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

Yale

Themes for this talk

How to think about reproducible research practices

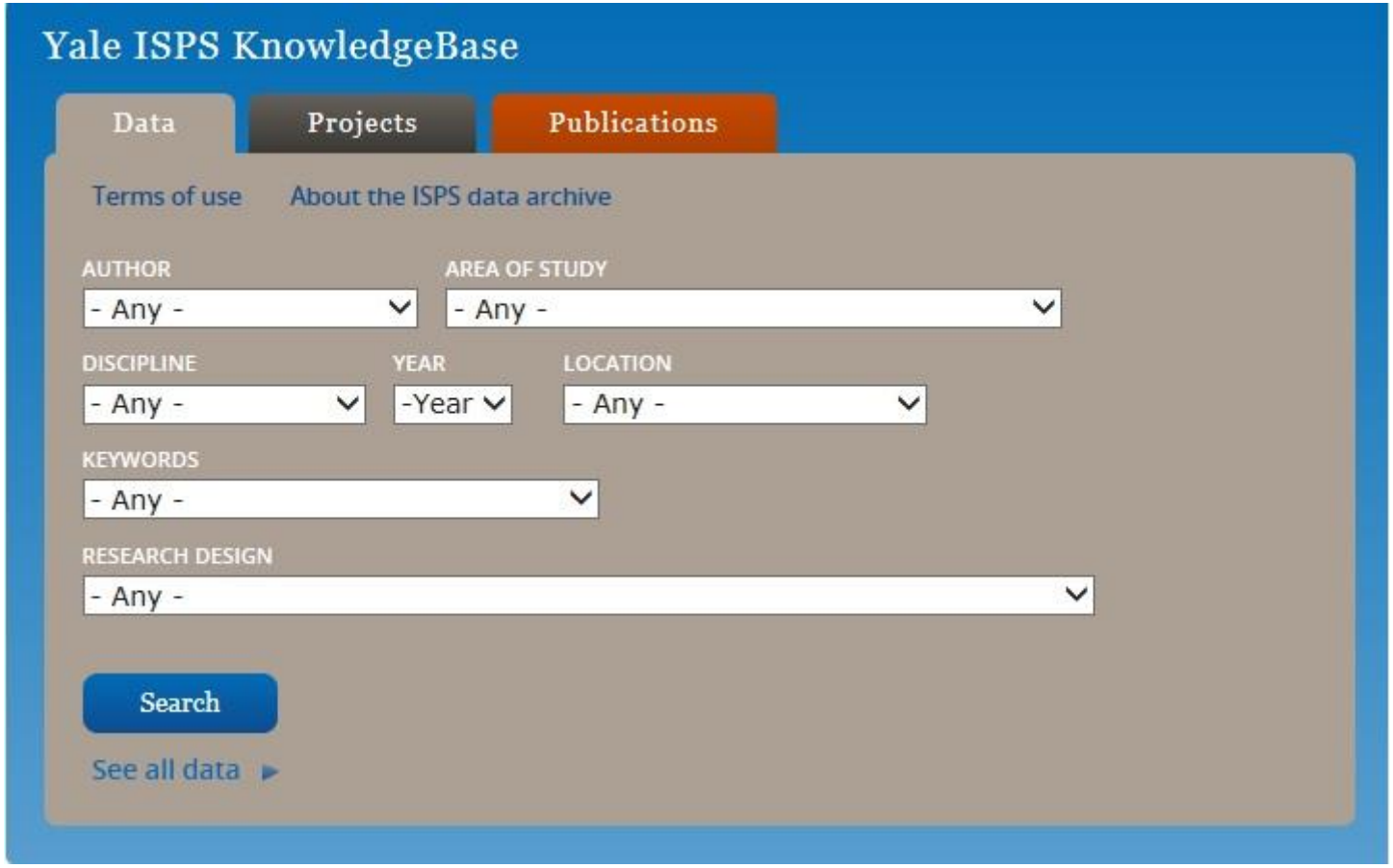
Practice	Teach	Preach
Transparency	Quality	Independence

