Ten simple rules for creating a replication package

Koren, Vilhuber, Csóka, Connolly, Llull

This is draft

Goal:

- detailed and actionable guide on creating a replication package
- relying on our expertise as data editors of leading journals (AEA, ReStud, EJ, CJE)

Feedback:

lars.vilhuber@cornell.edu

Rule 1: Computational Empathy

Rule 1: Computational Empathy

Consider the the audience of your replication package

You are sharing your research with others. Some of your assumptions, tools, or methods may be trivial to you, but not to them.

Rule 1: Computational Empathy

Consider the the audience of your replication package

The replicator may have none of the setup, packages, and data that you have.

Their computer may not run the same operating system.

Rule 1: Computational Empathy

Consider the the audience of your replication package

The replicator of your package is likely to be less qualified than you are.

- Assume that the replicator has basic knowledge in how to run your software package, if the software is commonly used in your field.
- Compiled or new computer languages are much less likely to be **widely** used

Rule 1: Computational Empathy

Consider the the audience of your replication package

Minimize the input you need from the replicator.

- You can assume that they can manipulate a top-level configuration file
- but not 25 different files.

Rule 1: Computational Empathy

Consider the the audience of your replication package

Don't waste the replicator's time!

Rule 1: Computational Empathy

Consider the the audience of your replication package

Pity the
poor replicator!

Rule 2: Make your data accessible

Rule 2: Make your data accessible

Reproducing your research requires access to the same data as you have.

Share as much of your research data as you **legally** can.

Rule 2: Make your data accessible

Reproducing your research requires access to the same data as you have.

Collected original data through surveys or experiments?

- include the original data unchanged, with the exception of anonymization and other privacy protection.

Rule 2: Make your data accessible

Reproducing your research requires access to the same data as you have.

Used secondary data?

- include the raw data used from other sources, IF your usage terms permit

Rule 2: Make your data accessible

Reproducing your research requires access to the same data as you have.

Data creator or publisher prohibits redistribution?

- direct the reader on how to access them
- Provide as much detail as possible ($$, time, application details, etc.)

Rule 2: Make your data accessible

Reproducing your research requires access to the same data as you have.

The data files can usually be provided in any format compatible with any commonly used statistical package or software.

- You are encouraged to provide data files in **open, non-proprietary** formats.

Rule 3: Cite data and describe how others can access it

Rule 3: Cite data

and describe how others can access it

Cite all data you use that was produced by someone else.

Data citations are the best way to direct readers to these resources and to give credit to the original authors/creators.

Rule 3: Cite data

and **describe how others can access it**

Provide a **Data Availability and Provenance Statement** for each dataset you used, whether or not you included it in your package.

Rule 4: Describe software and hardware requirements

Rule 4: Describe software and hardware requirements

Describe all **hard** requirements about your computational environment.

But do not impose any fake requirements!

