Research Replication: Practical Considerations

R. Michael Alvarez
California Institute of Technology

Ellen M. Key
Appalachian State University

Lucas Núñez
California Institute of Technology

R. Michael Alvarez is a professor of political science at California Institute of Technology and co-editor of *Political Analysis*. He can be reached at rma@hss.caltech.edu. Ellen M. Key is an assistant professor of political science at Appalachian State University. She can be reached at keyem@appstate.edu. Lucas Núñez is a graduate student at California Institute of Technology and the graduate editorial assistant at *Political Analysis*. He can be reached at lnunez@caltech.edu.

With the discipline's push toward DA-RT, journal replication archives are becoming more and more common. As researchers work to ensure replication materials are provided, they should also pay attention to the content—rather than just the provision—of journal archives. Based on our experience in analyzing and handling journal replication materials, we present a series of recommendations that can help make replication materials easier to understand and use. The provision of clear, functional, and well documented replication materials is key for achieving the goals of transparent and replicable research. Furthermore, good replication materials better boost the development of extensions and related research by making state-of-the-art methodologies and analyses more accessible.

Over two decades ago, Gary King published an important and then provocative paper, "Replication, Replication" (King 1995). In his paper, King presented a simple claim, *"the only way to understand and evaluate an empirical analysis fully is to know the exact process by which the data were generated and the analysis produced"* (King 1995, 444, emphasis in original). At the time, most social scientists tried to make it clear how their analyses were conducted, usually by providing important details in footnotes or appendices. But often there was insufficient space in journals or books for much important detail to be provided. Instead, King argued for a different approach, for authors to make available the actual data used to make an empirical claim, and all of the materials necessary to manipulate and analyze that data to reproduce the results underpinning the empirical claim.[i]

At the time King wrote that essay, the general principle of research replication was widely discussed, and many researchers began to think about ways to make their research materials available to other scholars. Two decades ago smartphones and the "cloud" did not exist, electronic storage and sharing of large datasets or other materials was not straightforward, and there were few mechanisms scholars could use to make their replication materials available. There were certainly some places where authors could store replication materials, in particular the ICPSR's "Publication-Related Archive,"[ii] which launched about the time that King published his paper. But few scholars at that time used facilities like the ICPSR to store their replication materials, and few journals or funding agencies required researchers provide easily-accessible replication materials.

In recent years, however, the situation has dramatically changed. First, cloud-based computing has made data and code sharing simple, some would say trivial. Researchers can easily share their materials from their own file-sharing archives (Dropbox, Box, or Google

Drive), from code-sharing archives (GitHub or Bitbucket), or by using a shared archive either at their institution or those provided for researchers (like Dataverse). Second, researchers in many graduate programs are now trained to build replication into their workflows, and many have come to recognize that providing data, code, and other research materials helps boost their visibility and increases citations. Third, many journals and funding agencies now ***require*** that research materials be made available upon publication. Finally, due to a number of highly publicized issues regarding research transparency in the social sciences, professional organizations, colleges and universities, and other advocates are pushing researchers to be more open regarding the details of their research.

Thus, research replication and transparency has become a central concern in the social sciences, which we argue strengthens the research done in political science and other disciplines. That another scholar can easily and quickly confirm the published results in a paper helps that scholar gain confidence in the integrity of the published result. When that scholar can use replication materials to test the robustness of those published results in various ways and possibly improve upon the methodology or analysis previously published, it helps build the type of cumulative knowledge that makes for a better social science.

Journals and Replication Policies

Political Analysis (the journal of the Society for Political Methodology, SPM) was among the first journals in social science to develop a replication policy for papers published in the journal, beginning after the publication of King's 1995 paper. For example, in 1996 some papers published in *Political Analysis* stored their replication materials in the ICPSR archive (Box-Steffensmeier and Lin 1996). However, as the policy was largely voluntary, other articles published in the journal at that time make no mention of replication materials. By 2000, the

journal began a more systematic collection of replication materials, storing them on the SPM website.[iii]

Beginning in 2012, *Political Analysis* developed a new replication policy—all papers reporting data analyses (including simulation modeling) would be ***required*** to store the materials necessary to replicate the results reported in each published paper in the journal's Dataverse, prior to publication.[iv] This made *Political Analysis* one of only a few journals prior to 2015 requiring the provision of replication materials prior to publication.[v] Presently, replication materials are requested before the final acceptance of the manuscript for publication, and are reviewed by both the editor overseeing the review of the manuscript and one of the journal's graduate editorial assistants. Only after the replication materials have been reviewed are they released to the public on the journal's Dataverse, and the paper is sent to the publisher for production. Although many issues arise during our review of replication materials, we have not had authors who have refused to meet the journal's current replication requirement.