Yale

ISPS Data Archive

Yale ISPS Institution for Social and Policy Studies

ADVANCING RESEARCH • SHAPING POLICY • DEVELOPING LEADERS



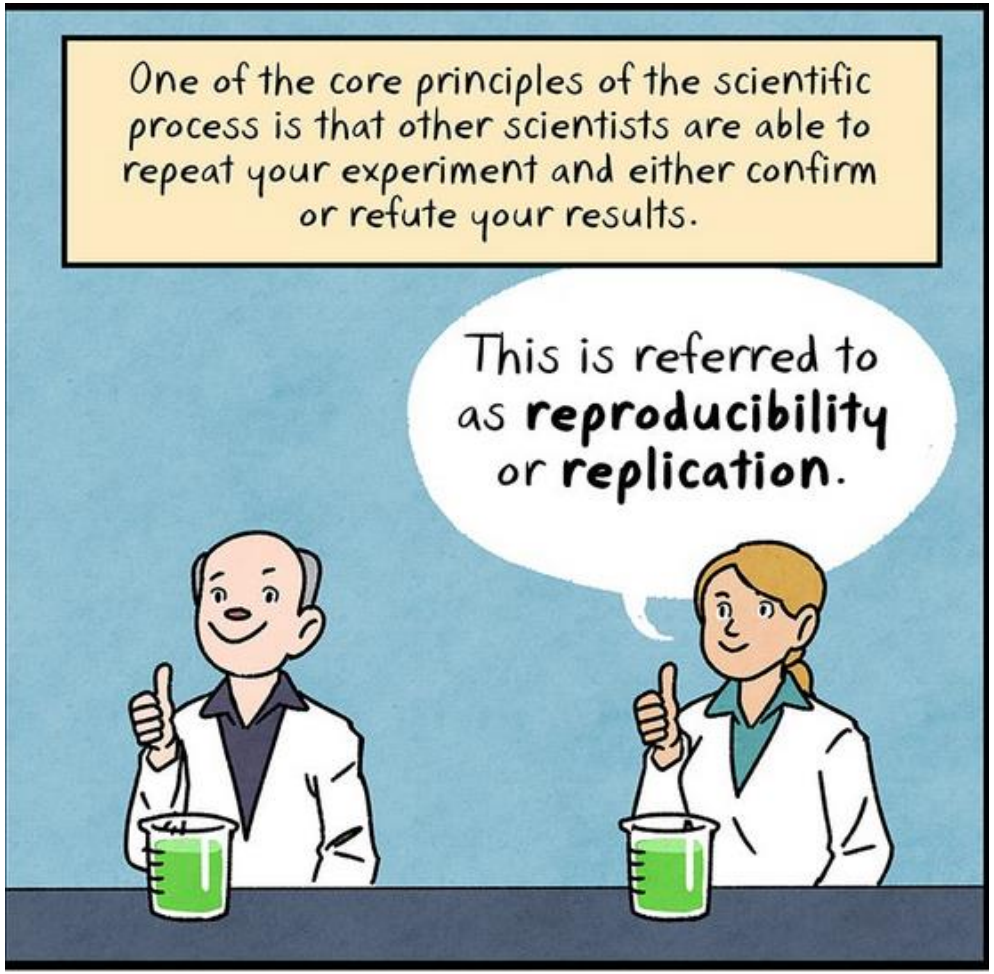
http://isps.yale.edu/research/data

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**.

