Reproducible Research Practices

Limor Peer, PhD
Associate Director for Research, Institution for Social and Policy Studies
Research and Data Specialist, Office of the Provost

Symposium on Improving Reproducible Research Practices in Schools of Public Health
Yale School of Public Health | 16 April 2018
Themes for this talk

How to think about reproducible research practices

Practice   Teach   Preach
Transparency   Quality   Independence
An open access digital collection of social science experimental data, metadata, code, and associated files produced by ISPS researchers, for the purpose of replication of research findings, further analysis, and teaching.
...when scientific claims are subjected to scrutiny

One of the core principles of the scientific process is that other scientists are able to repeat your experiment and either confirm or refute your results.

This is referred to as reproducibility or replication.

Repeat After Me, Maki Naro https://thenib.com/repeat-after-me
computational reproducibility

REPLICABILITY direct replication
direct replication

empirical reproducibility VALIDATION

REPEATABILITY conceptual replication

methodological reproducibility

VERIFICATION REPRODUCIBILITY

statistical reproducibility
Illuminating the black box:

“Transparency requires making visible both the empirical foundation and the logic of inquiry of research.”

Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors
Data sharing and reuse (#otherpeoplesdata)

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

“Because there are more ways to share data, and because the scholarly landscape supports and encourages that, there is a proliferation of data files on many different types of systems that do not meet the criterion of quality...”

“The replication standard holds that sufficient information exists with which to understand, evaluate, and build upon a prior work if a third party could replicate the results without any additional information from the author.”


“Could the published computational findings be reproduced on an independent system by using the data and code provided?”

ISPS Data Archive: first re-user

DATA QUALITY REVIEW

FILE REVIEW

DOC REVIEW

DATA REVIEW
ISPS Data Archive: first re-user

DATA QUALITY REVIEW

FILE REVIEW
DOC REVIEW
DATA REVIEW
CODE REVIEW

Building trust and expecting accountability
Curating for reproducibility

- Assign persistent identifier
- Create study citation and study-level metadata record
- Record file size details
- Check for presence of all files
- Verify content of files matches expected format
- Create non-proprietary versions of files
- Implement migration strategy for file formats
Curating for reproducibility

- Confirm presence of comprehensive descriptive information necessary for informed reuse
  - Data definitions
  - Variable construction
  - Methodology
  - Sampling information
  - Original data source citation
  - Analysis software version
- Link to related research products
Curating for reproducibility

- Check for undocumented variable and value information
- Examine data for inconsistencies and errors
  - Discrepancies in number of observations
  - Out-of-range or wild codes
  - Undefined null values
- Review data for confidentiality issues
Curating for reproducibility

- Convert absolute file paths to relative file paths
- Check code for presence of non-executable comments that document analysis processes
- Identify packages required to execute code
- Execute code to ensure code is error-free
- Compare code output to findings presented in article
Prying information from researchers

"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 you forget about after not thinking about something for two years!"

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

Archive staff

Researcher
Example 1:

```stata
/*Create variables used in regressions*/
gen mol_mcond=0
   replace mol_mcond=1 if mccain==1 & endorse==1
ngen mol_obend=0
   replace mol_obend=1 if mccain==1 & endorse==5
gen obl_mcond=0
   replace obl_mcond=1 if mccain==0 & endorse==1
gen differ=0
   replace differ=1 if mcl_obs==1 | obl_moe==1
```

```
.replace mol_mcond=1 if mccain==1 & endorse==1
(12 real changes made)
gen mol_obend=0
.replace mol_obend=1 if mccain==1 & endorse==5
(36 real changes made)
gen obl_mcond=0
.replace obl_mcond=1 if mccain==0 & endorse==1
(15 real changes made)
gen differ=0
.replace differ=1 if mcl_obs==1 | obl_moe==1
(51 real changes made)
```
Example 2:

\[ \text{data} + \text{code} = \text{reported results?} \]
Common replication problems