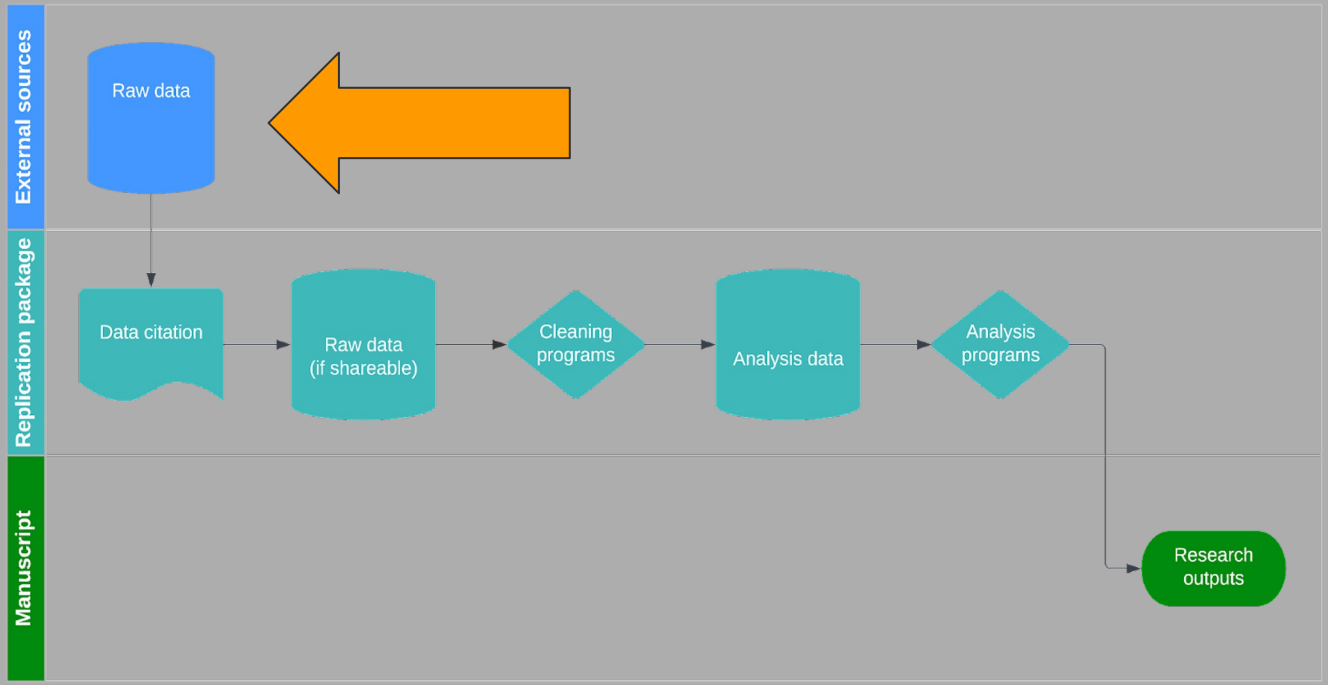
Rule 4: Describe software and hardware requirements

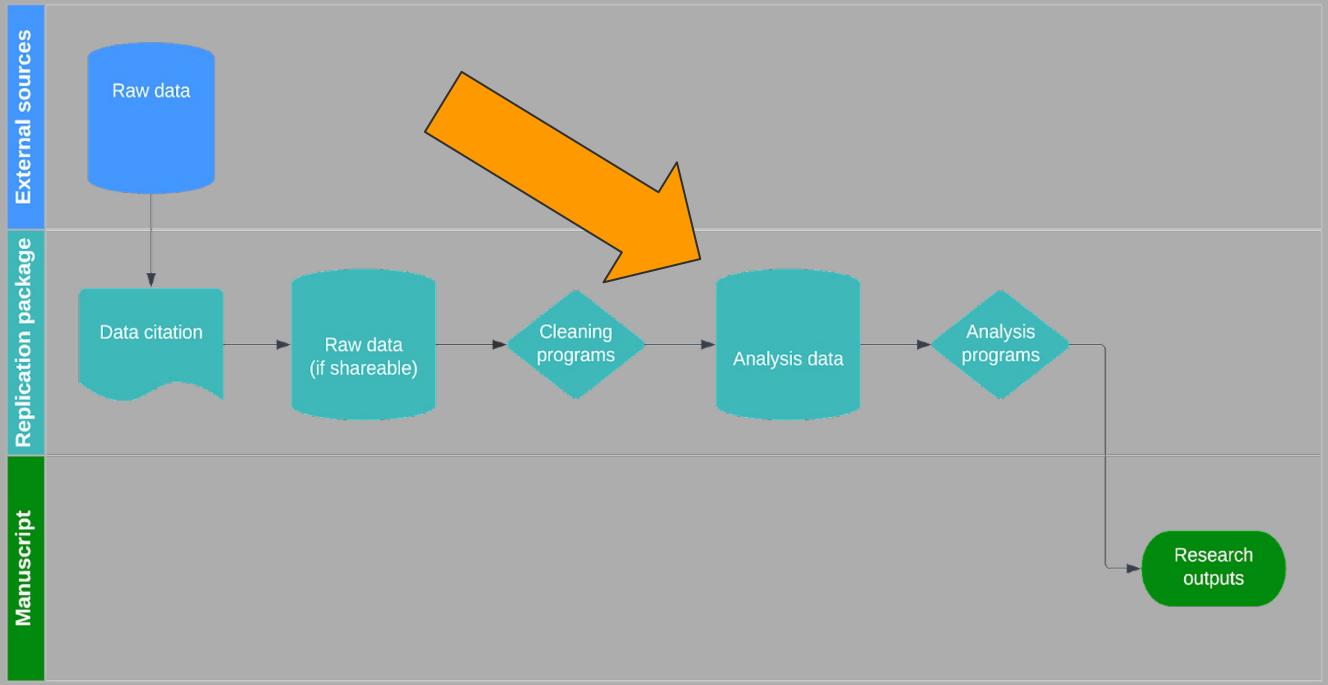
- Exact software versions used, including libraries, toolboxes, packages, etc.
- Hardware and OS on which last run

