Guideline on Ethics, Privacy & Research Data Management BMS

Version 3.0 6 May 2024

1. ROLES & RESPONSIBILITIES

It is the individual researcher/first author of the project who always bears ultimate responsibility for the **proper handling** of data. Before the start of a project, researchers must always ensure that they have the appropriate expertise and capacity available to carry out the research in line with the VSNU's guidelines and Code of Conduct. This is one aspect of a professional attitude in science. Furthermore, in the case of junior researchers and PhD students, the first supervisor of the project (with a PhD) bears **joint responsibility** for proper data management, including verification of the data packages involved. The faculty will monitor events at a distance, and will perform occasional data package audits.

2. COURSES, GUIDELINES & TEMPLATES

Courses

- All **PhD candidates** are obliged to follow the TGS course Research Data Management Bootcamp (online course + interactive session + DMP). This course is organized 5 times per year, registration should be done via this link: https://www.utwente.nl/en/ctd/courses/1000227/data-management-bootcamp/.
- **Other employees** can access the content of the RDM online course on Canvas via this link: https://canvas.utwente.nl/courses/2167.
- **Students** can check the micro-lectures to get to know the topic of research data management.
- **Handling Personal Data in Research**: This Canvas course, which is open to every UT staff member, will help you understand the range of perspectives needed to build and demonstrate compliance with privacy regulation in research. Please enroll via this link https://canvas.utwente.nl/enroll/LGWB6F.

Guidelines

- Guide with examples on writing a data management section in NWO proposals
- Guide on handling personal data: Appropriate Use of Personal Data in Research According to GDPR (https://www.utwente.nl/en/bms/research/ethics/informed-consent-procedure/)

Templates

- DMP template: a DMP template with guiding information is available in the UT DMP tool, which has been accepted by funders such as NWO, ZonMw and the EU. Therefore it is recommended to write your DMP in this DMP tool.
- Informed consent template: A informed consent form template which takes data publishing/sharing after research into account is available in BMS (Dutch version, English version)

3. PRIVACY REGULATIONS

Human privacy and personal data is protected by the European General Data Protection Regulation (GDPR). Due to the nature of research in BMS, many researchers process (e.g. collect, store, analyze, etc.) personal data in research. Personal data means any information that can be traced directly (clear link) or indirectly (when it is bit hidden) to a identifiable natural person ('data subject'), for example a name, (email) address, photograph, voice/video recording, etc. Note, that indirect identifiers (age, place of birth, occupation, family composition, salary) may not be traceable as separate variables, but linked to each other or with other information, can lead to a person's identification.

Personal data should be handled according to the GDPR. UT made a **flowchart for researchers** for more detailed guidance on the appropriate use of personal data in research. Some of the limitations and requirements for processing personal data are listed here:

- **There should be a legal basis for processing personal data.** Researchers must base on at least one of the 6 legal grounds to be able to process personal data in a lawful manner.
- **Data subjects need to be transparently informed** about the fact that their personal data is being processed, for what purposes and whether their personal data will be transferred to other parties. Please check the informed consent procedure at BMS in section 4 ETHICS in this document.
- **Registration:** For every research project, to comply with GDPR, reporting any new processing which uses personal data to the Data Protection Officers (DPO) team is required at UT. If you **collect personal data for your research**, you **must record this in the data processing register** via the GDPR registration tool, with which you can also write a DMP. The Privacy Contact Person (PCP) of BMS faculty is able to support you on this. Only anonymized data is exempt from reporting in that register since it is by law no longer personal data. If the **students** process personal data in their **bachelor/master-thesis**:
    - and the student participates in an **existing project**: the student does **not** need to take any further action, assuming that the **responsible employee has already registered the research project** in the GDPR registration tool.
    - If the **research leads to a new processing**, the **student must register the processing**, where the **supervisor is recorded as the contact person**.
- **Agreements:** If you bring in someone (e.g. external research partners, app-developer) who will be processing personal data for you, this person is not allowed to use this information for his or her own purposes. You need to formalize this in a **data processing agreement** (in Dutch: Verwerkersovereenkomst).
- **Proper safeguards:** During research, personal data must be anonymized/pseudonymized as quickly as possible. Here are some basic steps of pseudonymization which are summarized by the working group of LCRDM.
- **Storage limitation:** Personal data must be deleted or rendered anonymous as soon as identification of data subjects is no longer necessary, and personal data must be stored and managed in a secure ICT system (like the UT network storage: P-drive & BMS server). More information about storing and transferring personal data can be found in 5. RESEARCH DATA MANAGEMENT section in this document.
- **The tools used for processing personal data should be GDPR compliant**, e.g. an online conference platform, software for data analysis, a survey tool (eg. Qualtrics), etc.

