

Reproducible Results and the Workflow of Data Analysis

Scott Long

Departments of Sociology and Statistics
www.indiana.edu/~jslsc/ftp/

Workshop in Methods | August 2016

The reproducible results movement

- Replication and reproducible results
- Open science
- Transparency in science
- Teaching integrity in research

Changing expectations

- Journals require data and analysis files *before* acceptance
- Funding agencies strengthen requirements for data access
- Haverford College requires reproducibility for undergraduates

With access comes accountability

- retractionwatch.com
- For example...

Retraction due to 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 the work has ignored the impact of health on the risk of marital dissolution. In this study we use a 1992-2010 panel from the Health and Retirement Study (HRS) to examine the risk of divorce among those with (vs. 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 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 later life and the potential for differential and gendered social pathways.

Keywords
aging, chronic disease, gender, marital dissolution, widowhood

A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are divorced and widowed individuals (e.g., Lillard and Willis 1991; Casper 1992), but studies that find that divorce and widowhood are predictors for increased physical and mental health (e.g., Hughes et al. 2009; Williams and Uchino 2003) are common. However, less has been paid to the health consequences of marital status. This article has tried to focus on the positive outcomes of the health into marriage (e.g., Byrne et al. 1995; Smith and Smith 2010), but poor health may be an equally important issue for widow-

hood, and Johnson 2008). Illness may initiate changes to spouses' roles—in particular, increasing competing responsibilities for the healthy spouse—which can tax marital relationship dynamics (Wolff and Knapp 2005). Illness may also decrease household income due to the inability of one or both spouses to work (Tuchman 2010), which may increase marital stress.

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

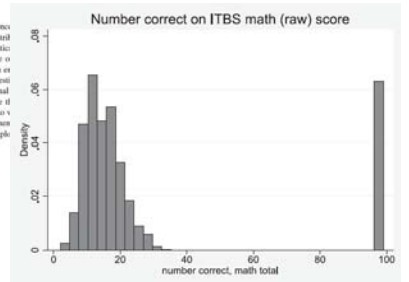
Finding errors in prior research

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

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



Assessing the fragility of published results

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 Regnerus' 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.

Science Isn't Broken by Christie Aschwanden

Peer review?

Circumvented peer review at prestigious journals

Scientific journals?

Two journals published Maggie Simpson & Edna Krabappel's "Fuzzy, Homogeneous Configurations"

Revolutionary findings?

Retraction at *Science* when data not found

Is science broken?

"I've learned that the headline-grabbing cases of misconduct and fraud are mere distractions. The state of our science is strong, but it's plagued by a universal problem: **Science is hard – really f*ing hard.**"

"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.**"

Replication and reproducible results

- Distinct but related concepts

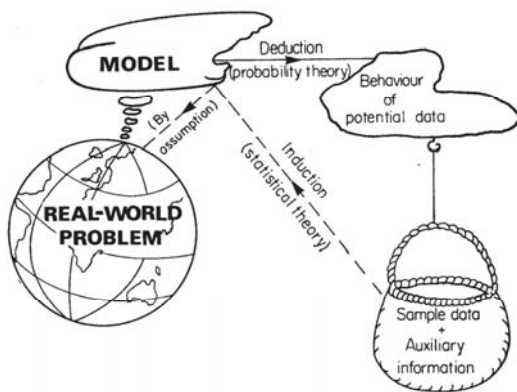
Replication of results

- Confirm published results with *new data*

Challenges to replication

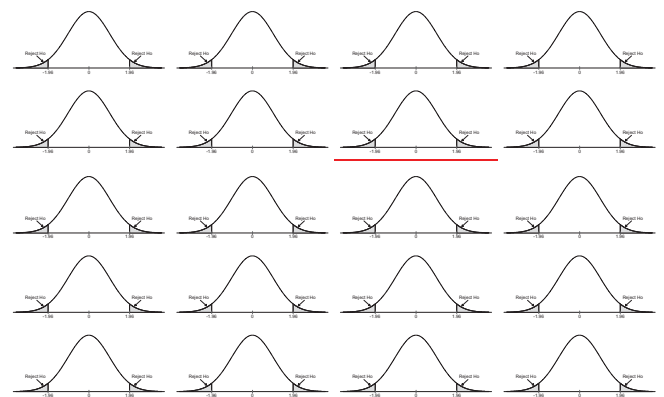
- Abuse of the uniqueness of the sample
- Data mining portrayed as theory testing
- Post analysis hypothesis construction
- "Cherry picking" the sample
- Undocumented specification searches

Classical model of inference



Vic Barnett *Comparative Statistical Inference*

Multiple tests at 5% level



Using the sample data to select a model

1. Observations are randomly assigned

- Exploration sample to find a model by stepwise regression
- Verification sample to confirm the results

