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.

Data Availability Statements

> INSTRUCTIONS: Every README should contain a description of the location and accessibility of the data used in the article. These descriptions are generally referred to as "Data Availability Statements" (DAS). This should include ALL data, regardless of whether they are provided as part of the replication archive or not, and regardless of size or scope. For instance, if using deflators, the source of the deflators (e.g. at the national statistical office) should also be listed here. DAS can be complex and varied. Examples are provided here, and below.
>
> INSTRUCTIONS: If providing a datafile per data source, list them here; if providing combined/derived datafiles, list them separately after the DAS.
>
> INSTRUCTIONS: 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** that 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.

Example for public use data

> 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]

Example for public use data with required registration:

> 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 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.

Data file	Source	Notes	Provided
`data/raw/lbd.dta`	LBD	Confidential	No
`data/raw/terra.dta`	IPUMS Terra	As per terms of use	Yes
`data/derived/regression_input.dta`	All listed	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, and Python are easy to set up and implement.

Software Requirements

- 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.readthedocs.io/en/1.1/requirements.html for further instructions on 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.

Description of programs

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/master.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/master.do` will run them all. Each program called from `master.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/master-appendix.do` will run them all.
- Ado files have been stored in `programs/ado` and the `master.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.

Memory and Runtime Requirements

INSTRUCTIONS: Memory and compute-time requirements may also be relevant or even critical. Some example text follows.

The code was last run on a **4-core Intel-based laptop with MacOS version 10.14.4**.

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**.

Instructions

> 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 master 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_master.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/master.do` will run them all in sequence, which should take about 2 hours.
- `programs/02_analysis/master.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/master-appendix.do`. The programs were last run top to bottom on July 4, 2019.

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.

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.

Acknowledgements

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