Political Analysis Replication Guidelines

(June 20, 2022)

Political Analysis requires conditionally accepted manuscripts to go through a rigorous replication process and demonstrate full reproducibility. The authors of conditionally accepted manuscripts must provide data, code, and other necessary materials and make them publicly available prior to publication. The editor should be notified at the time of submission if access to the data used in the paper is restricted or limited, or if, for some other reason, the requirements above cannot be met. The replication team at *Political Analysis* will assess compliance with these guidelines, and replicate and verify the information and results of the manuscript prior to approval and formal acceptance by the editor.

This document describes how to prepare a replication package, including both general principles and specific instructions. The objective is to establish a broad standard for the information that must be made available to demonstrate the reproducibility of the results that appear in *Political Analysis* and to allow the broad scientific community to evaluate and use the methodology proposed in the research. *Political Analysis* recognizes that the guidelines may not cover all the situations that may occur. Therefore, some adjustments can be made on a case-by-case basis. Nevertheless, the following principles should sufficiently address the vast majority of research contexts.

General Scope and Principles

The authors of conditionally accepted manuscripts should provide all necessary information that is required to reproduce and evaluate the methods, results, and arguments reported in the papers. *Political Analysis* especially encourages authors to prepare usable and helpful documentation and programs for eventual readers of their methodology and replication code. The authors are only required to provide enough information to replicate the results in the article. However, the more information the replication package provides, the more likely readers can follow upon the article. The replication materials generally include (a) the data set(s), (b) the code and programs used to conduct analysis, (c) descriptions and instructions sufficient to reproduce the results in the manuscript using the data and code provided.

For research that contains empirical data, either original analysis or replication of previous studies, in addition to providing the data file(s), the replication materials shall also include (d) description and instructions sufficient to obtain all data from their original source. For simulation works, the materials shall include (e) the code and programs that allow reproducing the simulation results from scratch.

For original data collection through surveys or experiments, the replication materials shall also include (e) survey or experiment instruments, (f) programs or descriptions of experiment assignment and survey collection mechanisms, (g) instructions and details on subject selection. If the experiment contains interactive features, the materials shall also include descriptions or records of the interactions, in particular the treatment mechanisms that subjects are exposed to.

For papers that involve adaptive, computer-assisted design and data collection (e.g. adaptive experiment design, online validation, web scraping, etc.), or any parts besides data analysis that contain code or computer program, the replication materials shall also include (h) the code and programs that can create the design mechanisms or data, and (i) the original data collected through the mechanisms. If the mechanisms contain human intervention in computer-assisted design and data collection, such as crowd-sourcing or human coding, the materials shall include (j) the instructions given to the subjects or coder, and (k) original data of human responses.

For non-public data, such as confidential, embargoed, copyrights-protected or other restricted-access data, the authors should notify the editor at the time of submission. Generally, all the code, including the ones that are used on the restricted-access data, shall be made public. The goal is, for readers who have access to the same data set(s), they should be able to use the provided code and reproduce the results. While the restricted-access data may be excluded from the published version of the replication package, they are still subject to replication with accommodations made on a case-by-case basis. Authors are also encouraged to seperate the restricted part of the materials into a different step (i.e. code scripts) in order to make most of the materials publicly available. *Political Analysis* also encourages authors to employ privacy protection methods at the beginning of the study to maintain confidentiality and privacy while ensuring the availability and accessibility of the methodology and data.

For qualitative studies, the replication materials should include all of the raw information and materials necessary to reproduce the analyses reported in the paper. For qualitative studies using matrix approaches, such as a dataset that is analyzed holistically, the replication materials should generally contain the source information and the collection and derivation process if such information is not provided in the paper (including supplement information). For qualitative studies using granular approaches, all fragments of orginal sources should be included in the replication materials.

The replication materials will be released and permanently archived on the journal's Dataverse. Thus, authors should bear in mind that code, documentation and data will be in the public domain, and thus all should be edited carefully. In particular, authors shall properly handle personally identifiable information and copyrights of the data and code. Generally, no individually-identifying information should be present in published replication datasets. Thus, the relevant variables or cell values should be anonymized before publishing. The reuse of third-party data should comply with the terms of service and policy of the original collector or distributor.

Instructions

The replication data and code will be eventually uploaded and published in a Dataset within the *Political Analysis* Dataverse at http://dvn.iq.harvard.edu/dvn/dv/pan, on the Harvard Dataverse, and cited in the final version of your manuscript. *Political Analysis* currently conducts replication both in-house and through computational research platform, Code Ocean. Authors are highly encouraged to use Code Ocean to prepare and run the replication materials as it gives authors more control of the process and will also allow readers to view the materials interactively.

