Tips from a Data Archive: Preparing and working with replication packages

October 21, 2025

Yale DISSC-Library ORR series

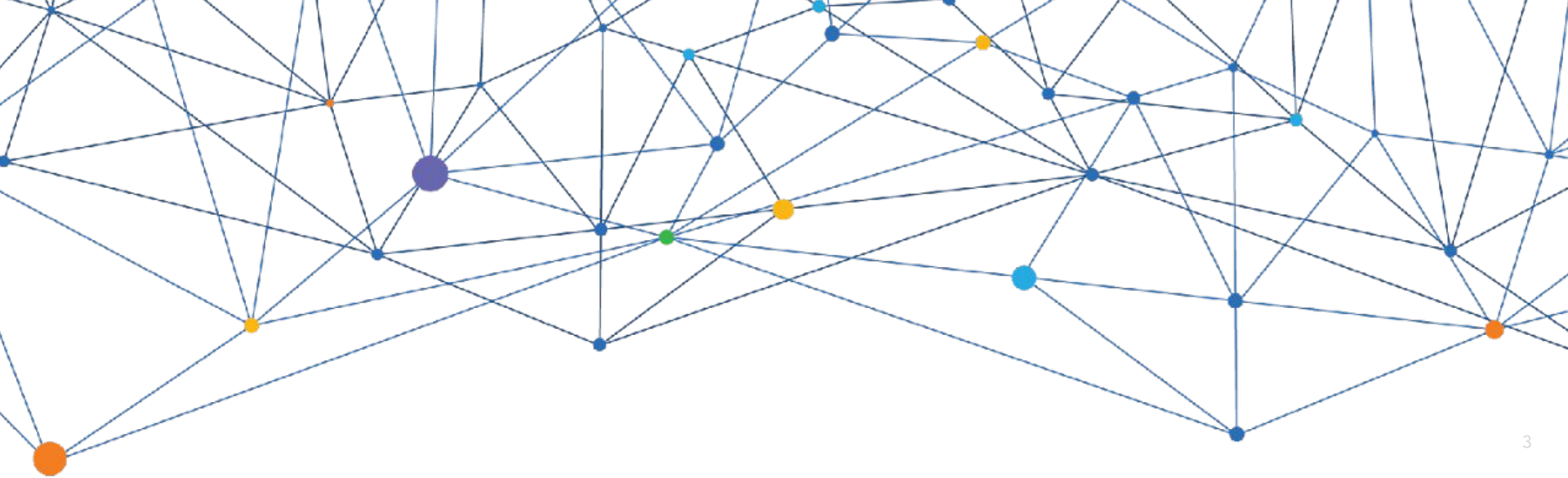
Limor Peer
Fanmei Xia
Zhouyan Liu

Yale *Data-Intensive Social Science Center*

Open & Reproducible Research

Welcome & Introduction

- Agenda for today
  - Reproducibility
  - ISPS Data Archive
  - Case studies
  - Tips & best practices
  - Q&A



On reproducibility at ISPS (Limor Peer)

What is reproducibility?

This may be confusing…

methodological reproducibility

REPLICABILITY

validation

direct replication

statistical reproducibility

COMPUTATIONAL REPRODUCIBILITY

REPEATABILITY

conceptual replication

empirical reproducibility

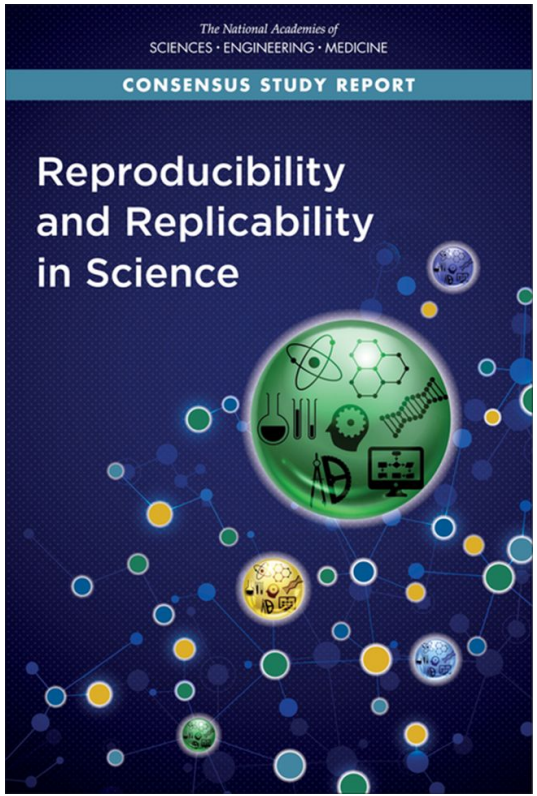
VERIFICATION

We follow the [The Turing Way](https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html)…

		Data: Same	Data: Different
Analysis	Same	Reproducible	Replicable
Analysis	Different	Robust	Generalisable

https://the-turing-way.netlify.app/reproducible-research/overview/overview-definitions.html

What is reproducibility?