2. Process repeated three times

Variable	explore1	verify1	explore2	verify2	explore3	verify3
bmi	1.067***	1.066***	1.004	1.074***	1.101***	0.971
white	0.518***	0.547***	0.521***	0.543***	0.505***	0.562***
age	1.262***	1.351***	1.324***	1.288***	1.282***	1.341***
agesq	0.999***	0.998***	0.998***	0.998***	0.998***	0.998***
hsdegree	0.720***	0.680***	0.662***	0.749***	0.780***	0.650***
weight	1.006***	1.006***	1.016***	1.004**		1.022***
height			0.936**			0.909***
female				0.854*	0.733***	
_cons	0.000***	0.000***	0.000***	0.000***	0.000***	0.000***
N	8036	8035	8036	8035	8036	8035
bic	7557.1	7479.6	7450.6	7594.2	7622.1	7405.3

legend: p<.1; ** p<.05; *** p<.01

Model Robustness

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

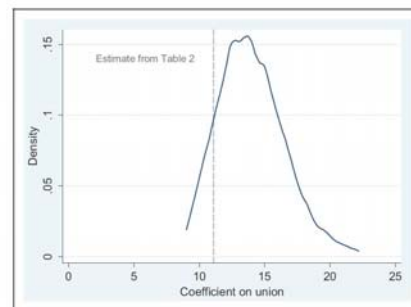


Figure 1. Modeling distribution of union wage premium. Note: Kernel density graph of estimates from 1,024 models. Vertical line indicates the preferred estimate of an 11 percent union wage premium as reported in Table 2.

Reproducible results

- Identical results with the *same data*

Demands for RR

- Journals require verification of results before a paper is published
- Data and script files are made publicly available

Challenges to RR

- Not as easy as it looks; not as hard as some fear
- Requires a systematic workflow based on reproducibility

My talk focuses on reproducible results

- The workflow of data analysis
- Even non-replicable results should be reproducible

Reproducible Results and Workflow | 12

What is the workflow of data analysis?

Workflow is a coordinated framework for data analysis that deals with all aspects of data analysis:

- Planning, organizing and documenting research
- Cleaning data
- Analyzing data
- Presenting results
- Backing up and archiving materials
- Reproducing results



Reproducible Results and Workflow | 13

Why must the workflow be coordinated?



Reproducible Results and Workflow | 14

You already have a workflow

1. Your WF might be:
 - **Planned**
 - **Ad hoc**
 - **Planned in an ad hoc way**
2. You can improve your WF with a modest investment of time.
 - The less experience you have, the easier it is
 - ✓ Undergraduates find it easier than faculty!
 - In the long run, it saves time
 - It makes you a better data analyst
 - It prevents retractions

Reproducible Results and Workflow | 15

Why workflow is essential

Three primary criteria for developing your workflow

Reproducibility

1. Reproducible results are essential for good science
2. Workflow is critical for reproducibility

Getting the right answer

1. You want your analysis to be correct
2. With open science others will find your mistakes

Efficiency

Science is a voracious institution. -- Harriet Zuckerman

Reproducible Results and Workflow | 16

Origins of the workflow project

1. Consulting on easy things instead of hard things
2. Incorrect results with clever explanations
3. A dissertation delayed 18 months to determine provenance
4. Unreproducible results from a 743 line do-file with no comments
5. Analyzing the wrong data set:
 - "The datasets are exactly the same except for the married variable."
6. Using the wrong variable when writing a report for the NAS
7. Mislabelled gene in a study of alcoholism
8. Collaborations that multiply the ways things go wrong
9. Misleading output such as...

Reproducible Results and Workflow | 17

Definitel in a \$3M study

```
. tabulate female sdchild_v1
```

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

How important is it...

```
. codebook tc1*, compact
```

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...

Which number is which?

```
. tab occ ed, row
```

Occupation	Years of education						
	3	6	7	8	9	10	11
Menial	0	2	0	0	3	1	3
38.71	0.00	6.45	0.00	0.00	9.68	3.23	9.68
6.45	100.00						
BlueCol	1	3	1	7	4	6	5
26	1.45	4.35	1.45	10.14	5.80	8.70	7.25
37.68	10.14	100.00					
Craft	0	3	2	3	2	2	7
39	0.00	3.57	2.38	3.57	2.38	2.38	8.33
46.43	8.33	100.00					
WhiteCol	0	0	1	0	1	2	
19	0.00	0.00	2.44	0.00	2.44	4.88	
46.34	9.76	100.00					

Why learning WF is difficult

Tacit knowledge

1. **Explicit knowledge** is the stuff of textbooks and articles.
2. **Tacit knowledge** is implicit and undocumented (Polanyi).
 - A. People are unaware of their essential tacit knowledge.
 - o Henry Bessemer's 1855 patent for steel did not work.
 - B. Tacit knowledge is transferred "**at the bench**".
 - o Personal computers impede the transfer of tacit knowledge.

Data analysis involves heavy lifting

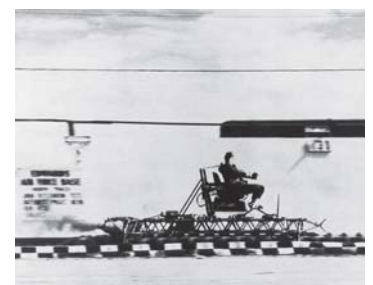
There's a lot of undifferentiated heavy lifting that stands between your idea and that success. -- Jeff Bezos, amazon.com

The Workflow of Data Analysis Using Stata

1. Makes tacit knowledge explicit
2. Deals with details for heavy lifting
3. Provides specifics on issues discussed today
4. While focusing on Stata, the principles apply broadly
 - o An ethnographer uses it for her research team
 - o An researcher in China found it crucial for getting his paper accepted by *Nature*
 - o A manager of health statistics for a European country said it "improved the quality of my life, not my data analysis, my life."

The foundation of WF is **ironical optimism**

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



40G's: From 0 to 995mph and back in 3 seconds...



"I was fine, only blind for a few days."