Yale

Science advances

...when scientific claims are subjected to scrutiny



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

Yale

"Reproducibility"

computational reproducibility

REPLICABILITY direct replication

empirical reproducibility VALIDATION

REPEATABILITY conceptual replication

methodological reproducibility

VERIFICATION **REPRODUCIBILITY**

statistical reproducibility

Yale

Transparency

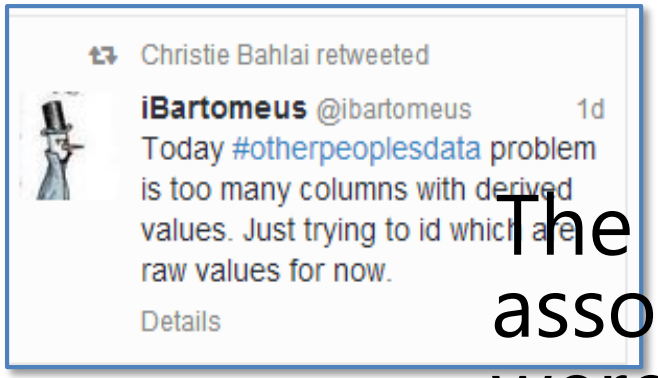
Illuminating the black box:

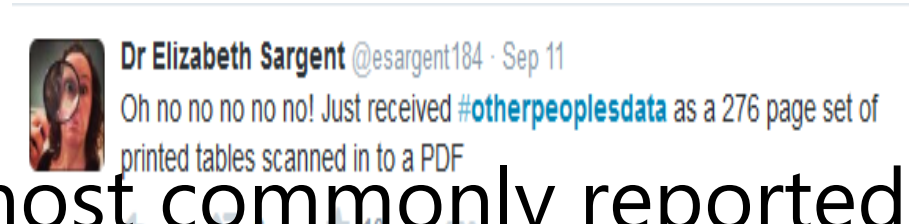
"Transparency requires making visible both the empirical foundation and the logic of inquiry of research."

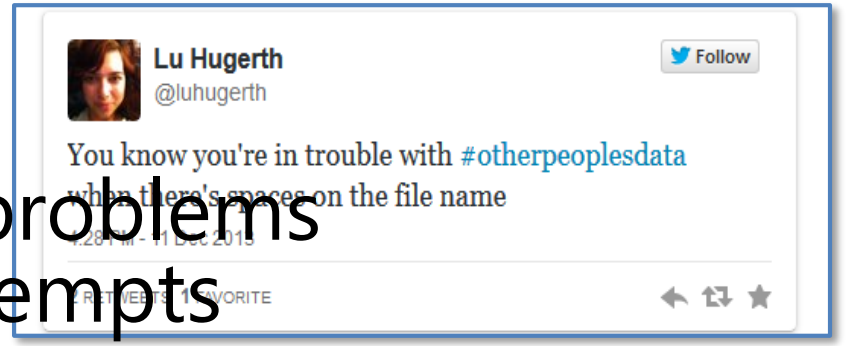
Data Access and Research Transparency (DA-RT): *A Joint Statement by Political Science Journal Editors*
https://www.dartstatement.org/2014-journal-editors-statement-jets

Yale

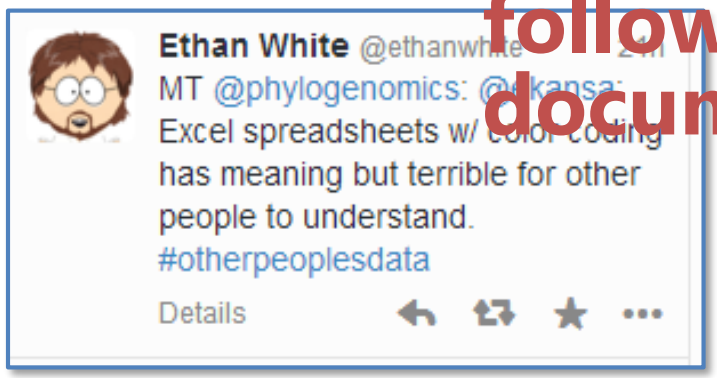
Data sharing and reuse (#otherpeoplesdata)

