

Template README and Guidance

INSTRUCTIONS: This README suggests structure and content that have been approved by various journals, see [Endorsers](#). It is available as [Markdown/txt](#), [Word](#), [LaTeX](#), and [PDF](#). In practice, there are many variations and complications, and authors should feel free to adapt to their needs. All instructions can (should) be removed from the final README (in Markdown, remove lines starting with > INSTRUCTIONS). Please ensure that a PDF is submitted in addition to the chosen native format. Please ensure that the README is called “README” plus the appropriate suffix, not some non-standard name. This helps replicators immediately locate the necessary document.

Overview

INSTRUCTIONS: The typical README in social science journals serves the purpose of guiding a reader through the available material and a route to replicating the results in the research paper. Start by providing a brief overview of the available material and a brief guide as to how to proceed from beginning to end.

Example: The code in this replication package constructs the analysis file from the three data sources (Ruggles et al, 2018; Inglehart et al, 2019; BEA, 2016) using Stata and Julia. Two main files run all of the code to generate the data for the 15 figures and 3 tables in the paper. The replicator should expect the code to run for about 14 hours.

Data Availability and Provenance Statements

INSTRUCTIONS: Every README should contain a description of the origin (provenance), location and accessibility (data availability) of the data used in the article. These descriptions are generally referred to as “Data Availability Statements” (DAS). However, in some cases, there is no external data used.

☐ This paper does not involve analysis of external data (i.e., no data are used or the only data are generated by the authors via simulation in their code).

If box above is checked and if no simulated/synthetic data files are provided by the authors, please skip directly to the section on [Computational Requirements](#). Otherwise, continue.

INSTRUCTIONS: - When the authors are **secondary data users** (they did not generate the data), the provenance and DAS coincide, and should describe the condition under which (a) the current authors (b) any future users might access the data. - When the data were generated (by the authors) in the course of conducting (lab or field) **experiments**, or were collected as part of **surveys**, then the description of the provenance should describe the data generating process, i.e., survey or experimental procedures: - Experiments: complete sets of experimental instructions, questionnaires, stimuli for all conditions, potentially screenshots, scripts for experimenters or research assistants, as well as for subject eligibility criteria (e.g. selection criteria, exclusions), recruitment waves, demographics of subject pool used. - For lab experiments specifically, a description of any pilot sessions/studies, and computer programs, configuration files, or scripts used to run the experiment. - For surveys, the whole questionnaire (code or images/PDF) including survey logic if not linear, interviewer instructions, enumeration lists, sample selection criteria.

The information should describe ALL data used, regardless of whether they are provided as part of the replication archive or not, and regardless of size or scope. The DAS should provide enough information that a replicator can obtain the data from the original source, even if the file is provided.

For instance, if using GDP deflators, the source of the deflators (e.g. at the national statistical office) should also be listed here. If any of this information has been provided in a pre-registration, then a link to that registration may (partially) suffice.

DAS can be complex and varied. Examples are provided [here](#), and below.

Importantly, if providing the data as part of the replication package, authors should be clear about whether they have the **rights** to distribute the data. Data may be subject to distribution restrictions due to sensitivity, IRB, proprietary clauses in the data use agreement, etc.

NOTE: DAS do not replace Data Citations (see [Guidance](#)). Rather, they augment them. Depending on journal requirements and to some extent stylistic considerations, data citations should appear in the main article, in an appendix, or in the README. However, data citations only provide information **where** to find the data, not **how to access** those data. Thus, DAS augment data citations by going into additional detail that allow a researcher to assess cost, complexity, and availability over time of the data used by the original author.

Statement about Rights

- ☐ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.
- ☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package. Appropriate permission are documented in the [LICENSE.txt](#) file.

(Optional, but recommended) License for Data

INSTRUCTIONS: Most data repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a LICENSE.txt file, separately from the README, possibly combined with the license for any code. Some data may be subject to inherited license requirements, i.e., the data provider may allow for redistribution only if the data is licensed under specific rules - authors should check with their data providers. For instance, a data use license might require that users - the current author, but also any subsequent users - cite the data provider. Licensing can be complex. Some non-legal guidance may be found [here](#). For multiple licenses within a data package, the LICENSE.txt file might contain the concatenation of all the licenses that apply (for instance, a custom license for one file, plus a CC-BY license for another file).

NOTE: In many cases, it is not up to the creator of the replication package to simply define a license, a license may be *sticky* and be defined by the original data creator.

