Indiana University

Reproducible Results and the Workflow of Data Analysis

Scott Long

Departments of Sociology and Statistics

www.indiana.edu/~jslsoc/ftp/

Workshop in Methods | January 2018

The reproducible results movement

- Open science
- Transparancy in science
- Teaching integrity in research

Changing expectations for researchers

- NAS 2018 Committee on Reproducibility and Replicability in Science
- Journals require data and analysis files
- Funding agencies strengthen requirements for data access
- Haverford College requires reproducibilty for undergraduates

With access comes accountability

- Retraction Watch (retractionwatch.com)
- Recent examples of flawed research…

Retraction due to coding error

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

[redacted]

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 701 marriages from the Health and Retirement Study (1992–2010) to examine the role of serious 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 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 via biological and gendered social pathways.

Keywords

aging, chronic disease, gender, [illegible], physical health

A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are the married healthier than the unmarried (e.g., Lillard and Waite 1995; Umberson 1992), but studies find that both divorce and widowhood are predictors of declines in physical and mental health (e.g., Hughes and Waite 2009; Williams and Umberson 2004). Less attention, however, has been paid to how health may be a determinant of marital status. Work in this area has tended to focus on the positive selection of the healthier into marriage (e.g., Byrne et al. 1989; Smith and Smith 2010), but poor health may be an equally important force for selection. Booth, and Johnson 2008). 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 (Teachman 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

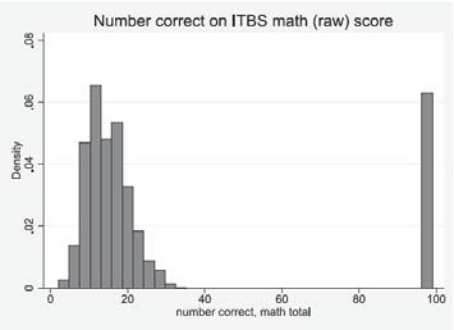
Incorrect data in published research

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 consequenc assignment New York City School Choice Scholarship Program across the distri achievement. Our analyses suggest that the program had negligible and statistic effects across the skill distribution. In addition to contributing to the literature o the article illustrates several ways in which distributional effects estimation can e research: First, we demonstrate that moving beyond a focus on mean effects est possible to generate and test new hypotheses about the heterogeneity of educational that speak to the justification for many interventions. Second, we demonstrate t effects can uncover issues even with well-studied data sets by forcing analysts to new ways. Finally, such estimates highlight where in the overall national achievemen scores of children exposed to particular interventions lie; this is important for expl validity of the intervention's effects.



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 [redacted] article are fragile—so fragile that they appear largely a function of these possible misclassifications and other methodological choices. Our replication and reanalysis of [redacted] 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.

Is science broken?

Misconduct, fraud, and retractions

- Peer review was circumvented at prestigious journals
- Two journals published Maggie Simpson & Edna Krabappel's "Fuzzy, Homogeneous Configurations"
- Retraction at *Science* when data not found

Science Isn't Broken by Christie Aschwanden

"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 reproduction of results

Reproducibility requires identical results with the *same data*.

Replicability requred confirmation of results with *new data*.

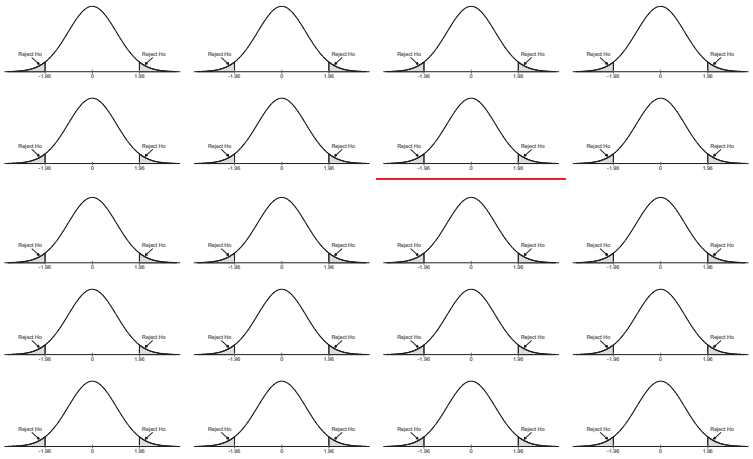
Replicability	*Reproducibility*	
	High	Low
High	*Scientific Ideal*	*Careless Research*
Low	*Fragile Findings*	*Unscientific Work*

Challenges to replicability

Sample driven analyses

1. Decisions based on unique characteristics of the sample invalidate statistical tests.
2. Examples
   - Data mining portrayed as theory testing
   - Post analysis hypothesis construction
   - Undocumented specification searches and p-hacking
   - "Cherry picking" the sample
3. Consider the effect on the sampling distribution of a test statistics…

If α=0, twenty tests of H_0: α=0 at the 5% level



Example of using sample to select a model

1. Randomly select six sub-samples.
2. Use stepwise logit to select a model predicting diabetes.
3. Seven different models were selected.

