proving that a paper is not replicable under any circumstance could require a very large investment, as you would have to show that all possible permutations of your decisions in your replication would lead to the conclusion that the paper is not replicable.

The context of the replication and your preferences will determine how much flowtime and budget you are willing to invest to get a "successful" replication. The Bill & Melinda Gates Foundation is willing to give the International Initiative for Impact Evaluation (or 3ie) a large budget to do aid impact replications both because the foundation has money and because the foundation's aid projects depend on accurate research. Your budget for a "successful" replication is, most likely, less than that of the Bill & Melinda Gates Foundation. Context should also determine what you define as a "successful" replication, in terms of both the set of results from the original paper that you are interested in and the precision of the replication of those results.

I know that "context matters" is an unsatisfactory answer for a special journal issue seeking a definitive answer of what defines a "successful" replication. But context *does* matter.

Consider screaming, *"The roof is on **fire!**"* When might the context determine the meaning of this outburst?

How about the following:

1. You are partying at a nightclub on New Year's Eve.

2. You are on the phone with fire and rescue dispatch.

3. You are a contestant on *Wheel of Fortune*.

Now consider three contexts where you might want to do a replication and where the context determines what a "successful" replication means:

1. You are verifying the original paper for the archival record.

2. You are writing a new paper that extends the original paper.

3. You are learning an econometric technique from the original paper.

How would you ***verify an original paper for preservation as an archival record?*** Presumably, you would want to replicate all estimates designated for preservation to machine precision, which most likely would be those estimates that appear in the main text and any published appendix.[1] You would be less concerned —if you would be concerned at all —about any non-published appendix to the paper that was not designated for preservation.

What about your ***new paper that extends the original paper?*** You probably would be interested in a subset of key results from the original paper, most likely the "main result" of the original paper. Robustness check #534 in footnote #81 would probably be irrelevant for you even if it appears in the main text of the original paper. A "successful" replication in this context would probably treat your replication of 0.41, when the original estimate was 0.43, as a "successful" replication.

And how about ***learning an econometric technique from the original paper?*** In this case, the technique you want to learn might be robustness check #534 scribbled out in footnote #81, making this footnote the most important part of the paper. But

[1]Of course, if you had a vendetta to debunk the original paper, then "success" for you would be getting estimates as far from the original paper as possible.

a "successful" replication for your human capital development is the procedure that you would go through to learn the technique, so long as you are confident that the procedure you would go through is correct. The sausage of econometric estimates might be irrelevant in this context.

In the specific context of replications, prespecification of stopping rules for flowtime and budget mitigates any inclinations a replicator might have for debunking the original paper. While you may think that a replicator would be able to unfairly claim that a paper is not replicable by prespecifying a low amount of flowtime or a small budget, because the consumers of the prespecified replication plan can also see the amount of flowtime and budget that the replicator dedicated to the replication, these consumers can also judge, in the context of that amount of specified flowtime and budget, whether or not to believe a failed replication.[2]

And much like preanalysis plans for new research, preanalysis plans for replications: (1) reduce our incentives to specification search and p-hack (that is, run models until your p-values are "significant"),[3] (2) ground our estimates in statistical theory because the specifications that we report are not pretested (that is, the estimates are not conditioned on observing some other unreported specification),[4] and (3) provide a results-free defense against potential criticism.[5]

A results-free defense is even more important for replications than for new research. At least some authors will not like you attempting to replicate their research. Should you find something contrary to the original paper, these authors may feel extra motivated to rebut your replication. Possibly with a less-than-civil rebuttal. A results-free defense will help absolve you from criticisms of replicator bias.

[2]I believe that replication preanalysis plans should be made public, although perhaps with a delay.

[3]Preanalysis plans can also help limit model overfitting, which can be a consequence of p-hacking. Model overfitting has particularly disastrous consequences in forecasting (Frye, 2017).

[4]See Poirier (1995) for a discussion on pretesting.

[5]The first economics research that I am aware of that used a preanalysis plan is Neumark (1999, 2001). Casey, Glennerster, and Miguel (2012) discuss the strengths and weaknesses of preanalysis plans. For a further discussion on the limitations of preanalysis plans, particularly for experiments, see Coffman and Niederle (2015).

Preanalysis plans have two main weaknesses: (1) they force researchers to ignore findings outside the scope of the preanalysis plan, so that unexpected but extraordinary findings cannot contribute to a study, and (2) researchers have difficulty prespecifying contingencies for all possible outcomes. But these weaknesses are less concerning for replications compared with new research. With respect to ignoring findings outside the scope of the preanalysis plan, when replicating you know the general scope of results that you are aiming for prior to running your replication. Therefore, finding something extraordinary that is outside the scope of the original paper that you are replicating is less likely compared with finding something unexpected but extraordinary with new research, so this feature of preanalysis plans is less of a drawback for replications.[6] Regarding the difficulty of prespecifying contingencies, you will still need to prespecify contingencies for when you find different estimates than the original paper. But the original paper, by narrowing the research question and outlining its methodology, sets bounds on reasonable contingencies.

