AJPS AMERICAN JOURNAL of POLITICAL SCIENCE

AJPS, South Kedzie Hall, 368 Farm Lane, East Lansing, MI 48824
ajps@msu.edu, (517) 884-7836

GUIDELINES FOR PREPARING REPLICATION FILES

Version 2.1, May 19, 2016

William G. Jacoby
Robert N. Lupton

Michigan State University

The *American Journal of Political Science* requires the authors of all accepted manuscripts to provide replication files before the article enters the production stage of the publication process. The replication files for each article must be made available as a Dataset (i.e., a collection of files) located in the *AJPS* Dataverse on the Harvard Dataverse Network. Instructions for getting started on the *AJPS* Dataverse can be found in the "Quick Reference for Uploading Replication Files," available on the *AJPS* website.

This document provides some guidelines, including both general principles and specific requirements, for preparing replication files.[1] The objective is to establish a broad standard for the information that must be made available showing how to reproduce and evaluate the work that appears in the pages of the *American Journal of Political Science*. This standard should facilitate and encourage active use of the replication files by interested members of the social science research community.

These guidelines are designed to fulfill the universal principle that evidence-based claims must be made with transparency. But, the *American Journal of Political Science* recognizes the heterogeneity of social inquiry. Accordingly, several sets of guidelines are offered below, each providing different instructions for different types of work. Scholars employing multiple methods should follow the guidelines for each type of method they use. The *AJPS* also recognizes that these guidelines still may not cover all data analysis situations that may occur. Therefore, small exceptions to the general rules may be necessary in order to accommodate some individual articles. Nevertheless, these guidelines should sufficiently address the vast majority of research contexts.

General Principles

All analyses reported in *AJPS* articles are open and available to the scientific research community. Authors are not permitted to "embargo," or withhold, information that has been used to perform an analysis featured in an *AJPS* article (except as described below, in the section on "Possible Exceptions to Data Access Requirements"). Instead, authors should provide all information that

[1]Staff, affiliates, and users of the Qualitative Data Repository (Syracuse University) provided the guidelines for qualitative analyses and made a number of other important contributions throughout this document. We are very grateful for their hard work, editorial skills, and strong support.

The Editor of the AJPS is at Michigan State University and the Editorial Office is supported by the MSU Department of Political Science and the MSU College of Social Science

is required to reproduce and evaluate any analytic result (in quantitative analyses) or central inferential/interpretive claim (in qualitative analyses) that is reported in their article.

Authors do not need to provide any additional information or data, beyond what is necessary to reproduce and evaluate the analyses reported in the *AJPS* article. So, for example, the source dataset may contain variables that are neither employed in any models, nor used to construct variables that are employed in reported models. If so, then these variables need not be included in the replication materials for the article.

Instructions and Recommendations for Quantitative Analyses

In most cases, the replication materials will include: a Readme file, information about the data source(s); the analysis dataset; and code for running relevant software. Each of these elements will be discussed in greater detail, below. Optionally, authors may include extensions to the analyses reported in the article. Note that the replication materials are expected to cover all analyses presented in the main article and in the article's Supporting Information.

Readme File:

Every Dataset in the *AJPS* Dataverse must include a plain-text file named "`readme.txt`". This file provides the names of all other files contained in the Dataverse Dataset, along with a brief description of each one. For relatively small Dataverse Datasets, this information can be presented as a simple list. Larger Dataverse Datasets should group similar files under subheadings, such as "Data files," "Stata `.do` files," or "Files to Reproduce Table 1," and so on.

Analysis Dataset(s):

Every Dataverse Dataset must include one or more files containing the data required to reproduce all tables, figures, and other analytic results reported in the *AJPS* article and its Supporting Information. Authors can choose their own data storage formats (e.g., rectangular text files; comma-delimited files; Stata `.dta` files; data objects within an R package; SAS files; SPSS files, etc.), as long as the files are readily accessible to researchers in the social science community. Files stored in arcane or proprietary formats generally are not acceptable.

Data should be arranged within each file to facilitate understanding of the contents. If possible, variables should be given meaningful names. And, a unique case identifier variable always should be included. If the data are extracted from another source (as often will be the case), then the case identifier should link the observation in the analysis dataset to its information in the original source.