Variable	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
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***	

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

Model variability versus sampling variability

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

- Estimates are sensitive to credible changes in model specification.
- Point estimates capture just ‘‘one ad-hoc route through the thicket of possible models’’ (Leamer 1985:308)
- For example, do higher income tax rates cause taxpayers to ‘‘vote with their feet’’ and migrate to states with lower taxes?

Effects of tax rate on migration

- Estimate is significant in only 1.5% of 24,567 models.
- The mean estimate is roughly zero.

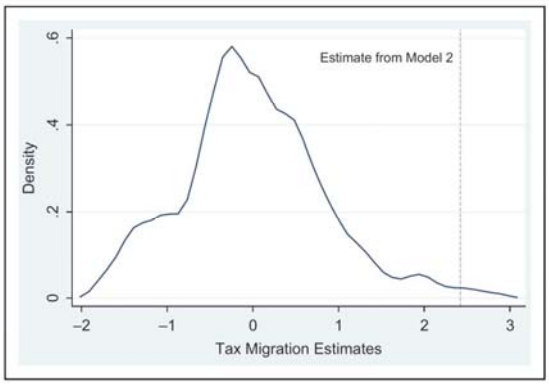


Figure 4. Modeling distribution of tax migration estimates.
Note: Kernel density graph of estimates from 24,576 models.

Reproducibility with same data

Changing expectations for reproducibility

- AJPS requires verification of results before a paper is published.
  - Only ~5 of 200 submissions succeeded.
- Many jounals required that data and script files are distributed.

Challenges to reproducibility

- Reproducibility requires a systematic workflow built around the requirement of reproduciblity.

My talk focuses on the workflow for reproducibility

What is a workflow for data analysis?

A workflow is a set of *coordinated* procedures for all aspects of data management, analysis, and presentation.

- Planning research
- Organizing and documenting
- Importing and cleaning data
- Analyzing data
- Presenting and publishing results
- Revising results
- Preserving files

You 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.
   - It takes time, but saves more time.
   - It prevents errors.
   - It makes you a better data analyst.
   - It is critical for reproducibility.

Origins of the workflow project

1. Incorrect results with clever explanations
2. Dissertation delayed 18 months to determine provenance
3. Unreproducible results from a 743 line do-file
4. Analyzing the wrong data set:

   "The datasets are exactly the same except for the married variable."
5. The wrong variable when writing a report for the NAS
6. Mislabeled gene in a study of alcoholism
7. Collaborations that multiply the ways things go wrong
8. Misleading output such as...

Definitel a problem

```
. tabulate female sdchild_v1

     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…

```
. 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...
-------------------------------------------------------------------------------
```

Confusing output

