

The Replication Crisis and the Workflow of Data Analysis

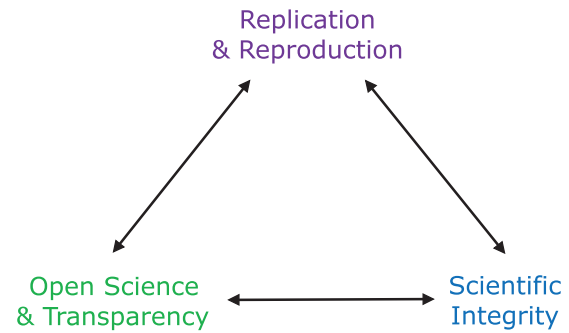
Scott Long

January 17, 2020

© 2020 J. Scott Long

wf wim 2020-01-16 rc1.docx

The Replication Movement



The Replication Crisis

Page 1

Institutional and organizational changes

- NAS Committee on Reproducibility and Replicability in Science
- Journals require submission of data and analysis files
- Requirements for data access by funding agencies
- Haverford College students deposit research on Dataverse
- 2020 Sociology Methods Meeting on replication (finally!)
- Retraction Watch -- retractionwatch.com

The Replication Crisis

Page 2

A simple coding error

RETRACTED: In Sickness and in Health? Physical Illness as a Risk Factor for Marital Dissolution in Later Life

Abstract
The health consequences of marital dissolution are well known, but little work has examined the impact of health on the risk of marital dissolution. In this study we use a sample of 2,201 marriages from the Health and Retirement Study (1992–2010) to examine the role of chronic physical illness onset (i.e., cancer, heart problems, lung disease, and/or stroke) in subsequent marital dissolution due to either divorce or widowhood. We use a series of discrete-time event history models with competing risks to estimate the impact of husband's and wife's physical illness onset on the risk of divorce and widowhood. We find that only wife's illness onset is associated with elevated risk of divorce, while either husband's or wife's illness onset is associated with elevated risk of widowhood. These findings suggest the importance of health as a determinant of marital dissolution in late life, and the potential for divergent social pathways.

Keywords
aging, chronic disease, gender, (ill)marital health

A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are divorced healthier after the remarried (e.g., Lillard and Willis 1995; Emerson 1992), but studies find that both divorce and widowhood are predictors to divergent physical and mental health (e.g., Kaplan and White 2000; Williams and Uchino 2003). Little attention, however, has been paid to how health may be a determinant of marital status. While this area has tended to focus on the positive selection of the healthier into marriage (e.g., Byrne et al. 1999; Smith and Smith 2010), but poor health may be an equally important force for selection (Booth and Johnson 2000). Illness may initiate changes to spouses' roles—in particular, increasing caregiving responsibilities for the healthy spouse—which can tax marital relationship dynamics (Wolff and Kasper 2006). Illness may also decrease household income due to the inability of one or both spouses to work (Tschann 2010), which may increase marital strain. Only a few studies have examined the role of poor health in subsequent divorce, and these studies are mixed in their findings, with some finding

The Replication Crisis

Page 3

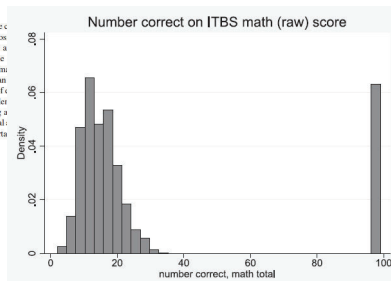
Earlier research treated 99s as valid

Distributional Analysis in Educational Evaluation: A Case Study from the New York City Voucher Program

Marianne Bitler, Thurston Domina, and Emily Penner
University of California, Irvine, Irvine, California, USA

Hilary Hoynes
University of California, Berkeley, Berkeley, California, USA

Abstract: We use quantile treatment effects estimation to examine the effects of the assignment New York City School Choice Scholarship Program across achievement. Our analyses suggest that the program had negligible effects across the skill distribution. In addition to contributing to the literature, the article illustrates several ways in which distributional effects estimation research: First, we demonstrate that moving beyond a focus on mean effects is possible to generate and test new hypotheses about the heterogeneity of effects that speak to the justification for many interventions. Second, we demonstrate effects can uncover issues even with well-studied data sets by forcing a new way. Finally, such estimates highlight where in the overall national scores of children exposed to particular interventions lie; this is important validity of the intervention's effects.



