Reproducible Research: What, Why, and How?

Thomas J. Leeper

Department of Political Science and Government
Aarhus University

October 28, 2014

- Reinhart and Rogoff

- Reinhart and Rogoff
- Psychology's "replication crisis"

- Reinhart and Rogoff
- Psychology's "replication crisis"
- "Most published research findings are false"

- Reinhart and Rogoff
- Psychology's "replication crisis"
- "Most published research findings are false"
- Diedrick Stapel

Why reproducible research?

- External reasons
- Internal reasons

External Reasons

External Reasons

- Philosophical perspective

External Reasons

- Philosophical perspective
- Journal requirements

External Reasons

- Philosophical perspective
- Journal requirements
- Funding agency requirements

External Reasons

- Philosophical perspective
- Journal requirements
- Funding agency requirements
- The coming revolution

Internal Reasons

Internal Reasons

- Confidence in your own work

Internal Reasons

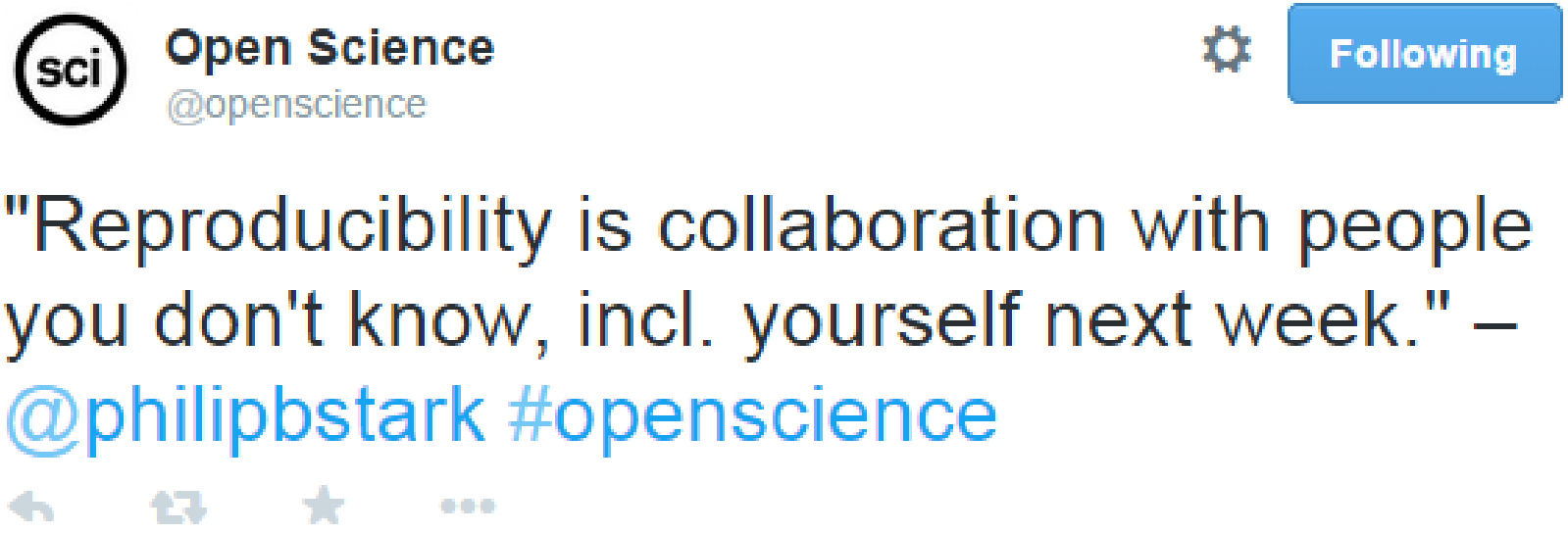
- Confidence in your own work
- Easier workflow

Internal Reasons

- Confidence in your own work
- Easier workflow
- Easier collaboration

So what does that mean?

So what does that mean?



So what does that mean?

1. Do it for *yourself* first!

2. Do it for *science* second.

Why is research still irreproducible?

Why is research still irreproducible?

Barriers to Data and Code Sharing in Computational Science

Survey of Machine Learning Community, NIPS (Stodden, 2010):

Code		Data
77%	Time to document and clean up	54%
52%	Dealing with questions from users	34%
44%	Not receiving attribution	42%
40%	Possibility of patents	-
34%	Legal Barriers (ie. copyright)	41%
-	Time to verify release with admin	38%
30%	Potential loss of future publications	35%
30%	Competitors may get an advantage	33%
20%	Web/disk space limitations	29%

Why is research still irreproducible?

Why is research still irreproducible?

1. Technology

Why is research still irreproducible?

1. Technology

Why is research still irreproducible?

1. Technology
2. Individual actions

Why is research still irreproducible?

1. Technology
2. Individual actions
3. Collective behavior and norms

So what is reproducible research?

So what is reproducible research?

- Evolving standards and technology

So what is reproducible research?

- Evolving standards and technology
- Discipline-specific meaning

American Association for Public Opinion Research[1]

Researchers must publish:

1. Research sponsor
2. Question wordings
3. Population, sampling frame, and sampling desng
4. Sample sizes and margins of error
5. Dates of data collection

[1] "Disclosure Standards"

American Psychological Assoc.[2]

- "After research results are published, **psychologists do not withhold the data on which their conclusions are based from other competent professionals who seek to verify the substantive claims through reanalysis and who intend to use such data only for that purpose**, provided that the confidentiality of the participants can be protected and unless legal rights concerning proprietary data preclude their release..." (8.14a)

[2] "Ethical Principles of Psychologists and Code of Conduct"

Assoc. for Psychological Science[3]

1. Sample sizes and exclusion criteria
2. Report all manipulations used
3. Report all outcomes analyzed

[3] "Submission Guidelines"

American Anthropological Association[4]

- "Anthropological researchers should seriously consider all reasonable requests for access to their data and other research materials for purposes of research. They should also make every effort to insure preservation of their fieldwork data for use by posterity."