```
. tab occ ed, row

           |                                   Years of education
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
-----------+-----------------------------------------------------------------------------
-----------------------+----------
```


Why is learning WF difficult?

Tacit knowledge

1. *Explicit knowledge* is the stuff of textbooks and articles.
2. *Tacit knowledge* is implicit and undocumented (Polanyi).
   - People are unaware of essential tacit knowledge.
     - Henry Bessemer's 1855 patent for steel
   - Tacit knowledge is transferred *at the bench*.
     - Personal computing makes this harder

Data analysis is hard

There's a lot of *undifferentiated heavy lifting* that stands between your idea and that success. – Jeff Bezos

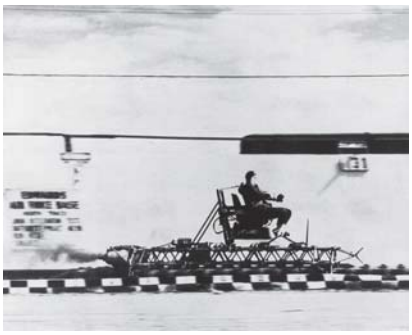
Resources for learning about workflow

- Stat 612: Reproducible Results and Managing Data Analysis
- *The Workflow of Data Analysis Using Stata*
- ICSPR Summer Program Workshop

These materials focus on Stata, but the principles work with any software.

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…

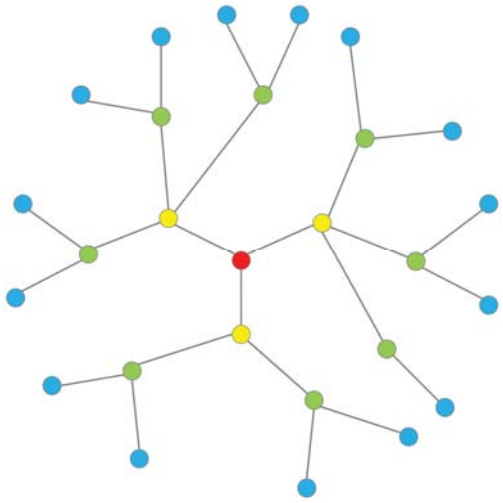
"I was fine, only blind for a few days."

Why are results hard to reproduce?

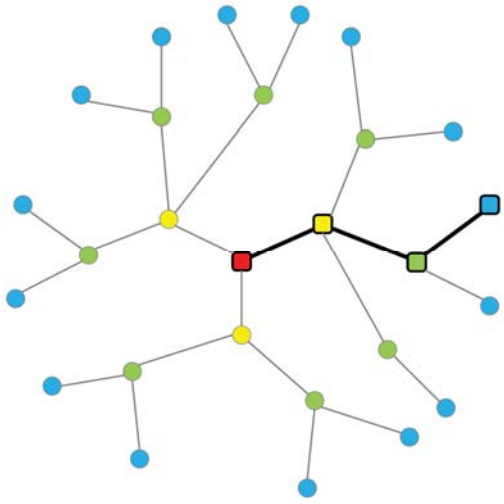
1. **The curse of dimensionality**: Research involves many decisions.
   - Where to truncate a va riable?
   - What seed for the RN generator?
   - How to scale with partially missing data?
   - Which cases to keep for analysis?
   - How to code education?
   - What values to assign to income greater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

With only 10 such decisions, there are 1,024 combinations.

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? (continued)

2. With missing documentation, you might not find the right path.
3. Changes in software can lead to different results.
   - A colleague spent weeks to reproduce results because he forgot `version 7` in a do-file.
4. Lost or changed files make reproducibility impossible.
   - Retraction in *Science* because of lost data
   - $2,000 to retrieve a file that was "backed up"
   - Virtual servers might have 30 day rolling backups

Criteria for choosing your workflow

Accuracy

- Given reproducibility, you want the correct result

Efficiency

- Completing work quickly