The Replication Crisis

Page 4

Fragile findings due to questionable coding decisions

Measurement, methods, and divergent patterns: Reassessing the effects of same-sex parents^{†‡}

Simon Cheng^{a,1}, Brian Powell^{b,1}

^a 344 Mansfield Rd., Department of Sociology, University of Connecticut, Storrs, CT 06269, United States
^b 744 Ballantine Hall, 1020 E. Kirkwood Ave., Department of Sociology, Indiana University, Bloomington, IN 47405-7103, United States

ARTICLE INFO

Article history:
Received 8 October 2013
Revised 24 March 2015
Accepted 8 April 2015
Available online 23 April 2015

Keywords:
Children
Family structure
Methodology
Same-sex parenting
Sexuality

ABSTRACT

Scholars have noted that survey analysis of small subsamples—for example, same-sex parent families—is sensitive to researchers' analytical decisions, and even small differences in coding can profoundly shape empirical patterns. As an illustration, we reassess the findings of a recent article by Regnerus regarding the implications of being raised by gay and lesbian parents. Taking a close look at the New Family Structures Study (NFSS), we demonstrate the potential for misclassifying a non-negligible number of respondents as having been raised by parents who had a same-sex romantic relationship. We assess the implications of these possible misclassifications, along with other methodological considerations, by reanalyzing the NFSS in seven steps. The reanalysis offers evidence that the empirical patterns showcased in the original article are fragile—so fragile that they appear largely a function of these possible misclassifications and other methodological choices. Our replication and reanalysis of the study offer a cautionary illustration of the importance of double checking and critically assessing the implications of measurement and other methodological decisions in our and others' research.

The Replication Crisis

Page 5

Retraction due to classification errors found by reader

RETRACTED 12 DECEMBER 2019; SEE LAST PAGE

SCIENCE ADVANCES | RESEARCH ARTICLE

SOCIAL SCIENCES

Police violence and the health of black infants

"A reanalysis of the data leads to revised findings that do not replicate the results in the original paper."

Quintet of retractions rocks criminology community

On May 5, 2019, [REDACTED] received an email from "John Smith" with subject line "Data irregularities and request for data":

Seven issues are listed from anomalies in standard errors, coefficients, and p-values to an unlikely survey design and data structure.

Five papers in *Criminology*, *Social Problems*, and *Law & Society Review* were retracted.

Thanks to Patrick Kaminski <pkamins@iu.edu>

The Replication Crisis

Page 6

Data lost in a flood

A chemical engineer claims his data were wiped out in a flood. This led to nine retractions for suspicious images and related issues. — *Retraction Watch* 2019-09-03

Replication study challenges work on fish and ocean acidification

"Does acidification affect fish brains? In these experiments, not so much."

— SK Johnson *arsTechnica* 13 Jan 2020



The Replication Crisis

Page 7

What is replication?

The type of replication depends on the data and methods used.[†]

	Using original methods	Using Improved methods
Using original data	Verification ★ / Reproduction	Robustness
Using similar data	Direct replication	Generalization

[†] - Christensen, Freese & Miguel. 2019. Transparent and Reproducible Social Science Research.

The Replication Crisis

Page 8

Findings are confirmed or disputed in different ways

Using original data	
	Findings Confirmed
Using similar data	Findings Confirmed
	Scientific ideal
Using similar data	Findings Disputed
	Contradictory findings
Findings Disputed	
Careless research	
Incompetent or fraudulent research	

The Replication Crisis

Page 9

Reasons for disputed findings (CFM 2019)

New data, theory, or methods

- o New findings replace old with scientific advances

Fraud

- o Deceptions by authors
- o Particularism in the review process

Sloppy science

- o Errors in the data
- o Incorrect methods

Incomplete documentation

- o Inaccurate descriptions
- o Tacit knowledge

The Replication Crisis

Page 10

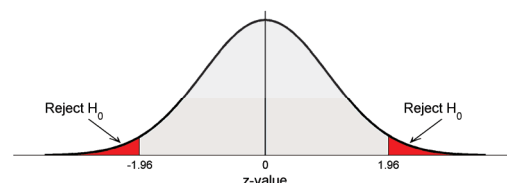
Invalid statistical tests

Issues of testing have received a great deal of attention.

- o p-hacking and post hoc hypothesis construction
- o Stepwise regression portrayed as model testing
- o Cherry picking the sample
- o Subgroup analyses that only report significant group differences

Why is data driving modeling misleading?