2 Application to Haurin and Rosenthal (2007)

I selected Haurin and Rosenthal (2007), a paper on housing demand and household formation, as an application of how I would conduct a replication.[7]

To select Haurin and Rosenthal (2007), I used two criteria to match the special issue's requirements:

- The issue's requirement for an "influential economics article," which I took to mean a paper with at least one Google Scholar citation as of March 14, 2017.

- The issue's requirement for a paper "not previously replicated," which I took to

[6] Unless you undertook a replication expecting to find something extraordinary, in which case you might have a vendetta or are extremely cynical about published research.

[7] I did not tell the editor that I had self-imposed any of the criteria in this section prior to selecting Haurin and Rosenthal (2007).

mean a paper without a replication registered on the Replication Wiki[8] as of March 14, 2017.

I used an additional three criteria, which the special issue did not mandate, to reduce my own biases in writing a replication plan:

- A paper not published from July 2008 to September 2013 in the following 13 well-regarded journals: *American Economic Journal: Economic Policy; American Economic Journal: Macroeconomics; American Economic Review; American Economic Review: Papers and Proceedings (or P&P); Canadian Journal of Economics; Econometrica; Economic Journal; Journal of Applied Econometrics; Journal of Political Economy; Review of Economic Dynamics; Review of Economic Studies; Review of Economics and Statistics;* and *Quarterly Journal of Economics;* which are the journal issues in the sampling frame of my own work on replication, Chang and Li (2015a, 2017, Forthcoming).[9]

- A paper not written by an author who, at the time of the paper's publication, was from the Board of Governors of the Federal Reserve System.

- A paper not written by an author with whom I had corresponded before March 14, 2017.

Finally, I used two characteristics to reduce my search costs and ensure feasibility of the replication plan.

- A paper that I read within a year prior to the special issue's call.

- An empirical paper that used no confidential or proprietary data.

[8] See Höffler (2017), http://replication.uni-goettingen.de/wiki/index.php/Main_Page

[9] The time frame and journals here are the sampling frame that I used, not the set of papers that I attempted to replicate, which was a subset of the articles from this sampling frame.

3 Replication Steps and Discussion of "Success"

I focus this discussion on the hypothetical scenario in which I, acting as a representative of the journal *Real Estate Economics*, would try to replicate the results of Haurin and Rosenthal (2007). In this scenario, I would seek to verify Haurin and Rosenthal (2007) for preservation as an archival record. Verification would not require that the authors' estimates are externally valid. This discussion presumes that the authors' estimates are correct. This replication would not be undertaking to purposefully debunk or uncover fraud by the authors.

Here are the steps that I would take:

1. I would want to replicate the authors' results using my code and their raw data.[10] I would want their raw data so that I could code everything from start to finish, following the procedures described in Haurin and Rosenthal (2007) and Maddala (1983), which is a reference cited by Haurin and Rosenthal (2007) that supports their estimation procedure. Their transformed data would also be useful for verification purposes.

2. As I would be planning on coding Haurin and Rosenthal (2007) but could still make a coding error in doing so, I still would want the authors' code, also for verification purposes.

3. As a representative of the journal, presumably I would have the full cooperation of the authors. If the authors had not yet already supplied the journal with replication files and the files were not available on their personal websites or through the Inter-university Consortium for Political and Social Research, then I would first request code, data, and readme files directly from the authors. For

[10] I would want to use their data because public-use datasets, which Haurin and Rosenthal (2007) use, are revised over time and authors rarely document which version of data they use (Chang and Li, 2015b), so I would want to eliminate data revisions as a source of uncertainty in the replication. In addition, even if you know which version of the public-use dataset to look for, locating a particular version of a public-use dataset can be time consuming or even impossible.

points of contact, I would email both authors (Haurin and Rosenthal (2007) does not list a corresponding author) using the email addresses listed in the following locations, moving to the next email address if an email were returned as undeliverable: (1) the addresses provided to the journal at the time of submission, (2) addresses listed on the authors' personal websites, and (3) addresses listed on their current institutions' websites. If emails to all of these locations were undeliverable, then I would attempt to phone both authors, leaving voicemails when necessary and available, and using phone numbers obtained from the same three-step procedure as were email addresses. If I were unable to speak with both authors, I would write both authors letters and send them to the postal addresses that the authors provided to the journal. In my correspondence, I would also ask the authors which operating system and software version their files ran on.

4. After contacting the authors or writing letters and leaving voicemails, I would wait three weeks to receive files. If the authors did not provide me with files after three weeks, I would then recontact the authors, using the same points of contact. I would attempt to solicit files up to four times (an initial contact plus three follow-ups) before reporting the status of files (procured or not procured) to the other journal staff.