Meeting Replication Requirements

In April 2016, one of us published a study in *PS*: "How Are We Doing? Data Access and Replication in Political Science'' (Key 2016a). This study examined the replication polices of six major journals, including *Political Analysis*, and looked to see how many of the papers published recently in those top journals had replication materials. The study found that during the period of the study, there was an important bifurcation in the percentages of papers published that had available replication materials. Three of the journals had a high percentage of studies with no replication materials, with 67.6% of the papers published in the *APSR* not having replication materials available, followed by the *JOP* (51.1%), and the *BJPS* (50.0%). At the other end of the distribution were *IO* (10.2%), *AJPS* (9.3%), and *Political Analysis* (1.9%).

Ordinarily, the fact that 98.1% of the papers published in this period in *Political Analysis* had available replication materials might be cause for celebration, especially given that the journal had the highest compliance rate among the major journals in political science in the study period. However, from the journal's perspective, *Political Analysis* has universal compliance with the replication requirement.[vi] This discrepancy highlights the importance of journal replication archives and raises the issue of accessibility.

Because Key (2016b) published a replication archive, it was possible to identify the *Political Analysis* article by Bowers, Fredrickson, and Panagopoulos (2013) that was coded as being unavailable. Although Bowers et al. had provided a replication archive on the journal's Dataverse, additional software was needed to extract the file. In other words, the issue was not with the availability of the replication material, rather, it was that the single file containing the materials was in a format that is not easily accessible using standard software available on Macs or PCs.[vii] The more steps involved in retrieving replication materials, the less useful these materials become.

Like a ramp that is too steep, journal replication archives that require users not only download the data they want to access, but also require users to download specialty software to open the replication file, make the archives inaccessible to a wider audience. Although many researchers at R1 institutions have access to a bevy of statistical packages and powerful machines, graduate students and those at smaller institutions may not be as a fortunate. This resource discrepancy highlights tension between the admirable goals of DA-RT and the realities faced by authors, editors, and replication archive users. This led us to understand that there are a number of important, pressing problems regarding replication materials—in particular, making sure that the materials are provided in usable and accessible formats.

Common Issues that Arise with Journal Replication Materials

By now, we have a lot of experience with replication policies and materials. Recently, *Political Analysis* has begun to devote more time and resources to the process of reviewing and using the replication materials prior to the final acceptance of a paper for publication. It is during the verification stage we often find many common issues that arise with replication materials. Here we discuss many of those issues, and later we present a number of recommendations for authors and other journals to improve the practice of replication and minimize the accessibility issues that arose with the Bowers et al. replication materials. The most common issues found in replication materials can be broadly grouped into three categories: organization, clarity, and usability.

Organization

Many replication packages include several files with data, codes, and codebooks. In some cases, these files are simply bundled together in a folder or a compressed file without any indication of what each of the files contains and how a user of the replication materials should proceed. The problem is usually compounded by the use of obscure files names that convey little information about the content.

Clarity

Clarity refers to problems understanding codes or scripts. It is relatively common for authors to provide codes with little to no annotation in them. Some codes are straightforward because the manipulations of the data and the estimation techniques used are fairly simple. In other cases, neither the manipulations nor the estimation techniques are standard. In these cases, the lack of guidance in the code itself can render the replication materials extremely hard to understand and use. This can significantly hinder the validation of the data manipulation and

analysis, as well as the ability of users to improve upon the analyses or techniques presented in the paper.

A second factor that affects the clarity of the replication materials is the way the outputs of the statistical software relate to the figures and tables presented in the paper. For example, tables containing the output from several regressions are sometimes not produced in the code, with the code instead producing separate outputs for each regression. This leaves the user to figure out what column in a table corresponds to which line of code. This can lead the author to make mistakes in (manually) transferring multiple outputs into a common display. This is made more challenging when they are not in the same order or do not always report the full outcome produced by the code.