[4] "Code of Ethics"

CONSORT Group[5]

- "The checklist includes the 25 items selected because empirical evidence indicates that not reporting the information is associated with biased estimates of treatment effect, or because the information is essential to judge the reliability or relevance of the findings."
- No requirement for open data or analyses

[5] "CONSORT Statement"

American Political Science Assoc.[6]

- "When statements that are challenged are based on reproducible data authors are obliged to facilitate replication." (5.5)
- "Researchers making evidence-based knowledge claims should reference the data they used to make those claims. If these are data they themselves generated or collected, researchers should provide access to those data or explain why they cannot." (5.6)
- "Production transparency" (6.2)
- "Analytic transparency" (6.3)

[6] "A Guide to Professional Ethics in Political Science"

European Research Council[7]

- "The European Research Council supports the basic principle of Open Access to research data. It therefore recommends to all its funded researchers that they follow best practice by retaining files of all the research data they have used during the course of their work, and that they be prepared to share this data with other researchers whenever it is not bound by copyright restrictions, by confidentiality agreements, or by contractual clauses."

[7] "Open Access Guidelines for researchers funded by the ERC"

PLoS[8]

- "Publication is conditional upon the agreement of the authors to make freely available any materials and information described in their publication that may be reasonably requested by others."

- Software created for use in publications must be open source

[8] "Editorial and Publishing Policies"

So what is reproducible research?

- Evolving standards and technology
- Discipline-specific meaning

So what is reproducible research?

- Evolving standards and technology
- Discipline-specific meaning
- Hard to define

*Ir*reproducibility

*Ir*reproducibility

- Fabrication

*Ir*reproducibility

- Fabrication
- Human error

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware
- Analysis unavailable

*Ir*reproducibility

- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware
- Analysis unavailable
- “Available from the author”

*Ir*reproducibility

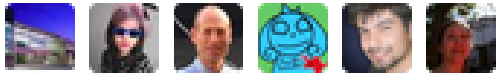
- Fabrication
- Human error
- Lack of methodological transparency
- Ambiguous data citations
- Proprietary data and file formats
- Unavailable data
- Analysis uses proprietary software/hardware
- Analysis unavailable
- “Available from the author (now deceased)”

Kaitlin Thaney
@kaythaney

Following

"'Reproducible research' is a redundant term. 'Irreproducible research' just used to be known as 'bullshit'." - @fperez_org
::slow clap::

RETWEETS 122 FAVORITES 61



6:11 PM - 8 May 2014

Distinguish from other concepts

Distinguish from other concepts

- *Reproducible* versus *Replicable*

Distinguish from other concepts

- *Reproducible* versus *Replicable*
- *Reproducible* versus *Automated*

Distinguish from other concepts

- *Reproducible* versus *Replicable*
- *Reproducible* versus *Automated*
- *Reproducible* versus *True*

Arrive at a definition

Stanford University's David Donoho:

"An article about computational science in a scientific publication is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures."

Reproducible research enumerates a complete set of physical actions needed to transforms transparent inputs into outputs.

What makes up the ideal reproducible research product?

Past

- Data and method description
- Closed data and analysis
- Use of proprietary software
- Paywalled publications

Present

- Detailed or full protocols
- Data and analysis sharing (on request)
- Mix of proprietary and open software
- “Green” open access

Future

- Study preregistration and “outcome-blind” review
- Open lab notebooks
- Persistent, archived, open-licensed data
- Open source software
- Open peer review
- Open access publication
- Literate, reproducible output

How do you make your work more reproducible?

How do you make your work more reproducible?

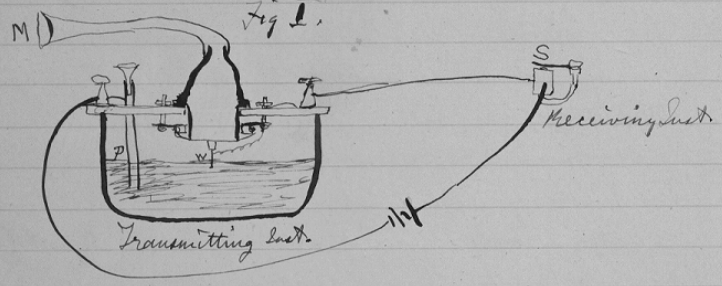
Always think about your future self!


DEAR FUTURE SELF,
#
YOU'RE LOOKING AT THIS FILE BECAUSE
THE PARSE FUNCTION FINALLY BROKE.
#
IT'S NOT FIXABLE. YOU HAVE TO REWRITE IT.
SINCERELY, PAST SELF
DEAR PAST SELF, IT'S KINDA CREEPY HOW YOU DO THAT.
ALSO, IT'S PROBABLY AT LEAST
2013. DID YOU EVER TAKE
THAT TRIP TO ICELAND?
STOP JUDGING ME!

(1) Write Everything Down

March 10th 1876

Fig 1.



1. The improved instrument shown in Fig. I was constructed this morning and tried this evening. P is a brass pipe and W the platinum wire M the mouth piece — and S the armature of the Receiving Instrument.