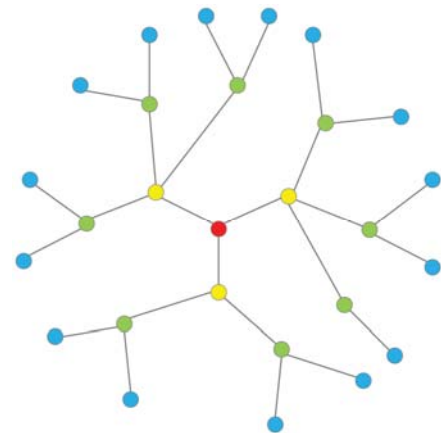
Reproducibility is the prime criterion for WF

1. WF facilitates reproducible results.
2. Ask yourself:
 - Can you reproduce **exactly** the results you published?
 - How long would it take?
3. Reproducible results requires planning from the start of a project.

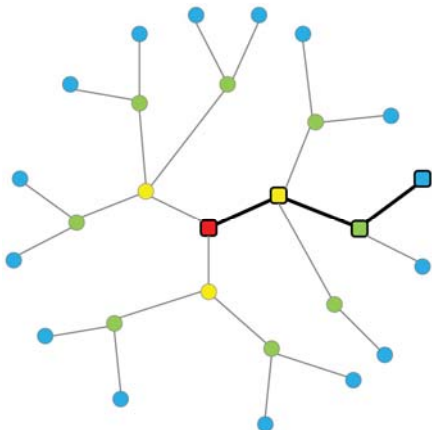
Why are results hard to reproduce?

1. **The curse of dimensionality:** 10 decisions, 1,024 possibilities
 - Where to truncate a variable?
 - What seed for the RN generator?
 - How to scale with partially missing data?
 - Which cases to keep for analysis?
 - How to code education?
 - What values to assign to income greater than \$200,000?
 - And so on...

Decisions in the path to analysis: *the choices that could be made*



Decisions in the path to analysis: *the choices made*



Why are results hard to reproduce?

2. **Missing documentation:** Replication should involve retrieving documentation, not trying to remember.
3. **Changing software:** New software can give different results.
 - A colleague spent painful weeks failing to reproduce results because he forgot **version 7** in a do-file.
4. **Lost files:** corrupted, lost, unreadable, obsolete, or ambiguous files
 - Do you have \$2,000 to retrieve the file that was "backed up"?
 - Do virtual servers archive your data?

Given reproducibility, criteria for choosing WF

1. Accuracy

If your program is not correct, then nothing else matters.

--Oliveira and Stewart

2. Efficiency

- o Complete work quickly
- o Working quickly competes with working accurately

3. Standardization

- o Avoiding repeatedly, inconsistently deciding how to do things
- o Standardization makes it easier to find mistakes

4. Automation

- o Automated procedures prevent mistakes
- o Time invested learning automation can save time

5. Simplicity

- o Unnecessarily complicated procedures are abandoned

6. Usability

- o Your workflow should reflect the way **you** like to work
- o If you won't do it, it is not a good workflow

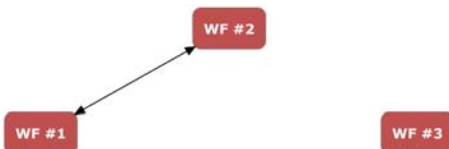
Collaboration and workflow

1. Collaboration makes it more difficult to have an effective workflow.
2. Disciplines with a history of collaboration emphasize an explicit workflow.
3. Why is workflow harder when you collaborate?

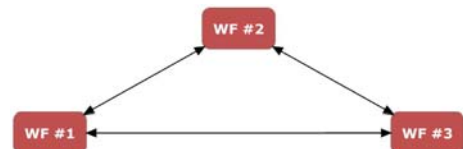
Coordinating multiple workflows



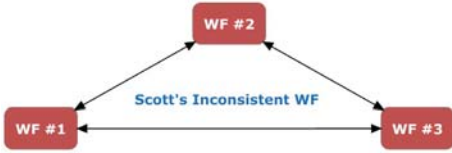
Coordinating multiple workflows



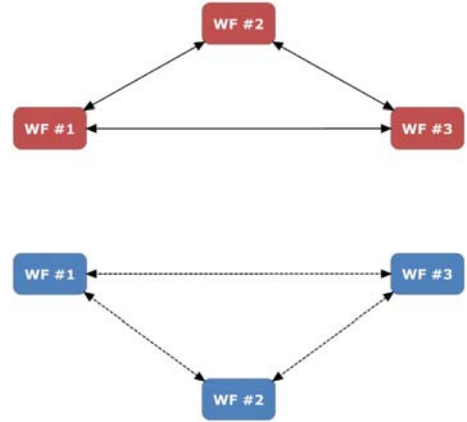
Coordinating multiple workflows



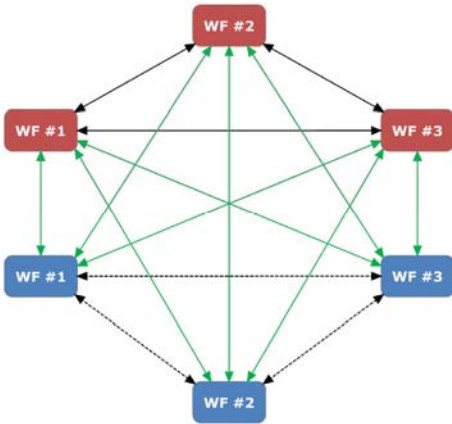
Coordinating multiple workflows



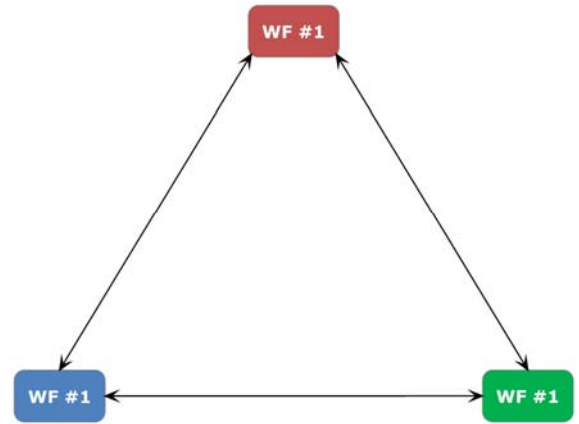
Coordinating multiple workflows starts here