4. ETHICS

Ethical review

To ensure an ethically responsible research practice, it is **mandatory for staff and students from the Faculty of BMS** to submit their research project for ethical assessment **before the start of the research,** regardless of where it is conducted. This is in principle for all intended research **involving human participants** in an **indirect (i.e. file or social media research)** or **direct manner (i.e. experiments, surveys, interviews)**, and/or **using potentially sensitive data** about and/or **from individuals, groups, or organizations**.

The UT offers ethical review by one of the **4 domain-specific committees** which are facilitated by faculties, BMS runs the domain Humanities and Social Sciences (HSS). A small part of the research by our students/staff may better fit the domain CIS (e.g. cyber-related research) or NES (technology-related health research).

Currently at BMS, the BMS Ethics Web App is used for ethical review. It is recommended to submit the ethical review request **6 weeks before the start of the data collection.** Please check the ethical committee domain HSS page for more information about the ethical review procedure.

Informed consent procedure

Informed participation is an **ethical and legal requirement** for research involving human participants. **Consent for research ethics** is composed of providing information beforehand regarding study, purpose, risks, benefits, voluntary participation, as **consent as a legal basis** is used for the processing of personal data under GDPR. A '**Informed consent procedure**' consists of **an information sheet** and **an informed consent form**. There are different types of informed consent, how to choose which type to use in your research? Who are capable of giving consent? What information should be included in the information sheet and the informed consent form? Please check this page for detailed guidance on the informed consent procedure, at the bottom of this page, you can also find **example templates of informed consent forms**.

5. RESEARCH DATA MANAGEMENT

Data storage

Data storage concern all storage during the research. After a research project the term archiving applies. During the research data should be stored in such a way to minimize risk of data loss and to maintain data integrity. Research groups take measures to avoid loss of research data during the course of a research project, due to e.g. theft of laptops, fire and water damage, or a sudden leave of a researcher without the group having access to the data.

All collected research data, including related materials (e.g. protocols, models or questionnaires), must be stored in an ISO 27001- and NEN 7510-certified directory such as the Project and Organization directory (P-drive) including backups hosted by or offered through LISA, unless exceptions apply.

Where to store?
Common recommendations are listed below, complete option lists can be found in this decision tool.

Storage options	**Storage type**	**Suitable for personal/sensitive data** *	**Data sharing**	**Available to students**
P-drive	UT Network storage* (ISO 27001- and NEN 7510-certified)	Yes, by restricting access to a specific folder via ICT servicedesk	Yes: via ICT servicedesk	Yes, with invitation from supervisor via ICT servicedesk
BMS-server	UT Network storage* (ISO 27001- and NEN 7510-certified)	Yes	Yes, via BMS lab	Yes, with pre-requisite to register your project
M-drive	UT Network storage* (ISO 27001- and NEN 7510-certified)	Yes	No	No
SurfDrive	Cloud storage	Yes	Yes	No
Google drive & One drive	Cloud storage	No	Yes	Yes

* UT network storage requires VPN connection, please check the VPN setup manuals if needed.
* Personal data is not allowed to store in private devices. In case of portable storage, storage must be encrypted (see manuals here).