5. After finishing step #4, which should take a maximum of 12 weeks, if I had the original files, then I would allocate six months to do this replication (eight months if I did not have the files, assuming that I would have access to sufficient computing power and the appropriate software to run the original files).

6. With the authors' files, I would estimate about 540 hours of work (6 productive hours per working day x 90 working days). Without the files, I would estimate

720 hours.[11]

7. As I would be verifying the paper for the archival record as a member of the journal's staff, if something were to happen to me that would halt my ability to work on the replication, then the replication could be transferred to another member of the journal's staff. But if such a transfer were required, then I would add an additional 32 hours of work to the project's budget. I would also add to the project's flowtime a week plus the amount of time it would take for the new representative to transfer onto the replication.

8. I would follow the steps outlined in Haurin and Rosenthal (2007) to re-download their data. I would check to make sure that the relevant variables for the analysis, including those on their page 423, are in the downloaded dataset, and that the observation counts match what they report on page 424. Haurin and Rosenthal (2007) does not have a table of detailed summary statistics that I could use to compare the moments from my downloaded version of data with theirs, though if I were able to obtain the dataset from the authors, then I would directly compare first and second unconditional moments for all variables listed on page 423 from the two datasets.

9. If the dataset that I downloaded were missing a variable that was reported as included in the analysis on page 423 or the dataset contained a different number of observations than the authors reported, then I would attempt to contact the authors again to resolve the differences, waiting one week between contacts, with a maximum of five initiated contacts. If I were unable to resolve differences between the downloaded dataset and what the authors reported in their paper, then I would mark the paper as not replicable.

[11] I am basing these estimates on how long it has taken me to replicate previous papers under the same criteria.

10. If the dataset that I downloaded matched Haurin and Rosenthal (2007), then I would proceed to attempt to recode and replicate all figures in the main text and also in online published appendixes A1 and A2 using the same operating system and software version that the authors used (if the authors did not provide me with this information, then I would use the most recently available operating system and software versions). I would be successful with this replication if there were no visual difference between any replicated figure and published figure and if the observation count cited on page 424 (190,778 observations) matched the observation count in the replication.

11. Regardless of the replication results, I would file a report alongside my preanalysis replication plan that would state the results and whether my replication was successful or not. If I was unsuccessful, the report would also include why I was unsuccessful, which steps I went through before deciding that I was unsuccessful, and what (if any) output I produced.[12]

References

Casey, Katherine, Rachel Glennerster, and Edward Miguel. "Reshaping Institutions: Evidence on Aid Impacts Using a Preanalysis Plan," Quarterly Journal of Economics 127:4 (2012), 1755-1812.

Chang, Andrew C., and Phillip Li. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Usually Not"," Finance and Economics Discussion Series 2015-083. Washington: Board of Governors of the Federal Reserve System (2015a).

[12]Filing this report, regardless of replication success or failure, would help mitigate file-drawer (Rosenthal, 1979) concerns over incomplete replications and would also assist future replicators.

Chang, Andrew C., and Phillip Li. "Measurement Error in Macroeconomic Data and Economics Research: Data Revisions, Gross Domestic Product, and Gross Domestic Income," Finance and Economics Discussion Series 2015-102. Washington: Board of Governors of the Federal Reserve System (2015b).

Chang, Andrew C., and Phillip Li. "A Preanalysis Plan to Replicate Sixty Economics Papers that Worked Half of the Time," American Economic Review 107:5 (2017), 60-64.

Chang, Andrew C., and Phillip Li. "Is Economics Research Replicable? Sixty Published Papers from Thirteen Journals Say "Often Not"," Critical Finance Review (Forthcoming).

Coffman, Lucas C., and Muriel Niederle. "Pre-Analysis Plans Have Limited Upside, Especially Where Replications Are Feasible," Journal of Economic Perspectives 29:3 (2015), 81-98.

Frye, Jon, "The Etiology of the Overfit Model," Working Paper (2017).

Haurin, Donald R., and Stuart S. Rosenthal. "The Influence of Household Formation on Homeownership Rates Across Time and Race," Real Estate Economics 35:4 (2007), 411-450.

Höffler, Jan H. "ReplicationWiki: Improving Transparency in Social Sciences Research," D-Lib Magazine 23:3/4 (2017). https://doi.org/10.1045/march2017-hoeffler

Maddala, G.S., Limited Dependent and Qualitative Variables in Econometrics. New York: Cambridge University Press.

Neumark, David. "The Employment Effects of Recent Minimum Wage Increases: Evidence from a Pre-specified Research Design," NBER Working Paper No. 7171 (1999).

Neumark, David. "The Employment Effects of Minimum Wages: Evidence from a Prespecified Research Design," Industrial Relations: A Journal of Economy and Society 40:1 (2001), 121-144.

Poirier, Dale J., Intermediate Statistics and Econometrics: A Comparative Approach. MIT Press (1995).

Rosenthal, Robert, "The File Drawer Problem and Tolerance for Null Results," Psychological Bulletin 86:3 (1979), 638-641.