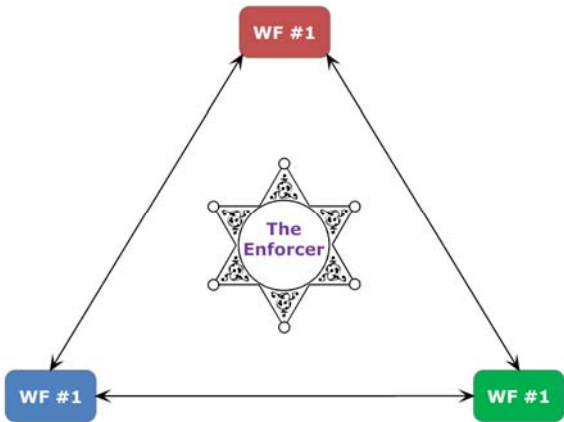
Coordinating 30 pairs of workflows



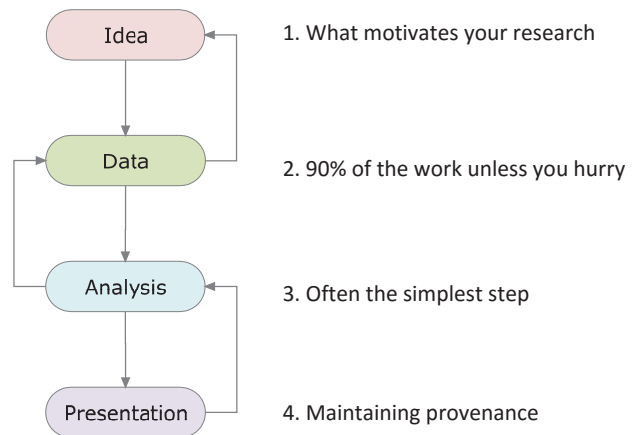
Coordinating multiple workflows: Agree on a WF