- Working quickly competes with accuracy
- Requires investing time to save time

Scalability

- Adapts to projects of different sizes
- Works with individuals and teams

Standardization

- Uniform decisions for how to do things
- Increases efficiency and accuracy

Automation

- Saves times and prevents errors
- Time learning automation saves time executing

Usability

- If you won't do it, it is not a good workflow

Transferability

- Can someone else continue your work?

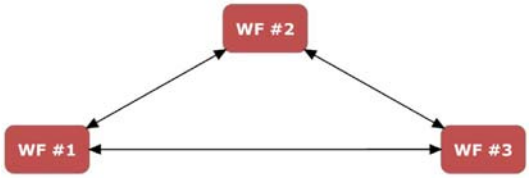
Collaboration and workflow

1. Collaboration makes it harder to have an effective workflow.
2. Why is workflow harder when you collaborate?

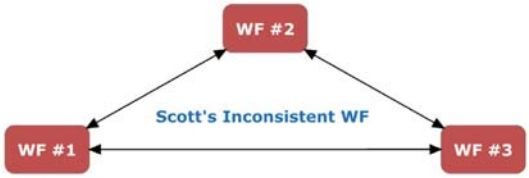
Coordinating multiple workflows

Coordinating multiple workflows

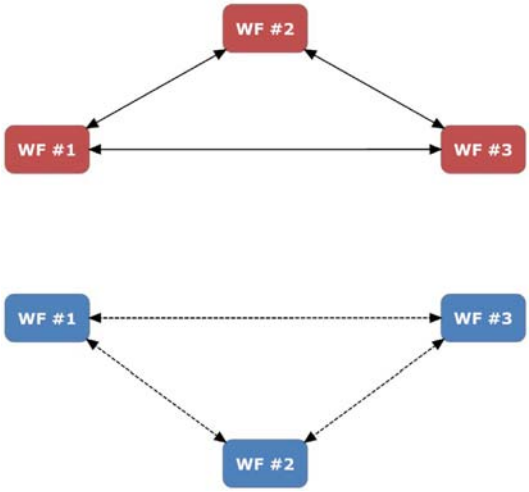
Coordinating multiple workflows



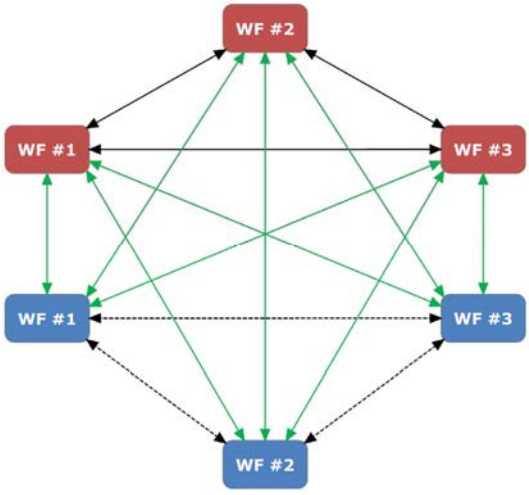
Coordinating multiple workflows



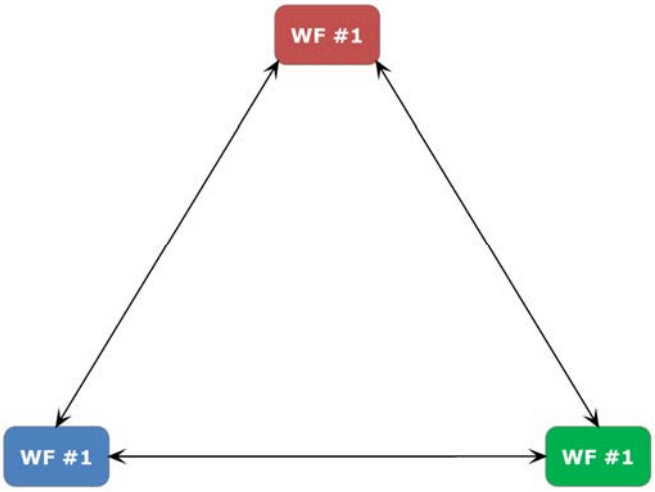
Coordinating multiple workflows starts here



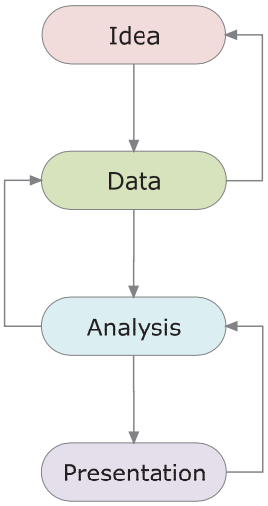
Coordinating 30 pairs of workflows



Coordinating multiple workflows: Agree on a WF



Steps in your workflow



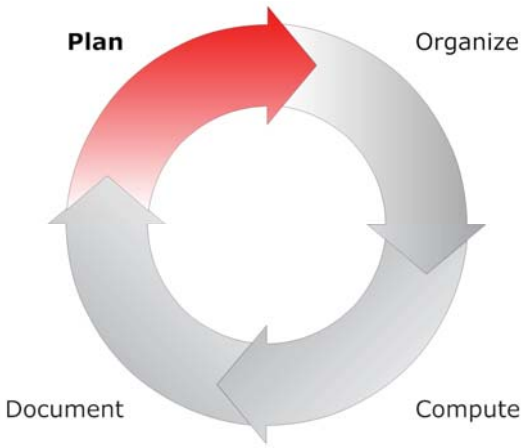
Substantive goals for your research

Can be 90% of the work

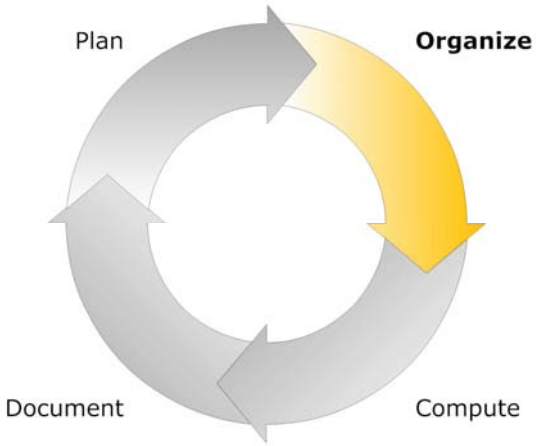
Often the simplest step

Writing and maintaining provenance

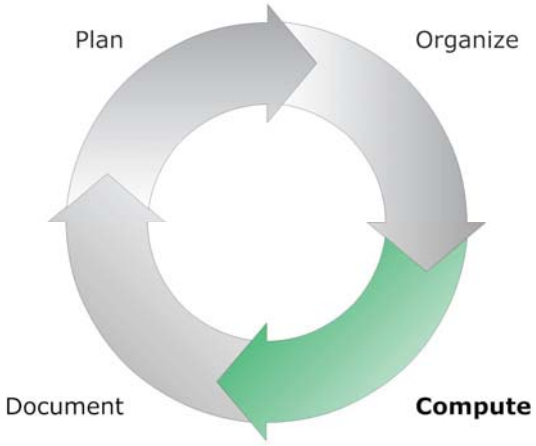
Tasks within each step



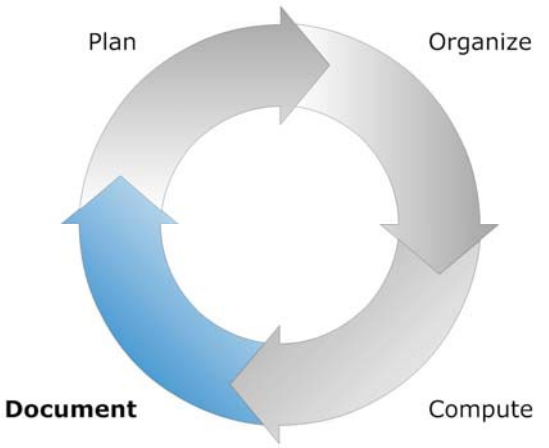
Tasks within each step



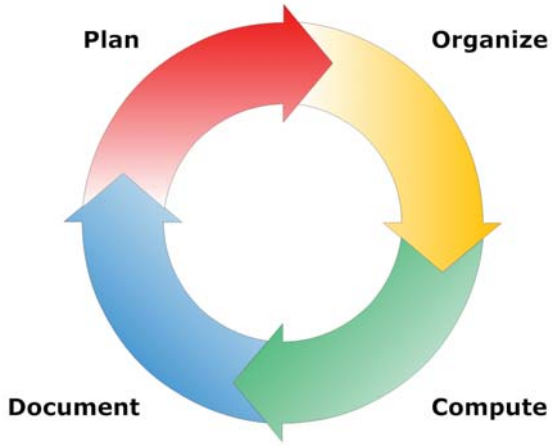
Tasks within each step



Tasks within each step



Tasks within each step



Planning

- The project timeline
- Division of labor
- Scheduling your time
- How to document and organize research
- Variable names, labels and metadata
- Procedures for missing data
- Analyses
- Writing
- Preserving files
- And more…

Blau and Duncan's *The American Occupational Structure*

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

Michael Faraday's advice

Work. Finish. Publish.

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

Organizing

1. Organization is motivated by two goals
   - o *Finding things*
   - o *Avoiding duplication*
2. Organization
   - o Lets you work faster
   - o Rewards consistency and uniformity
   - o Is contagious (and so is disorganization)