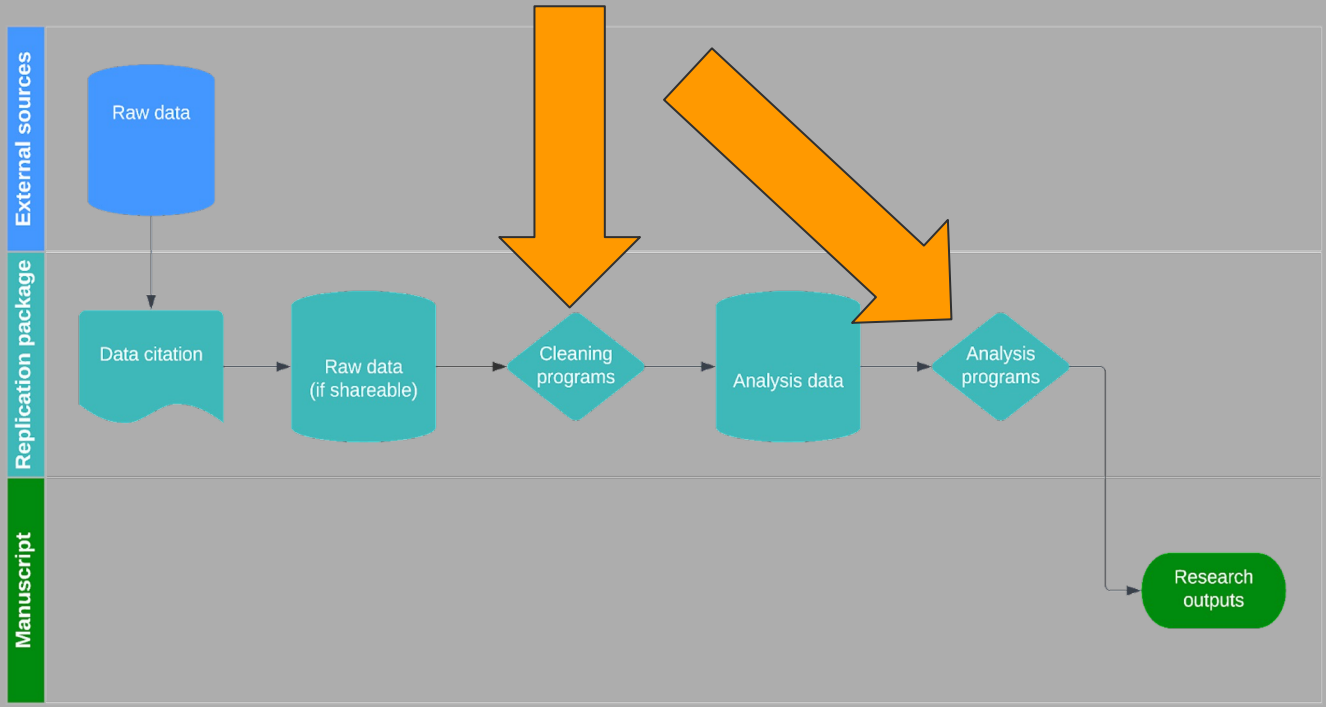
Rule 4: Describe software and hardware requirements

Software citations

- StataCorp. 2021. *Stata Statistical Software: Release 17.* College Station, TX: StataCorp LLC.
- Ben Jann, 2004. "ESTOUT: Stata module to make regression tables," *Statistical Software Components* S439301, Boston College Department of Economics, revised 12 Feb 2023.