Mr. Watson was stationed in one room with the Receiving Instrument. He pressed one ear closely against S and closed his other ear with his hand. The Transmitting Instrument was placed in another room and the doors of both rooms were closed.

I then shouted into M the following sentence: "Mr. Watson — Come here — I want to see you". To my delight he came and declared that he had heard and understood what I said.

I asked him to repeat the words — He said He answered "You said 'Mr. Watson — come here — I want to see you'."

We then changed places and I listened at S while Mr. Watson read a few passages from a book into the mouth piece M. It was certainly the case that articulate sounds proceeded from S. The effect was loud but indistinct and muffled.

If I had read beforehand the passage given by Mr. Watson I should have recognized every word. As it was I could not make out the sense — but an occasional word here and there was quite distinct. I made out "to" and "out" and "further"; and finally the sentence "Mr. Bell Do you understand what I say? DO-you-understand-what-I-say" came quite clearly and intelligibly. No sound was audible when the armature S was removed.

(1) Write Everything Down

(1) Write Everything Down

1. Mark up your analysis files

(1) Write Everything Down

1. Mark up your analysis files
2. Write (and maintain) your research protocols

(1) Write Everything Down

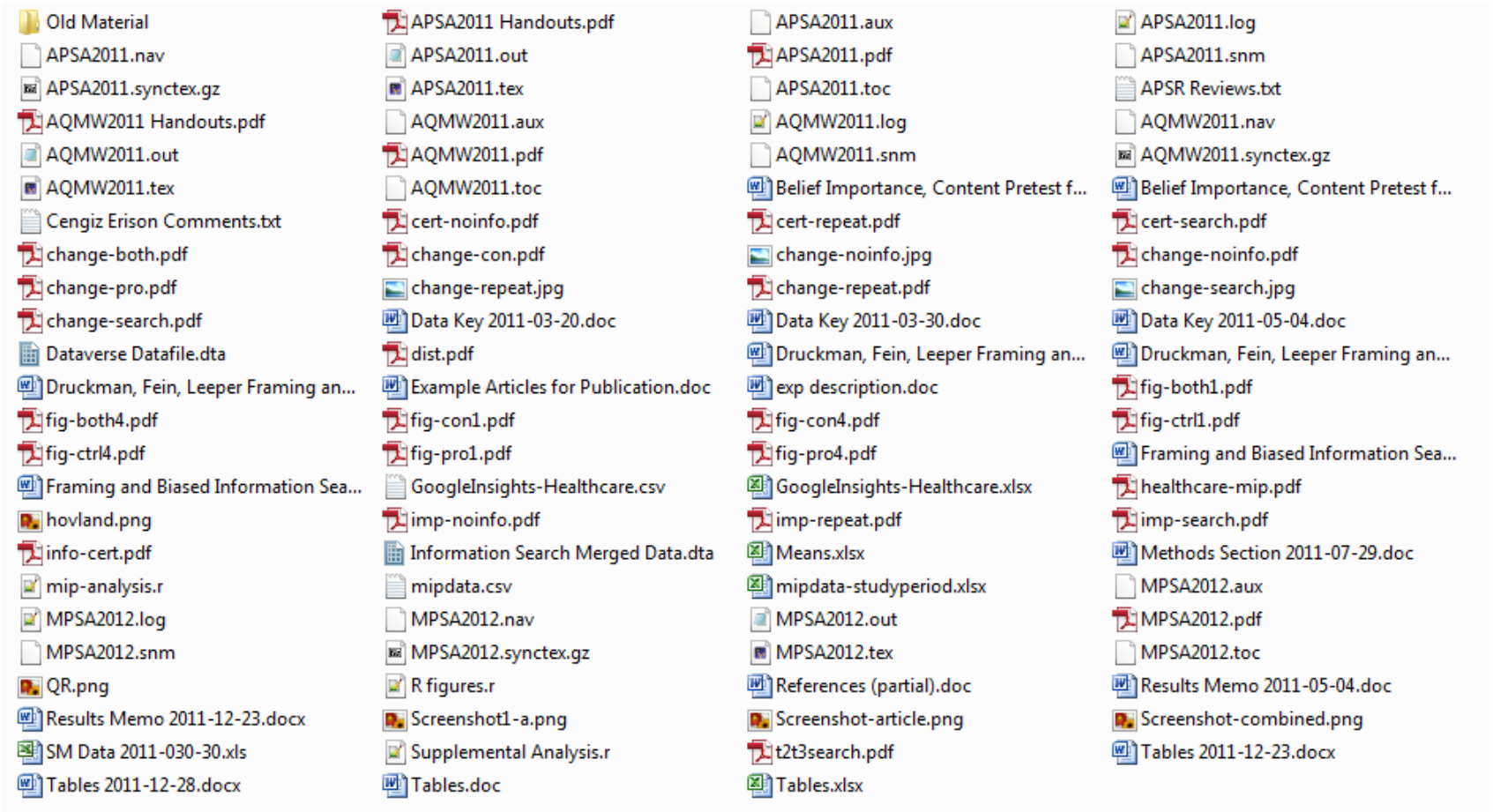
1. Mark up your analysis files
2. Write (and maintain) your research protocols
3. Keep codebooks, questionnaires, and stimulus materials

(1) Write Everything Down

1. Mark up your analysis files
2. Write (and maintain) your research protocols
3. Keep codebooks, questionnaires, and stimulus materials
4. Try version control