Finally, simulation studies and many estimation techniques use randomly generated numbers or samples (e.g. some types of bootstrap). When replication materials fail to provide the random sequence (the seed) and pseudorandom number algorithm used to generate the results in the paper, it makes the comparison of the output obtained by the user and those available in the paper harder to assess.[viii]

Usability

The final broad class of issues relate to the usability of the replication materials. Many replication packages involve the use of multiple interdependent files. The dependencies between these files are not always clear. Moreover, in many cases they require a folder structure that is not provided by the author. Often there is no clear indication of what the user should modify in the codes for the dependencies to work.

A second usability problem refers to software and packages in general. Some statistical software and statistical packages work well in one operating system, but not others.[ix] In other

situations different versions of software and statistical packages can produce different results because of updates, bug fixes, different optimization methods, etc. Lack of information about the software version used by the author, and missing information about software and package dependencies needed to replicate the results in the paper, can be barriers to replication.

Finally, with the proliferation of Markov Chain Monte Carlo and other related estimation techniques, some replication codes require multiple hours (or days) to produce the estimates. In many cases this is not indicated by the authors, which can lead to users believing that the code is not functioning or that there is some other problem. Additionally, authors are increasingly making use of parallel computing, whose operationalization can differ across operating systems and typically depends on the capabilities of the computer being used. For users unfamiliar with parallel computing, this can create a significant barrier for replicating and using the materials provided by the author(s).

Recommendations

Based on our experiences with journal replication archives, we offer the following recommendations to authors and other journal editors to improve organization, clarity, and usability. While the widespread practice of providing replication materials is relatively recent, we hope in the near future journals, professional societies, and publishers might come together to develop common standards for replication materials.[x]

1. ***Archive.***
    a) Journal replication materials should be stored in permanent archives that ensure their availability after long periods of time. Although there are many data storage options available, some are more durable than others. Personal websites are susceptible to link rot due to website reconfiguration, changing institutions, or

other maintenance issues (Key 2016a). At present, the best archives are those with an institutional guarantee, such as Dataverse or other university or consortium-backed archives.

b) Authors should try to store their materials using file formats that are likely to be accessible in the future (for example, use comma-delimited or flat text files for their data, rather than specific proprietary software files).[xi] Besides the preservation advantage of using these types of files, they can be easily read on currently available statistical software, simplifying their use across different platforms. If it is important that files be compressed or stored in some particular format, they should provide some documentation for users who may not be familiar with how to un-compress or access these files in the future.

2. ***Readme Files.*** The creation of a clear and sufficiently (but not overly) detailed readme file is a key element of every replication package. There are several items that a readme file should include:

   a) A reference to the associated paper or publication.

   b) A short description of the files and file types that are included in the replication package: raw data, processed data, scripts to manage the data, scripts to produce estimates, etc.

   c) An indication of the order in which the scripts are to be run as well as noting where the different tables and figures found in the associated publication are generated and stored. The authors should also indicate whether there are intermediate outputs generated by one script and used by another, so that it is

clear to users how to proceed (e.g. if a script handles the raw data to produce a processed dataset to be used in another script, this should be noted in the readme).

d) A list of software and software packages (as well as the dependent packages they rely upon), and the operating system used to produce the results in the paper. Technical information on the hardware used to produce the results is also helpful (e.g. the number of cores), especially if computationally demanding techniques were applied. Authors should also indicate which versions of the packages were used, as packages are frequently altered to fix bugs or other issues. These version changes can produce different results when running the same code with a different version of the package or software. Some models also require additional software like JAGS or a C++ compiler to be available in the computer. Indicating this in the readme file can prove useful for a less familiar or inexperienced user.

e) If unusual file extensions are used, authors should clearly indicate, to the extent possible, how to proceed with these files.

3. ***File Names.*** File names should be easy to understand and provide information about the content of the file. This is particularly important for replication materials that include multiple scripts and data files. If multiple scripts are to be run in a certain order, including the order in the file name is useful (e.g. "1_DataProcessing", "2_Estimation", etc.). Alternatively, authors can create a master file that calls on the different scripts.