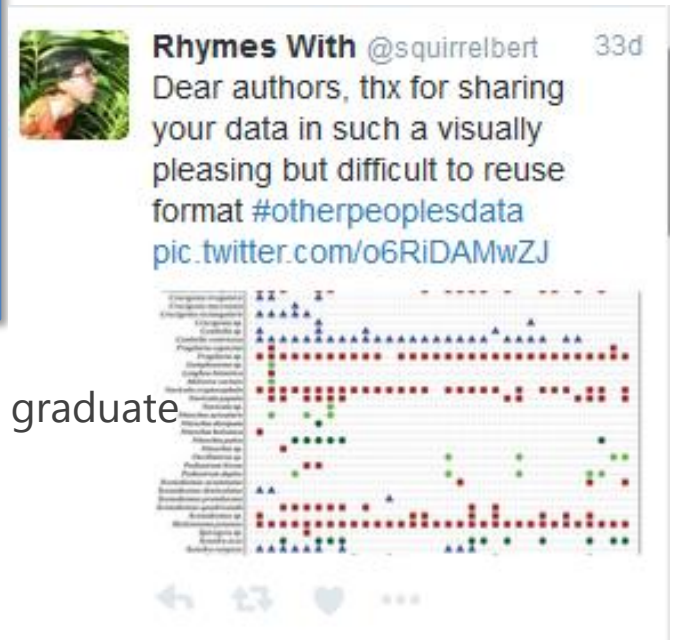


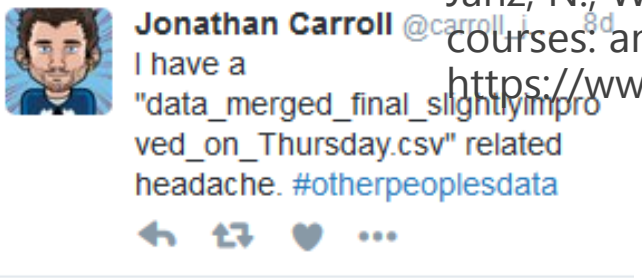


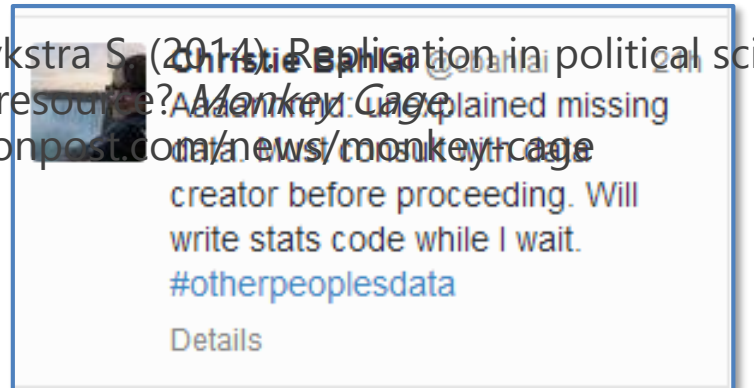


The most commonly reported problems associated with [replication] attempts were the lack of… data and code, **followed by insufficient documentation.**









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

Yale

Quality

"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**…"

Peer, L., Green, A., & Stephenson, E. (2014). Committing to data quality review. *International Journal of Digital Curation*, *9*(1). http://doi.org/10.2218/ijdc.v9i1.317

Yale

Independence

"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**."

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452. http://doi.org/10.2307/420301

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

Stodden, V., McNutt, M., Bailey, D. H., Deelman, E., Gil, Y., Hanson, B., … Taufer, M. (2016). Enhancing reproducibility for computational methods. *Science*, *354*(6317), 1240–1241. https://doi.org/10.1126/science.aah6168