(2) Get Organized

My dissertation folder



(2) Get Organized

1. Use a folder structure than can be shared

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
  - RawData.csv
  - CleanData.csv
  - Codebook.txt
- Analysis
- Figures
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
  - GatherAndMerge.R
  - DataCleaning.R
  - Descriptives.R
  - Regression.R
  - Figures.R
- Figures
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
  - Distributions.png
  - MarginalEffects.png
  - PredictedValues.png
- Tables
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
  - Descriptives.tex
  - Regression.tex
  - MarginalEffects.tex
- Paper
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
  - Draft.tex
  - References.bib
- Presentation
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
- Presentation
  - Slides.tex
- Materials
- README

Project Directory Structure

- Data
- Analysis
- Figures
- Tables
- Paper
- Presentation
- Materials
  - Protocol.tex
  - StimulusMaterials.pdf
  - Questionnaire.txt
- README

Cataloging Information | **DATA & ANALYSIS** | Comments (0) | Versions

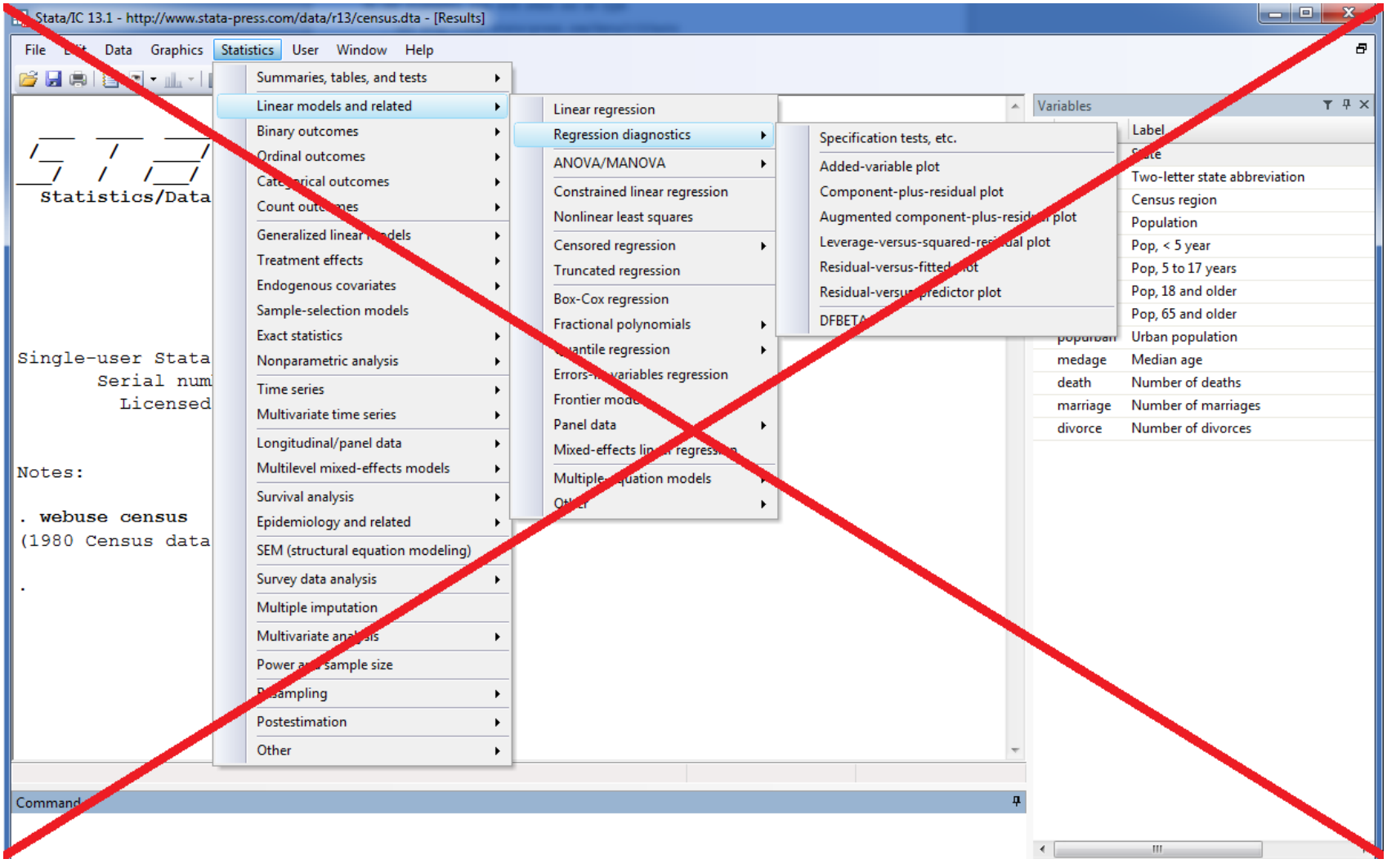
Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Select all files | Download Selected Files | Total Number of Files: 10 | Total Downloads: 48 | Downloads of Files in This Version: 44

