PROJECT TIER
TEACHING INTEGRITY
IN EMPIRICAL RESEARCH

www.projecttier.org

The DRESS Protocol (version 1.0): Documenting Research in the Empirical Social Sciences

The DRESS Protocol is a set of standards for documenting empirical research in the social sciences. It specifies the content and organization of the replication documentation for a study that reports results obtained by manipulating and analyzing statistical data.

Section I of this document gives an overview of the DRESS Protocol in qualitative terms. It discusses the purposes for which replication documentation can be used, and then enumerates several principles to which the documentation for a study must adhere if it is to serve those purposes. The purposes and principles described in Section I constitute the spirit of the DRESS protocol.

The letter of the DRESS Protocol is detailed in Sections II and III, which present formal specifications for the various items that should be included in the replication documentation for an empirical study and how those items should be organized. Replication documentation that meets these specifications will satisfy the principles enumerated in Section I, and will be sufficient to serve the underlying purposes.

I. OVERVIEW

Because of the diversity in the kinds of data, computer software and hardware, and analytical methods used in empirical social science research, unanticipated situations will inevitably raise questions that are not addressed in the DRESS Protocol, or in which following the letter of the Protocol would be awkward or impossible. The author preparing the replication documentation will then need to exercise independent judgment about how to resolve the situation. In such cases, the author's judgment should be guided by the spirit of the Protocol: any deviations from the formal specifications of the Protocol should be made in such a way that the completed replication documentation still satisfies the principles and can be used for the purposes described in this section.

A. *Purposes.* Replication documentation should be sufficient to be used for any or all of several purposes:

(i) Confirmation: verifying that there were no errors in the computations conducted for the study.

(ii) Robustness checking: assessing the robustness of the results reported in the paper.

(iii) Extension: initiating new research that extends or builds on the paper.

(iv) Communication: recording concisely and unambiguously the ways in which the data were manipulated in preparation for analysis, and the procedures that generated the reported results.

B. *Principles.* The design of the protocol was guided by three principles:

(i): Complete replicability: The documentation should make it possible to conduct a replication of the study that begins with original data files identical to those with which the author started the research, processes them as necessary to prepare them for analysis, and finally executes the commands that generate the results reported in the study.

(ii) Independent replicability: All the information necessary to replicate the study should be included in the documentation. In particular, it should not be necessary to request any additional information from the author.

(iii) Realism: The documentation should be clearly enough organized and presented that it is realistic to expect a reasonably competent researcher to be able to conduct a complete and independent replication of the study without undue difficulty.

II. COMPONENTS OF THE DOCUMENTATION

A. *A list of the results reported in the paper that the documentation is intended to reproduce.*

Content: Every statistical result reported in the paper that the documentation is intended to reproduce should be assigned a reference number, and described in way that identifies it unambiguously. For example:

Result 1: Figure 3
Result 2: Table 2, column 4
Result 3: The simulation reported in section 4.3
Result 4: The income elasticity of demand for beef (1.86) reported on page 58 of the paper

Presentation: This list of results should be included in the Read Me file that accompanies the documentation, as described below in Item II.G.

B. *Information on the necessary software.*

Content: This information should include a list of every program used for any part of the data processing and analysis conducted for the study. For each program, there should be an indication of the version number, any necessary add-ons to the standard package, and any other special information about the software a reader would need to know to be able to replicate the study.

Presentation: This information should be included in the Read Me file that accompanies the documentation, as described below in Item II.G.

C. ***Information about the original data.***

Content: The term "original data file" refers to any data file the author initially obtained for use in the project, before the author processed or modified it in any way.

The documentation for the paper should include a list of the names of all the original data files used for the project.

> ***Note:*** Every file from which any of the statistical data used for the project were extracted should be included in this list. Conversely, every statistical datum required for the data processing and analysis conducted for the project should have been taken from one of the listed original data files.

For every original data file, the following additional information should be provided:

(i) A bibliographic citation in a standard style (e.g., Chicago or American Psychological Association).

(ii) The date the file was created or accessed by the author for use in the study.

(iii) A verbal explanation of how a reader can obtain a copy of the file. This description should be detailed and precise enough to allow the reader to obtain a file identical to the one from which the author extracted the data used for the paper.

(iv) A list of the variables the author extracted from the file for use in the study, along with a precise definition or complete coding information for each of these variables.

(v) Any additional information a researcher would need to be able to understand and make use of the data in the file, such as sampling methods and weights, a description of the population from which the sample was drawn, or the structure of the data in the file (e.g., each row a country/year pair, with each variable containing values of a particular indicator, versus each row a country/indicator pair, with each variable containing values for a particular year).

A note about confidential or proprietary data: If access to an original data file is restricted for confidentiality or proprietary reasons, the explanation of how a reader can obtain the file (part (iii) of this section) should include instructions on how to apply for access to the data.

Other than this consideration, the information that should be provided about restricted data does not differ from the information required for public data.