Replication package

A replication package should generally contain (a) a README file that provides the descriptions of the files that are part of the replication package and instructions for how to use the replication package, (b) code scripts for conducting the analyses and producing output figures/tables, (c) the data sufficient to reproduce the results reported in the paper. The following describes these elements in greater detail:

Readme: *Political Analysis* recommends the README to follow the schema provided by the Social Science Data Editors' template at https://social-science-data-editors.github.io/template_README/. The README file should not require proprietary software to view. Common formats are TXT, RTF, PDF, Markdown, and HTML.

The README shall contain the following items:

- Descriptions about data and code availability that contain information about the sources of data used in the replication package, in addition to such statements in the manuscript. It should provide necessary information for how to obtain the original, raw data. These may include required registrations, memberships, application procedures, monetary cost, or other qualifications.
- Computational requirements that include (1) the software and programming language used, including version info, (2) additional packages/libraries and their viersion info or similar, as used, (3) the hardware specification as used by running the code in terms of OS, CPU generation and quantity, memory and necessary disk space (If multiple computers were used, the specification for each should be identified), and (4) the computing time given the provided hardware platform, expressed in appropriate units (minutes, days).
- Instructions that describe the requirements and procedures for reproducing the results. They should specify how to use the provided files and run the code.
- List or table of content that specifies the file names and their usages. For output figures/tables, it should also include the corresponding references in the manuscript.

Data: The data files may be provided in any format compatible with any commonly used statistical package or software. Authors are encouraged to provide data files in open, non-proprietary formats. Authors should ensure that a meaningful name or description (label) is available for every variable in the provided datasets. Codebooks or similar metadata should describe the allowed values and their meaning for each variable. It is acceptable to reference publicly available documentation for these items.

Code: Code and programs should be provided in formats compatible with commonly used statistical packages or softwares. Should costly proprietary or unusual software be required, authors are required to notify the editor before the replication process starts. *Political Analysis*, along with partner Code Ocean, currently supports R, Python, MATLAB, Stata, Julia, etc. Authors are also encouraged, sometimes required, to adapt their code that was originally developed in proprietary softwares into an open-source software for broader audience and usage. The code also should not require a specific type of operating system, platform (e.g. HPC/cloud), or IDE to run unless it's absolutely unavoidable. If the installation of software or packages is beyond the general procedure, such as packages/libraries that are not available through the default installation methods, please provide a setup script to install those. The installation should also be a separate step (i.e. in a dedicate code script) instead of during runtime. The replication package should use a minimal number of automated scripts. A master script is strongly encouraged when there are multiple code scripts. The code should generally be able to run in a non-interactive fashion, with no manual interventions required within a script unless unavoidable. For

the sake of user-friendliness, the programs and code should avoid absolute/user-specific path.

Output: The output for all code-created figures (.pdf, .png, jpg, etc.) and tables (.csv, .txt, .tex, etc.) in the main text and the appendix should be saved. All output figures/tables are strongly recommended to be well-named, such as using the corresponding numbers from the manuscript (e.g. “figure1.pdf”, “table2.csv”). The saved output needs to be included in the replication package. The goal is to run the code and then overwrites the provided files. In the occasion where a complete reproducible run requires intensive computing resources and running time, it is recommended to provide intermediate objects, data, and results that can be used to start at different stages of the analysis.

Folder structure: The replication files should be well-organized for the sake of user-friendliness. A flat structure should only be used when there are only a limited number of files and the content of the files are straightforward. For most cases, replication package should contain sub folders according to different types of the files (\data, \code, \figure, \table, etc.) and/or to different steps of the study (\1.analysis1, \2.analysis2, etc.). README and master script should be in the root directory. Manipulation of files and folder structure, including creating or removing folders, unzip files, etc., should be done through code.

- A simple folder structure might be:

  `README.md`

  `master.R`

  `data/`

    `raw/`

      `anes.csv`

    `analysis/`

      `sim1.csv`

      `sim2.csv`

  `code/`

    `01_data_processing.R`

    `02_simulation.R`

    `03_analysis.R`

  `results/`

    `table1.csv`

    `table2.csv`

    `figure1.pdf`

    `figure2.pdf`

Miscellaneous issues:

- The authors may contact the replication team when preparing the materials regarding what information, data, and code should be provided, especially if (some of) the elements or procedures described in this

document don't apply to the particular study.
- For studies that involve simulation and stochastic process, random seed(s) should be set up in the script to ensure reproducibility. The authors are also encouraged to test the sensitivity of the results to a specific seed.
- If the replication material takes a long time to run (i.e. weeks), we encourage the author to optimize the code, use smaller data or simulations, or adopt other relevant approaches, to reduce the computing time if possible. Especially for methodological works, a reasonable computing time will also help the proposed method to be widely received and adopted.