4. ***User Created Software and Packages.*** Sometimes researchers create their own statistical packages to produce the results in the paper. These packages should ideally be archived in a stable location, like CRAN for R. If possible, authors should also include a copy of the package or software in the replication package.

5. ***Operating System Compatibility.*** Ideally, authors should make sure that their replication materials function under common operating systems. If this is not possible, the authors should indicate on what operating system(s) they function.
6. ***Script Documentation.***
    a) The scripts should include a short description of what the code does (e.g. "data recoding"), what dependencies it has, what package(s) it requires, and what the outputs are.
    b) Authors should avoid scripts that include several pages of commands with little or no indication of what each line (or group of lines) is doing. Code should be commented where necessary to indicate the purpose of each line or group of lines. This makes identifying specific parts of the data management or estimation process easy and also helps in the identification of potential issues or mistakes.
    c) Authors should also strive to organize their scripts well. To this end, authors should avoid interspersing data management and recoding with estimations, unless absolutely necessary for the analysis.
7. ***Outputs.*** Replication scripts should have as a clearly identified output (either a file or an object in the statistical software) the tables and figures that are included in the paper, exactly as they appear in the paper (to the largest extent possible). There are many ways authors can achieve this. For example, for tables with several models estimated in Stata authors can use a combination of *eststo* and *esttab*, among many other commands available. In R, authors can use *stargazer*. For tables that are more ad hoc, authors can store the outputs of multiple commands into an object (typically a matrix) in the statistical software that can then be printed into a file.

8. ***Excluded Outputs.*** Many replication materials include scripts (or parts of scripts) and data for outputs that are not included in the final paper (typically excluded at some point in the review process) or relegated to an online appendix. When this is the case, authors should clearly differentiate between what is included in the paper and what is not. This is particularly important for excluded robustness checks that tend to look very similar to the main results included in the paper and can be a cause for confusion.
9. ***Intermediate Outputs.*** In many cases, scripts generate intermediate outputs that are then further processed to produce the final outputs included in the paper. In certain circumstances, these intermediate outputs require a significant amount of time to be generated. Among many others, these include the generation of simulated datasets in simulation studies and the Markov Chains when estimating via MCMC. When this is the case, it is useful to include these intermediate outputs in the replication materials.
10. ***Random Sequences.*** For simulation studies or estimation strategies that use randomization, authors should always include the random sequence of numbers used to produce the results in the paper (the seed) and the pseudorandom number algorithm. Although the exact seed used should not be critical in terms of the results obtained in the paper, the availability of the seed simplifies the replication of the exact numbers in the paper. The location of the seed should be clearly identified at the beginning of the code in case the user wishes to change it.
11. ***Directories and Folders.*** Directory paths should be easily identifiable in the scripts so users can change them accordingly. Ideally, the directory setting should only need to be specified at the beginning of the code; replacing multiple directories in different parts of the code can turn into a cumbersome task. If the replication materials are organized into

multiple folders, authors should indicate at the beginning of each script where to set the main directory. All directory changes within the scripts should be automatic.

12. ***Computing Time.*** Authors should indicate (both in the code and readme file) whether a particular script or part of a script requires a considerable amount of time to compute. This is important, because unaware users facing an estimation that takes a long time, might incorrectly assume that there is an issue with the materials.
13. ***Parallel Computing.*** Authors are increasingly making use of parallel computing in their scripts as it can speed up computations significantly. Unfortunately, not all parallelized scripts work on all computers and operating systems. For this reason, authors should take care in highlighting the part of the script that is being parallelized and suggest how to proceed in case of incompatibilities with the user's computer or operating system.
14. ***Warnings and Errors.*** It is not unusual to find scripts that correctly reproduce the output in the paper but for some reason return warnings or errors (usually related to inappropriate use of a package, or differences between operating systems). When this is the case, authors should note the reasons why these warnings and errors are not a concern.

All replication materials are different, and creating a set of recommendations that applies to each and every one of them is a monumental task. A simple rule to for authors to follow when creating their replication materials is to put themselves in shoes of another person whose only knowledge about the replication materials in question is what the author provides. Looking into the future, the construction of well organized, clear, and usable replication materials hinges upon the training of scholars so that research replication is effectively incorporated into their

workflows. Researchers can find further advice, particularly in relation to coding practices, in the articles by Nagler (1995) and Bowers and Voors (2016).