Data transfer & sharing

To safely transfer research data, the SurfFileSender is recommended. In case of personal data, please tick ‘Encryption’ before sending the data in SurfFileSender.

Data documentation

Data documentation is ‘information about the research data’. Sufficient documentation information about the data and analysis scripts are essential for the data to be understandable and therefore the research verifiable. Documentation can be in various formats. A more structured way of documenting data is to create metadata for your data. Metadata is ‘Data about data’. Check these pages about metadata to see what metadata need to be captured and how to add metadata to a dataset in Excel. Documentation can also be in plain text, e.g. in a README form, please check Appendix 1 for two adjustable templates of a README file.

Data archiving

Data archiving concerns data storage after a research project ends. In the light of open science and scientific integrity, sustainably archiving of static data and providing access is crucial. Data archiving aims in the first place at preventing physical data loss or destruction and securing the authenticity of data. Besides, it contributes to the quality and impact of your scientific work by enabling verification and possible reuse. For instance by allowing further analysis or follow-up research, or as a contribution to a data resource for the scientific community.

In order for the data to survive for the long term, an active preservation regime has to be applied because data automatically gets lost over time due to e.g. digital sources degrade over time (‘bit rot’) or file formats and software become outdated. Therefore preparations are necessary before data gets archived. One of the preparations is to convert data file formats to non-proprietary (open) and persistent formats, so data files can always be accessed (opened). Please consult the preferred file formats before you preserve your data for the long term.

It is recommended that research data is archived together with other related materials (e.g. analysis scripts, documentation materials) at the UT data archive Areda for at least 10 years. Especially personal data should not be archived in third parties unless contracts/agreements allow it and secure archive is guaranteed. However, what files need to be archived depends on the purpose of archiving, e.g. reuse or verification/reproduction, and also on the type of research, e.g. qualitative or quantitative research. Please check Appendix 2 to decide what files should be archived and what should be better destroyed.

For more information about Areda and how to prepare data for archiving, please see the UT Areda page. However, if secondary data is used in the project, please be aware that contracts and/or other written agreements between involved parties in a project may contain information about rights, limitations and licences related to these data, which sometimes may prevent secondary data to be archived.

Data publishing

To make your data and research more visible to the scientific community, in addition to archiving your data in Areda, you can also use trusted repositories to publish your data. By publishing/depositing your data set to a trusted repository, your data set gets a **persistent digital identifier** (e.g. DOI) which allows your data to be widely findable, accessible, and easily cited by others. Similar to data, your analysis syntax (e.g. R scripts/python code) may be re-used by others as well. Therefore, in addition to depositing your data to repositories, we strongly advise that you also make your code available. For making code available, you can use GitHub or upload a copy of your code to a trusted repository as you do with your data. For UT researchers, we recommend using DANS Easy repository and 4TU.ResearchData repository to make your data or analysis syntax available to others. DANS Easy focuses on humanities, life and health sciences, social and behavioral sciences, oral history and spatial sciences. A UT researcher can upload up to 50 GB of data (including code) free of charge on the DANS Easy repository. The 4TU.ResearchData repository focuses on disciplines of natural sciences, engineering and design. A UT researcher can upload up to 1 TB of data (including code) per year on the 4TU.ResearchData repository. After you decide which repository you are going to deposit your data and code to, you should also consider what license you want to have to accompany your data and/or code. A license will define what others may do or may not do with your data and/or code. Check this license selector to find the most suitable license for your data and/or code.