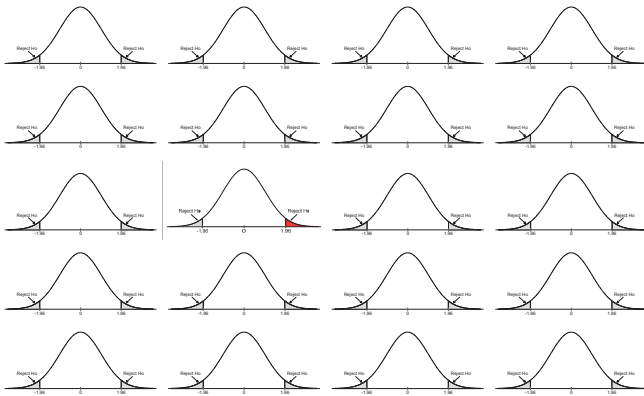
If $H_0: \beta=0$ is true, 5% of random samples will reject the true H_0 .



The Replication Crisis

Page 11

In 20 random samples, a true H_0 should be rejected once



What if only the red study gets published?

Publication bias: an implicit specification search

Publication bias is when a paper's results affect its probability of publication.

- o Correctly finding no effect is less interesting than incorrectly finding an effect.

Which study is published? The jelly bean problem.

Consider stepwise regression with five random samples.

Models for Diabetes

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
female	0.733***	NS	NS	NS	NS
bmi	1.101***	1.067***	1.066***	1.004	0.971
white	0.505***	0.518***	0.547***	0.521***	0.562***
age	1.282***	1.262***	1.351***	1.324***	1.341***
agesq	0.998***	0.999***	0.998***	0.998***	0.998***
hsdegree	0.780***	0.720***	0.680***	0.662***	0.650***
weight	NS	1.006***	1.006***	1.016***	1.022***
height	NS	NS	NS	0.936**	0.909***

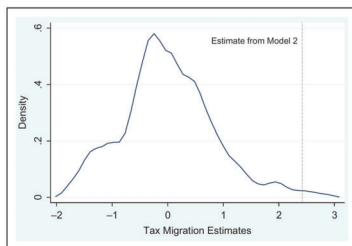
Legend: * p<.1; ** p<.05; *** p<.01 NS dropped from model

Proposals to improve replicability

Prepublication robustness checks

Young and Holsteen. 2015. Model Uncertainty and Robustness. *SMR*.

- o Point estimates capture "one ad-hoc route through the thicket of possible models" -- Leamer 1985
- o For example, do higher income tax rates cause taxpayers to "vote with their feet" and migrate to states with lower taxes?



Tax migration estimates from 24,576 models.

New standards for reporting findings

- o Deposit data and scripts to replicate findings
- o Use standardized definitions to document the sample, data collection, variable construction, model selection, etc.

Using evidence from prior research

- o Study registries with research results
- o Meta-analysis to combine results from similar research

Pre-analysis plans (PAPs) / pre-registration

- o Analyses are registered before analysis begins

Results-blind review process (CFM 114)

Peer Review



Design → Collect → Analyze → Write → Publish

The "replication crisis"

The focus on replication has grown dramatically in the last decade.

Wikipedia (edited)

The replication crisis or replicability crisis or reproducibility crisis is a methodological crisis in which many scientific studies are difficult or impossible to replicate or reproduce.

Science Isn't Broken

After months of investigating the failures of science, Aschwanden concluded:

- o The state of our science is strong, but it's plagued by a universal problem: *Science is hard – really f...ing hard.*
- o If we're going to rely on science as a means for reaching the truth - and it's still the best tool we have - it's important that we understand and respect just how difficult it is to get a rigorous result.

-- Christie Aschwanden, 2015
fivethirtyeight.com/features/science-isnt-broken/

Workflow for replication

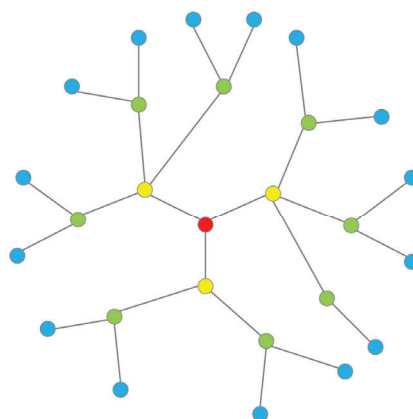
Workflow is a coordinated set of procedures to efficiently produce accurate results than are reproducible.