Codebook		
Study 1 (Lab) Key.docx application/octet-stream - 21 KB - 7 downloads	Download	Study 1 data key
Study 2 (Exit Poll) Key.docx application/octet-stream - 16 KB - 3 downloads	Download	Study 2 data key
Data		
Study 1 (Lab) Data.tab Tab Delimited - 54 KB - 18 downloads + analyses TABULAR DATA | 647 Cases | 25 Variables	Download as... Access Analysis + Subsetting	Study 1 Data View Data Citation [+]
Study 2 (Exit Poll) Data.tab Tab Delimited - 42 KB - 4 downloads + analyses TABULAR DATA | 765 Cases | 22 Variables	Download as... Access Analysis + Subsetting	Study 2 Data View Data Citation [+]
Experimental Materials		
Con Pretreatment Articles.doc MS Word - 28 KB - 1 download	Download	Con Pretreatment Articles
Control Pretreatment Articles.doc MS Word - 23 KB - 1 download	Download	Control Pretreatment Articles
Pro Pretreatment Articles.doc MS Word - 28 KB - 2 downloads	Download	Pro Pretreatment Articles
Questionnaire.doc MS Word - 97 KB - 4 downloads	Download	Questionnaire
Replication code		
Study 1 (Lab).do Stata Syntax - 5 KB - 3 downloads	Download	Study 1 Replication code
Study 2 (Exit Poll).do Stata Syntax - 2 KB - 1 download	Download	Study 2 Replication code

(2) Get Organized

1. Use a folder structure than can be shared
2. Never use absolute file paths in code


Stata/IC 13.1 - http://www.stata-press.com/data/r13/census.dta - [Results]
File Data Graphics Statistics User Window Help
Statistics/Data
Single-user Stata
Serial num
Licensed
Notes:
. webuse census
(1980 Census data
.
Summaries, tables, and tests
Linear models and related
Binary outcomes
Ordinal outcomes
Categorical outcomes
Count outcomes
Generalized linear models
Treatment effects
Endogenous covariates
Sample-selection models
Exact statistics
Nonparametric analysis
Time series
Multivariate time series
Longitudinal/panel data
Multilevel mixed-effects models
Survival analysis
Epidemiology and related
SEM (structural equation modeling)
Survey data analysis
Multiple imputation
Multivariate analysis
Power and sample size
Resampling
Postestimation
Other
Linear regression
Regression diagnostics
ANOVA/MANOVA
Constrained linear regression
Nonlinear least squares
Censored regression
Truncated regression
Box-Cox regression
Fractional polynomials
Quantile regression
Errors-in-variables regression
Frontier models
Panel data
Mixed-effects linear regression
Multiple-equation models
Other
Specification tests, etc.
Added-variable plot
Component-plus-residual plot
Augmented component-plus-residual plot
Leverage-versus-squared-residual plot
Residual-versus-fitted plot
Residual-versus-predictor plot
DFBETA
Variables
Label
Two-letter state abbreviation
Census region
Population
Pop, < 5 year
Pop, 5 to 17 years
Pop, 18 and older
Pop, 65 and older
Urban population
medage Median age
death Number of deaths
marriage Number of marriages
divorce Number of divorces
Command

(3) Abandon Point-and-Click

1. Don't clean data by hand
2. Use scripts rather than menus for graphics
3. Record your OS and software (and their versions)

(4) Publicly Archive Your Research

1. Use persistent, public archives, not your website or "on request"

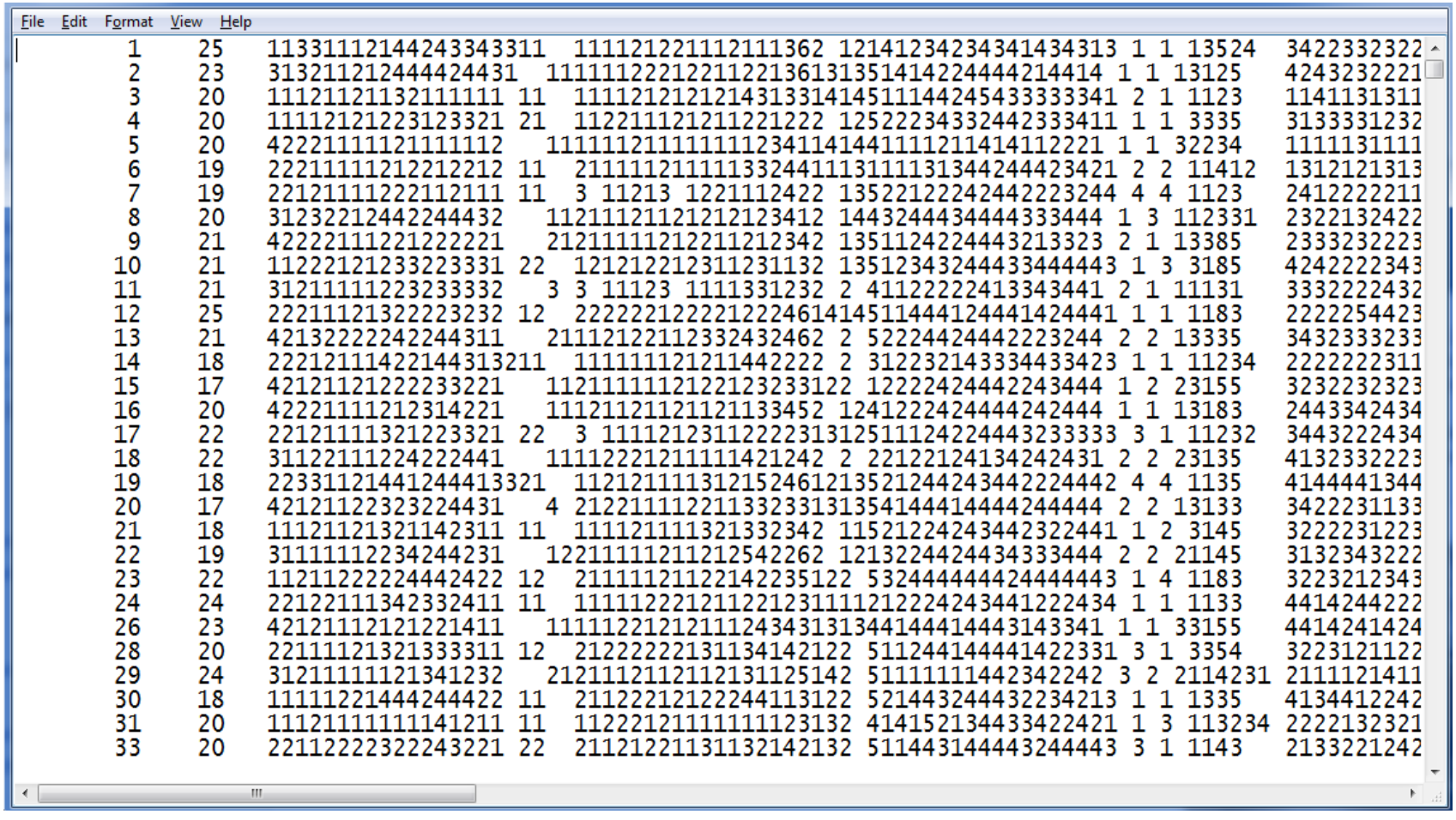
Where do you archive your research?