Reproducibility is obtaining consistent results using the same input data, computational steps, methods, and code, and conditions of analysis. This definition is synonymous with "computational reproducibility."

National Academies of Sciences, Engineering, and Medicine. (2019). Reproducibility and Replicability in Science. National Academies Press. https://doi.org/10.17226/25303

Why reproducibility?

Good practice*

- Improve your productivity
- Verify your own results
- Enables others to extend your work
- Verify/disprove other's results
- Survive the tech evolution
- Enable community maintenance & support

Required by funders and journals

Scientific norm

Ethical and credible science

What is Gold Standard Science?

As detailed in Executive Order 14303, Gold Standard Science refers to science conducted in a manner that is:

- Reproducible.
- Transparent.
- Communicative of error and uncertainty.
- Collaborative and interdisciplinary.
- Skeptical of its findings and assumptions.
- Structured for falsifiability of hypotheses.
- Subject to unbiased peer review.
- Accepting of negative results as positive outcomes.
- Without conflicts of interest.

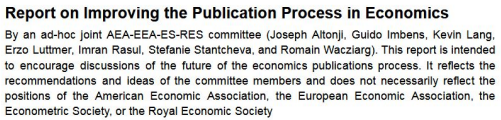
https://www.nsf.gov/policies/gold-standard-science
August 22, 2025

*Ivie & Thain https://doi.org/10.1145/3186266

What do researchers think about journals' data and code policies?

- They do enhance the credibility of economic research and the credibility of economics as a discipline overall
- They do allow catching inadvertent mistakes in their own work or caused them to adopt better ways of conducting their empirical research
- They won't prevent ill-intentioned researchers from committing misconduct,
- They don't ensure the code is correct or corresponds to the methods described in the paper
- All in all, these policies enable others to uncover such mistakes

2024 survey of members of four Economic associations



Report on Improving the Publication Process in Economics

By an ad-hoc joint AEA-EEA-ES-RES committee (Joseph Altonji, Guido Imbens, Kevin Lang, Erzo Luttmer, Imran Rasul, Stefanie Stantcheva, and Romain Wacziarg). This report is intended to encourage discussions of the future of the economics publications process. It reflects the recommendations and ideas of the committee members and does not necessarily reflect the positions of the American Economic Association, the European Economic Association, the Econometric Society, or the Royal Economic Society

February 2024, report by the American Economic Association (AEA), the European Economic Association (EEA), the Econometric Society (ES), and the Royal Economic Society (RES)
https://www.econometricsociety.org/uploads/documents/editorial/Improving%20Publication%20Process%20in%20Economics%20Report_2025.pdf

Reproducing other's research

…this may initially sound like a trivial task but experience has shown that it's not always easy to achieve this seemingly minimal standard.

American Statistical Association (2017). Recommendations to Funding Agencies for Supporting Reproducible Research https://www.amstat.org/asa/files/pdfs/POL-ReproducibleResearchRecommendations.pdf (accessed September 18, 2025)

The most commonly reported problems associated with [replication] attempts were the lack of… data and code, followed by insufficient documentation.

Janz, N., Werfel, S., Wykstra S. (2014). Replication in political science graduate courses: an untapped resource? Monkey Cage https://www.washingtonpost.com/news/monkey-cage/wp/2014/02/12/replication-in-political-science-graduate-courses-an-untapped-resource/ (accessed September 18, 2025)

Reproducibility at ISPS (Institution for Social and Policy Studies)

ISPS standard: Can a 3rd party reproduce the results… without any additional information from the author?

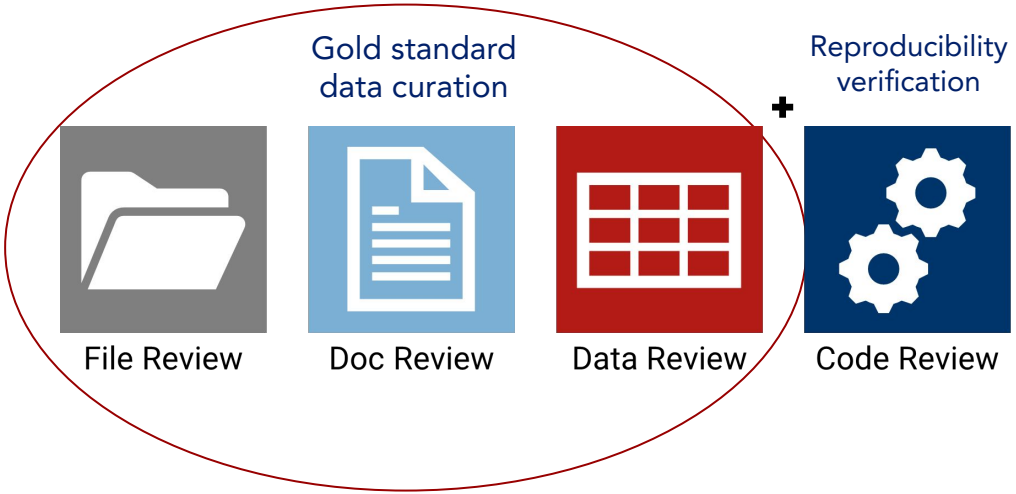
We review the code and confirm that,

- The code executes. We sometimes have problems running the code and need to diagnose the issue e.g., different versions of commands, syntax errors, user-generated commands needed, missing code for some published tables.
- The output matches the reported results. We sometimes catch discrepancies e.g., differences between output and tables in a manuscript.