Step-by-Step Guide

1. **Prepare**: Prepare your data and code replication package, including README, data, and code scripts. This step can be done at any time no later than the submission deadline for the final manuscript. It is recommended the author to prepare the package as early as possible, even before submitting the manuscript to *Political Analysis*. Once the replication package is finalized, the author should re-execute the entire process, and reproduce the tables and figures in your manuscript faithfully. For Code Ocean, this step can be done during the submission.
2. **Upload**: Provide metadata and upload the replication package. This step simultaneously prepares the materials for the replication process. For Dataverse submission, the replication should be uploaded to the *Political Analysis* Dataverse at http://dvn.iq.harvard.edu/dvn/dv/pan. For Code Ocean, the author needs to create a capsule on Code Ocean at https://codeocean.com/ and upload the replication files.
3. **Submission**: Submit the replication dataset on Dataverse or share the Code Ocean capsule with the replication team (political.analysis.replication@gmail.com) after a successful reproducible run. The citation of the replication package should be added to the manuscript's references. When this process completed, please email confirmation to politicalanalysis@cambridge.org.
4. **Replication & Verification**: The replication package will go through a reproducible run and be verified by the replication and editorial team. For Dataverse submissions, the replication team downloads the material from Dataverse and conducts replication. The replication team will contact the author during the replication team if any additional information or changes are needed. For Code Ocean replication, the author will conduct a reproducible run themselves after setting up the replication package on Code Ocean. Once the reproducible run is successfully finished, the author can share the capsule with the replication team for verification.
5. **Approval and Forthcoming**: After the replication is finished and verified, the replication team will report the results to the editor. The editor will approve the replication if the replication package is deemed satisfied. Once the replication is approved, the replicator of the paper will publish the Dataverse and/or Code Ocean entry and contact the author with the reference information of the replication package. This starts the copy-editing process, and at this time, the manuscript becomes "**forthcoming**."

Instructions for Code Ocean

- Create a capsule for the replication package on the Code Ocean website (https://codeocean.com/). Code Ocean account can be signed up for free. Email address with .edu will receive 10 hours of free computing

time.

- Set up the computing environment. Code Ocean offers several options with pre-configured software packages including R (RStudio), Python (and with GPU support), MATLAB, Stata, TensorFlow, Julia, etc. (please refer to https://help.codeocean.com/en/articles/1120266-which-programming-languages-does-the-platform-support for more details). Additional packages/libraries can also be added. All the dependencies should be installed at this stage rather than runtime.
- Enter metadata to describe the paper, including name (paper title), research field (Social Sciences), description (the abstract of the paper), authors and affiliations, corresponding contributor and contact info, license for code (MIT is recommended) and data, and associated publication (yet to be published, title of the paper, Political Analysis).
- Add code and data files to respective folders. This may need adjustments for folder structure and paths in code scripts to be compatible with Code Ocean's setup. Every folder except /results is reset after each run. All the runtime results that are needed to be saved, especially output figures and tables, should be directed to `/results` folder. Intermediate products passed between scripts that don't need to be verified can be saved to `/data` or another folder. Additional information about the is available at Code Ocean help documentation (https://help.codeocean.com/).
- Run the material in the capsule by clicking on Reproducible Run. This will run the material in the capsule non-interactively. Any issue/error will stop the reproducible run. The material may need to be adjusted a couple of times until it runs without errors.
- After a successful reproducible run without any issue or error and the output results are identical to the manuscript, please share the capsule with the replication team (political.analysis.replication@gmail.com) and email a confirmation to pdxeditorial@gmail.com.
- Once the replication is verified and approved, reference information of Code Ocean capsule and Dataverse will be available. During the copy-editing stage, the authors can insert the complete citation in the manuscript's references, and refer to this in both the Data Availability Statement in the manuscript and in a footnote in the manuscript around where it first mentions the data or analysis.

Instructions for Dataverse

- Go to: https://dataverse.harvard.edu/dataverse/pan and click "Add Data" link on upper right box.
- Enter cataloging fields to describe the paper and replication package. The minimum information should include (a) title ("Replication Data for: [paper title]"), (b) author name(s) and (c) contact information, (d) description (abstract of the paper and/or description of the replication package), (e) subject (Social Sciences), (f) related publication ("Forthcoming, Political Analysis" for the initial submission).
- Save the cataloging information, click Add Files, and upload the data files, code, documentation, and a brief description of what each of the files are. We recommend using tabular data files in one of the formats Dataverse presently recognizes, in which case it will process the files and provide additional formats to the end user.
- After all the files are uploaded, click "Send study for review"
- The citation information will be available upon the creation of dataset entry. Please insert the complete citation in the manuscript's references, and refer to this in both the Data Availability Statement in the manuscript and in a footnote in the manuscript around where it first mentions the data or analysis.
- Having completed this process, please email confirmation to pdxeditorial@gmail.com.