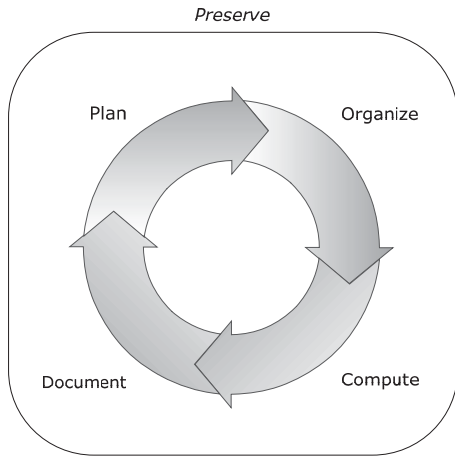
Get an enforcer



Steps in your workflow

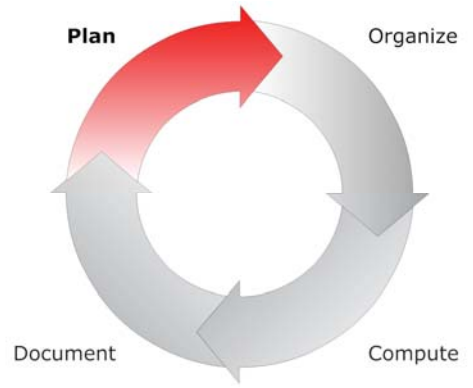


Tasks within each step



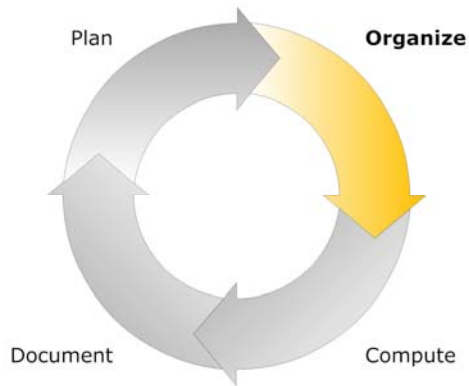
Reproducible Results and Workflow | 42

Tasks within each step



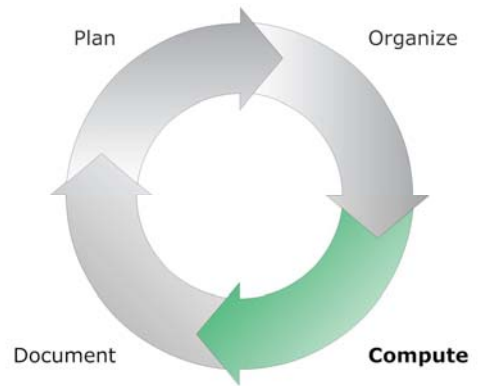
Reproducible Results and Workflow | 43

Tasks within each step



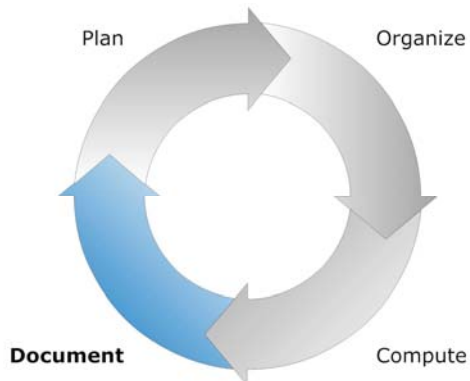
Reproducible Results and Workflow | 44

Tasks within each step



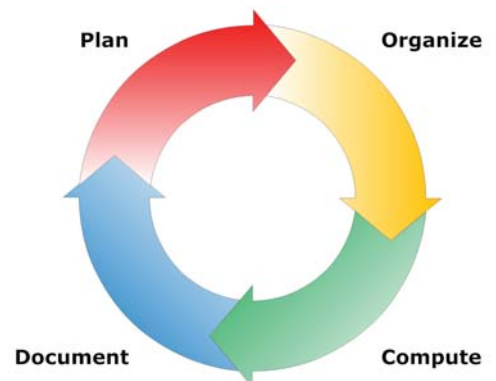
Reproducible Results and Workflow | 45

Tasks within each step



Reproducible Results and Workflow | 46

Tasks within each step



Reproducible Results and Workflow | 47

Planning

- Your time
- Publishing plans and deadlines
- Division of labor
- Data construction: names, labels, formats
- Procedures for missing data
- Anticipated analyses
- Documentation and organization
- Preserving files
- And more...

Blau and Duncan's *The American Occupational Structure*

- Analyses were specified 9 months before output was received.
- Book was written based entirely on a single set of output.
- Later books with full access to the data were not better.

Michael Faraday's famous sign

Work. Finish. Publish. --Michael Faraday's sign in his lab

What is a plan

A plan is a reminder to stay on track, finish the project, and publish results.

Organizing

1. Organization is motivated by two goals:

- **Finding things**
- **Avoiding duplication**

2. Organization...

- Helps you work faster
- Rewards consistency and uniformity
- It is contagious

Signs of poor organization

1. **Can't find a file** and think you deleted it.

2. **Multiple versions** of a file and don't know which is which.

- You and a colleague are working on different versions of the same paper. You changed what she changed and now there are three versions of the paper.

- You need the final version of the paper the was submitted for review, but you have two (or 16) files with "final" in the name.

This: final_report_v16.docx

Or this: NSF_gsssci_report 2010-10-21.docx

3. **Finally:** After this talk a student showed me this text:

Urgent: don't analyze **final.dta**, use **lastversion.dta** for our presentation tomorrow."

Organization should be like a Model T



"Any color you want as long as it is black."

Too often it is more like this



With predictable consequences



Digital assets and the curse of cheap storage

1. It is easier to create a file than to find a file.
2. It is easier to find a file than to know what is in a file.
3. It is easy to create lots of files.
 - o 115,000 files on a research center's LAN
 - o 2,000,000 files accumulated in 10 years

Where are your files?

1. Laptop
2. LAN
3. Dropbox
4. Box
5. USB sticks
6. Old laptop
7. Friend's laptop
8. External drives
9. Mom's computer

Operating systems focus on entertainment

Win	Mac
Desktop	Desktop
Music	Music
Pictures	Pictures
Videos	Movies
Documents	Documents

Digital asset management (DAM)

How important is this?

How much time do you waste dealing with files?

What can you do?

1. Suppose I put my files in **\Dropbox**.
2. Start with general categories of files.

For example...

Primary directories

\Active	Research projects I am actively working on
\Admin	Administrative files, templates, etc.
\Bookshelf	Books, articles, reprints, etc.
\Inactive	Incomplete projects that are on hold
\Programs	Files that customized installed programs
\Service	Documents related to service work
\Shared	Files shared with others
\Students	Files from students
\Teaching	Class materials
\Templates	Sample files used as templates
\Vault	Completed work that will never change

Within a primary directory, make subdirectories

- \Bookshelf**
 - \Articles**
 - \Books**
 - \Computing**
 - \Figures**

Where to put David Allen's "stuff"

- \- Hold then delete**
- \- To shelve**
- \- To transfer**

A structure for projects

\Group Differences

\- History starting 2016-01-20

\- Hold then delete

\- To shelve

\Admin

\Preposted

\Posted

\Resources

\Work

\Write

Organization: uniform formats for do-files

```
capture log close
log using wftalk01-example, replace text
version 14.1
clear all
set linesize 80

// project: wf talk
// task:
local pgm wftalk01
local dte 2016-01-20
local who scott long
local tag "`pgm'.do `who' `dte'"

// #1 description of task 1

// #2 description of task 2

log close
exit
```

Documentation

1. **Long's Law:** It is faster to document it today than tomorrow.

Addendum 1: Nobody likes to write documentation.

Addendum 2: Nobody regrets having documentation.

How often do you hear: *"Drat, I have too much documentation."*

2. Without documentation, replication is virtually impossible, mistakes are likely, and work takes longer.

3. The more codified the field the greater the emphasis on documentation

The Research Log by the American Chemical Society

Suggestions for writing documentation

1. Use reinforcing logs, metadata, comments, names
2. Do it today
3. Check it next week (it always makes sense today)
4. Review it at key stages of your work, like finishing a draft
5. Include full dates and names

The core of your documentation: the research diary

1. The diary is a road map connecting activities and files.
2. Script files precisely describe what is done.
3. Metadata in datasets points to script files.