1. Your workflow encompasses the entire process of research.
 - o Planning research
 - o Importing data
 - o Constructing variables
 - o Analyzing data
 - o Documenting work
 - o Presenting results
 - o Revising analyses
 - o Reproducing results
 - o Preserving files
2. Replication depends on the art, craft, and science of conducting research.

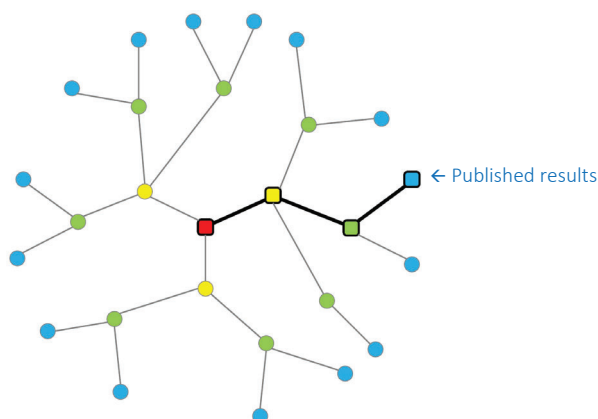
Verifying your own findings is hard

1. *Verification / reproduction* is the easiest form of replication.
 - o You use your own data to confirm your results.
2. Research involves many decisions.
 - o Which cases to keep
 - o How to code education
 - o Value to assign if income is greater than \$200,000
 - o Which transformation to use for a skewed variable
 - o Where to top code count variables
 - o What seed for the RN generator
 - o The type of standard errors for a RE model
 - o Where to dichotomize a 7 point scale
 - o And many more...
3. With 20 decisions, there are 1,000,000 paths to results.

Possible paths to your findings



Actual path to findings



You already have a workflow

Your workflow is how you navigate the path to your results.

Your workflow might be

- o Planned
- o Ad hoc
- o Planned in an ad hoc way?

Is your workflow good enough?

1. Think of a project completed a year ago.
2. How long will it take...
 - o To find the scripts and datasets?
 - o To rerun the scripts and get identical results?
 - o To change one variable, rerun analyses, and update your paper?
3. It should take an hour or two.

Why worry about workflow?

1. Workflow is essential for replication.
2. It prevents errors—both foolish mistakes and black swans.
3. Errors made are easier to find and fix.
4. It saves a lot of time.

Example of why workflow matters

1. A dissertation delayed 18 months to replicate results
2. A 743-line do-file that didn't reproduce a paper's results
3. Conflicting results from the "same" dataset
 - "The datasets are exactly the same except for the married variable."
4. NAS report delayed months by careless variable construction
5. Misabeled gene in a study of alcoholism
6. Misleading output such as...

Definitel a problem

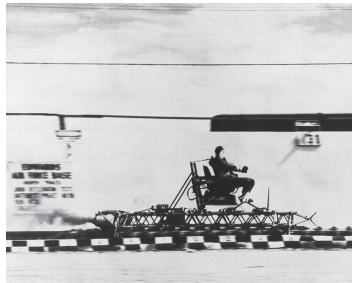
R is	Q15 Would let X care for children				
female?	Defintel	Probably	Probably	Definitel	Total
Male	41	99	155	197	492
Female	73	98	156	215	542
Total	114	197	311	412	1,034

How important is it to...

Variable	Obs	Unique	Mean	Min	Max	Label
tcldoc	1074	10	8.714153	1	10	Q46 How important is it to go to ...
tcifam	1074	10	8.755121	1	10	Q43 How important is it to turn t...
tcifriend	1073	10	7.799627	1	10	Q44 How important is it to turn t...
tcimhprof	1045	10	7.58756	1	10	Q48 How important is it to go to ...
tcipsy	1050	10	7.567619	1	10	Q47 How important is it to go to ...
tcirelig	1039	10	5.66025	1	10	Q45 How important is it to turn t...

The foundation for your workflow

The *universal aptitude for ineptitude* makes any human accomplishment an incredible miracle. — John Paul Stapp

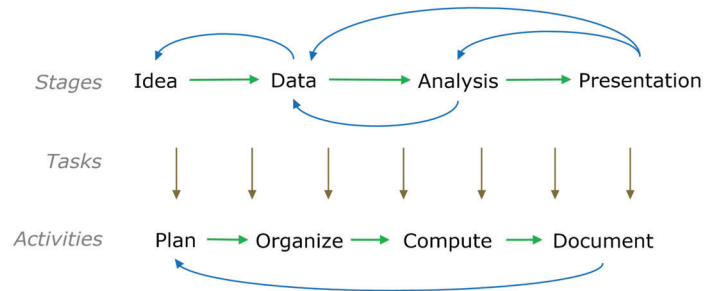


From 0 to 995mph & back in 3 seconds.
"I was fine, only blind for a few days."