Yale

ISPS Data Archive: first re-user

DATA QUALITY REVIEW

Yale

ISPS Data Archive: first re-user

DATA QUALITY REVIEW

Building trust and expecting accountability

Yale

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

Yale

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

Yale

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

Yale

Curating for reproducibility

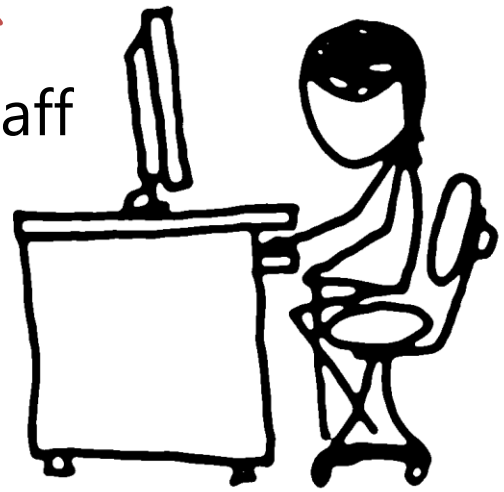
- Convert absolute file paths to relative file paths
- Check code for presence of non-executable comments that document analysis processes
- Identify packages required to execute code
- Execute code to ensure code is error-free
- Compare code output to findings presented in article

Yale

Prying information from researchers

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

Archive staff



http://xkcd.com/662/ Creative Commons Attribution-Noncommercial

"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!"

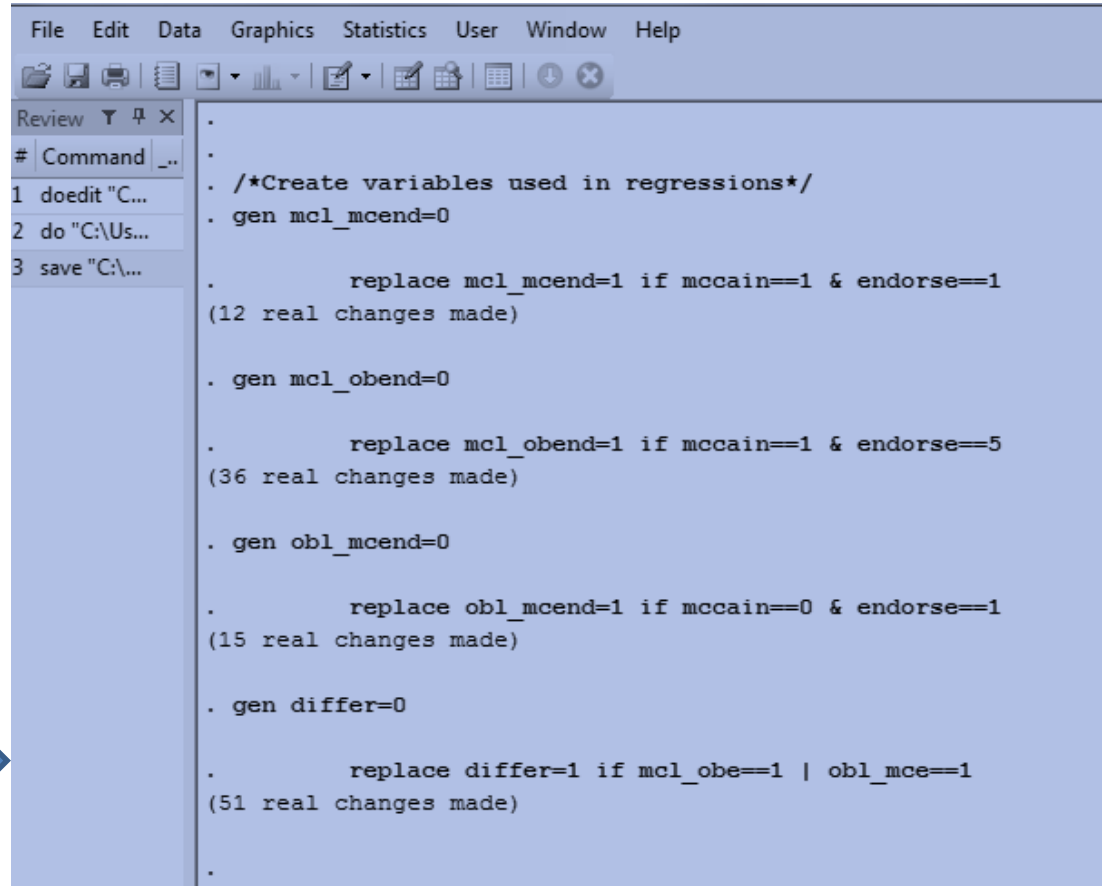
Researcher

Yale

Does the Code Fully Execute?