- Dataverse Network
- Data Dryad
- figshare

(4) Publicly Archive Your Research

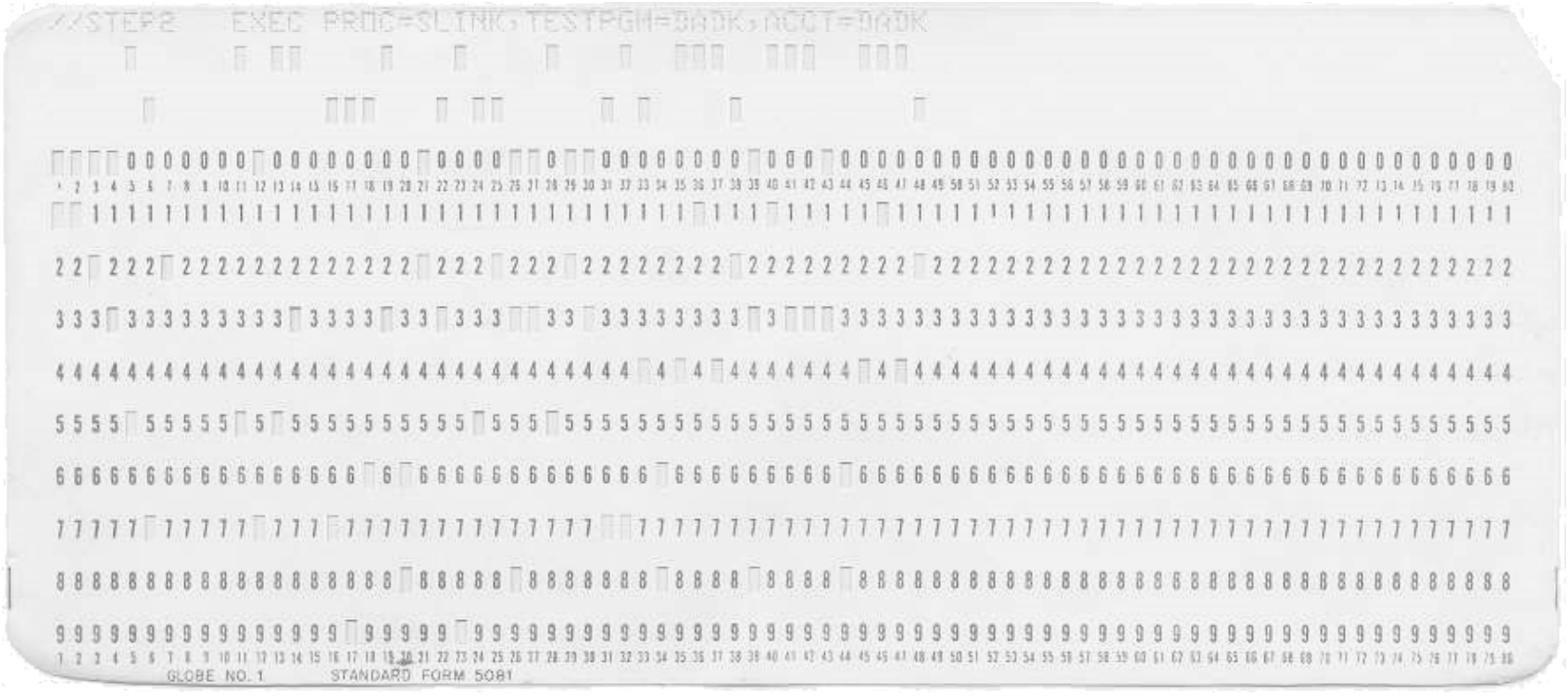
1. Use persistent, public archives, not your website or "on request"