External sources
Raw data
Replication package
Data citation
Raw data
(if shareable)
Cleaning
programs
Analysis data
Analysis
programs
Manuscript
Research
outputs


External sources
Raw data
Replication package
Data citation
Raw data
(if shareable)
Cleaning
programs
Analysis data
Analysis
programs
Manuscript
Research
outputs


External sources
Raw data
Replication package
Data citation
Raw data
(if shareable)
Cleaning
programs
Analysis data
Analysis
programs
Manuscript
Research
outputs

Rule 5: Provide code

Rule 5: Provide code

for data transformation and analysis

- Include all programs you wrote to **create any final and analysis** data sets from raw data.
- Include programs used to produce the final **computational results**

Rule 5: Provide code

for data transformation and analysis

- The programs may be provided in any format compatible with statistical packages or software commonly used in your discipline (**native format)**

- Do NOT provide code in Word, PDF, or some other transformed format!

Side-note

Many researchers conduct computational research in an interactive way

- Load data manually
- Run parts of code interactively ("highlight-and-run")

This is perfectly fine while developing or debugging your code.

Side-note

Many researchers conduct computational research in an interactive way

- Load data manually
- Run parts of code interactively ("highlight-and-run")

This is fine while developing or debugging your code.

It is an incredibly inefficient way to reproduce code

- For the replicator (see Rule 1, "Pity the poor replicator")
- For the researchers themselves

Side-note

Many researchers conduct computational research in an interactive way

- Load data manually
- Run parts of code interactively ("highlight-and-run")

This is fine while developing or debugging your code.

It is an incredibly inefficient way to reproduce code

- For the replicator (see Rule 1, "Pity the poor replicator")
- For the researchers themselves

Corollary: Try to automate as much of the code as possible, without however obscuring what is happening in the code (transparency)

Rule 6: Explain how to reproduce your work

Rule 6: Explain how to reproduce your work

Readers may be expert quantitative scientists (or students wishing to become such experts)

but still need **specific** instructions on how to reproduce your analysis.

Rule 6: Explain how to reproduce your work

Provide a document outlining

- where the data comes from
- what data are provided
- what requirements are needed
- how to run the code,
- what results to expect
- where to find the results.

Rule 6: Explain how to reproduce your work

Rule 6: Explain how to reproduce your work

Ideal:

single main script that runs all analyses

Rule 6: Explain how to reproduce your work

Process should be as simple as necessary.

But: if there are any manual steps in the process, state them explicitly

- readers may need to copy the confidential dataset they obtained to a particular folder,
- a small number of configuration parameters before running your code
- Manual steps that cannot be implemented in a script (e.g. most frequent implementations of ArcGIS)

write **clear step-by-step instructions** that allow users to reproduce the results.

Rule 7: List all exhibits that can be reproduced

Rule 7: List all exhibits

that can be reproduced
and
save them via scripts

Create a list of exhibits and state which one is produced by which script.

If a script creates multiple exhibits, point to the exact line number.

Rule 8: Include all supporting materials

Rule 8: Include all supporting materials

Surveys documents, experiment code, etc.

For papers collecting original data through **surveys**:

- Survey instruments
- Computer code for survey collection
- Original instructions to survey personnel
- Original instructions to survey respondents
- Details on subject selection.

Rule 8: Include all supporting materials

Surveys documents, experiment code, etc.

For papers running (lab or field) **experiments**:

- experiment instructions for lab personnel,
- computer code for experiment mechanisms
- original instructions for l details on subject selection.

Rule 8: Include all supporting materials

Surveys documents, experiment code, etc.

Include details about

- ethical approval and
- pre-registration of the research.

(If not already present in the manuscript)

Rule 9: Use a permissible license

Rule 9: Use a permissible license

What's a license?

Rule 9: Use a permissible license

- A license specifies the terms of use of code and data in the replication package.

Rule 9: Use a permissible license

- By default, code and data is under **copyright** protections
- A license **relaxes** some of those protections without needing to ask the copyright owners explicitly for permission.

Rule 9: Use a permissible license

An appropriately liberal license allows for replication by researchers unconnected to the original parties.

- Ideally, full re-use in their own research (with attribution = citation!)
- At a minimum, usage for reproducibility and replication

Rule 9: Use a permissible license

Common open licenses for data:

- **CC-BY** or
- public domain,
- "open data licenses" (some stats agencies).

For code,

- MIT
- **BSD**

Rule 10

Re-run everything

The End