In analyses based upon highly data-intensive procedures (e.g., Bayesian simulations, bootstrap resampling, etc.), it is not necessary to provide the full contents of each replicated dataset. However, the full set of relevant results (e.g., the simulated parameter values in MCMC estimation, the bootstrap replications of a sample statistic, etc.) should be provided in a coherent data file. And in such cases, providing software command files (see next subsection) to reproduce the entire data-intensive analysis is particularly important.

Each file containing an analysis dataset must be accompanied by a PDF file containing a codebook for the dataset. The codebook always should contain variable definition information for all variables used in the analysis. If the dataset is stored in a proprietary format (e.g., a Stata `.dta` file), then the codebook must include variable names. If the data are stored in a text file, then formatting information must be provided.

In some rare cases (see the section below on "Possible Exceptions to Data Access Requirements"), the analyses in an *AJPS* article may be based upon restricted data that cannot be posted in a publicly-accessible location. As noted below, any such exceptions to the general *AJPS* replication policy must receive explicit approval from the Editor upon initial submission of the manuscript. Once this permission is received, the analysis dataset need not be posted to the *AJPS* Dataverse. But, the author still must provide instructions that interested researchers can use to access the data (see the section on "Information to Reproduce the Analysis Dataset" below), as well as formatting and variable definition information for the data that are analyzed in the *AJPS* article.

Software Commands:

Every Dataverse Dataset must include one or more files containing the software commands that can be applied to the analysis dataset in order to reproduce all tables, figures, and other analytic results presented in the *AJPS* article. Typically, these will be plain text files, but the exact format of the file contents depends upon the software used to carry out the original analyses. Authors can provide Stata `.do` files, R command scripts, or text files appropriate for submission to other software systems and environments.

Regardless of the format used for the command files, comment statements should be used extensively throughout the files to explain the steps of the analysis. Authors can assume that users are familiar with the software system used for the analysis (i.e., there is no need to explain how specific commands function). But, authors should explain how the various commands produce results that are relevant for the analyses reported in the article (e.g., "The following commands recode variables X and Y in preparation for the logistic regression model"; "The following commands create Figure 1 in the article"; etc.).

In some cases, conducting the analyses in an *AJPS* article may require software tools that are not readily available to the research community. Examples include (but are not limited to) Stata `.ado` files or R packages written by, or specially available to, the author. Any such software resources required to replicate an analysis from the *AJPS* article must be included in the Dataverse Dataset for the article, along with relevant documentation and instructions for installing (if necessary) and using them.

Authors always should provide clear and specific information about the version of the software system used to conduct the analyses reported in their *AJPS* article. This requirement is critically important because algorithms, procedures, and functions can (and do) change across software versions! For example, an R command file might begin with the comment statement, "The following analyses were carried out using R version 3.2.2," or the `Readme.txt` file for the Dataverse Dataset could contain the following statement: "All data analyses in this article were carried out using Stata/MP 14.1 for Windows (64-bit x86-64)."

As mentioned earlier, software command files are particularly important for data-intensive analyses in which the "intermediate" datasets used to obtain the final results (e.g., MCMC simulations or bootstrap replications of the original data) are not, themselves, included among the Dataverse Dataset's replication files. In such cases, users would be recreating the original analyses. Thus, the command file needs to provide especially clear instructions for doing so. On a related point, any commands that generate random numbers (e.g., for Monte Carlo simulations, bootstrap resampling, jittering points in a graphical display, etc.) should include a seed value in order to ensure consistent results.

Authors can provide either a single command file that covers all analyses reported in the *AJPS* article, or separate files for the specific analyses contained within the article. In the former case, comment statements should be used within the command file to distinguish the commands used for different figures, tables, or other analytic results. The process is best served when authors use meaningful, rather than "generic," file names. For example, files named "`Stata commands for performing logistic regressions.do`" or "`R functions to reproduce Figure 1.R`" are better than "`Commands.do`" or "`R_scripts.R`."