Example: The data are licensed under a Creative Commons/CC-BY-NC license. See LICENSE.txt for details.

Summary of Availability

- ☐ All data **are** publicly available.
- ☐ Some data **cannot be made** publicly available.
- ☐ **No data can be made** publicly available.

INSTRUCTIONS: If *No or some data can be made* **publicly** available, the journal may require that data be preserved for a specified time.

☐ Confidential data used in this paper and not provided as part of the public replication package will be preserved for ___ years after publication, in accordance with journal policies.

Details on each Data Source

INSTRUCTIONS: For each data source, list the file that contains data from that source here; if providing combined/derived datafiles, list them separately after the DAS. For each data source or file, as appropriate,

- Describe the format (open formats preferred, but some software-specific formats OK if open-source readers available): .dta, .xlsx, .csv, netCDF, etc.
- Provide a data dictionary, either as part of the archive (list the file name), or at a URL (list the URL). Some formats are self-describing *if* they have the requisite information (e.g., .dta should have both variable and value labels).
- List availability within the package
- Use proper bibliographic references in addition to a verbose description (and provide a bibliography at the end of the README, expanding those references)

A summary in tabular form can be useful:

Data.Name	Data.Files	Location	Provided	Citation
“Current Population Survey 2018”	cepr_march_2018.dta	data/	TRUE	CEPR (2018)
“Provincial Administration Reports”	coast_simplepoint2.csv; rivers_simplepoint2.csv; RAIL_dummies.dta; railways_Dissolve_Simplify_point2.csv	Data/maps/	TRUE	Administration (2017)
“2017 SAT scores”	Not available	data/to_clean/	FALSE	College Board (2020)

where the Data.Name column is then expanded in the subsequent paragraphs, and CEPR (2018) is resolved in the References section of the README.

Example for public use data collected by the authors

The [DATA TYPE] data used to support the findings of this study have been deposited in the [NAME] repository ([DOI or OTHER PERSISTENT IDENTIFIER]). [1]. The data were collected by the authors, and are available under a Creative Commons Non-commercial license.

Example for public use data sourced from elsewhere and provided

Data on National Income and Product Accounts (NIPA) were downloaded from the U.S. Bureau of Economic Analysis (BEA, 2016). We use Table 30. Data can be downloaded from <https://apps.bea.gov/regional/downloadzip.cfm>, under “Personal Income (State and Local)”, select CAINC30: Economic Profile by County, then download. Data can also be directly downloaded using <https://apps.bea.gov/regional/zip/CAINC30.zip>. A copy of the data is provided as part of this archive. The data are in the public domain.

Datafile: CAINC30__ALL_AREAS_1969_2018.csv

Example for public use data with required registration and provided extract

The paper uses IPUMS Terra data (Ruggles et al, 2018). IPUMS-Terra does not allow for redistribution, except for the purpose of replication archives. Permissions as per <https://terra.ipums.org/citation> have been obtained, and are documented within the “data/IPUMS-terra” folder. > Note: the reference to “Ruggles et al, 2018” would be resolved in the Reference section of this README, **and** in the main manuscript.

Datafile: data/raw/ipums_terra_2018.dta

Example for free use data with required registration, extract not provided

The paper uses data from the World Values Survey Wave 6 (Inglehart et al, 2019). Data is subject to a redistribution restriction, but can be freely downloaded from <http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>. Choose WV6_Data_Stata_v20180912, fill out the registration form, including a brief description of the project, and agree to the conditions of use. Note: “the data files themselves are not redistributed” and other conditions. Save the file in the directory data/raw.

Note: the reference to “Inglehart et al, 2018” would be resolved in the Reference section of this README, **and** in the main manuscript.

Datafile: data/raw/WV6_Data_Stata_v20180912.dta (not provided)

Example for confidential data

INSTRUCTIONS: Citing and describing confidential data, in particular when it does not have a regular distribution channel or online landing page, can be tricky. A citation can be crafted ([see guidance](#)), and the DAS should describe how to access, whom to contact (including the role of the particular person, should that person retire), and other relevant information, such as required citizenship status or cost.

The data for this project (DESE, 2019) are confidential, but may be obtained with Data Use Agreements with the Massachusetts Department of Elementary and Secondary Education (DESE). Researchers interested in access to the data may contact [NAME] at [EMAIL], also see www.doe.mass.edu/research/contact.html. It can take some months to negotiate data use agreements and gain access to the data. The author will assist with any reasonable replication attempts for two years following publication.