Example 1:

```
/*Create variables used in regressions*/
gen mcl_mcend=0
    replace mcl_mcend=1 if mccain==1 & endorse==1
gen mcl_obend=0
    replace mcl_obend=1 if mccain==1 & endorse==5
gen obl_mcend=0
    replace obl_mcend=1 if mccain==0 & endorse==1
gen differ=0
    replace differ=1 if mcl_obe==1 | obl_mce==1
```



Yale

data + code = reported results?

Example 2:

```
. /*Table 2*/
.         probit interest_in_letter mccain

Iteration 0:   log likelihood = -57.305692
Iteration 1:   log likelihood =  -56.19505
Iteration 2:   log likelihood = -56.193827
Iteration 3:   log likelihood = -56.193827

Probit regression                                 Number of obs   =        100
                                                  LR chi2(1)      =       2.22
                                                  Prob > chi2     =     0.1359
Log likelihood = -56.193827                       Pseudo R2       =     0.0194

------------------------------------------------------------------------------
interest_in_letter |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+----------------------------------------------------------------
           mccain |   .4061989   .2736818     1.48   0.138    -.1302077    .9426054
            _cons |  -.8557124     .20097    -4.26   0.000    -1.249606   -.4618184
------------------------------------------------------------------------------

.                mfx

Marginal effects after probit
      y  = Pr(interest_in_letter) (predict)
         =  .25569496

------------------------------------------------------------------------------
variable |      dy/dx    Std. Err.      z    P>|z|  [    95% C.I.   ]       X
---------+--------------------------------------------------------------------
  mccain*|   .1304522     .08706     1.50   0.134  -.040174  .301079       .49
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1
```

(table)

Table 2. The Effect of Candidate Support on Interest in Letter to the Editor

Dependent Variable = Interested in Publishing Letter Independent Variable	Coefficient (Standard Error) [Change in Probability]	Coefficient (Standard Error) [Change in Probability]
Pro-McCain letter	0.41 (0.27) [13.0%]	0.58** (0.29) [16.3%]
Circulation (in units of 10,000)		−0.028** (0.013) [−16.4%]
Unemployment rate in metro area		−0.009 (0.13) [−0.3%]
Intercept	−0.86** (0.20)	−0.23 (0.75)
N	100	100
Pseudo R^2	.02	.09
Log-likelihood	−56.2	−51.9

Note: The dependent variable is a binary variable that takes the value of 1 if the newspaper either tried to contact the alias for verification purposes or if it published the letter and 0 otherwise. Standard errors are given in parentheses. The estimated predicted probabilities are given in brackets. For the binary variables, the predicted probabilities report the change in the predicted probability when the value of the variable goes from 0 to 1 while holding other variables constant. For the continuous variables, the predicted probabilities report the change in predicted probability when increasing the value of that variable from the mean value to one standard deviation above the mean.
*p < .10. **p < .05.

Yale

Common replication problems

- Insufficient documentation
- Missing variables
- Deviations in number of observations
- Unavailable software extensions
- Omitted code
- Incompatible datasets

Yale

Archives curating for reproducibility

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.

