Economics
The Open-Access, Open-Assessment E-Journal

Discussion Paper
No. 2019-5 | January 10, 2019 | http://www.economics-ejournal.org/economics/discussionpapers/2019-5

Takeaways from the Special Issue on The Practice of Replication

W. Robert Reed

Abstract
In July 2017, issued a call for papers for a special issue on "The Practice of Replication." In that call, the journal explained that there was no generally accepted procedure for how to do a replication. Likewise, there was no generally accepted standard for determining whether a replication "confirms" or "disconfirms" an original study. Accordingly, the journal called for papers to identify principles for how to do a replication and how to interpret its results; and to apply those principles in crafting a replication plan for a study of the author's choosing. The hope was that this exercise would produce some progress on "the practice of replication". The special issue is now complete with a total of eight journal articles. This commentary places the respective articles within a common framework and identifies observations and lessons learned from the respective studies.

(Published in Special Issue The practice of replication)

JEL C10 C18 C50
Keywords Replication; pre-analysis plan; reproduction; repetition; extension; robustness

Authors
W. Robert Reed, Department of Economics and Finance, University of Canterbury, Christchurch, New Zealand, bob.reed@canterbury.ac.nz

Citation W. Robert Reed (2019). Takeaways from the Special Issue on The Practice of Replication. Economics Discussion Papers, No 2019-5, Kiel Institute for the World Economy. http://www.economics-ejournal.org/economics/discussionpapers/2019-5

Received November 26, 2018 **Accepted as Economics Discussion Paper** December 7, 2018
Published January 10, 2019

1 Introduction

In July 2017, *Economics: The Open Access, Open Assessment E-Journal* issued a call for papers for a special issue on "The Practice of Replication." In that call, the journal explained that there was no generally accepted procedure for how to do a replication. Relatedly, there was no generally accepted standard for determining whether a replication "confirms" or "disconfirms" an original study.

Accordingly, the journal called for papers to identify principles for how to do a replication and how to interpret its results; and to apply those principles in crafting a replication plan for a study of the author's choosing. Submitted papers were to consist of four parts: (i) a general discussion of principles about how one should do a replication, (ii) an explanation of why the candidate paper was selected for replication, (iii) a replication plan that applied these principles to the candidate article, and (iv) a discussion of how to interpret the results of the replication. Authors were free to choose the article they wanted to replicate, but no two articles could be the same. Authors were not given any instruction or guidance about how they should do their replication. Finally, all authors worked independently, and did not see the others' work until the papers were simultaneously published as discussion papers.

The hope was that this exercise would produce some progress on "the practice of replication"; or, at the very least, identify some reasons why progress was unlikely to be achieved. The special issue is now complete with a total of eight journal articles. Therefore the time is right to sit back and ask what has been learned.

2 The articles

Six Types of Replications. Reed (2018) provides a 6-part typology for classifying replications that offers a useful framework for organizing the articles of the Special Issue (see the top panel of Figure 1 in the appendix). "Reproduction" replications use the same data and same methods as the original study. If the original author's data and code are available, then this is a simple matter of checking that the author's data and code produce the numbers reported in the original study. An alternative approach consists of the replicating author drawing the data directly from primary sources, and/or writing his/her own code according to the original study's description. In both cases, this is a verification exercise to confirm the original author's findings while closely sticking to their data and methods.

A "Robustness Analysis – Same Dataset" replication also restricts itself to the same data as the original study, but checks to see if the original study's results hold up to reasonable perturbations to the data and methods. In the "garden of forking paths" (Gelman and Loken, 2014), there are many subjective choices that a researcher must make in deciding which observations to keep or omit ("outliers"); which variables to include in equation specifications; which forms of the variables to use (e.g., linear, log, quadratic, ratios); and which estimation methods to use. As brilliantly highlighted by (Silberzahn et al., 2018), "defensible, yet subjective analytic choices" can generate vastly different results from the same dataset. This

type of replication is designed to assess the robustness of the original study's results to these subjective choices.

A "Repetition" replication uses the same measures and methods as the original study, but applies them to a new sample drawn from the same population as the original study. A textbook example of this would be an experiment where the replicating author undertakes the exact same experiment as the original study, but applies it to a different set of subjects drawn from the same "population".

A conceptual grey area arises in determining whether a sample comes from the same or a different population. If a sample comes from a different, but related, population, and the replicating author uses the same measures and methods as the original study, it becomes an "Extension" replication. Most empirical studies make an implicit claim to identifying relationships that extend beyond the immediate subjects of their analysis. "Extension" replications are important for delineating the extent of an original study's external validity.