Conclusion

Replication materials are becoming more and more common. As political scientists (and social scientists in general) work to ensure journal replication archives are provided, researchers also need to pay attention to the content—rather than just the provision—of those archives. Based on our experience in analyzing and handling journal replication materials, we have presented a series of recommendations that can help make replication materials easier to understand and handle. By improving organization, clarity, and usability, authors will bolster the accessibility of their replication materials. While the list of recommendations we make is by no means intended to be exhaustive, we believe it can prevent many important problems that arise with replication materials. The provision of clear, functional, and well documented replication material is key for achieving the goals of transparent and replicable research. Furthermore, good replication materials better boost the development of extensions and related research by making state-of-the-art methodologies and analysis more easily accessible.

References

Bowers, Jake, Mark M. Fredrickson, and Costas Panagopoulos. 2013. "Reasoning About Interference Between Units: A General Framework." *Political Analysis* 21(1), 97-124.

Bowers, Jake, and Maarten Voors. 2016. "How to Improve Your Relationship with Your Future Self." *Revista de Ciencia Política* 36 (3): 829-848.

Box-Steffensmeier, Janet M. and Tse-min Lin. 1996. "A Dynamic Model of Campaign Spending in Congressional Elections." *Political Analysis* 6(1), 37-66.

Eubank, Nicholas. 2014. "A Decade of Replications: Lessons from the *Quarterly Journal of Political Science*." *The Political Methodologist* 22(1): 18-19.

Key, Ellen M. 2016a. "How Are We Doing? Data Access and Replication in Political Science." *PS: Political Science and Politics* 49(2), 268-272.

Key, Ellen M. 2016b. "Replication Data for: How Are We Doing? Data Access and Replication in Political Science" doi:10.7910/DVN/5LJAMC, Harvard Dataverse, V2

King, Gary. 1996. "Replication, Replication." *PS: Political Science and Politics* 28: 444-452, September.

Nagler, Jonathan. 1995. "Coding Style and Good Computing Practices." *PS: Political Science and Politics* 28 (3): 488-492.

[i] In this paper, we will generally assume that the replication materials are from quantitative research, not qualitative research. We take that approach as the principles associated with research replication and transparency for quantitative materials are generally agreed-upon in the quantitative research community. The statements we make in this paper can (and should) generally apply to qualitative research.

[ii] Accessible at https://www.icpsr.umich.edu/icpsrweb/deposit/pra/index.jsp

[iii] This has led to some confusion, as SPM's website has much of this replication material, but the official website for the journal (run by the publisher) does not have that replication material. So those interested in finding and using those materials need to be diligent in their search. We also suspect some authors have stored replication materials

on their own personal or research websites, but we have not made a systematic search to determine how many of those websites still exist.

[iv] Accessible at https://dataverse.harvard.edu/dataverse/pan

[v] The *Quarterly Journal of Political Science* is another journal in the discipline which has had a rigorous replication policy (Eubank 2014).

[vi] Some research articles or letters published in the journal are commentaries, reviews, or critiques; as they do not contain simulations nor quantitative analyses, they have no materials subject to journal's replication policy.

[vii] The file was a tar archive file, which had been further compressed using the GNU gz format. While these formats are well-known to Unix and Linux users, they may require the installation of special software by users of other platforms.

[viii] This issue should not present with major complications, as any robust simulation study or estimation technique should not see its results affected by the random sequence used; the results obtained with any random sequence should be qualitatively the same, and quantitatively extremely similar.

[ix] In some cases, the operating system used has an impact in how the statistical software interacts with other software. For example, the using the package Rcpp, that allows for C++ operations in R, requires the user follow different instructions in Windows and Mac to allow C++ and R to communicate correctly. In other situations, parallel computing requires different configurations for different operating systems.

[x] Although providing a universally accessible archive is the gold standard, we hope researchers who have difficulty meeting all the following criteria do not use that as an excuse to avoid providing replication materials at all. In other words, something is always better than nothing.

[xi] While codebooks may be developed internally (e.g. as part of a Stata .dta file), this will be lost when files are saved in a flat file format. For this reason we also recommend the inclusion of a separate codebook file.