The Replication Crisis

Page 8

Overview of a project

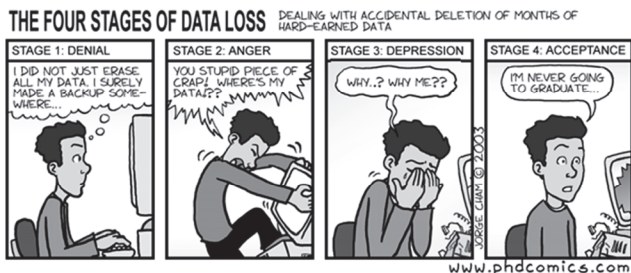


The Replication Crisis

Page 9

Preserving files: an overarching activity

Expect things to go wrong, expect to delete a file at the worst time, expect a hose left on in the room above your computer, expect your partner to delete your files, expect your computer to be stolen.



The Replication Crisis

Page 10

Lost files cause retractions

You must preserve files and their meaning to replicate results.

Preserve the bits on physical media

Make 3 copies on 2 media with 1 copy off-site.

Have equipment to read the media

As a device becomes obsolete, migrate files to new media.

Use file formats you can decode

"These files were saved 6 years ago as GAUSS FMT files. We need them to revise the paper, but old versions of GAUSS no longer run and the format is no longer supported. Any ideas?"

The Replication Crisis

Page 11

Planning

Work. Finish. Publish. —Michael Faraday

1. Planning is the most important thing you can do to improve productivity.
2. You compute, every day. Do you plan every day?
3. Reserve uninterrupted time to plan -- turn off devices and control distractions.
4. A plan is a reminder to stay on track, finish, and publish results.

The Replication Crisis

Page 12

What to plan

- o Project timeline
- o What to publish, where and when
- o Who does what when
- o Procedures for documentation
- o Standards for organization
- o Names and labels for variables
- o What analyses
- o How files are preserved

Plan on different levels

- Grand How does the project fits into your research program?
- Big What are the objectives, goals, and limitations of the project?
- Middle What manageable, distinct tasks are needed to complete the project?
- Small What are the "nitty gritty details" for executing the plan?

The Replication Crisis

Page 13

Organize

Organization is the low hanging fruit in your workflow.

Signs of poor organization

1. You can't find a file and think you deleted it.
2. You have multiple files with the same name but different content.
3. You and a co-author simultaneously edit different versions of a paper.
4. Two authors think they are analyzing the same data.
5. Is **FinalReportV16.docx** the final draft?
6. You have multiple copies of the same reprint but can't find any of them.
7. You receive a text:
URGENT: don't analyze **final.dta**, use **lastversion.dta** for presentation tomorrow.
8. Surely a rookie mistake....

The final paper



Principles of organization

1. Organization has two primary objectives:
 - o Finding things quickly
 - o Avoiding duplication
2. Organization
 - o Rewards consistency and uniformity
 - o Is contagious, as is disorganization
 - o Requires maintenance to overcome entropy

Organization of files

It is easier to create a file than to find a file.

It is easier to find a file than to know what is in a file.

It is easy to create lots of files.

Digital asset management (DAM)

A file's name and directory document the file's content.

- o Choose names carefully and systematically.
- o Use a *planned directory structure* so every file belongs in exactly one place.

File naming examples

Manuscripts with topic, date and author

groups 2017-11-07 js1.docx ← draft by JS Long on 2017-11-07
groups 2018-01-17 sam.docx ← draft by SA Mustillo on 2018-01-17

Reprints

Long 1978 ASR prod position.pdf
not 02839211.pdf

Datasets

groups-hrs1.dta ← not **final1.dta**
groups-hrs2.dta ← not **final2.dta**



A simple directory structure

\- To shelf	Files to move to the correct directory
\Active	Active projects and classes
\GroupDif	: Each project has its own project directory
\Soc650	
\Active - hold	Projects that are on hold
\RM2nd	
\Admin	Administration and service
\Bookshelf	Books, articles, reprints, etc.
\Shared	Files shared on the cloud
\Mason, Fred	
\Templates	Sample documents
\Vault	Files that never change (e.g., published papers)