An example...

First complete set of analysis for FLIN measures paper

f2alt01a.do - 24May2002

Descriptive information on all rhs, lhs, and flm measures

f2alt01b.do - 25May2002

Compute bic' for each of four outcomes and all flm measures.

```
** Outcome: Can Work          global lhs "qsmen05"
** Outcome: Work in three categories global lhs "dhitna05"
** Outcome: math trouble       global lhs "mathtr05"
** Outcome: adlsum05 - sum of adle global lhs "adlsum05"
```

f2alt01c.do - 25May2002

Compute bic' for each of four outcomes and with only these restricted flm measures.

```
* 1. ln(r*5) and ln(r*1)
* 2. 9 counts: >=60 <=70? (50% and 75%)
* 3. 8 counts: >=4 <=6 (50% and 75%)
* 4. 18 counts: >=9 <=11 (50% and 75%)
* 5. probability splits at .5? these don't work well in prior tests
```

f2alt01d.do - 25May2002

bic' for all four outcomes in models that include all raw flm measures (fla*pb; fill*pb); pairs of w/l measures; groups of LCA measures

f2alt01e.do - all LCA probabilities - 25May2002

f2alt01j.do - use three probability measures from LCA - 29May2002

f2alt02a.do - 29May2002

use three binary variables, not just LC class numbers.
: dummies work better than the class numbers.
: effects of lower and severe are not significantly different.

Redo f2 analyses - error in adlsum - 3Jun2002

ARGH! adlsum is incorrect -- it included going to bed twice.
All of the f2alt analyses need to be redone using the corrected dataset.

f3alt_gflim07.do: create gflim07.dta 3Jun2002

```
1) Correct adlsum: adlsum05b
2) Add binary indicators of lmaxp5: lmaxnonep5, etc.
```

f3alt01a (redo f2alt01a.do) - 3Jun2002

f3alt01b.do (redo f2 job) - 3Jun2002

Execution and computing

1. Execution involves carrying out tasks within each step.
2. Effective execution requires the right tools.

o Software

- a. *File manager: Explorer and Finder do not work well*
- b. *Macro program: don't retype the same thing*
- c. *Text editor: use just one for all programs*
- d. *Word processor: page feed, outlines, headings*
- e. *Statistical software*

o Hardware: display, storage, memory, CPU

3. Planning is more important than computing power.

- Consider the changes in computing...

Cornell 1975: the entire computing infrastructure



IBM 370 with 240K memory



Winchester drives with 3MB storage

- Cost of computing \$1,000,000.
- Mean time to degree 7.6 years.

Indiana 2009: a disposable PC



Asus 1000HE with 2GB memory
10,000 times more



Free Agent with 1TB storage
350,000 times more...

- Cost of computing \$400.
- Mean time to degree 7.6 years.

A thought experiment on planning and computing

1. Divide yourselves into two groups:

- **Computers** can compute whenever they want to (i.e., all the time).
- **Planners** can compute for three four-hour sessions a week.

2. Who finishes first?

Principles for a computing workflow

1. Posting files
2. Dual workflow
3. Run order naming

The essential posting principle

Posting is defined by two simple rules.

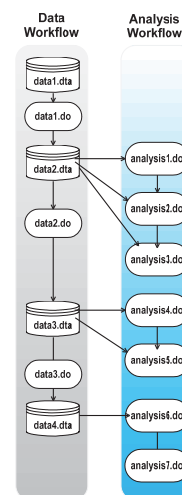
The share rule

Only share results after the files are posted.

The no change rule

Once a file is posted, never change it.

Dual workflow



Run order and a dual workflow

Name files so if re-run in alphabetical order, you produce *exactly* the same results.

Data management

```
data1.do
data2.do
data3.do
data4.do
data5.do
data6.do
```

Data analysis

```
desc1.do
desc2.do
desc3.do

graph1.do
graph2.do
graph3.do

logit1.do
logit2.do
logit3.do
logit4.do
logit5.do
```

Data analysis: use script files

Robust script files

Simply put: Your programs should run on another computer at a later date without requiring *any* changes.

1. Self-contained
2. Version control
3. Exclude directory information (which might change)
4. Explicitly set seeds for random numbers
5. Archive user written ado-files