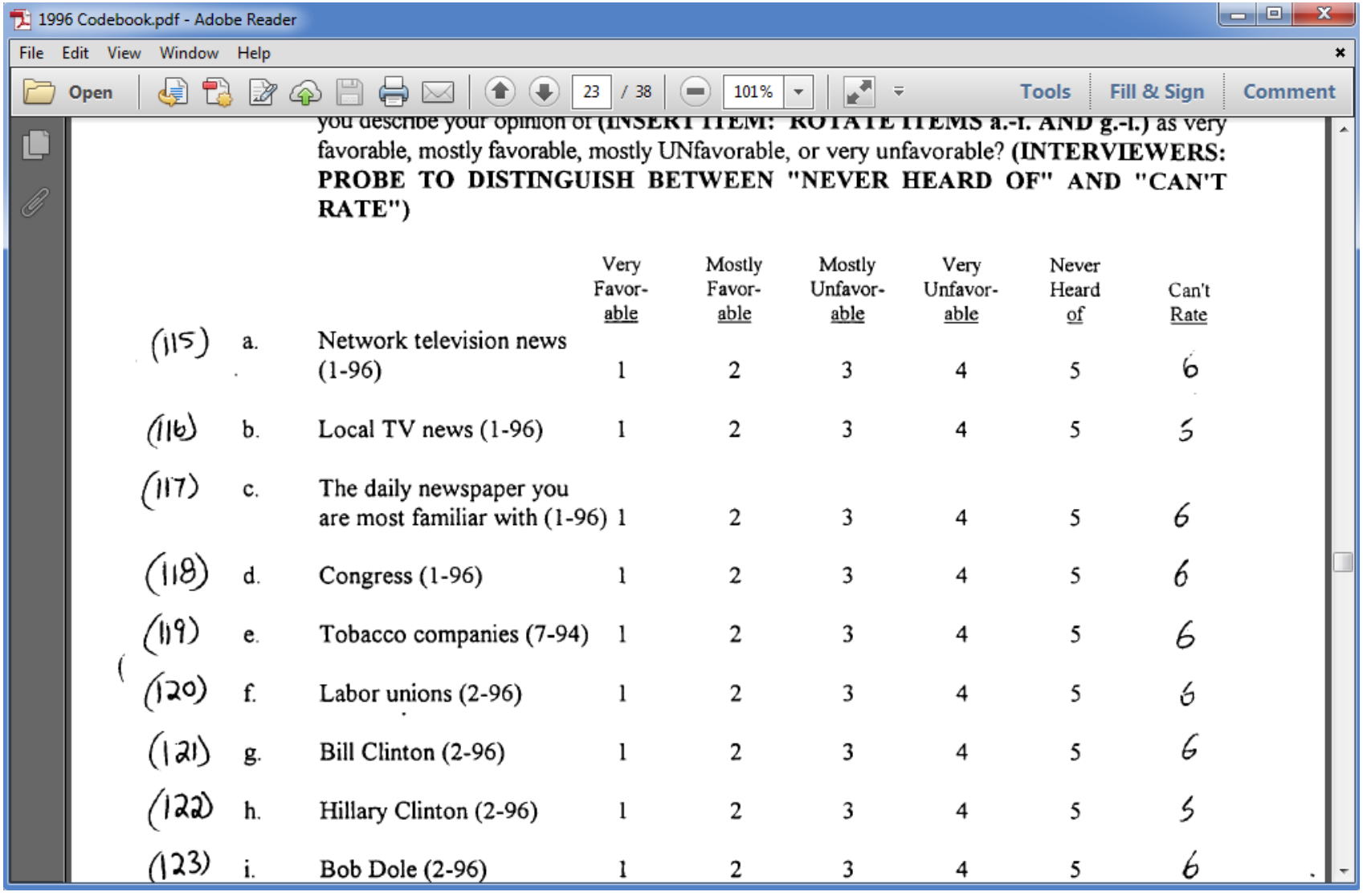
2. Use Simple, Structured, and Semantic open file formats



File Edit Format View Help

1	25	113311121442433433 11	111121221112111362	12141234234341434313	1 1	13524	3422332322
2	23	313211212444424431	11111122212211221361313514142244444214414	1 1	13125	4243232221	
3	20	111211211321111111 11	11112121212143133141451114424543333334 1	2 1	1123	1141131311	
4	20	111121212231233 21	112211121211221222	125222343324423334 11	1	3335	3133331232
5	20	422211111211111112	11111121111111123411414411112114141122 21	1 1	32234	1111131111	
6	19	222111112122122 11	211111211111332441113111131344244423421	2 2	11412	1312121313	
7	19	221211112221121 11	3 11213 1221112422	135221222424422232 44	1123	2412222211	
8	20	312322124422444 32	112111211212123412	14432444344443334 44	1 3	112331	2322132422
9	21	422221112212222 21	212111112122112123 42	135112422444321332 3	2 1	13385	2333232223
10	21	112222121233223331 22	121212212311231132	135123432444334444 43	1 3	3185	4242222343
11	21	312111112232333 32	3 3 11123 1111331232	2 411222224133434 41	2 1	11131	3332222432
12	25	222111213222232 32	12 22222212222122246141451144412444144	1 1	1183	2222254423	
13	21	421322222422443 11	21112122112332432462	2 52224424442223244	2 2	13335	3432333233
14	18	222121114221443132 11	111111121211442222	2 312232143334433423	1 1	11234	2222222311
15	17	421211212222332 21	112111111212212323	3122224244422434 44	1 2	23155	3232232323
16	20	422211112123142 21	111211211211211334	52 124122242444424244	1 1	13183	2443342434
17	22	221211113212233 21 22	3 111121231122223131251112422444323	3333 3	1	11232	3443222434
18	22	311221112242224 41	111122212111114212	42 221221241342424	31 2	2 23135	4132332223
19	18	223311214412444 13321	112121111312152461213521244243442224442	4 4	1135	4144441344	
20	17	421211223232244 31	4 212211112211332331313541444144442444	44 2	2 13133	3422231133	
21	18	111211213211423 11 11	111121111321332342	11521224243442322	441 1	2 3145	3222231223
22	19	311111123424423 1	122111112112125422	62 12132244244343	33444 2	2 21145	3132343222
23	22	112112222244424 22 12	211111211221422351	22 532444444244444	43 1	4 1183	3223212343
24	24	221221113423324 11 11	111112221211221231111212224243441	222434 1	1 1133	4414244222	
26	23	421211121212214 11	111112212121112434313134414441444	3143341 1	1 33155	4414241424	
28	20	221111213213333 11 12	212222221311341421	22 51124414444142	2331 3	1 3354	3223121122
29	24	312111111213412 32	212111211211213112	5142 5111111144234	2242 3	2 2114231	2111121411
30	18	111112214442444 22 11	211222121222441131	22 521443244432234	213 1	1 1335	4134412242
31	20	111211111111412 11 11	112221211111111231	32 414152134433422	421 1	3 113234	2222132321
33	20	221122223222432 21 22	211212211311321421	32 511443144443244	443 3	1 1143	2133221242