https://isps.yale.edu/research/data

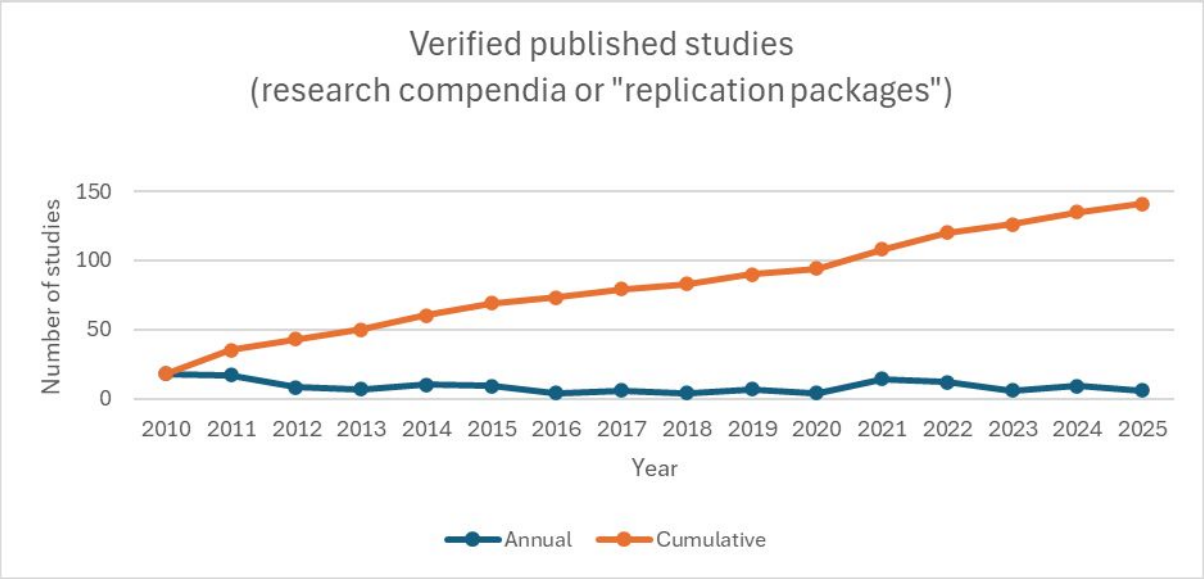
Reproducibility at ISPS

- We evaluating scientific claims using the underlying data and code (aka "replication package" or "replication file" or "research compendium")
- "Data Quality Review" Framework: Data curation to enhance usability and interpretability + code review to verify reproducibility



Peer et al., (2014). Committing to Data Quality Review https://doi.org/10.2218/ijdc.v9i1.317 (accessed October 18, 2025).

Reproducibility at ISPS



- → 142 published studies
- → 3,225 files/handles
- → 19 GB

(September 2025)

Reproducibility at ISPS - additional reading

- Code Review, Reproducibility, and Improving the Scholarly Record (2025) https://doi.org/10.1162/99608f92.f9d748d4
- Why and How We Share Reproducible Research at Yale University's Institution for Social and Policy Studies (2024) https://doi.org/10.1162/99608f92.dca148ba
- Active Maintenance: A Proposal for the Long-term Computational Reproducibility of Scientific Results (2021) https://doi.org/10.1017/S1049096521000366
- New Curation Software: Step-by-Step Preparation of Social Science Data and Code for Publication and Preservation (2016) https://doi.org/10.29173/iq902
- Mind the Gap: Data They Share May Not Be Data You Can Use (2014) https://isps.yale.edu/news/blog/2014/03/mind-the-gap
- The Role of Data Repositories in Reproducible Research (2013) https://isps.yale.edu/news/blog/2013/07/the-role-of-data-repositories-in-reproducible-research
- Building an Open Data Repository for a Specialized Research Community: Process, Challenges, and Lessons (2012) http://www.ijdc.net/index.php/ijdc/article/view/212