Example for confidential Census Bureau data

All the results in the paper use confidential microdata from the U.S. Census Bureau. To gain access to the Census microdata, follow the directions here on how to write a proposal for access to the data via a Federal Statistical Research Data Center: <https://www.census.gov/ces/rdcresearch/howtoapply.html>. You must request the following datasets in your proposal: 1. Longitudinal Business Database (LBD), 2002 and 2007 2. Foreign Trade Database – Import (IMP), 2002 and 2007 [...]

(adapted from [Fort \(2016\)](#))

Example for preliminary code during the editorial process

Code for data cleaning and analysis is provided as part of the replication package. It is available at <https://dropbox.com/link/to/code/XYZ123ABC> for review. It will be uploaded to the [JOURNAL REPOSITORY] once the paper has been conditionally accepted.

Dataset list

INSTRUCTIONS: In some cases, authors will provide one dataset (file) per data source, and the code to combine them. In others, in particular when data access might be restrictive, the replication package may only include derived/analysis data. Every file should be described. This can be provided as a Excel/CSV table, or in the table below.

INSTRUCTIONS: While it is often most convenient to provide data in the native format of the software used to analyze and process the data, not all formats are “open” and can be read by other (free) software. Data should at a minimum be provided in formats that can be read by open-source software (R, Python, others), and ideally be provided in non-proprietary, archival-friendly formats.

INSTRUCTIONS: All data files should be fully documented: variables/columns should have labels (long-form meaningful names), and values should be explained. This might mean generating a codebook, pointing at a public codebook, or providing data in (non-proprietary) formats that allow for a rich description. This is in particular important for data that is not distributable.

INSTRUCTIONS: Some journals require, and it is considered good practice, to provide synthetic or simulated data that has some of the key characteristics of the restricted-access data which are not provided. The level of fidelity may vary - it may be useful for debugging only, or it should allow to assess the key characteristics of the statistical/econometric procedure or the main conclusions of the paper.

Data file	Source	Notes	Provided
data/raw/lbd.dta	LBD	Confidential	No
data/raw/terra.dta	IPUMS Terra		Yes

data/derived/regression_input.dta All listed

As per terms of use

Combines multiple data sources, serves as input for Table 2, 3 and Figure 5. Yes

Computational requirements

INSTRUCTIONS: In general, the specific computer code used to generate the results in the article will be within the repository that also contains this README. However, other computational requirements - shared libraries or code packages, required software, specific computing hardware - may be important, and is always useful, for the goal of replication. Some example text follows.

INSTRUCTIONS: We strongly suggest providing setup scripts that install/set up the environment. Sample scripts for [Stata](#), [R](#), [Julia](#) are easy to set up and implement. Specific software may have more sophisticated tools: [Python](#), [Julia](#).

Software Requirements

INSTRUCTIONS: List all of the software requirements, up to and including any operating system requirements, for the entire set of code. It is suggested to distribute most dependencies together with the replication package if allowed, in particular if sourced from unversioned code repositories, Github repos, and personal webpages. In all cases, list the version *you* used. All packages should be listed in human-readable form in this README, but should also be included in a setup or install script.

- ☐ The replication package contains one or more programs to install all dependencies and set up the necessary directory structure. [HIGHLY RECOMMENDED]
- Stata (code was last run with version 15)
 - estout (as of 2018-05-12)
 - rdrobust (as of 2019-01-05)
 - the program “0_setup.do” will install all dependencies locally, and should be run once.
- Python 3.6.4
 - pandas 0.24.2
 - numpy 1.16.4
 - the file “requirements.txt” lists these dependencies, please run “pip install -r requirements.txt” as the first step. See https://pip.pypa.io/en/stable/user_guide/#ensuring-repeatability for further instructions on creating and using the “requirements.txt” file.
- Intel Fortran Compiler version 20200104
- Matlab (code was run with Matlab Release 2018a)
- R 3.4.3
 - tidyr (0.8.3)
 - rdrobust (0.99.4)
 - the file “0_setup.R” will install all dependencies (latest version), and should be run once prior to running other programs.

Portions of the code use bash scripting, which may require Linux.

Portions of the code use Powershell scripting, which may require Windows 10 or higher.