CURATING for REPRODUCIBILITY

https://cure.web.unc.edu/

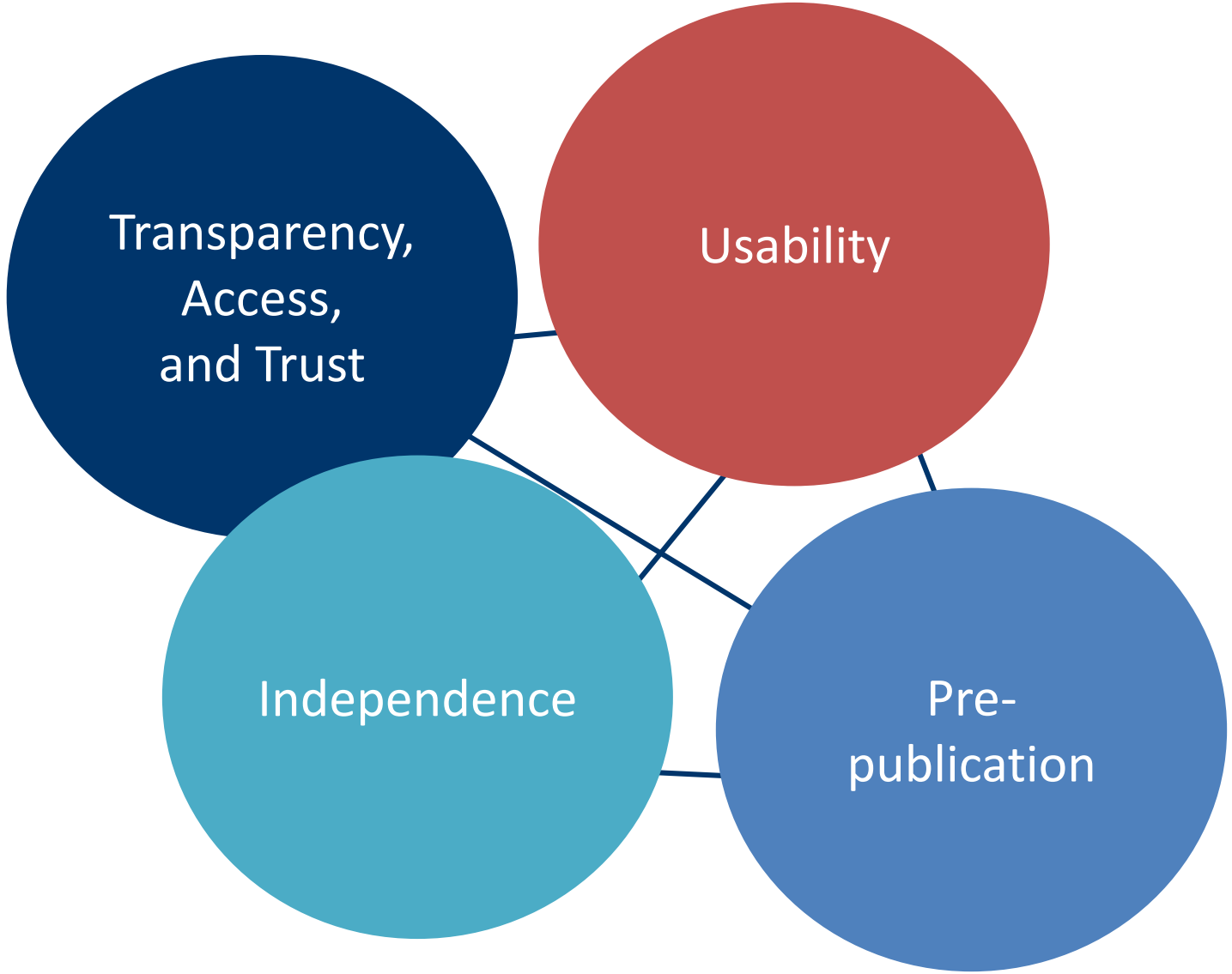
THE ODUM INSTITUTE
FOR RESEARCH IN SOCIAL SCIENCE

CISER

Yale ISPS

Yale

Curating for reproducibility



- Establish Standards
- Share Practices
- Promote Data Quality Review

CURATING for REPRODUCIBILITY

https://cure.web.unc.edu/

Yale

Reproducible research: practice

Recommendations for statistical studies

1. Do all data preparation and analysis in code.
2. Adopt best practice for coding.
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.