ISPS is likely the first re-user of your data & code*

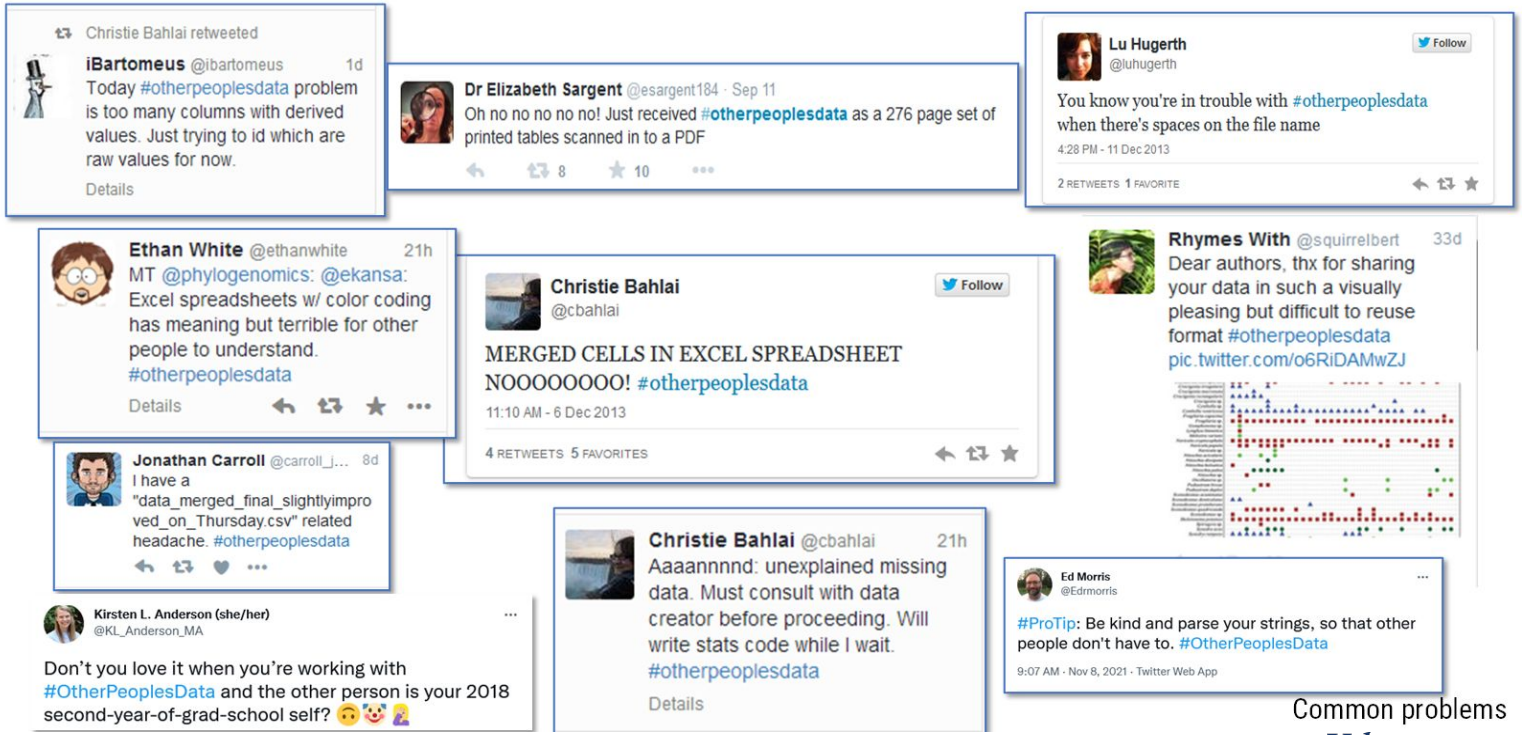
RA: "We are missing labels for the following variables:
_n1, _n0, V1 and V0."

Researcher: "Here are the labels:
_n1 is the number of observations in the treated strata before matching
_n0 is the number of observations in the comparison strata before matching
v1 = turnout for treated observations
v0 = turnout for comparison observations

… this reminds me that I needed to include the .ado code in the Matching Code folder. I just did that and updated the readme file. Boy, the things your forget about after not thinking about something for two years!"