   - o Requires regular maintainance to overcome entropy

Signs of poor organization

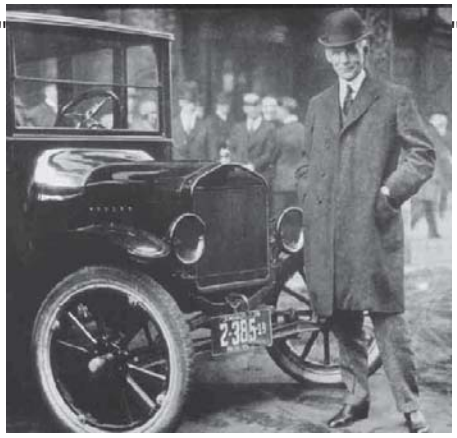
1. *Can't find a file* and you think you deleted it.
2. *Multiple versions of a file* and you don't know which is which.
   - o You and a co-author edit different versions of a paper. You have incompatible, incomplete drafts.
   - o You need the file for draft submitted for review, but you have two (or 16) files with "final" in the name.

     This: final report v16.docx

     Or this: NSF report 2010-10-21.docx
3. *Finally*, after this talk a student showed me a text:
   - o Urgent: don't analyze **`final.dta`**, use **`lastversion.dta`** for presentation tomorrow."
   - o Surely this is a rooky mistake….

The final paper

Organization should be like a Model T



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.
   - 115,000 files on a research center's LAN
   - 2,000,000 files accumlated in 10 years

Files are scatter across multiple, overlapping locations

1. Office computer	2. Home computer
3. Laptop	4. LAN
5. Dropbox	6. Box
7. USB sticks	8. Old laptop
9. External drives	10. Mom's computer

Operating systems organize files for 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?
- How many PDFs do you have of the same article?

How to manage files

1. Name files carefully and systematically.
2. Use a planned directory structure.
   - Every file has one place it belongs.
   - A file's location documents the file.

For example...

A planned set of primary directories

\- To shelve	Files to put in the correct directory
\Active	Active projects
\Admin	Administration and service
\Bookshelf	Books, articles, reprints, etc.
\Inactive	Projects that are on hold
\Shared	Files shared with others on the cloud
\Teaching	Teaching materials
\Templates	Files used as templates
\Vault	Completed work that will *never* change

A structure for projects in \Active, \Inactive and \Teaching

\Group Differences

- **\- Hold then delete**
- **\- To shelve**
- **\Admin**
- **\Posted**
- **\Resources**
- **\Work**
- **\Write**

File naming

Writing

- groups 2017-11-07.docx
- groups 2018-01-17.docx

PDF files

- Long 1978 ASR productivity position.PDF

Datasets

- groups-hrs1.dta
- groups-hrs2.dta

Script files

- groups-data03-recoding.do
- groups-data04-scales.do

Organization: uniform formats for robust script files