The last two types of replications are "Robustness Analysis – Same Population" and "Robustness Analysis – Different Population". These, like "Robustness Analysis – Same Dataset", attempt to assess the original study's findings to reasonable modifications of data and methods. By the time one gets to the lower, right hand corner of the top panel of Figure 1 in the appendix ("Robustness Analysis – Different Population"), the line becomes quite blurry between a replication study and an independent study that stands on its own. In the remainder of this section, I classify each of the articles in the Special Issue according to this six-part typology (see the bottom panel of Figure 1). Note that two of the articles, McCullough (2018) and Brown and Wood (2018), do not explicitly include replication plans, but do provide directions for how a replication should be done.

"Reproduction" replication. Both Chang (2018) and McCullough (2018) call for replications whose sole purpose is to confirm that they can reproduce the results from original studies. Chang proposes replicating Haurin and Rosenthal (2007), "The influence of household formation on homeownership rates across time and race." He positions himself as a representative of the journal whose concern is to verify H&R's paper for "preservation as an archival record." He proposes using their data, but writing his own code, following the description provided in their study. McCullough (2018) does not target a specific study, but is also concerned with confirming the archival record.[1] He focuses on the AER's data-code archive, arguing that anything less than "push button reproducibility" is unacceptable. In other words, if journals such as the AER require data and code to accompany published papers, the supplied data and code should allow perfect reproduction of the original study's results.

Note that in both cases, the purpose of the replication is not to determine whether the findings of the original study are "true". Rather, it is simply to confirm that one can obtain the same numbers reported in the original study using the authors' data.

"Reproduction" and "Robustness Analysis – Same Dataset" replications. Wood and Vasquez (2018), Hannum (2018), Brown and Wood (2018), and Owen (2018) also include a "Reproduction" replication, but they propose to go further and perform robustness analyses.

[1] In his original submission (see http://www.economics-ejournal.org/economics/discussionpapers/2017-78?searchterm=mccullough), McCullough did include a plan for replicating a report done by the AER on the status of its data archive, but this was dropped in response to a reviewer's comment.

They are very much concerned with determining whether the findings of an original study are "true".

Wood and Vasquez are interested in replicating Santos et al. (2014), "Can government-allocated land contribute to food security? Intrahousehold analysis of West Bengal's microplot allocation program". The Santos et al. paper was chosen because it found that allocating small plots of lands to impoverished households had a substantial impact on their welfare. Wood and Vasquez want to determine the reliability of these findings for guiding development policy.

They follow a four-part replication format (more on that below), consisting of "validity of assumptions", "data transformations", "estimation methods" and "heterogeneous impacts". Some of the questions they want to address are: Are the results robust to alternative estimation procedures (difference-in-differences), is there a minimum plot size necessary to produce beneficial results, and is there confirmatory evidence that longer access to land resulted in better outcomes?

Hannum proposes replicating the study "Does social media reduce corruption?" by Jha and Sarangi (2017). Using data provided by the authors, Hannum first would check whether he can produce results identical to those of the original study. He would also go back to the primary data sources to confirm that the authors correctly merged the data from multiple sources. Having done that, he would explore some "forking paths". For example, the original study omitted Argentina, and did not include a variable for the share of the population that was Protestant, a common explanatory variable in corruption studies. Further, the original paper used Facebook penetration as a measure of social media. Hannum would like to see if other measures of social media correlate with lower corruption. Hannum's ultimate goal is to determine whether Jha and Sarangi's finding is "true" – that increased social media use is negatively correlated with corruption.

Brown and Wood (2018) do not present a replication plan, but do present a diagnostic checklist for carrying out a replication. The checklist comes out of their experience with supervising 3ie's replication program. As mentioned above, it focuses on four aspects: "validity of assumptions", "data transformations", "estimation methods" and "heterogeneous impacts". What makes this particularly germane for the purposes of the Special Issue is that in addition to a checklist, it provides both examples and a list of resources for learning more about the issues related to each aspect.

The final paper in this category is Owen (2018). Owen proposes investigating a study's underlying statistical assumptions. He notes that reproducing a study's findings is of little value if the associated assumptions are invalid. His study of choice is Acemoglu, Gallego and Robinson's (2014), "Institutions, human capital, and development". AGR argue that institutional quality is a fundamental cause of historical development, reflected in current stocks of physical and human capital and levels of GDP per capita. They use instrumental variable procedures to identify the effects of institutional quality and human capital. However, the associated tests of instrument relevance and validity make a number of assumptions regarding model specification, parameter constancy, and error behavior in the reduced form models. If these assumptions do not hold, the subsequent empirical work may be invalid.