*Actual email exchange

#otherpeoplesdata



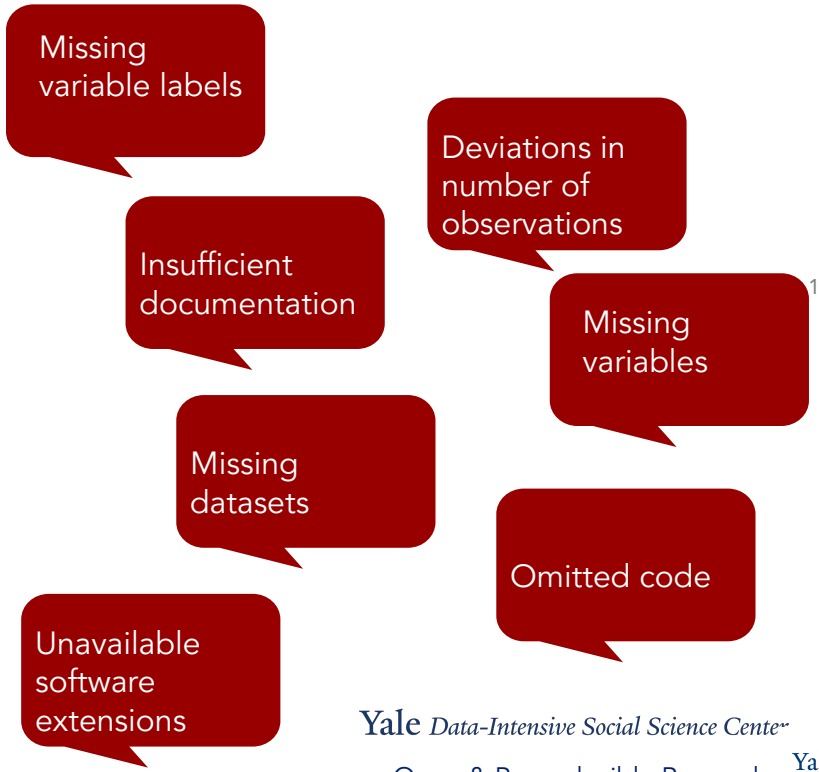
#otherpeoplescode

Common problems

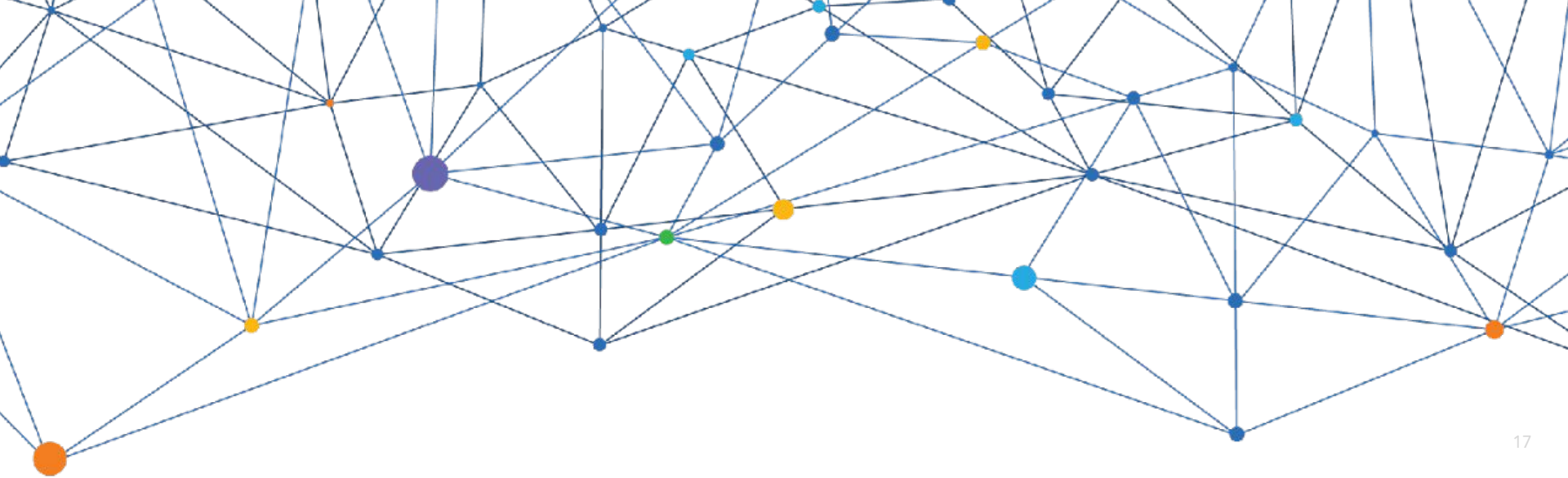
Common problems (i.e., attempting to reproduce results you often find…)

An Institute for Replication (I4R) meta paper found,

- 25% of studies have a coding error:
  - Range from minor to MAJOR
    - » Ex. 75% of observations are duplicates
    - » Not cleaning raw data (e.g., St. Louis, St Louis, StLouis, …)
    - » Not fully interacting DID model
    - » Not specifying GMM function
- Mentioning something in the paper, but doing something else in the code
  - Rare, but happened twice for inference
- Important coding decisions buried in footnote or appendix



From Brodeur, Abel et al. (2024) https://hdl.handle.net/10419/289437



Example (Zhouyan Liu)

A Case Study

Manuscript

DE GRUYTER Stat Polit Pol 2017; 8(1): 29–40

Alexander Coppock*

Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

DOI 10.1515/spp-2016-0005

+

Replication package

Name	Date modified	Type	Size
shy_trump_cleaned	6/28/2017 4:49 PM	Microsoft Excel C...	471 KB
shy_trump_analysis.R	6/28/2017 4:49 PM	R File	11 KB
README	6/28/2017 4:49 PM	Text Document	1 KB
.DS_Store	6/28/2017 4:49 PM	DS_STORE File	7 KB

=

HOME > RESEARCH > DATA

Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

Suggested citation:

Alexander Coppock (2017). replication materials for, 'Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment.' http://hdl.handle.net/10079/zw3r2f9. ISPS Data Archive.