Dafoe, A. (2014). Science Deserves Better: The Imperative to Share Complete Replication Files. *PS: Political Science & Politics*, 47(1), 60-66. doi:10.1017/S104909651300173X

Yale

Reproducible research: teach

PLSC 500: STATISTICS

Fall 2016

Course Personnel:

- Instructor: Alex Coppock alex.coppock@yale.edu 87 Trumbull Street (ISPS) Office D322 Office Hours: Tuesdays 9am - 12pm. Ple … wait.
- Teaching Assistant: Jonath… …pm-5pm, RKZ 104.
- Teaching Assistant: Steph… …30pm-7:30pm, RKZ 104.

in the disciplines; in data science

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: PSLC 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.

http://alexandercoppock.com/papers/PLSC_500_syllabus.pdf

Yale

Reproducible research: preach

Individuals
Practice;
Expect;
Hold accountable

Academic societies
Joint statements;
Standards and guidelines;
Declarations

Academic centers
Prizes;
Events;
Communication

Publishers
Policies;
Review process

Community enforcement
Reproducibility projects

Archives & repositories
Review process

Demand investment in infrastructure and workforce

Yale

"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]."

Claerbout, J.F. and Karrenbach, M. (1992). *Electronic Documents Give Reproducible Research a New Meaning*. SEG Expanded Abstracts 11, 601. https://library.seg.org/doi/pdf/10.1190/1.1822162

Yale

Thank you!

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

Yale

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.

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

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**.

Yale

Reproducible research: practice

Examples: How to

- Open Science Framework. **Transparency and Openness Promotion (TOP) Guidelines**. https://cos.io/top/
- **TIER Documentation Protocol** 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/
- Christensen, Garret (2016). **Manual of best practices in transparent social science research** https://github.com/garretchristensen/BestPracticesManual
- Stodden, Victoria et al. (2016). **Enhancing reproducibility for computational methods**. Science http://science.sciencemag.org/content/354/6317/1240.full
- Markowetz, Florian (2015), **Five selfish reasons to work reproducibly**. Genome Biology https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7
- 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

Yale

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.

	LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3
Citation standards	Journal encourages citation of data, code, and materials—or says nothing.	Journal describes citation of data in guidelines to authors with clear rules and examples.	Article provides appropriate citation for data and materials used, consistent with journal's author guidelines.	Article is not published until appropriate citation for data and materials is provided that follows journal's author guidelines.
Data transparency	Journal encourages data sharing—or says nothing.	Article states whether data are available and, if so, where to access them.	Data must be posted to a trusted repository. Exceptions must be identified at article submission.	Data must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Analytic methods (code) transparency	Journal encourages code sharing—or says nothing.	Article states whether code is available and, if so, where to access them.	Code must be posted to a trusted repository. Exceptions must be identified at article submission.	Code must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Research materials transparency	Journal encourages materials sharing—or says nothing	Article states whether materials are available and, if so, where to access them.	Materials must be posted to a trusted repository. Exceptions must be identified at article submission.	Materials must be posted to a trusted repository, and reported analyses will be reproduced independently before publication.
Design and analysis transparency	Journal encourages design and analysis transparency or says nothing.	Journal articulates design transparency standards.	Journal requires adherence to design transparency standards for review and publication.	Journal requires and enforces adherence to design transparency standards for review and publication.
Preregistration of studies	Journal says nothing.	Journal encourages preregistration of studies and provides link in article to preregistration if it exists.	Journal encourages preregistration of studies and provides link in article and certification of meeting preregistration badge requirements.	Journal requires preregistration of studies and provides link and badge in article to meeting requirements.
Preregistration of analysis plans	Journal says nothing.	Journal encourages preanalysis plans and provides link in article to registered analysis plan if it exists.	Journal encourages preanalysis plans and provides link in article and certification of meeting registered analysis plan badge requirements.	Journal requires preregistration of studies with analysis plans and provides link and badge in article to meeting requirements.
Replication	Journal discourages submission of replication studies—or says nothing.	Journal encourages submission of replication studies.	Journal encourages submission of replication studies and conducts blind review of results.	Journal uses Registered Reports as a submission option for replication studies with peer review before observing the study outcomes.

B. A. Nosek, et al., Promoting an open research culture, *Science* 26 Jun 2015 : 348(6242): 1422-1425. DOI: 10.1126/science.aab2374

Yale

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

Yale

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

Yale