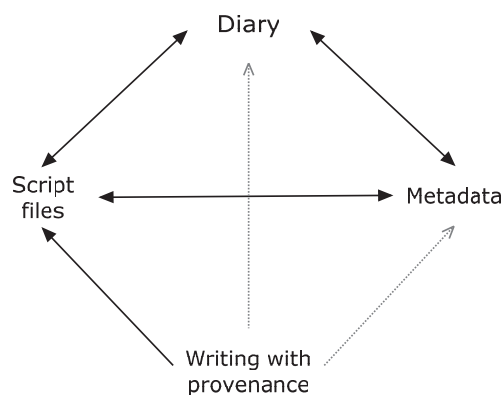
Documentation

1. Nobody likes to write it nor regrets having it.
2. Without documentation,
 - o Replication is hard or impossible.
 - o Mistakes are more likely.
 - o Revisions take longer.

Rules for documentation

1. Write it today.
2. Revise it later.
3. Review it at waypoints in your project, like finishing a draft.
4. Use reinforcing modes of documentation...

Reinforcing, nonredundant modes of documentation



Diary

The diary records the plans, progress and agreements.

WF example based on SRM submit files

Zack, Kennedy & Long (2019). Can Nonprobability Samples be Used for Social Science Research? Survey Research Methods, 13(2), 215-227.

2019-10-21 Create reproduction scripts

Create reproduction dataset

SRM_data01.dta has all variables and observations. This dataset used to extract reproduction dataset with only variables and observations used in paper.

wf-srm-data1-extract.do 2019-10-21

Create wf-srmrep1.dta from SRM_MTURK_data01.dta

wf-srm-data2-checks.do 2019-10-22

List notes for variables as needed. Not critical; can be deleted.

Analyses reported in paper

Analyses using wf-srmrep1.dta

wf-srm-stat1-descriptive.do 2019-10-21

wf-srm-stat2-DCR.do 2019-10-22

DC at representative values with full and restricted samples.

wf-srm-stat3-ADC.do 2019-10-22

Average DC with full and restricted sample.

2019-10-22 add provenance notes

Add provenance notes to wf-srm-with provenance 2019-10-22.docx.

Metadata: data about data

Internal documentation about variables and datasets

Metadata about the dataset

_dta:

1. srm_mturk_data01.dta / turk-data08-keepV3.do Elizabeth Zack 2018-05-15
2. wf-srmrep1.dta / wf-srmrep-data1.do Scott Long 2019-10-22

Variable names and labels

```
lfp  In paid labor force?
k5   # kids < 6
k618 # kids 6-18
age  Wife's age in years
wc   Wife attended college?
```

Metadata about variables

advfront2:

1. source variable scigov with reverse coding
2. renamed advfront2 / wf-srmrep-data1.do Scott Long 2019-10-22

Script files

1. Scripts document exactly how variables are constructed, observations selected, and analyses are conducted.
 - o They are the BEST description of exactly what you did.
2. The diary, metadata, and provenance tags point to these scripts.
3. The diary explains, justifies, and reviews subtle issues about the scripts.
 - o For example, provide the rationale for a method of variable construction.
 - o Provide references justifying results.
 - o Discuss insights from the results.

Papers and provenance tags

1. The provenance of a result is the script that produced it.
2. Maintaining provenance is essential for replication and efficiency.
 - o If you don't know where a number came from, how do you reproduce it?
 - o Revisions are easier since you know what needs to be changed.

1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively ($z=2.55$, $p<.01$ (cwhrr-fig03c-hrmemp4.do #4 jsl 17May06)).

However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Computing



1975 IBM 370: \$1,000,000
228K of memory
3MB disk storage
Time to PHD: 7.6 years



2020 Laptop: \$450
8GB of memory
4TB disk storage
Time to PHD: 7.6 years

A thought experiment

1. Divide graduate students randomly into two groups.
Computers compute as much as they like.
Planners can compute only 12 hours a week.
2. Which students finish their dissertation first?

A workflow for computing

1. Use of robust and legible script files.
2. Never change files that have been shared.
3. Distinguish data management from data analysis.
4. Use effective variable names and labels.

Robust and legible script files

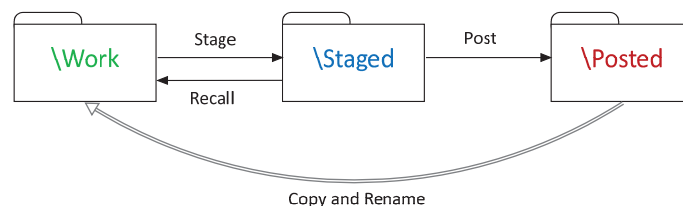
1. Use scripts since interactive results are hard to reproduce.
2. Use robust scripts that run on another computer with no changes.
 - o To tell if a script is robust, run it on another computer.
3. Use thoughtful comments and uniform formatting.