Legible script files: output that is easy to read

1. Include thoughtful comments
2. Formatted with alignment, indentation, and spacing
3. Text that does not wrap

Legible output files

```
-----+-----
| Key                                     |
|-----+-----|
| frequency                               |
| row percentage                          |
|-----+-----|
Occupation |      3      6      7      8      9      10
            |-----+-----|
11 12 13 | Total
-----+-----
      Menial |      0      2      0      0      3      1
3     12 | 2 | 31
      | 0.00 6.45 0.00 0.00 9.68 3.23
9.68 38.71 | 6.45 | 100.00
-----+-----
      BlueCol |      1      3      1      7      4      6
5     26 | 7 | 69
      | 1.45 4.35 1.45 10.14 5.80 8.70
7.25 37.68 | 10.14 | 100.00
-----+-----
      Craft |      0      3      2      3      2      2
7     39 | 7 | 84
      | 0.00 3.57 2.38 3.57 2.38 2.38
8.33 46.43 | 8.33 | 100.00
-----+-----
```

```
WhiteCol |      0      0      0      1      0      1
2     19 | 4 | 41
      | 0.00 0.00 0.00 2.44 0.00 2.44
4.88 46.34 | 9.76 | 100.00
-----+-----
      Prof |      0      1      1      0      0
2     13 | 10 | 112
      | 0.00 0.00 0.89 0.89 0.00 0.00
1.79 11.61 | 8.93 | 100.00
-----+-----
      Total |      1      8      4      12      9      10
19    109 | 30 | 337
      | 0.30 2.37 1.19 3.56 2.67 2.97
5.64 32.34 | 8.90 | 100.00
-----+-----
Occupation |      3      6      7      8      9      10
            |-----+-----|
11 12 13 | Total
-----+-----
      Menial |      0      2      0      0      3      1
3     12 | 2 | 31
      | 0.00 6.45 0.00 0.00 9.68 3.23
9.68 38.71 | 6.45 | 100.00
-----+-----
```

```
-----+-----
BlueCol |      1      3      1      7      4      6
5     26 | 7 | 69
      | 1.45 4.35 1.45 10.14 5.80 8.70
7.25 37.68 | 10.14 | 100.00
-----+-----
      Craft |      0      3      2      3      2      2
7     39 | 7 | 84
      | 0.00 3.57 2.38 3.57 2.38 2.38
8.33 46.43 | 8.33 | 100.00
-----+-----
```

Automation

1. Data analysis involves repetitive tasks
2. Repetition invites errors due to boredom and fatigue
3. Automation is less error prone and ultimately faster

macros to represent strings of text or numbers

loops to repeat the same commands

returned results to avoid typing the value of a statistical result

matrices to summarize results

Data cleaning, including names and labels

Planning labels

Bad labels

```
. codebook tc1*, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
tc1doc	1074	10	8.714153	1	10	Q46 How important is it to go to ...
tc1fam	1074	10	8.755121	1	10	Q43 How important is it to turn t...
tc1friend	1073	10	7.799627	1	10	Q44 How important is it to turn t...
tc1mhprof	1045	10	7.58756	1	10	Q48 How important is it to go to ...
tc1psy	1050	10	7.567619	1	10	Q47 How important is it to go to ...
tc1relig	1039	10	5.66025	1	10	Q45 How important is it to turn t...

Better labels

```
. codebook tc2*, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
tc2doc	1074	10	8.714153	1	10	Q46 How Impt: Go to a gen med doc...
tc2fam	1074	10	8.755121	1	10	Q43 How Impt: Turn to family for ...
tc2friend	1073	10	7.799627	1	10	Q44 How Impt: Turn to friends for ...
tc2mhprof	1045	10	7.58756	1	10	Q48 How Impt: Go to a mental heal...
tc2psy	1050	10	7.567619	1	10	Q47 How Impt: Go to a psych for Help
tc2relig	1039	10	5.66025	1	10	Q45 How Impt: Turn to a religious...

Even better labels

```
. codebook tc3*, compact
```

Variable	Obs	Unique	Mean	Min	Max	Label
tc3doc	1074	10	8.714153	1	10	Q46 Med doctor help important
tc3fam	1074	10	8.755121	1	10	Q43 Family help important
tc3friend	1073	10	7.799627	1	10	Q44 Friends help important
tc3mhprof	1045	10	7.58756	1	10	Q48 MH prof help important
tc3psy	1050	10	7.567619	1	10	Q47 Psychiatric help important
tc3relig	1039	10	5.66025	1	10	Q45 Relig leader help important

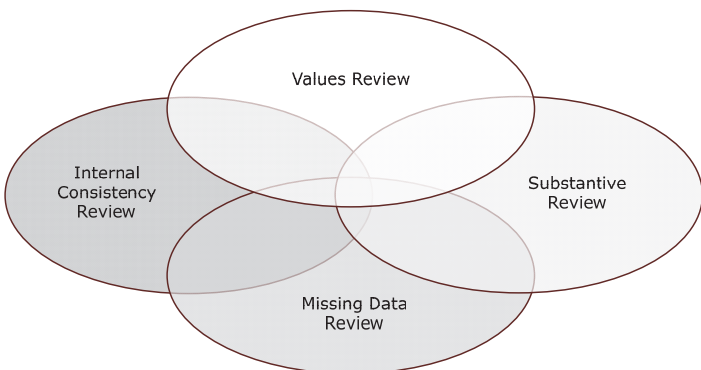
Planning labels

	A	B	C	D
1	Number	Name	Value label	Variable labels
2	1	id_iu		Respondent Number
3	2	cnytry_iu	cnytry_iu	IU Country Number
4	3	vignum	vignum	Vignette
5	4	serious	serious	Q1 How serious would you consider Xs situation to be?
6	5	opfam	Ldummy	Q2_1 What X should do:Talk to family
7	6	opfriend	Ldummy	Q2_2 What X should do:Talk to friends
8	7	tospi	Ldummy	Q2_7 What X should do:Go to spiritual or traditional healer
9	8	tonpm	Ldummy	Q2_8 What X should do:Take non-prescription medication
10	9	oppme	Ldummy	Q2_9 What X should do:Take prescription medication

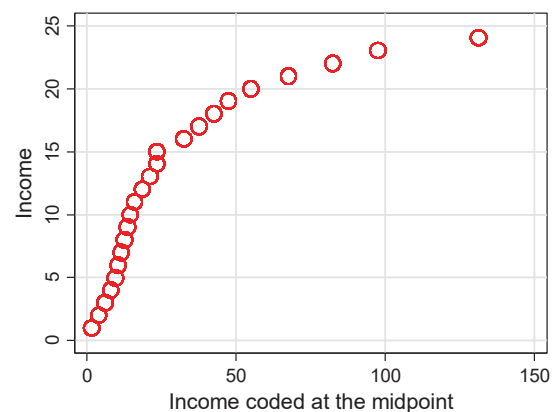
Planning variables names

1. **ownsex** and **ownsexu** caused weeks of delay.
2. Do you want **R003189** or **R001389**?
3. **timetophd** was elapsed time not enrolled time.