Information to Reconstruct the Analysis Dataset:

Every Dataverse Dataset must contain complete information for constructing the analysis dataset(s) from the original data sources. The exact materials for doing so will depend upon the nature and sources of the analysis data. But regardless of the specific details, interested researchers always must be able to follow the author's instructions in order to reproduce the precise data values used for any analyses reported in the *AJPS* article.

The analysis dataset often is created by extracting variables and observations from another, larger, source dataset, such as an entry in the American National Election Study (ANES) series, the Comparative Study of Electoral Systems (CCES), the International Correlates of War (COW) project, or the General Social Survey (GSS). In such cases, the author must provide a software command file for doing so. Although the format will vary depending on the software that the author uses, the file always should contain commands for selecting the relevant variables, extracting subsets of observations if necessary, performing any data transformations that are carried out prior to the analysis itself, and assigning missing values. Again, comment statements should be used extensively throughout the file to explain the commands that are used. As stated earlier, there always should be a case identification variable that links observations in the analysis dataset to their original records in the source dataset. And, it is important to identify the specific version of the source dataset and the date that it is accessed in order to construct the analysis dataset.

The analysis dataset sometimes is created by merging information extracted from several other sources. For example, an analysis of the American states may use information obtained from both the states themselves and from the federal government. In such cases, the author must provide the relevant software commands for extracting the data from the separate sources, and for merging the separate subsets of data into the overall analysis dataset. As always, comment statements should be used extensively to explain the procedures.

Complete reference information must be provided for all source datasets used to construct the analysis dataset. Some useful guidelines about the practice of data citation can be found on the website of the Inter-university Consortium for Political and Social Research (ICPSR).

Instructions and Recommendations for Qualitative Analyses

The potential diversity of qualitative analyses implies that the information employed by researchers is likely to involve more varied formats than the data used in quantitative analyses. For example, automated content analysis and qualitative comparative analysis typically require that qualitative data are organized in matrix form and approached as an aggregate body of information. Other qualitative approaches require information in "granular" form, with individual sources considered separately or in small groups. Scholars who employ the granular approach to qualitative data draw on the content of each cited source (e.g., book, interview, newspaper article, video clip, etc.) as a distinct input to the analysis.

Replication and evaluation procedures vary for different forms of qualitative data and different types of analysis. When the data are organized in matrix form, authors must provide all of the materials necessary to reproduce the analyses reported in the *AJPS* article (just as with quantitative data). With granular data, exact reproduction of an analysis can be more difficult. For example, textual interpretations might be contested or scholars may place varying emphasis on different elements of data when drawing a conclusion. While the *American Journal of Political Science* does not mandate strict replication where doing so is infeasible, authors still must make their scholarship understandable and amenable to systematic evaluation. Authors must describe their research processes as explicitly and precisely as possible, and provide the materials necessary to elucidate how they arrived at their findings and conclusions.

Readme File:

Just as with quantitative analyses, every Dataset on the *AJPS* Dataverse must include a plain-text file named "`readme.txt`". This file provides the names of all other files contained in the Dataverse Dataset, along with a brief description of each one. For relatively small Dataverse Datasets, this information can be presented as a simple list. Larger Dataverse Datasets should group similar files under subheadings.

Matrix Approach to Qualitative Data:

The matrix approach to data in qualitative research typically implies that a dataset is being analyzed holistically. The analysis does not focus on the fragments of information that were measured to produce the overall dataset. Accordingly, authors employing a matrix-form qualitative dataset to develop their evidence-based conclusions must: ensure that the parts of the dataset they used (i.e., the analysis dataset) are uploaded to the *AJPS* Dataverse; provide a citation to the dataset at an appropriate point in the text; and list the dataset in the reference section of their article, using a persistent identifier such as a Digital Object Identifier (DOI).

Authors whose evidence-based conclusions draw on the analysis of a dataset that they created from source information they collected themselves must clearly describe the context in which the elements of the analysis dataset were generated; cite the sources from which those elements are drawn; and describe the procedures used to (1) collect those sources, (2) generate the data (for instance, to code information drawn from those sources), and (3) create the dataset. Authors must also indicate the logic through which the analysis dataset was extracted or generated from the broader source(s) of information, as well as provide all relevant data collection instruments (e.g., survey questionnaires, interview protocols, etc.). Authors whose evidence-based conclusions draw on the analysis of a qualitative dataset created by another scholar must cite the location where the procedures used to create the dataset are explicated.