The essential posting principle

Posting is defined by two rules

1. Before you share results, post the files used to produce the results.
2. After a file is posted, never changed it.

File movement with posting



Why posting?

Posting may be the most important thing you can do to improve your workflow.

- o Posting preserves the files that produced your results.
- o Posting prevents changes that break replication.
- o Posting prevents two files with the same name and different content.
- o It simplifies returning to a project after an interruption.

Dual workflow

Data workflow

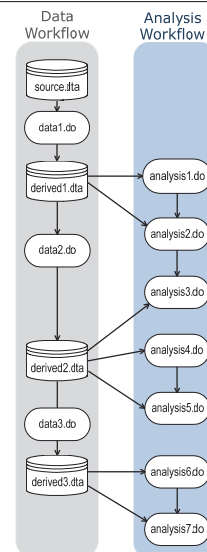
Variables are created and datasets are saved.

Analysis workflow

Datasets are loaded and variables are analyzed, but variables are not created and datasets are not saved.

Why?

1. Errors are prevented.
2. It encourages careful data management.
3. Posting is much simpler.
4. Documentation is easier.
5. Revisions are faster.
6. Collaboration is easier.



Use effective names and labels

Poor variable names

```
logit R0051400 R0000100 R0002203 R0081000
logit lfp age educ kids
```

Misleading variable labels

Variable	Obs	Unique	Mean	Min	Max	Label
tcldoc	1074	10	8.714153	1	10	Q46 How important is it to go to ...
tcilmhprof	1045	10	7.58756	1	10	Q48 How important is it to go to ...
tcipsy	1050	10	7.567619	1	10	Q47 How important is it to go to ...
tcifam	1074	10	8.755121	1	10	Q43 How important is it to turn t...
tcifriend	1073	10	7.799627	1	10	Q44 How important is it to turn t...

Worthless value labels

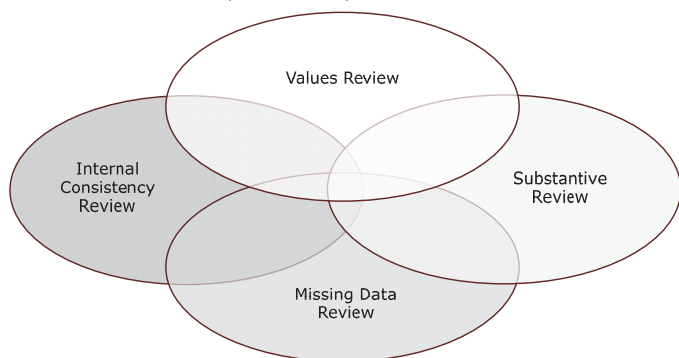
R is	Q15 Would let X care for children	Total
female?	Defintel Probably Probably Defintel	
Male	41 99 155 197	492
Female	73 98 156 215	542
Total	114 197 311 412	1,034

Rules for computing

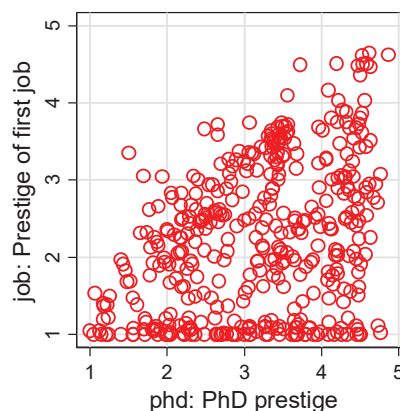
1. Plan more, compute less.
2. Spend more time on data management than statistical modeling.
3. New content always gets a new name.
 - o Change a variable, change the name.
 - o Change a dataset, change the name.
 - o Change the document, change the name.
4. Never change shared results—they were posted!.