Controlled Randomness

INSTRUCTIONS: Some estimation code uses random numbers, almost always provided by pseudorandom number generators (PRNGs). For reproducibility purposes, these should be provided with a deterministic seed, so that the sequence of numbers provided is the same for the original author and any replicators. While this is not always possible, it is a requirement by many journals’ policies. The seed should be set once, and not use a time-stamp. If using parallel processing, special care needs to be taken. If using multiple programs in sequence, care must be taken on how to call these programs, ideally from a main program, so that the sequence is not altered. If no PRNG is used, check the other box.

- ☐ Random seed is set at line ____ of program ____
- ☐ No Pseudo random generator is used in the analysis described here.

Memory, Runtime, Storage Requirements

INSTRUCTIONS: Memory and compute-time requirements may also be relevant or even critical. Some example text follows. It may be useful to break this out by Table/Figure/section of processing. For instance, some estimation routines might run for weeks, but data prep and creating figures might only take a few

minutes. You should also describe how much storage is required in addition to the space visible in the typical repository, for instance, because data will be unzipped, data downloaded, or temporary files written.

Summary

Approximate time needed to reproduce the analyses on a standard (CURRENT YEAR) desktop machine:

- ☐ <10 minutes
- ☐ 10-60 minutes
- ☐ 1-2 hours
- ☐ 2-8 hours
- ☐ 8-24 hours
- ☐ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days

Approximate storage space needed:

- ☐ < 25 MBytes
- ☐ 25 MB - 250 MB
- ☐ 250 MB - 2 GB
- ☐ 2 GB - 25 GB
- ☐ 25 GB - 250 GB
- ☐ > 250 GB
- ☐ Not feasible to run on a desktop machine, as described below.

Details

The code was last run on a **4-core Intel-based laptop with MacOS version 10.14.4 with 200GB of free space.**

Portions of the code were last run on a **32-core Intel server with 1024 GB of RAM, 12 TB of fast local storage.** Computation took **734 hours.**

Portions of the code were last run on a **12-node AWS R3 cluster, consuming 20,000 core-hours, with 2TB of attached storage.**

INSTRUCTIONS: Identifying hardware and OS can be obtained through a variety of ways: Some of these details can be found as follows:

- (Windows) by right-clicking on “This PC” in File Explorer and choosing “Properties”
- (Mac) Apple-menu > “About this Mac”
- (Linux) see code in [linux-system-info.sh](#)

Description of programs/code

INSTRUCTIONS: Give a high-level overview of the program files and their purpose. Remove redundant/obsolete files from the Replication archive.

- Programs in `programs/01_dataprep` will extract and reformat all datasets referenced above. The file `programs/01_dataprep/main.do` will run them all.
- Programs in `programs/02_analysis` generate all tables and figures in the main body of the article. The program `programs/02_analysis/main.do` will run them all. Each program called from `main.do` identifies the table or figure it creates (e.g., `05_table5.do`). Output files are called appropriate names (`table5.tex`, `figure12.png`) and should be easy to correlate with the manuscript.
- Programs in `programs/03_appendix` will generate all tables and figures in the online appendix. The program `programs/03_appendix/main-appendix.do` will run them all.
- Ado files have been stored in `programs/ado` and the `main.do` files set the ADO directories appropriately.
- The program `programs/00_setup.do` will populate the `programs/ado` directory with updated ado packages, but for purposes of exact reproduction, this is not needed. The file `programs/00_setup.log` identifies the versions as they were last updated.
- The program `programs/config.do` contains parameters used by all programs, including a random seed. Note that the random seed is set once for each of the two sequences (in `02_analysis` and `03_appendix`). If running in any order other than the one outlined below, your results may differ.

(Optional, but recommended) License for Code

INSTRUCTIONS: Most journal repositories provide for a default license, but do not impose a specific license. Authors should actively select a license. This should be provided in a `LICENSE.txt` file, separately from the

README, possibly combined with the license for any data provided. Some code may be subject to inherited license requirements, i.e., the original code author may allow for redistribution only if the code is licensed under specific rules - authors should check with their sources. For instance, some code authors require that their article describing the econometrics of the package be cited. Licensing can be complex. Some non-legal guidance may be found [here](#).

The code is licensed under a MIT/BSD/GPL [choose one!] license. See [LICENSE.txt](#) for details.

Instructions to Replicators

INSTRUCTIONS: The first two sections ensure that the data and software necessary to conduct the replication have been collected. This section then describes a human-readable instruction to conduct the replication. This may be simple, or may involve many complicated steps. It should be a simple list, no excess prose. Strict linear sequence. If more than 4-5 manual steps, please wrap a main program/Makefile around them, in logical sequences. Examples follow.