"Reproduction" and "Repetition" replications. The focus of Daniels and Kakar's (2018) replication plan is a study by Klump, McAdam, and Willman (2007) entitled "Factor substitution and factor-augmenting technical progress in the United States: A normalized

supply-side system approach". KMW provide evidence that the elasticity of substitution is significantly below unity and that there is an asymmetric pattern in the growth rates of technological progress. Thus, the Constant Elasticity of Substitution (CES) production function with nonunitary elasticity between capital and labor is a superior representation of the US economy than the Cobb-Douglas production function that is ubiquitous in empirical macroeconomic studies.

KMW report a number of important findings, the most important of which is that the elasticity of substitution between capital and labor is significantly less than one. Daniels and Kakar (2018) propose to reproduce KMW's findings using the same variables over the same time period (1953–1998), applying the same estimation procedure and tests. They note that one difficulty with a strict reproduction is that macroeconomic variables, particularly capital stock measures, are constantly being updated. They further propose updating the time period to the most recent data available. In the typology of Figure 1, this could be classified as a "Repetition" if one believed that the data generating process for the subsequent years was the same as that underlying KMW's data. Daniels and Kakar's ultimate goal is to determine if KMW are correct; namely, that the CES production function provides a superior way of modelling US production.

"Reproduction", "Repetition", and "Extension" replications. The last of the eight articles in the Special Issue is Coupé's (2018) replication of Choi and Varian's (2012) article, "Predicting the present with Google trends". Choi and Varian argue that forecasts of macroeconomic variables can be improved by including variables that measure the search intensity of words related to those variables on Google. They illustrate their claim with four time series: (i) sales of motor vehicles and parts in the US, (ii) US unemployment benefit claims, (iii) visitor arrivals to Hong Kong, and (iv) consumer confidence in Australia.

Coupé proposes a three-part replication of Choi and Varian's results for sales of motor vehicles and parts, which uses data from January 2004–June 2011. The first part reproduces their results using the data and code posted on Varian's website. The second part is a "Repetition" replication that covers the same time period as Choi andVarian, but draws a new sample of search intensities. Coupé, quoting Choi and Varian, notes that Google trends data is based on sampling procedures that produce different measures of search intensity depending on when the data are sampled. He also writes his own code based on the description in the original study. Finally, he extends Choi and Varian's analysis by applying it to New Zealand car sales. More generally, Coupé wants to determine if Google search data has the potential to improve forecasting across a wide variety of applications.

3 Takeaways

The following are my takeaways from the eight articles of the Special Issue.

Replication studies have different purposes. Not all studies are concerned with determining whether the original studies are true. Some replications merely want to establish the archival record (Chang, McCullough). Even studies that share a common interest in establishing the "truth" of an original study can have different purposes. For example, Owen wants to investigate the assumptions underlying a study to see if the inferences in that study are valid. In

contrast, Coupé is interested in knowing how far Choi and Varian's (2012) insight about Google trends and forecasting can be applied.

There can be no single procedure for doing replications. As Chang notes, "context matters". A replication study that is solely focused on verifying the numbers from a published study (Chang, McCullough) will proceed in a manner different than a study that, say, wants to test whether the assumptions underlying a study are valid (Owen), or wants to see what other data environments a particular model can be applied in (Coupé). That being said, Brown and Wood's checklist for performing replications provides some useful guidelines without requiring a one-size-fits-all approach.

There can be no single measure of replication "success". How one defines replication "success" depends on the goal of the replication. If the goal is to double check that the numbers in a published study are correct, then, as McCullough emphasizes, anything less than 100% reproduction is a failure: "For linear procedures with moderately-sized datasets, there should be ten digit agreement, for nonlinear procedures there may be as few as four or five digits of agreement" (McCullough, 2018, page 3). If the goal is to see if the original study is "true", it one needs to define the measure, or measures, of successful replication? The various approaches of the different studies is telling.

For Hannum, success depends on the significance of the estimated coefficient for the key variable (Facebook). Owen suggests a battery of tests based upon significance testing, but acknowledges "fallacies of acceptance and rejection" as challenges to interpreting test results. Coupé proposes counting all the parameters that are reproduced exactly and calculating a percentage correct index, perhaps weighted by the importance of the respective parameters. As this binary approach fails to distinguish between near and far misses, he proposes supplementing this with a measure of the "replicability gap" that sums the differences between the original study's estimates and their replicated values. Daniels and Kakar would identify success if the replicated parameters have "the same size and significance for all specifications", though they do not define what constitutes "the same".