• Insufficient documentation
• Missing variables
• Deviations in number of observations
• Unavailable software extensions
• Omitted code
• Incompatible datasets
Supporting research data curation and code review for the purpose of facilitating the digital preservation of the evidence base necessary for future understanding, evaluation, and replication of scientific claims.

https://cure.web.unc.edu/
Curating for reproducibility

- Transparency, Access, and Trust
- Usability
- Independence
- Pre-publication

Establish Standards
Share Practices
Promote Data Quality Review

https://cure.web.unc.edu/
Reproducible research: practice

Recommendations for statistical studies

1. Do all data preparation and analysis in code.
3. Build all analysis from primary data files.
4. Fully describe your variables.
5. Document every empirical claim.
6. Archive your files.
7. Encourage coauthors to adopt these standards.

Reproducible research: teach

PLSC 500: STATISTICS
Fall 2016

Course Personnel:

- Instructor: Alex Coppock
  alexcop@yale.edu, 87 Tweedall Street (ISPS), Office D227, Office Hours: Tuesdays 9am - 12pm. Please come and say hello. No appointment to wait.
- Teaching Assistant: Jonathan Gubser
  jonathan@yale.edu, Office 104.
- Teaching Assistant: Stephanie de la Roque
  roqueds@yale.edu, Office 104.

Course Meeting Times:

- Lecture: Tuesdays and Thursdays 1:30pm to 2:45pm in ISPS Room A001.
- Section: Fridays 10:30am to 11:20am in RKZ Room 102
- All course meetings are like a Liz Lemon party – mandatory.

Objectives: PLSC 500 is the first course in the graduate-level statistical methods sequence for political science students. It is nominally an introduction to statistics and linear regression with special emphasis on the nonparametric analysis of real-world data. We also have loftier goals. We hope to inspire:

1. An intuition for what data can and can’t tell us about the world.
2. A love of code.

3. A habit of creating beautiful, reproducible documents.

Reproducible research: preach

Individuals
- Practice;
- Expect;
- Hold accountable

Academic centers
- Prizes;
- Events;
- Communication

Academic societies
- Joint statements;
- Standards and guidelines;
- Declarations

Publishers
- Policies;
- Review process

Community enforcement
- Reproducibility projects

Demand investment in infrastructure and workforce
“We are nearing a time when it will simply be the author's choice whether to keep detailed means to results confidential with the use of traditional publication or to communicate fully [by using reproducible documents or other means].”

Thank you!

limor.peer@yale.edu
@l_peer
https://isps.yale.edu/
Yale values

Yale Mission: Improving the world for future generations through outstanding research, education, and practice. Yale educates aspiring leaders worldwide who serve all sectors of society. We carry out this mission through the free exchange of ideas in a diverse community of faculty, staff, students, and alumni.

Yale Goals: ...share more broadly Yale’s intellectual assets with the world.

Yale regards making data resulting from academic research available to the public within regulatory and legal constraints as a natural extension of its mission.

Yale regards appropriate stewardship of research data as fundamental to both high-quality research and academic integrity.

Yale supports researchers’ academic freedom which comes with the responsibility of researchers to disseminate their research findings to the scientific and academic community.

Yale supports the academic community’s standard that the principle of reproducibility is essential to the advancement of science.

https://research.yale.edu/research-data
Reproducible research: practice

Examples: How to

• Open Science Framework. **Transparency and Openness Promotion (TOP) Guidelines.** [https://cos.io/top/](https://cos.io/top/)

• **TIER Documentation Protocol** [https://www.haverford.edu/project-tier/protocol-v2](https://www.haverford.edu/project-tier/protocol-v2)

• Janz, Nicole & Figueiredo, Dalson (2017, March 13). **Workshop: The gold standard of reproducible research** [https://osf.io/2fqnw/](https://osf.io/2fqnw/)


• Stodden, Victoria et al. (2016). **Enhancing reproducibility for computational methods.** Science [http://science.sciencemag.org/content/354/6317/1240.full](http://science.sciencemag.org/content/354/6317/1240.full)


• Brandt et al. (2014) **The replication recipe: What makes for a convincing replication?** Journal of Experimental Social Psychology [https://doi.org/10.1016/j.jesp.2013.10.005](https://doi.org/10.1016/j.jesp.2013.10.005)
### Reproducible research: TOP guidelines

#### Summary of the eight standards and three levels of the TOP guidelines

Levels 1 to 3 are increasingly stringent for each standard. Level 0 offers a comparison that does not meet the standard.