Appendix 1 README file templates

Template 1 README template from Cornel University
Link: https://data.research.cornell.edu/content/readme

AUTHOR INFORMATION

This readme file was generated on [YYYY-MM-DD] by [NAME]

<help text in angle brackets should be deleted before finalizing your document>

<[text in square brackets should be changed for your specific dataset]>

GENERAL INFORMATION

Title of Dataset:

<provide at least two contacts>

Author/Principal Investigator Information

Name:

ORCID:

Institution:

Address:

Email:

Author/Associate or Co-investigator Information

Name:

ORCID:

Institution:

Address:

Email:

Author/Alternate Contact Information

Name:

ORCID:

Institution:

Address:

Email:

Date of data collection: <provide single date, range, or approximate date; suggested format YYYY-MM-DD>

Geographic location of data collection: <provide latitude, longiute, or city/region, State, Country>

Information about funding sources that supported the collection of the data:

SHARING/ACCESS INFORMATION

Licenses/restrictions placed on the data:

Links to publications that cite or use the data:

Links to other publicly accessible locations of the data:

Links/relationships to ancillary data sets:

Was data derived from another source?

If yes, list source(s):

Recommended citation for this dataset:

DATA & FILE OVERVIEW

File List: <list all files (or folders, as appropriate for dataset organization) contained in the dataset, with a brief description>

Relationship between files, if important:

Additional related data collected that was not included in the current data package:

Are there multiple versions of the dataset?

If yes, name of file(s) that was updated:

Why was the file updated?

When was the file updated?

METHODOLOGICAL INFORMATION

Description of methods used for collection/generation of data: <include links or references to publications or other documentation containing experimental design or protocols used in data collection>

Methods for processing the data: <describe how the submitted data were generated from the raw or collected data>

Instrument- or software-specific information needed to interpret the data: <include full name and version of software, and any necessary packages or libraries needed to run scripts>

Standards and calibration information, if appropriate:

Environmental/experimental conditions:

Describe any quality-assurance procedures performed on the data:

People involved with sample collection, processing, analysis and/or submission:

DATA-SPECIFIC INFORMATION FOR: [FILENAME]

<repeat this section for each dataset, folder or file, as appropriate>

Number of variables:

Number of cases/rows:

Variable List: <list variable name(s), description(s), unit(s) and value labels as appropriate for each>

Missing data codes: <list code/symbol and definition>

Specialized formats or other abbreviations used:

Template 2 A template README for social science replication packages

> 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.

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 master 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. 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** 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.

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.

(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.

The code is licensed under a Creative Commons/CC-BY-NC/CC0 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.

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 dictionairy, 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).

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.

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, Python, Julia are easy to set up and implement.

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.

- 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://datagy.io/python-requirements-txt/ 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.

Memory and Runtime 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.

Summary

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

- ☐ <10 minutes
- ☐ 10-60 minutes

- ☐ 1-8 hours
- ☐ 8-24 hours
- ☐ 1-3 days
- ☐ 3-14 days
- ☐ > 14 days
- ☐ 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**.

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: Identifiying 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 tools/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/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.

(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/Creative Commons license.

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 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.
- 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.

Appendix 2 What data should be archived?

Archiving research outputs, such as results and the underlying research data, analysis and records of essential procedures, methods, notes, etc., is a very important step in the research life cycle. It is not always clear to every researcher what data should be retained when they prepare the archive package. Here we try to provide guidelines on data archiving for the BMS faculty for verification purposes. In terms of scientific integrity, individual researchers are responsible for keeping their research verifiable. In principle, anyone with relevant domain knowledge should be able to verify your research findings with the provided data archive package. To prepare such an archive package, we recommend considering including (at least) the 6 listed categories if applicable:

1. **The (published) manuscript/report** which demonstrates the research question, research design, data collection, processing and analysis methods and results