Wood and Vasquez shy away from even using the words "success" or "failure". Instead, they see the purpose of replication as contributing to a "research dialogue". They advocate a holistic approach, "looking for similar coefficient sizes, direction of coefficients, and statistical significance". To aid interpretation in their "push button replication", they categorize differences between the original and replicated study's estimates: p-value differences less than 0.05 are "comparable"; differences between 0.05 and 0.10 are "minor"; and differences larger than 0.10 are "major". For coefficient estimates, differences less than 15% are "comparable"; those between 15% and 30% are "minor"; and those larger than 30% are "major".

Clearly, with so many "forking paths" available for defining success, replication researchers need to be credible that the measure of success they choose is above reproach. One way to do this is to specify the measure in a pre-analysis plan (see below). In this respect, researchers may find it useful to take their lead from the original study: Had the original study produced the replicated results, would it still have reached the same conclusion?

Most of the replication plans stick closely to the original study. As is apparent from the bottom panel of Figure 1, the replication plans, or implied replication plans, of the Special Issue mostly focus on re-working the same dataset used by the original study; either to confirm the reproducibility of the original study's results, or to determine their robustness in the face of

modest changes from the original study's methods or data. With due appreciation of the risk of extrapolating from a small sample, this may represent a general view that "replication" is primarily about confirming the immediate claims of an original study (Clemens, 2017). If so, I think that would be unfortunate. Hubbard and Lindsay (2013) make a compelling argument for "significant sameness"; namely, that research should focus more on identifying the boundaries within which reported findings are applicable. In the Special Issue, only Coupé gave this as a major motivation for his replication plan.

All replications should include a "Reproduction" replication. All of the replication plans in the Special Issue propose to attempt to reproduce the original study's results. My own view is that this should be required in all replication studies.[2] Effort should be made to exactly match the original study's results. If the first attempts fail to do that, the replicating author should try, insofar as the original author is willing to help, to reconcile any differences. The reason for this is that it gives credibility to the subsequent replication results. For if the replicating author is unable to exactly reproduce the original results, how can the reader be confident that the replication study has correctly handled the original study's data, or implemented the original study's procedures? Without this confidence, subsequent replication results are suspect.

Researchers should justify their choice of studies to replicate. A point made by several authors is that replicators need to avoid the appearance that they have chosen a study for the purpose of discrediting it, or are in some other way biased towards it (Hannum, Chang). One way to minimize this suspicion is make the reasons for selecting a particular study as objective as possible. For example, Chang goes to lengths to explain his selection for replication. Brown and Wood note that the replication program at 3ie first selects candidate studies for replication, then matches researchers to do the replication. Nevertheless, this suspicion is difficult to shake given the predilection of journals to publish replications that overturn the results of an original study, thus incentivizing replication authors to find faults (Gertler, Galiani, and Romero, 2018). This motivates the last takeaway.

Researchers should formulate a "pre-analysis" plan and make it publicly available before doing their replication. Because of concern that replication researchers will cherry pick negative findings, Wood and Vasquez, and especially Chang, emphasize the importance of developing a replication plan and making it publicly available prior to undertaking a replication. This serves both to discourage seek and destroy missions on the part of the replicator, and to establish credibility that subsequent results were not cherry picked to emphasize negative findings. This need not handcuff the replicating researcher in their research. As Nosek et al. (2018) emphasize, deviations from a pre-analysis plan are allowed, it's just that they need to be noted. It is then up to the replicating author to convince readers of the motivation for the deviation, that they were not the result of a fishing expedition. Nosek et al. (2018) provide details about how, and where, one can post pre-analysis plans.

I will close this summary with a confession: I have carried out approximately a dozen replication studies. As the editor of the Special Issue, I was the one who specified that submissions consist of replication plans. While I have been associated with replication plans, I have never written a pre-analysis plan myself. One result of this exercise is that I am now

[2] This point was also made by Brown & Wood in their "Don'ts" section.

resolved to do that. It is hoped that readers of this Special Issue will likewise find useful observations and guidance for their own replication research.

Acknowledgments I thank Annette Brown, Andrew Chang, Tom Coupé, Gerald Daniels, Randy Hannum, Venoo Kakar, Bruce McCullough, Dorian Owen, and Ben Wood for helpful comments and feedback. Acknowledgment of their input should not be interpreted to mean they agree with the views expressed herein.