ISPS ID: D149

Related publications:

Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment

Keyword(s): List experiment
Election Polling

Research design: Survey experiment

Data type: Survey/interview (e.g., ANES)

Data source(s): Author

Field date: September 1, 2016

Field Date: 2016-09

Location: United States

Location details: United States

Unit of observation: Individual

Sample size: 5290

Inclusion/exclusion: 18+

Randomization procedure: Simple random assignment

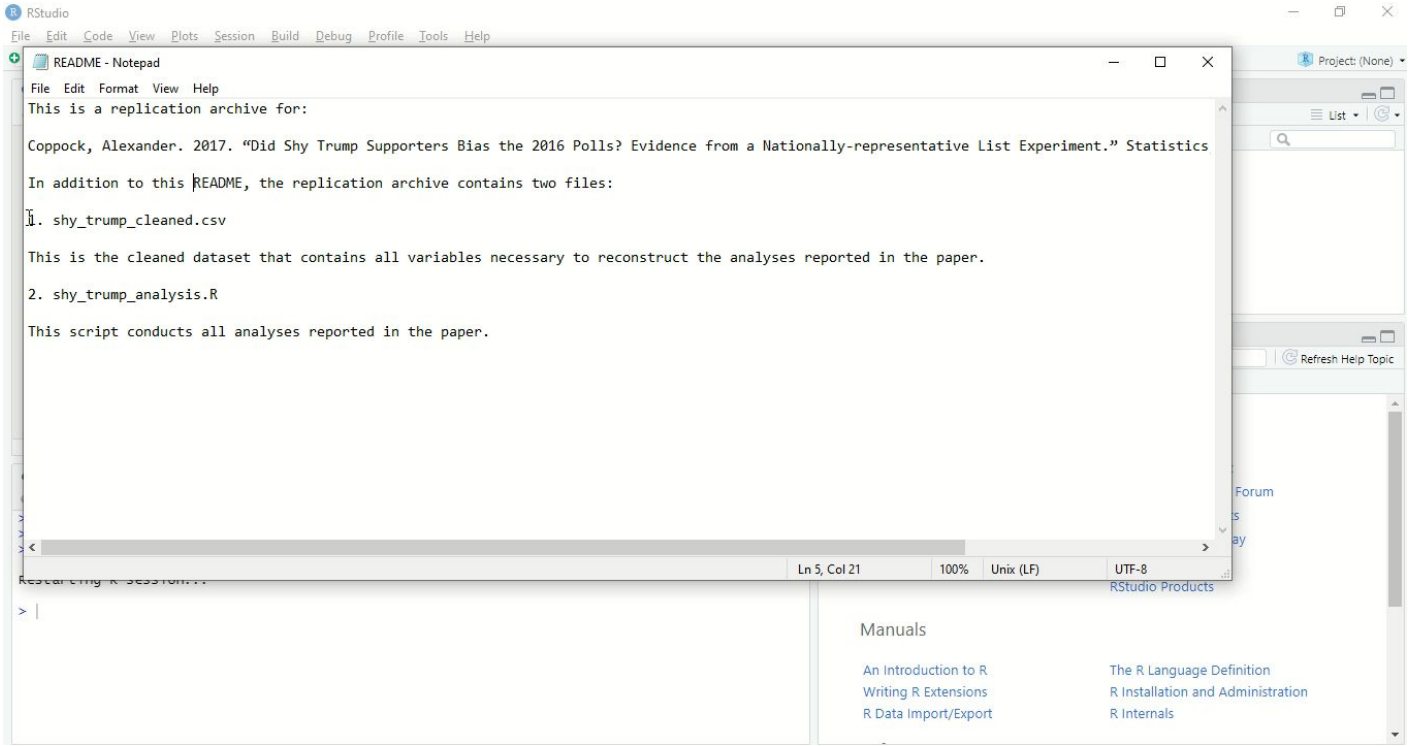
Treatment: treatment and control list

Treatment administration: Web delivered

Outcome measures: Number of things a respondent "would do"