2. **Research data**
   - **Primary data** (data that is collected for the first time within the project)
     - **Raw (interview/focus group) video and audio recordings:** if analysis can be done completely on transcripts, keeping the raw recordings for the long term is generally not advised because of anonymization challenges;
     - **(De-identified) raw data files*** that you have not yet processed or analysed**,** such as interview/focus group transcripts or recordings if needed, answered Qualtrics questionnaires, raw time series for an EEG measurement, e-dat files for an E-Prime behaviour experiment, descriptions/notes of observations, raw scraped social media data, count of heartbeats/footsteps from wearable sensors, data extraction forms for systematic literature review and meta-analysis, etc.: these are the raw data that further processing and analysis are based on, therefore we recommend keeping them in an archive package;
     - **Simulation data** which is the input data that you feed to your simulation models;
     - **Final results** including the final versions of processed/analysed data that your research results are based on (e.g. an SPSS data file after transforming variables), (published) manuscript/report, figures, tables, etc.;
     - **Intermediate versions of processed/analysed data:** for verification purposes, it is preferably advised to keep the intermediate versions with clear and detailed description about the different versions in a version control table/change log. However in some cases that the research results can be reproduced without having the intermediate versions, they can also be excluded from the archive package. For re-use purposes only, intermediate versions are normally not necessary to be archived.

   - **(De-identified) secondary/third party data** (data that is normally collected for another purpose and usually before the start of the project)
     - In many cases, you are not the owner of the data because you receive this data from others/third-parties), such as patient data from hospitals, student score from schools, CBS micro data, public government data, etc.: keep this data in the archive package if the data owner/provider allows you to keep a local copy, otherwise try to provide information on how to access this data in the documentation files, e.g. in the README file, or generate derived data and keep sufficiently detailed documentation to enable other researchers to reproduce your findings.

3. **Computer code** (for example Atlas.ti, SPSS/JASP syntax file, MATLAB analysis scripts, R code) describing the steps taken to process the raw data into analysed data and/or to process the analysed data into results in the manuscript, including brief explanations of the steps in English.

4. **Supporting materials/notes** during research if applicable
   - **Ethical approval;**
   - **Signed informed consent forms and information sheets;**
   - **Code book/dictionary;**

- **The instructions, procedures, the design of the experiment and stimulus materials (e.g., interview guide, questionnaires, surveys, tests)** that can reasonably be deemed necessary in order to replicate the research;
- **A record of detailed information about data collection**: e.g. start and end date of data collection, any hard evidence that can relate to that (e.g. light reservations, train tickets, email conversations with the research subjects, etc.)
- **Etc**.

5. Documentation materials

- **A brief description** of the problem definition, research design, data collection (sampling, selection and representativeness of informants), data processing and analysis procedures and methods used if these are not included in the manuscript/report or the manuscript/report is not available.
- **A readme* file** (templates can be found in Appendix 1) describing which documents and files can be found where and how they should be interpreted. The readme file must also contain the following information:

a. Name of the person who stored the documents or files
b. Division of roles among authors, indicating at least who analysed the data
c. Date on which the manuscript was accepted, including reference
d. Date/period of data collection
e. Names of people who collected the data
f. If relevant: addresses of field locations where data were collected and contact persons (if any)
g. Whether the data is made open or not and if not, a valid reason for not opening up the data

6. A data management plan

** The readme file must be sufficiently clear. A relevant fellow researcher must be able to replicate the results discussed in the publication based on the components of the publication package.*

** Especially for convenient reuse, consider aggregating data into fewer, larger files, rather than many small ones. It is more difficult and time consuming to manage many small files and easier to maintain consistency across data sets with fewer, larger files. It is also more convenient for other users to select a subset from a larger data file than it is to combine and process several smaller files. On the other hand, very large files may exceed the capacity of some software packages. Some examples of ways to aggregate files include by data type, location, time period, measurement platform, investigator, method, or instrument.*

Resources:

Guidelines for the archiving of academic research for faculties of behavioural and social sciences in the Netherlands, March 2022, link available here: https://www.utwente.nl/en/bms/datalab/datasharing/guideline-faculties-of-behavioural-sciences-def.pdf

What data should be archived? Radboud University, link available here: https://www.ru.nl/rdm/archiving-data/what-data-should-archived/

Working with data: Weeding data. University of Bath Library, link available here: https://library.bath.ac.uk/research-data/working-with-data/weeding-data#s-lg-box-wrapper-17448856