```
capture log close
log using wftalk01-example, replace text
version 15.1
clear all
set linesize 80

// project: wf talk
local pgm wftalk01
local dte 2018-01-18
local who Scott Long
local tag "`pgm'.do `who' `dte'"
di "`tag'" // for provenance

// #1 describe task

// #2 describe task

log close
exit
```


Documentation

1. Without documentation,
   - Reproduction is much more difficult.
   - Mistakes are more likely.
   - Work takes longer.
2. *Long's Law*: It is faster to document it today than tomorrow.
   - Nobody likes to write documentation.
   - Nobody regrets having documentation.
3. More codified fields demand documentation.
   - *The Research Log* (American Chemical Society)

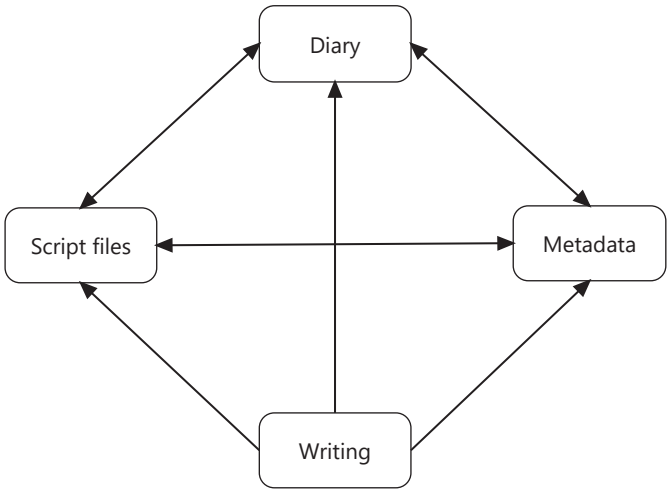
Suggestions for writing documentation

1. Do it today.
2. Check it next week even if it makes sense today.
3. Review it at key stages of your work, like finishing a draft.
4. Include full dates and names.
5. Use reinforcing, non-redundant forms of documentation.
6. Start with a research diary for each project.

A simple research diary

```
First complete set of analysis for FLIM measures paper

f2alt01a.do - 24May2002

 Descriptive information on all rhs, lhs, and flim measures

f2alt01b.do - 25May2002

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

 **  Outcome: Can Work                  global lhs "qcanwrk95"
 **  Outcome: Work in three categories  global lhs "dhlthwk95"
 **  Outcome: bath trouble              global lhs "bathdif95"
 **  Outcome: adlsum95 - sum of adls    global lhs "adlsum95"

f2alt01c.do - 25May2002

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

 *     1.  ln(x+.5) and ln(x+1)
 *     2.   9 counts: >=5=5  >=7=7   (50% and 75%)
 *     3.   8 counts: >=4=4  >=6=6   (50% and 75%)
 *     4.   18 counts: >=9=9 >=14=14  (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 flim measures
 (fla*p5; fll*p5); pairs of u/l measures; groups of LCA measures

f2alt01e.do - all LCA probabilities - 25May2002

:::

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

:::

f2alt02c.do - 29May2002

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


Reinforcing forms of documentation



Execution and computing

1. Execution involves carrying out tasks within each step.
2. Effective execution requires mastering *tools*.
   - Software
     - File manager
     - Macro program
     - Text editor
     - Statistical software
   - Hardware
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.**

Laptop 2009

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 compute any time they want to.

   Planners compute only 12 yours a week.

2. Who will finish their dissertation first?

Principles for a computing workflow

1. Legible and robust script files
2. Posting files
3. Dual workflow for data management and analysis
4. Run order naming of scripts

Robust and legible script files

1. Programs must run on another computer without *any* changes.
   - Self-contained
   - Version control
   - No hard coded directory information
   - Explicit seeds for random numbers
   - Archived user written programs
2. Careful internal documentation of what the script does.
3. Formatting to improve legibility.