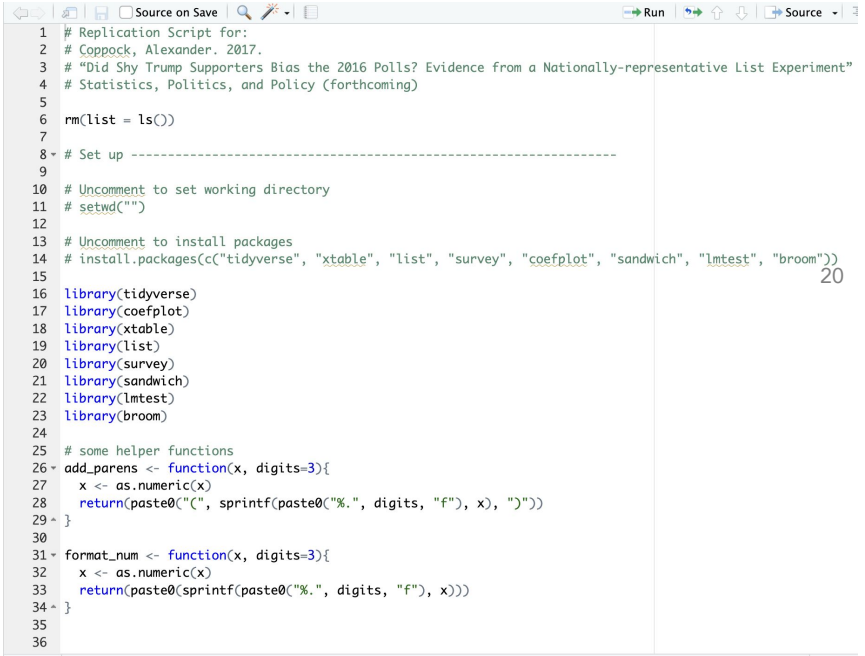
Archive date: 2017

A Case Study


RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
README - Notepad
File Edit Format View Help
This is a replication archive for:
Coppock, Alexander. 2017. "Did Shy Trump Supporters Bias the 2016 Polls? Evidence from a Nationally-representative List Experiment." Statistics
In addition to this README, the replication archive contains two files:
1. shy_trump_cleaned.csv
This is the cleaned dataset that contains all variables necessary to reconstruct the analyses reported in the paper.
2. shy_trump_analysis.R
This script conducts all analyses reported in the paper.
Ln 5, Col 21
100%
Unix (LF)
UTF-8
Project: (None)
Refresh Help Topic
Forum
RStudio Products
Manuals
An Introduction to R
Writing R Extensions
R Data Import/Export
The R Language Definition
R Installation and Administration
R Internals

Well-prepared Replication Packages

- Complete replication package: data, code, README
- Clear documentation facilitates quick orientation.
- Code maps cleanly to paper sections and tables.



When Perfect Packages Still “Fail”

- Package relies on outdated **dplyr** grouping format
- bootstrap( ) removed from **broom**

```
Error in `group_data()`:
! `.data` must be a valid <grouped_df> object.
Caused by error in `validate_grouped_df()`:
! Corrupt `grouped_df` using old (< 0.8.0) format.
i Strip off old grouping with `ungroup()`.
Run `rlang::last_trace()` to see where the error occurred.
Warning message:
In bootstrap(., m = 10) :
  `bootstrap()` is deprecated and will be removed in an upcoming
release of broom. See the rsample package instead.
```

Another example: when the original data cannot be shared

```
**********************************************************************************************

Version 1.0
2025/01/22

This is the replication archive for "Seeing the state in action: Public preferences about and
judgments of common police-civilian interactions." Forthcoming, Criminology.

Paige E. Vaughn and Gregory A. Huber

For questions about this replication archive, please feel free to contact gregory.huber@yale.edu

**********************************************************************************************

The source data for all three studies, which includes potentially identifying information, are
not included in this archive.

00_00_ExecuteAllPrepareDataScripts.do is a Stata .do file that executes the three cleaning files
(00_01_PrepareStudy1Data.do; 00_02_PrepareStudy2Data.do; and 00_03_PrepareStudyR1Data.do) that
process the raw exports from the experiments (Qualtrics files) into three anonymized datasets
included in this archive:

Study1data_publicdatafile_cleaned.dta
StudyR1data_publicdatafile_cleaned.dta
Study2data_publicdatafile_cleaned.dta

Log files from this cleaning are included for auditing purposes:

LogFile_PreparedPublicReplicationFileStudy1.txt
LogFile_PreparedPublicReplicationFileStudyR1.txt
LogFile_PreparedPublicReplicationFileStudy2.txt

01_Criminology_FinalAnalysisCombined.do is the core analysis file that loads all three study
datasets and produces this output:
```

Even though the raw data cannot be share, the authors provided:

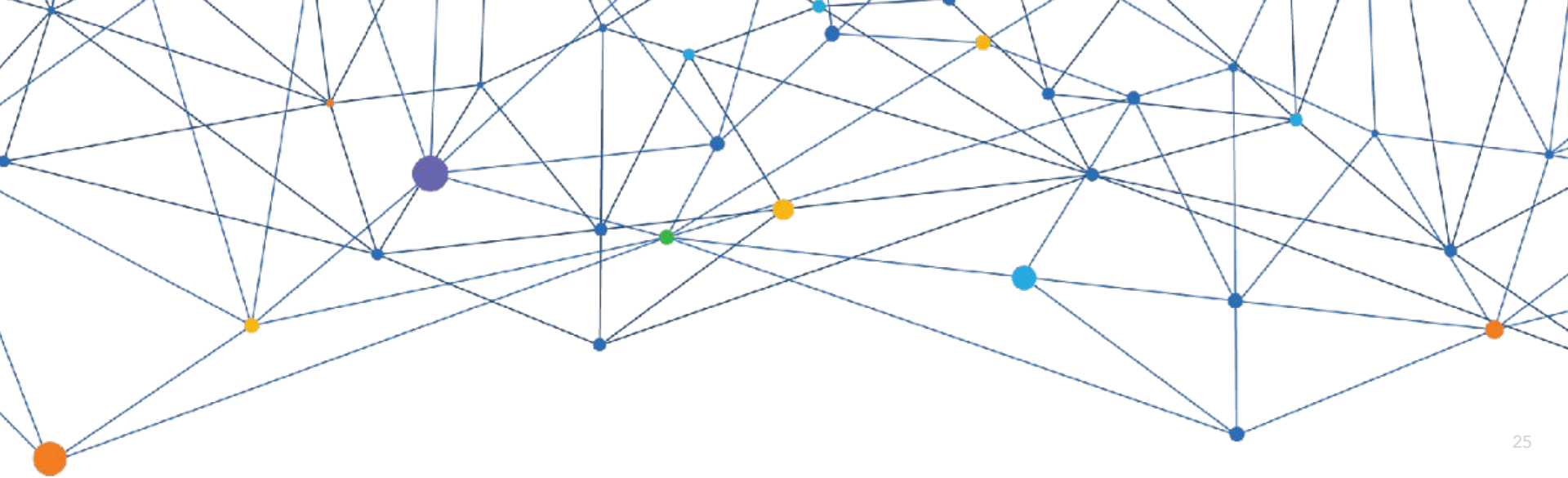
- the cleaned data;
- the code used to process the raw data into the cleaned data (for readers interested in technical details)
- the log files (for auditing purposes).