Presentation: This information should be presented in a document composed by the author, saved in .txt or .pdf format, and titled "Original-Data.txt" or "Original-Data.pdf."

The list of variables included in this document, and the additional information specified in (i), (ii), (iii), should be composed by the author of the paper.

For simple original data files that contain just a few variables, it may be feasible for the author of the paper to compose additional language in "Original-Data.txt/pdf" that provides the information specified in (iv) and (v). But any parts of this of this information that can be found in an existing resource such as a codebook, users' guide or data dictionary available with the original data file need not be reproduced in "Original-Data.txt/pdf." Instead, it suffices to include a reference to the resource in which the information can be found, along with an explanation of how a user can access the resource (e.g., via a link to a .pdf codebook or a .html users' guide). (When the information specified in (iv) and (v) is extensive, providing a reference to an existing resource in which the information can be found is usually preferable to composing it all *de novo* in "Original-Data.txt/pdf.")

D. *Command files.*

Content: The documentation should include one or more command files, each written in the syntax of the software that runs it, containing code that extracts the variables used for the study from the original data files; cleans, combines and otherwise processes the data as necessary to create the final dataset(s) used for the analysis; and then executes the procedures that generate the results.

Comments that explain each step of the data processing and analysis should be inserted throughout the command files.

Every command that generates one of the results that the documentation is intended to reproduce should be preceded by a comment indicating the reference number assigned to that result (in Item II.A. above). For example,

```
/*The following command generates Result 1 (Table 3 in the
paper)*/
```

Presentation: These command files should be stored in a directory called "Command Files." If convenient, the author may choose to organize the command files into two or more sub-directories within the "Command Files" directory (e.g., two sub-directories called "Data Processing" and "Analysis;" or three subdirectories called "Argentina," "Brazil," and "Colombia").

E. *Processed data files.*

Content: In most cases, the command files for a project contain commands that save "processed data files," which can be of two types:

> "Intermediate data files" contain partially processed data, and are reopened at some later point for further processing and/or combining with other data files.
>
> "Analysis data files" contain data that have been fully cleaned and processed in preparation for analysis; the tests or procedures that generate the results of the study are performed on the analysis data files.

When an interested reader has access to the command files for a study (as described in Section II.D) and is able to obtain copies of the original data files (using the information described in Section II.C), s/he can generate all the intermediate and analysis data files simply by running the command files that create and save them. In principle, it is therefore unnecessary to include any processed data files in the documentation.

Nonetheless, there may be cases in which including certain processed data files in the documentation, though redundant, would be helpful to a reader interested in understanding and replicating the empirical procedures used for the paper. In those cases, there is certainly not any reason not to include them. It is important to recognize, however, that including processed data files in the documentation is not an adequate substitute for providing all the information about obtaining the original data and all the command files necessary to generate them.

Presentation: If an author chooses to include any processed data files in the documentation, they should be stored in a directory called "Processed Data Files." If convenient, the author may choose to organize the processed data files into two or more sub-directories within the "Processed Data Files" directory (e.g., two sub-directories called "Intermediate Data Files" and "Analysis Data Files;" or three subdirectories called "Processed Argentina Data," "Processed Brazil Data," and "Processed Colombia Data").

F. *Instructions for replicating the study.*

Content: These instructions should be written for a user who has access to all the necessary software, and wishes to replicate the study on her/his own computer.

The instructions should indicate which original data files and which command files need to be copied onto the users' computer, and the name(s) that should be given to the folder(s) these files should be stored in. The instructions should then indicate the order in which the command files should be executed to replicate the data processing and analysis and reproduce the results of the paper.

In many cases, it is helpful to include a note for each command file indicating the inputs it uses—e.g., what data files it opens and/or other command files it calls—and what outputs it produces—e.g., what intermediate data files it creates and saves for later use. This information can help the user understands the steps involved in the replication.

Presentation: These instructions for replicating the study should be included in the Read Me file that accompanies the documentation, as described below in Item II.G.

G. *A Read Me file*.

Content: The documentation should be accompanied by a Read Me file that includes:

(i) The list of results the documentation is intended to reproduce, as described in Item II.A.

(ii) The information on the software necessary to conduct a replication of the study, as described in Item II.B.

(iii) Instructions for using the documentation to replicate the study, as described in Item II.F.

Presentation: A document containing the information specified above should be composed by the author, and saved in.txt or .pdf format, with the name "Read-Me.txt" or "Read-Me.pdf."

III. ORGANIZATION OF THE DOCUMENTATION

All of the documentation should be stored in one main directory called "Replication Documentation."

The top level of the "Replication Documentation" directory should contain:

(i) the readme file ("Read-Me.txt/pdf"), as described in Item II.G.

(ii) an electronic copy of the complete research paper (i.e., the paper for which the documentation has been prepared).

(iii) the document with information about the original data files ("Original-Data.txt/pdf"), as described in Item II.C.

(iv) the "Command Files" directory, as described in Item II.D.

(v) the "Processed Data Files" directory, as described in Item II.E. (if the author chooses to include any processed data files in the documentation).