Authors must also make the analytic materials necessary to evaluate their findings available in the *AJPS* Dataverse. Specifically, for authors who employ qualitative comparative analysis (QCA), the analysis dataset refers to the calibrated data— that is, each case's set membership score in all the condition sets and the outcome set. Authors must provide the code for the calibration of sets, construction of truth tables, logical minimization of truth tables, and

(graphical) representation of the results. The materials that are provided should include information regarding:

a. *Calibration function*: Which empirical information leads to which membership score in which set.
b. *Consistency threshold*: The minimum consistency value to be achieved in order for a truth table to be considered sufficient for the outcome.
c. *Frequency threshold*: The minimum number of cases to be contained in a truth table in order to consider it not as a logical remainder row.
d. *Treatment of logical remainders*: At least whether non, only easy counterfactual, or also difficult counterfactuals on logical remainder rows have been included in the logical minimization of the truth table; preferable would be the requirement to provide a (summary) list of those remainder rows that have been included in the logical minimization.
e. *Any other consideration influencing the decision*: These may be specific case knowledge (e.g., on the standardized indicator used for calibration if a specific case appears to be mis-measured). Authors who use case knowledge to rectify this measurement error by adjusting that case's membership score in the set should report doing so.
f. *Use of PRI measure*: Authors who use the PRI measure to exclude a truth table row from the logical minimization which, based on the consistency score, would otherwise have been included, should report doing so.

The preceding requirements are based on the assumption that authors use the R statistical computing environment to perform QCA. Authors who use different software packages should supply the analogous information, and should contact the *AJPS* Editor with any questions regarding the materials and information that must be provided.

Authors who employ automated content analysis should describe the text collection procedure and (where possible) provide digital files of the original full texts they are analyzing. When digital files of the full texts cannot be made available for intellectual property, legal, or ethical reasons, authors should explain the restrictions and describe in detail how the text collection can be reconstructed by other researchers. In all cases, authors should include a document term matrix representation of the text collection in the replication material. Authors should furthermore describe any text processing steps undertaken prior to the analysis and provide code for the particular text analysis method used. This may also include a codebook or a dictionary file, if applicable. Authors who analyze textual data using computer assisted qualitative data analysis software (CAQDAS) must provide a list of coding categories/criteria, a description of the coding process, final coding trees, and reports/output.

Granular Approach to Qualitative Data:

Authors who analyze individual sources must cite each source directly used in their analysis and list each source in the reference section of their article (including persistent identifiers when available).[2] In addition, for central or contested empirical claims in their article, authors are expected to provide the relevant fragments from the sources used to formulate those claims— for example, providing a brief redaction from individual textual sources or a transcript of the relevant section of an individual or focus group interview. Whenever feasible,

[2]Authors are not expected to cite every source that they encountered while conducting the research, but only those that are immediately implicated in the analysis, and directly support published empirical claims.

authors should also make the original source (and not just the relevant fragment) available. All source fragments and original sources should be uploaded to the *AJPS* Dataverse. If source materials are already stored at another trusted digital repository, however, then the article's *AJPS* Dataverse Dataset may instead include specific information directing researchers to the relevant locations.

Authors who analyze granular data generated from individual sources they collected themselves are required to provide documentation clearly describing the context in which those data were generated and the procedures used to (1) collect the sources; (2) generate and prepare the data; (3) select data for citation. Authors must also provide all relevant data collection instruments (e.g., survey questionnaires, interview protocols, etc.).

In order to show how they arrived at the conclusions in the article, authors who use granular data generated from individual sources must describe the analytic operations conducted on the data. In doing so, they must explain how the data map onto, and support, the central or contested claims in the text. The specific ways that authors carry out these steps will depend upon the inferential rules and structures implied by the author's underlying epistemology and employed in the type of qualitative research they are conducting. For example, authors may discuss how they evaluated the relative persuasiveness and consistency of evidence, the logic and steps that they followed while engaging in process tracing, how they evaluated the plausibility of counterfactuals, or how they developed a systematic scheme for weighting data. This requirement is analogous to authors providing the code and other supplemental materials in quantitative analyses.