Set clear expectations for users (what is included and what is not)

Meet the needs of users at different levels.

Broader lesson

- Even exemplary replication packages need active maintenance; Reproducibility is a continuous process, not one-time achievement
- From the reader's perspective, provide as much detailed information as possible, and include clear, step-by-step, and well-structured usage instructions — especially when:
  - raw data cannot be shared,
  - results come from multiple sources,
  - the environment or API setup is complex,
  - the code structure is intricate.



Tips & best practices (Fanmei Xia)

Guidelines for authors

- Computational compassion
- Any average user should be able to replicate your results using the files provided in the reproducible package
  - Consider the average user to be an undergrad who is NOT proficient in coding/your field of study
- Before submitting to the Archive, double-check:
  - Does the link to the data source/depository still work?
  - Is the same version of the data used in the study still available?
  - Does the code run? Check with a clean computing environment!
- It is okay to make mistakes; publication does not mean perfection

What to include in the replication package

Minimum:

- Data File(s)
- Program File(s)
- Link to publication

Strongly encouraged:

- README file
- Output File(s)
- Codebook(s)
- Questionnaire(s)
- Study metadata
- Treatment Materials
- Supplementary Materials

Tips and best practices

Before you start:

- Arrange the files in the order in which they need to be operated & clearly marked for their functions:
  - E.g. 00_README, 04_analysis_table_1.r
- Document all of the installed library/package/ado files (before your forget!)
- Arrange your code chunks/files corresponding to the order of the output
- Consider keeping a log file

Tips and best practices

README

- Project overview
- Setup instructions (e.g. computing environment, software versions, source datasets, etc.)
- File structure (list the files by name and describe their function)
- State any data access restrictions, licenses and rights for data and software, etc.
    - If data is obtained elsewhere, provide data citation
- Contact information

Data

- Label all variables and values
- Data files called by the code need to be included in the corresponding file path

Tips and best practices

Code

- Comment code to describe processes and map to paper sections
- Order code outputs in the same order as they appear in paper
- Anonymize file paths (use relative file paths)
- Use reproducibility-enabling commands (e.g., groundhog.library(), sessionInfo())
- Execute code with a clean computing environment
- Simulation/stochastic processes
  - Random seed(s)
  - Test the sensitivity
- If the replication material takes a LONG time to run
  - Optimize the code

Tips and best practices

Codebook

- Include all the information about the variables
  - E.g. variable name, label, weight, transformation, etc.

Resources on preparing and working with replication packages

- Best Practices for Replication Packages, Social Science Data Editors https://ejdataeditor.github.io/best.html
- Checklist for replication packages https://www.econometricsociety.org/uploads/reports%20Editorial/ES_Data_Editor_Website/Checklist.pdf
- 10 Things for Curating Reproducible and FAIR Research, Research Data Alliance https://curating4reproducibility.org/10things/
- The DIME Analytics Data Handbook, World Bank https://worldbank.github.io/dime-data-handbook/publication.html
- Handbook for Reproduction and Replication Studies https://forrt.org/replication_handbook/Handbook-for-Reproduction-and-Replication-Studies.pdf

Also may be of interest…

- Rokem A (2024) Ten simple rules for scientific code review. PLoS Comput Biol 20(9): e1012375. https://doi.org/10.1371/journal.pcbi.1012375
- TiSEM Guideline Replication Package https://www.tilburguniversity.edu/research/economics-and-management/replication-package

General guidelines for reviewers/general users

- Good faith effort in the best tradition of scientific inquiry; this is about constructive criticism not being adversarial
- Avoid “replication anxiety”:
  - It’s like following a recipe!
- When encountering research artifacts that do not computationally reproduce results, future users—including the original authors—can use the experience to contribute to the scholarly record by extending the prior work.
- Replication Compassion:
  - Replicate others as you would like to be replicated yourself
  - Making mistakes is human

General guidelines for organizations

- Active maintenance
- Obtain access to data
- Any future users can report errors and remedies to the original authors by
  - sharing a curator/replicator note
  - add comments
  - modify (and version) the code

DISSC

Please give us your feedback!

Questions?

Thank You!

https://dissc.yale.edu/

Yale *Data-Intensive Social Science Center*

Open & Reproducible Research