//STEP2 EXEC PROC=SLINK,TESTPGM=BADK,ACCT=BADK
GLOBE NO. 1
STANDARD FORM 5081



you describe your opinion of **(INSERT ITEM: ROTATE ITEMS a.-i. AND g.-i.)** as very favorable, mostly favorable, mostly UNfavorable, or very unfavorable? **(INTERVIEWERS: PROBE TO DISTINGUISH BETWEEN "NEVER HEARD OF" AND "CAN'T RATE")**

			Very Favor-able	Mostly Favor-able	Mostly Unfavor-able	Very Unfavor-able	Never Heard of	Can't Rate
(115)	a.	Network television news (1-96)	1	2	3	4	5	6
(116)	b.	Local TV news (1-96)	1	2	3	4	5	5
(117)	c.	The daily newspaper you are most familiar with (1-96)	1	2	3	4	5	6
(118)	d.	Congress (1-96)	1	2	3	4	5	6
(119)	e.	Tobacco companies (7-94)	1	2	3	4	5	6
(120)	f.	Labor unions (2-96)	1	2	3	4	5	6
(121)	g.	Bill Clinton (2-96)	1	2	3	4	5	6
(122)	h.	Hillary Clinton (2-96)	1	2	3	4	5	5
(123)	i.	Bob Dole (2-96)	1	2	3	4	5	6

(4) Publicly Archive Your Research

1. Use persistent, public archives, not your website or "on request"
2. Use Simple, Structured, and Semantic open file formats
3. Be explicit about data licensing

How to license data?

Attribution
CC BY

Attribution-NoDerivs
CC BY-ND

Attribution-NonCommercial-ShareAlike
CC BY-NC-SA

Attribution-ShareAlike
CC BY-SA

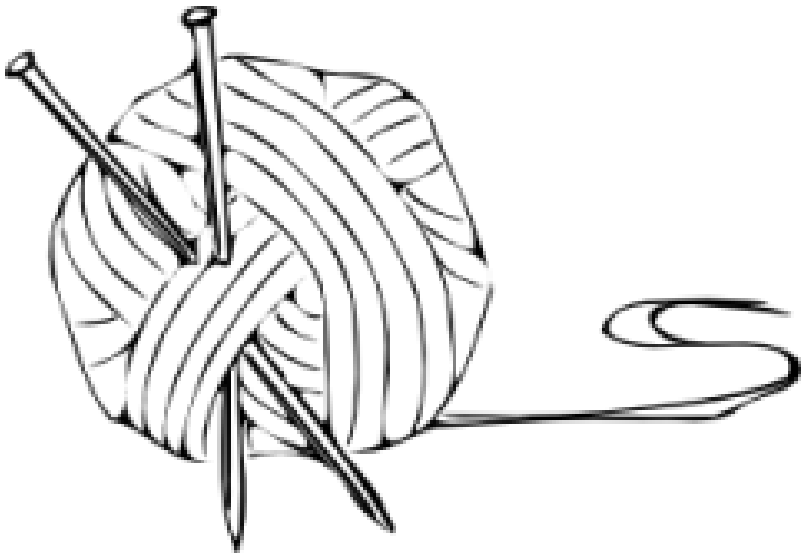
Attribution-NonCommercial
CC BY-NC

Attribution-NonCommercial-NoDerivs
CC BY-NC-ND

(4) Publicly Archive Your Research

1. Use persistent, public archives, not your website or "on request"
2. Use Simple, Structured, and Semantic open file formats
3. Be explicit about data licensing
4. Create useful metadata

(5) Learn Literate Programming



Learn to knit after lunch!

Where to go next?

- rOpenSci
- "Challenges in Irreproducible Research"
- Karl Broman's resources
- 2011 "Reproducible Research" conference slides
- "Six steps to a Better Relationship with Your Future Self."
- "Ten Simple Rules for Reproducible Computational Research."
- *Reproducible Research with R and RStudio*.
- Software Carpentry
- Johns Hopkins Data Science Certificate on Coursera

People to follow?

- @victoriastodden
- @carlystrasser
- @l_peer
- @OSFramework and @BrianNosek
- @RetractionWatch
- @UCBITSS
- @OpenScience

Reproducibility isn't everything

- Data archiving and data citation
- Open protocols and materials
- Methodological transparency
- Free and open-source software (FOSS)
- Open access

In the end...

- Be reproducible *for you*
- Science will benefit as a result