The preceding requirements can be satisfied by preparing a transparency appendix (TRAX). A TRAX typically consists of two elements, an overview section and a set of annotations. For more information about the logic and practice of annotation, see Andrew Moravcsik, Colin Elman, and Diana Kapiszewski, "A Guide to Active Citation" Version 1.8, November 2013.

The overview section of the TRAX provides a brief summary of the author's research trajectory, enhancing as necessary the description of the context, design, and conduct of research offered in the main text. The first part should outline the source collection and data generation processes employed in the article. The second part should demonstrate how the research attends to the inferential rules and structures underlying the type of analysis the author is carrying out.

The bulk of the TRAX consists of annotations that enhance as necessary the discussion of the micro-connections between data, analysis, and empirical claims offered in the main text. As noted above, authors need only annotate central or contested empirical claims in their article. Annotations that elucidate the use of sources cited in references must include:

a. Short references to the cited sources (e.g., "Skocpol 1979"), and additional information need to locate the relevant information within a cited source (e.g., a page number). Authors also must ensure that the article's reference section includes a precise and complete reference, with all information that scholars would need to locate the cited source.

b. Excerpts from cited sources, typically 100 to 150 words from a textual source. For handwritten material, audiovisual material, or material generated through interviews or focus groups, include an excerpt from the transcription. Excerpts of texts or transcripts not in English must include a translation of the relevant passage, including the source of the translation. In circumstances where no excerpt can be provided due to human

subjects, copyright or logistical constraints, provide an explanation of those restrictions and (if feasible) a redaction of the relevant text.

While authors are only required to provide excerpts from textual sources, they are strongly encouraged (where it is legal, ethical, and practical) also to provide the underlying source. If the source is available via a persistent link (e.g., a DOI) authors should include that information in the annotation. If the author will be uploading the source file as part of their data submission, they should include the file name in the annotation.

c. Analytic Notes that illustrate how the data generated from the sources support empirical claims in the text.

Occasionally, annotations may be used to expand on material in the main text, (for example, to elucidate analysis or otherwise clarify an inferential step) rather than to illustrate micro-connections among data, analysis, and empirical claims. Annotations of this type include only a note.

For technical guidance on how to construct a TRAX suitable for uploading to the AJPS Dataverse, see the "Instructions to Generate a Transparency Appendix" on the Qualitative Data Repository website.

Possible Exceptions to Data Access Requirements

The objective of sharing some or all of the data underlying the analysis in an *AJPS* article may be in tension with other critical goals. Under these circumstances, the *American Journal of Political Science* requires authors to articulate explicitly how they balance these competing concerns, and how that decision affects the availability of the data used to generate conclusions in the article. This explanation must be included as part of the article's Dataset in the *AJPS* Dataverse.

The default rule for articles published in the *American Journal of Political Science* is that authors must provide full access to the analysis dataset (or, for qualitative work using granular data, to the data cited in the article that support central or contested claims). "Blanket" denial of access to the analysis dataset used in an *AJPS* article generally will not be permitted. In some circumstances, however, an author may request an exemption allowing the author to withhold or limit public access to some or all of those data. Any such exemptions must be explicitly approved by the *AJPS* Editor, who retains final authority to decide whether the requested exemption will be granted. The corresponding author must inform the Editor that he or she will be requesting an exemption and explain the reasons for the request when the manuscript initially is submitted. The following three types of situations could justify an exemption.

Retaining Data for Future Research:

The author may want to prevent outside access to the source dataset in order to preserve the opportunity for conducting further research. As explained earlier, the analysis dataset— that is, the actual data used in the *AJPS* article— cannot be "embargoed" in this manner. However, the author can request an embargo for any additional information that is contained in the source dataset(s). Authors are encouraged to impose any such restrictions for a limited amount of time only, and to include a statement in the materials uploaded to the *AJPS* Dataverse specifying when the source data will be made available to the general research community. The *AJPS* Editor must explicitly grant permission for any such data embargo.