<table>
<thead>
<tr>
<th>Standard</th>
<th>Level 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citation standards</td>
<td>Journal encourages citation of data, code, and materials—or says nothing.</td>
</tr>
<tr>
<td>Data transparency</td>
<td>Journal encourages data sharing—or says nothing.</td>
</tr>
<tr>
<td>Analytic methods (code) transparency</td>
<td>Journal encourages code sharing—or says nothing.</td>
</tr>
<tr>
<td>Research materials transparency</td>
<td>Journal encourages materials sharing—or says nothing.</td>
</tr>
<tr>
<td>Design and analysis transparency</td>
<td>Journal encourages design and analysis transparency—or says nothing.</td>
</tr>
<tr>
<td>Preregistration of studies</td>
<td>Journal says nothing.</td>
</tr>
<tr>
<td>Preregistration of analysis plans</td>
<td>Journal says nothing.</td>
</tr>
<tr>
<td>Replication</td>
<td>Journal encourages submission of replication studies—or says nothing.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal describes citation of data in guidelines to authors with clear rules and examples.</td>
</tr>
<tr>
<td>Article states whether data are available and, if so, where to access them.</td>
</tr>
<tr>
<td>Article states whether code is available and, if so, where to access them.</td>
</tr>
<tr>
<td>Article states whether materials are available and, if so, where to access them.</td>
</tr>
<tr>
<td>Journal articulates design transparency standards.</td>
</tr>
<tr>
<td>Journal encourages preregistration of studies and provides link in article to preregistration if it exists.</td>
</tr>
<tr>
<td>Journal encourages submission of replication studies—or says nothing.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article provides appropriate citation for data and materials used, consistent with journal’s author guidelines.</td>
</tr>
<tr>
<td>Data must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
</tr>
<tr>
<td>Code must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
</tr>
<tr>
<td>Materials must be posted to a trusted repository. Exceptions must be identified at article submission.</td>
</tr>
<tr>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td>Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.</td>
</tr>
<tr>
<td>Journal encourages submission of replication studies and conducts blind review of results.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Level 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Article is not published until appropriate citation for data and materials is provided that follows journal’s author guidelines.</td>
</tr>
<tr>
<td>Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.</td>
</tr>
<tr>
<td>Journal requires adherence to design transparency standards for review and publication.</td>
</tr>
<tr>
<td>Journal requires preregistration of studies with analysis plans and provides link in article and certification of meeting preregistration badge requirements.</td>
</tr>
<tr>
<td>Journal requires preregistration of replication studies and provides link and badge in article to meeting requirements.</td>
</tr>
</tbody>
</table>

---

Reproducible research: teach

Examples...

Training
• COS
• BITSS
• ICPSR
• Project TIER
• NIH Rigor & Reproducibility

Online
short course
• EGUGA
Full course
• Johns Hopkins
• BITSS

University course syllabi
• Open and Reproducible Methods

More...
Nicole Janz
• Solving the Reproducibility Crisis, a teaching perspective
• Bringing the Gold Standard Into the Class Room: Replication in University Teaching

King, Gary
• How to Write a Publishable Paper as a Class Project
Reproducible research: preach

Examples...

**Journals**
- Data and code sharing policies
- TOP guidelines
- AJPS third-party analysis replication and verification

**Academic societies**
- e.g., APSA DA-RT

**Academic centers**
- Prizes, e.g., BITSS

**Community enforcement**
- Reproducibility projects Psychology, Cancer
- Impact Evaluation Replication Programme
- Curate Science
- The XPhi Replicability Project

**Repositories**
- Curating for Reproducibility