Data cleaning

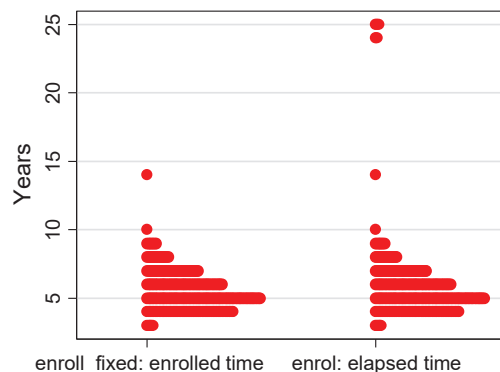
1. Statistical analysis assumes variables are clean.
2. Checking variables only when things look wrong is dangerous.
3. Clean is a critical first step in data analysis.



Cleaning leads to substantive insights



Cleaning prevents retraction worthy blunders



Informal imputation of missing values

1. **timemarried** had the most missing values in a survey of sexual behavior.
 - o I'll tell you how many affairs I had, but not when I got married.
2. The survey computed time married as:
$$\text{timemarried} = 12 * \text{years} + \text{months}$$
3. Newlyweds ignored the years question, while older couples skipped the months question.
4. I imputed missing values as:
 - o If **years** is missing and **months** is not, then **years** is 0.
 - o If **months** is missing but **years** is not, **months** is 6.
5. Is my imputation reasonable?
 - o Is multiple imputation better?
 - o Listwise deletion?

Data analysis

Plan your analyses

Without a plan:

- o You run analyses you don't need.
- o You forget key objectives to pursue distractions.

Planning a paper (pre-registration)

1. Write the abstract before analysis begins.
2. Create dummy tables and figures that guide analyses.

Can this be done? Doesn't it get in the way of creativity?

Blau and Duncan's *American Occupational Structure* was written from a single set of output without additional analyses.

Conducting the analyses

1. Use script files.
2. Do not create variables in analysis scripts.
3. Save new variables in a new dataset.

Create a replication package while writing the paper

1. The package contains datasets and scripts that reproduce all data-based tables, figures, and conclusions.
2. The paper, includes metadata pointing to scripts from the replication package.

1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively ($z=2.55$, $p<.01$ [cwhrr-fig03c-hrmemp4.do #4 jsl 17May06]).

However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Changing your workflow

Your goal is to develop a workflow that efficiently produces results that are accurate and reproducible.

Assessing your workflow

1. What problems have you noticed in your work?
 - o I forget which file is most recent.
 - o I don't remember how a variable was created.
 - o Revisions take a long time.
2. Replicate an earlier paper and keep track of problems.
3. Prioritize changes:
 - o Which problems affect reproducibility?
 - o Which waste the most time?

Selecting procedures for your new workflow

Chose methods that optimize these related, sometimes conflicting, criteria.

Essential criteria

Prime criterion

Reproducible: If you cannot reproduce your results, nothing else matters.

Core criteria

Accurate: wrong answers can be replicated, but right answers are better.

Efficient: you need more time to write.

Coordinated: procedures are consistent and reinforcing.

Useful criteria

Supporting criteria that improve efficiency

Standardized: do things the same way each time.

Automated: prevents errors and saves time.

Scalable: methods that work for every project.

Usability criteria to encourage use of your workflow

Simple: complex will be abandoned.

Congenial: you have to be willing to use your workflow.

Transferable: others need to use and understand your methods.

Whose workflow?

1. Sadly, my book was titled: *The Workflow of Data Analysis...*
It should have been: *A Workflow for Data Analysis...*

2. Is *my* workflow the *best* workflow?

3. Do you want to spend time finding something better?

- o Since workflow is as effective as the weakest link, will you anticipate the implications of your revisions?

Should you redo all of your projects?

1. If a project is nearly complete, finishing it using the "old way" ...
unless results cannot be replicated.
2. Fully adopt your new workflow for new projects.

Implementing a new workflow

1. Your workflow can be improved with a modest investment of time.
 - o The less experience you have, the easier it is to improve.
2. Initially, changes are hard, inefficient, and uncomfortable.

Vic Bradon: “How does that feel?”

Student: “Awful!”

Vic Bradon: “Remember that awful—it will make you famous.”
3. Stick with what you learn until the new procedures become routine.
 - o It takes a week or two of consistent use to become natural.
4. Make changes slowly, systematically and fully.
 - o Finish the last 5%.
5. Do not make changes under a deadline.

Resources

BITSS: Berkeley Initiative for Transparency in the Social Sciences.
Christensen, Freese and Miguel. 2019. *Transparent and Reproducible Social Science Research*.
Long. 2009. *The Workflow of Data Analysis Using Stata*.
Long. 2020. ICPSR Workshop on Reproducible Results.
Project TIER: Teaching Integrity in Empirical Research.

Thank you for listening