References

Acemoglu, D., Gallego, F.A., and Robinson, J.A. (2014). Institutions, human capital, and development. *Annual Review of Economics*, 6(1): 875–912. https://doi.org/10.1146/annurev-economics-080213-041119

Brown, A.N. and Wood, B.D.K. (2018). Which tests, not which hunts: A diagnostic approach for conducting replication research. *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-53): 1–26. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-53

Chang, A.C. (2018). A replication recipe: List your ingredients before you start cooking. *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-39): 1–8. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-39

Choi, H. and Varian, H. (2012). Predicting the present with Google trends. *The Economic Record*, 88(s1): 2–9. https://doi.org/10.1111/j.1475-4932.2012.00809.x

Clemens, M.A. (2017). The meaning of failed replications: A review and proposal. *Journal of Economic Surveys*, 31(1): 326–342. https://doi.org/10.1111/joes.12139

Coupé, T. (2018). Replicating "Predicting the present with Google trends" by Hyunyoung Choi and Hal Varian (The Economic Record, 2012). *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-34): 1–8. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-34

Daniels, E.D. Jr. and Kakar, V. (2018). Normalized CES supply-side system approach: how to replicate Klump, McAdams, and Willman (2007). *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-19): 1–13. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-19

Gelman, A. and Loken, E. (2014). The statistical crisis in science. *American Scientist*, 102(6): 460-465. https://search.proquest.com/openview/a83b0c4daa7508482a6b7e47eb7b8a3e/1?cbl=40798&pq-origsite=gscholar

Gertler, P., Galiani, S., and Romero, M. (2018). How to make replication the norm. *Nature*, 554: 417–419. https://www.nature.com/articles/d41586-018-02108-9

Hannum, R.J. (2018). A replication plan for "Does social media reduce corruption?" (Information Economics and Policy, 2017). *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018–49): 1–7. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-49

Haurin, D. and Rosenthal, S.S. (2007). The influence of household formation on homeownership rates across time and race. *Real Estate Economics*, 35(4): 411–450. https://doi.org/10.1111/j.1540-6229.2007.00196.x

Hubbard, R. and Lindsay, R.M. (2013). From significant difference to significant sameness: Proposing a paradigm shift in business research. *Journal of Business Research*, 66(9): 1377–1388. https://doi.org/10.1016/j.jbusres.2012.05.002

Jha, C.K. and Sarangi, S. (2017). Does social media reduce corruption? *Information Economics and Policy*, 39: 60–71. https://doi.org/10.1016/j.infoecopol.2017.04.001

Klump, R., McAdam, P., and Willman, A. (2007). Factor substitution and factor-augmenting technical progress in the United States: A normalized supply-side system approach. *The Review of Economics and Statistics*, 89(1): 183–192. https://www.mitpressjournals.org/doi/pdf/10.1162/rest.89.1.183

McCullough, B.D. (2018). Quis custodiet ipsos custodes?: Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive. *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-52): 1–13. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-52

Nosek, B.A., Ebersole, C.R., DeHaven, A.C. and Mellor, D.T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 115(11): 2600–2606. https://doi.org/10.1073/pnas.1708274114

Owen, P.D. (2018). Replication to assess statistical adequacy. *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-60): 1–16. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-60

Reed, W.R. (2018). A primer on the reproducibility crisis and ways to fix it. *Australian Economic Review*, 51(2): 286–300. https://doi.org/10.1111/1467-8462.12262

Santos, F., Fletschner, D., Svath, V., and Peterman, A. (2014). Can government-allocated land contribute to food security? Intrahousehold analysis of West Bengal's microplot allocation program. *World Development*, 64: 860–872. https://www.sciencedirect.com/science/article/abs/pii/S0305750X14002265

Silberzahn, R., Uhlmann, E.L., Martin, D.P., Anselmi, P., Aust, F., Awtrey, E., . . . Nosek, B.A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3): 337‑356. https://doi.org/10.1177/2515245917747646

Wood, B.D.K. and Vasquez, M. (2018). Microplots and food security: encouraging replication studies of policy relevant research. *Economics: The Open Access, Open Assessment E-Journal*, 12 (2018-50): 1–12. http://dx.doi.org/10.5018/economics-ejournal.ja.2018-50

Appendix: Figure 1

A: Six Different Types of Replications

Measurement and/or Analysis	***Source of Data***		
	Same dataset	***Same population***	***Different population***
Same	*(1) Reproduction*	*(3) Repetition*	*(4) Extension*
Different	*(2) Robustness Analysis – Same Dataset*	*(5) Robustness Analysis – Same Population*	*(6) Robustness Analysis – Different Population*

Source: *Reed* (2018)

Classification of the Special Issue Articles

Measurement and/or Analysis	***Source of Data***		
	Same dataset	***Same population***	***Different population***
Same	Chang McCullough		
	Wood & Vasquez Hannum Brown & Wood Owen		
	Daniels & Kakar Coupé	Daniels & Kakar Coupé	Coupé
Different	Wood & Vasquez Hannum Brown & Wood Owen		