- Edit programs/config.do to adjust the default path
- Run programs/00_setup.do once on a new system to set up the working environment.
- Download the data files referenced above. Each should be stored in the prepared subdirectories of data/, in the format that you download them in. Do not unzip. Scripts are provided in each directory to download the public-use files. Confidential data files requested as part of your FSRDC project will appear in the /data folder. No further action is needed on the replicator’s part.
- Run programs/01_main.do to run all steps in sequence.

Details

- programs/00_setup.do: will create all output directories, install needed ado packages.
 - If wishing to update the ado packages used by this archive, change the parameter update_ado to yes. However, this is not needed to successfully reproduce the manuscript tables.
- programs/01_dataprep:
 - These programs were last run at various times in 2018.
 - Order does not matter, all programs can be run in parallel, if needed.
 - A programs/01_dataprep/main.do will run them all in sequence, which should take about 2 hours.
- programs/02_analysis/main.do.
 - If running programs individually, note that ORDER IS IMPORTANT.
 - The programs were last run top to bottom on July 4, 2019.
- programs/03_appendix/main-appendix.do. The programs were last run top to bottom on July 4, 2019.
- Figure 1: The figure can be reproduced using the data provided in the folder “2_data/data_map”, and ArcGIS Desktop (Version 10.7.1) by following these (manual) instructions:
 - Create a new map document in ArcGIS ArcMap, browse to the folder “2_data/data_map” in the “Catalog”, with files “provinceborders.shp”, “lakes.shp”, and “cities.shp”.
 - Drop the files listed above onto the new map, creating three separate layers. Order them with “lakes” in the top layer and “cities” in the bottom layer.
 - Right-click on the cities file, in properties choose the variable “health”... (more details)

List of tables and programs

INSTRUCTIONS: Your programs should clearly identify the tables and figures as they appear in the manuscript, by number. Sometimes, this may be obvious, e.g. a program called “table1.do” generates a file called table1.png. Sometimes, mnemonics are used, and a mapping is necessary. In all circumstances, provide a list of tables and figures, identifying the program (and possibly the line number) where a figure is created.

NOTE: If the public repository is incomplete, because not all data can be provided, as described in the data section, then the list of tables should clearly indicate which tables, figures, and in-text numbers can be reproduced with the public material provided.

The provided code reproduces:

- ☐ All numbers provided in text in the paper
- ☐ All tables and figures in the paper
- ☐ Selected tables and figures in the paper, as explained and justified below.

Figure/Table #	Program	Line Number	Output file	Note
Table 1	02_analysis/table1.do		summarystats.csv	
Table 2	02_analysis/table2and3.do	15	table2.csv	
Table 3	02_analysis/table2and3.do	145	table3.csv	
Figure 1	n.a. (no data)			Source: Herodus (2011)
Figure 2	02_analysis/fig2.do		figure2.png	
Figure 3	02_analysis/fig3.do		figure-robustness.png	Requires confidential data

References

INSTRUCTIONS: As in any scientific manuscript, you should have proper references. For instance, in this sample README, we cited “Ruggles et al, 2019” and “DESE, 2019” in a Data Availability Statement. The reference should thus be listed here, in the style of your journal:

Steven Ruggles, Steven M. Manson, Tracy A. Kugler, David A. Haynes II, David C. Van Riper, and Maryia Bakhtsiyarava. 2018. “IPUMS Terra: Integrated Data on Population and Environment: Version 2 [dataset].” Minneapolis, MN: *Minnesota Population Center, IPUMS*. <https://doi.org/10.18128/D090.V2>

Department of Elementary and Secondary Education (DESE), 2019. “Student outcomes database [dataset]” *Massachusetts Department of Elementary and Secondary Education (DESE)*. Accessed January 15, 2019.

U.S. Bureau of Economic Analysis (BEA). 2016. “Table 30:”Economic Profile by County, 1969-2016.” (accessed Sept 1, 2017).

Inglehart, R., C. Haerpfer, A. Moreno, C. Welzel, K. Kizilova, J. Diez-Medrano, M. Lagos, P. Norris, E. Ponarin & B. Puranen et al. (eds.). 2014. World Values Survey: Round Six - Country-Pooled Datafile Version: <http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp>. Madrid: JD Systems Institute.

Acknowledgements

Some content on this page was copied from [Hindawi](#). Other content was adapted from [Fort \(2016\)](#), Supplementary data, with the author’s permission.