Furthermore, the author still must make the source data available to the *AJPS* Editorial Staff and to the *AJPS* contractor tasked with verifying the content of replication materials (currently, the Archive Staff at the Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill, for quantitative data, and the Qualitative Data Repository staff at the Center for Qualitative and Multi-Method Inquiry, Syracuse University, for qualitative data). The *AJPS* Editorial Staff, the Archive Staff at the Odum Institute, and the staff at the Center for Qualitative and Multi-Method Inquiry guarantee that the embargoed data will be used strictly to verify the integrity of the articles replication materials, and will not be retained after that has been established.

Restricted Access Datasets:

The second situation for which an exemption to the general replication requirements may be granted occurs when some or all of the analysis dataset (or cited data) are under legal constraints. Examples include the use of proprietary datasets, contractual arrangements that restrict data sharing, or information that is classified due to government regulations. Even in such situations, authors will be encouraged to negotiate general access to the analysis dataset with the holders of the restricted source dataset. When access is not permitted, the author must include as part of the request to the *AJPS* Editor for an exemption to the general replication requirements:

a. A statement of the terms under which access will be granted for qualified researchers.
b. An explanation of the conditions under which the data will be made accessible.
c. The qualifications and any applicable fees that a researcher must present in order to be granted access to the data.

In the context of qualitative data, copyright restrictions may impose significant constraints. Here the "Fair Use" exemption outlined in the U.S. 1976 Copyright Act may apply. Per this exemption, under many circumstances limited portions of different types of copyrighted materials can be shared for noncommercial use, including private study, teaching, or criticism/review.

Human Subjects Protection:

Public access to an analysis or source dataset may be restricted due to human subject concerns. The *American Journal of Political* Science recognizes that replication requirements and research transparency more generally are bound by ethical constraints. Foremost among these is the requirement that the safety, dignity, and general well-being of all human research subjects is maintained and protected. An exemption to the general replication requirements can be requested when release of the analysis or source data would be detrimental to this general objective.

In some cases, data that would otherwise be restricted can be made available by applying mechanisms that address human subjects concerns. For example, authors may be able to de-identify and thereby render anonymous the observations in their data. Or, authors may be able to impose access restrictions on who can view the data or where the data can be viewed (e.g., in a protected physical or digital data enclave). The mechanisms that are available for this purpose will depend in part on the type of data that is involved and the circumstances in which the data were generated. In some situations, mechanisms of this type may not be available at all. But, when they are a viable option, authors are encouraged to use them wherever possible.

If the *AJPS* Editor grants permission to withhold some or all of the analysis dataset, the source dataset, or cited data, then the exempted information does not need to be uploaded to the *AJPS* Dataverse. The author will be required to include a note at the beginning of the published article explicitly acknowledging the limitations on data availability and describing the restrictions that prevent public access to the exempted data. A Dataset still must be created in the *AJPS* Dataverse, containing materials that specify the procedures through which an interested researcher can apply for access to the analysis dataset for replication purposes (including the construction of the analysis dataset from the original source dataset) from the holders of the source data. For articles that analyze matrix format data (quantitative or qualitative), the author also must provide relevant information about variable definitions and formatting, as well as software command files for carrying out the analyses and constructing the analysis dataset. More generally, authors remain obligated to pursue all other dimensions of research transparency so that readers are fully informed about the data analysis, even if the data, themselves, are not immediately available through the Dataset on the *AJPS* Dataverse.

Conclusion

The preceding guidelines describe the minimum requirements for a Dataset on the AJPS Dataverse. They implement the principles of data access and research transparency (DA-RT) to which the *American Journal of Political Science* is a founding signatory. But, authors certainly are not limited to providing the files described in this document. In fact, authors are encouraged to provide as much information as possible. Additional contents of the Dataset on the *AJPS* Dataverse might include supplemental reports, pre-analysis plans, additional data, and extensions of the analysis beyond those reported in the *AJPS* article and published Supporting Information. Authors are encouraged to keep in mind that scientific research is an ongoing stream. Hopefully, the research reported in the pages of the *American Journal of Political Science* will facilitate and encourage further efforts to establish powerful theories of political and social phenomena. Any materials that facilitate this process are welcome elements of a Dataset on the *AJPS* Dataverse!