Data cleaning (prevents retractions)



Find errors with a graph



Documenting provenance

The provenance of every number must be fully documented.

1. The circled text contains results I may need to confirm later:

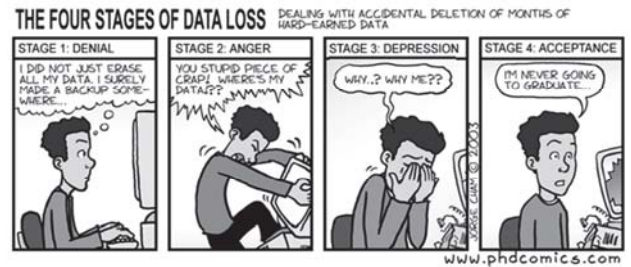
1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons, (.48 and .73, respectively ($z=2.55, p<.01$)) However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (.76 for non-

2. Turning on "show/hide ¶" reveals the provenance:

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-hrtemp4.do #4 jsl 17May06])) However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

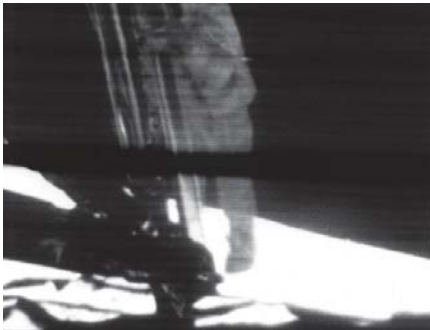
Preserving your data

When it comes to saving your work, expect things to go wrong, expect that you will delete the wrong file at the worst possible time, and expect a hose to be left on in the room above your computer. If you expect the worst, you might be able to prevent it.



Examples of data loss

1. Kennedy assassination on November 22, 1963 and the 9/11 survey
2. 508K volumes in obsolete formats at British Museum
3. Neil Armstrong seen as "a **fuzzy gray blob** wading through an inkwell".



What NASA saw and lost...



But, two tapes were archived by Pink Floyd's video producer!



Dark Side of the Moon

Tactics: Peer to Peer syncing

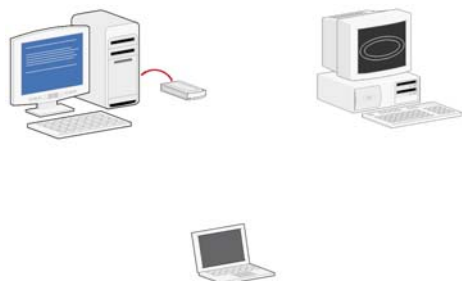


1. Cloud backup and ready availability
2. Easily share files with collaborators
3. Be aware of security issues and what you are syncing

A recent disaster and the advantages of the cloud

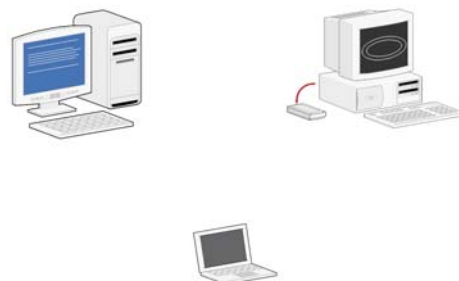
1. A graduate student's computer and backup drives were stolen.
2. He dropped out.

Tactics: Portable drives at home



Reproducible Results and Workflow | 96

Tactics: Portable drive at work



Reproducible Results and Workflow | 97

Preserving bits and preserving content

These files were generated six or seven years ago using Gauss and saved as Gauss FMT files. We need to revise a paper and need the data in these files, but I don't seem to be able to open them. We only have a very old version of Gauss that might not run anymore. Any ideas?

Reproducible Results and Workflow | 98

Conclusions

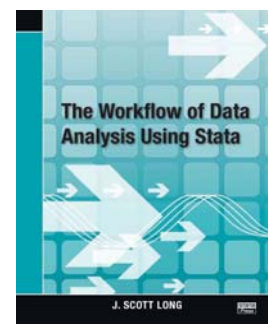
- Expectations for replication and reproduction are growing
- This positive development “raises the bar”

Changing your workflow

- Slowly, systematically, thoughtfully
- Finish the last 5% of the change
- Do not do it under deadline

Whose workflow

- There are **many** viable workflows, but it is nice to have your workflow written down.



Reproducible Results and Workflow | 99

Questions?

Let me know what's happening in your field.

Thank you!

Source: wf icptr block 2016-07-27.docx
Provenance: 2009-07-12; 2009-07-31; 2009-11-12; 2010-07-29; 2010-09-03; 2010-11-04; 2011-04-18; 2011-09-09; 2012-09-04; 2013-03-04; 2013-10-21; 2014-06-30; 2014-07-02; 2014-08-25; wf wim 2014-08-25.docx; wim 2015-08-21.docx; wim 2016-01-18v1.docx; wim 2016-01-21.docx; icptr 2016-07-24 updated figures 2016-08-26 wim.

Reproducible Results and Workflow | 100