The *essential* posting principle

1. Posting is defined by two simple rules.

   The share rule

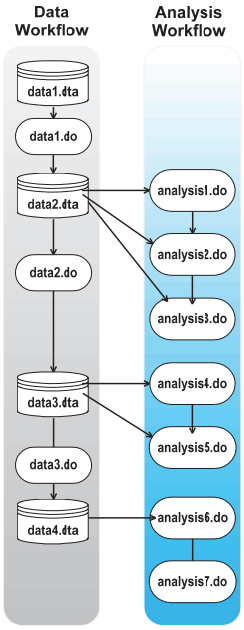
   Only share results after the files are posted.

   The no change rule

   Once a file is posted, *never* change it.

2. Without posting, you cannot reproduce your results.
   - If you don't have the dataset, how can you confirm the results?
   - If you don't have the scripts, how were the results produced?

Dual workflow and run order naming



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

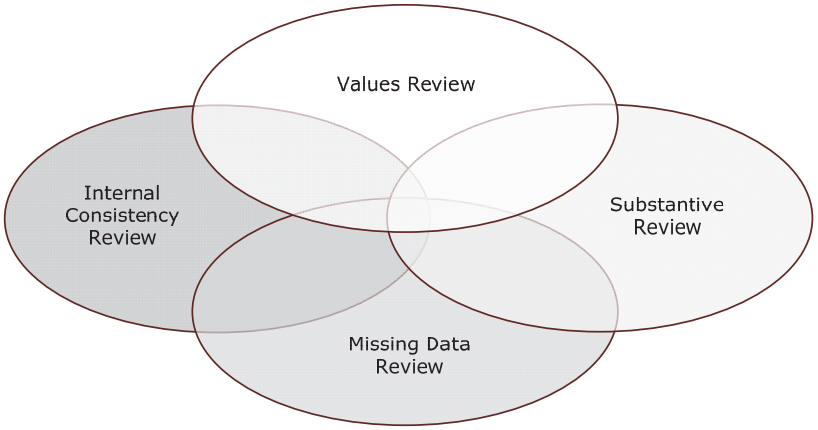
	A	B	C	D
1	Number	Name	Value label	Variable labels
2	1	id_iu		Respondent Number
3	2	cntry_iu	cntry_iu	IU Country Number
4	3	vignum	vignum	Vignette
5	4	serious	serious	Q1 How serious would you consider Xs situation to be?
6	5	opfam	Ldummy	Q2_1 What X should do:Talk to family
7	6	opfriend	Ldummy	Q2_2 What 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

Cost of not planning variables names

1. Confusion between **`ownsex`** and **`ownsexu`** caused weeks of delay.
2. Do you want **`R003189`** or **`R001389`**?
3. Is **`timetophd`** elapsed time or enrolled time?

Data cleaning and preventing retractions

Statistical analysis assumes the variables are clean.



A two-way table would have detected the problem

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

[redacted]

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 7,701 marriages from the Health and Retirement Study (1992–2010) to examine the role of serious 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 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 via both financial and gendered social pathways.

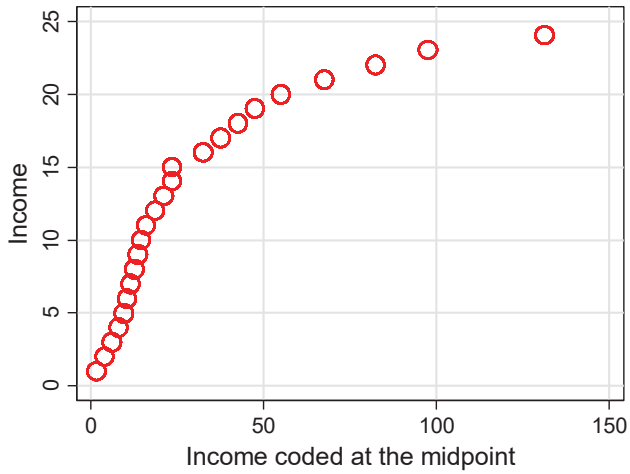
Keywords

aging, chronic disease, gender, [illegible] physical health

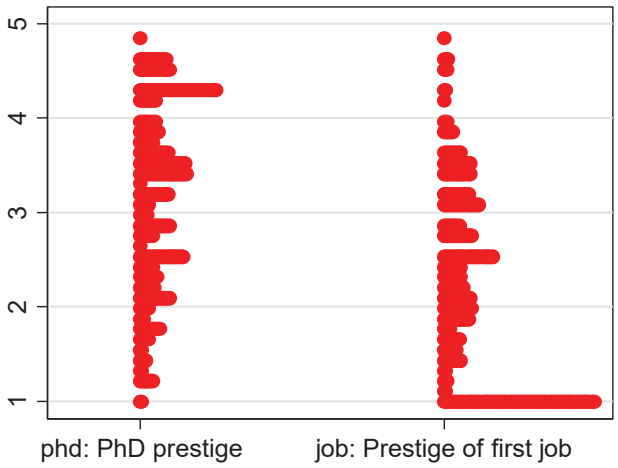
A large body of literature has identified marital status as a strong predictor of health and well-being. Not only are the married healthier than the unmarried (e.g., Lillard and Waite 1995; Umberson 1992), but studies find that both divorce and widowhood are predictors of declines in physical and mental health (e.g., Hughes and Waite 2009; Williams and Umberson 2004). Less attention, however, has been paid to how health may be a determinant of marital status. Work in this area has tended to focus on the positive selection of the healthier into marriage (e.g., Byrne et al. 1989; Smith and Smith 2010), but poor health may be an equally important force for selection. Booth, and Johnson 2008). 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 (Teachman 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

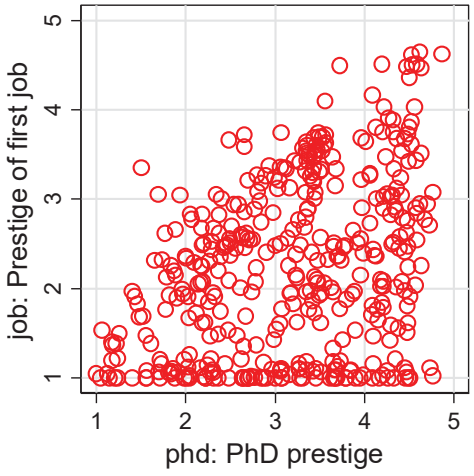
Use graphs to find errors



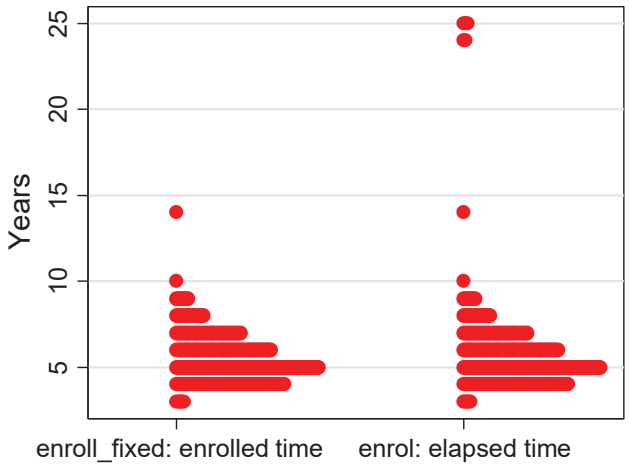
Graphs highlight forgotten coding decisions



Locating outliers and gaining substantive insights



Avoiding expensive mistakes from misread documentation



Statistical analysis

This can be the simplest part of the project.

1. *Take classes* and go to talks on data analysis.
2. *Find exemplars* in the best journals.
3. *Use automation* and script files.
4. Maintain a *dual workflow* to prevent errors.

Presentations and provenance

1. *Content and methods* are disciplinary decisions.
2. *Standards for presentations* vary by discipline.
   - Bad presentations transcend disciplines
3. *Maintaining provenance* is critical for reproducibility.

Documenting provenance

The provenance of every number must be documented.

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

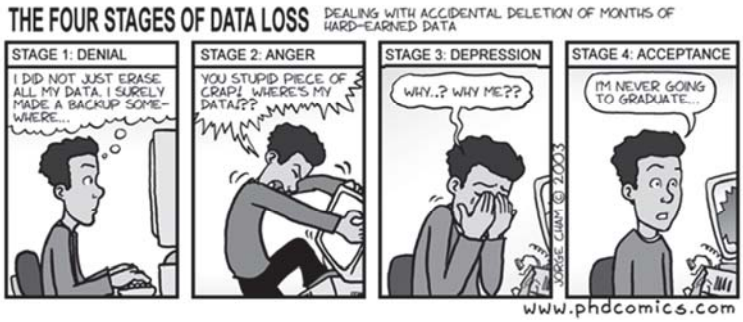
1922-1926 cohort, employed women have fewer limitations than those who are out for family reasons,(.48 and .73, respectively (z=2.55, p<.01).)However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (.76 for non-

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-hrmemp4.do #4 jsl 17May06}).) However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have

Preserving your files

Expect things to go wrong, expect to 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 prevent it.

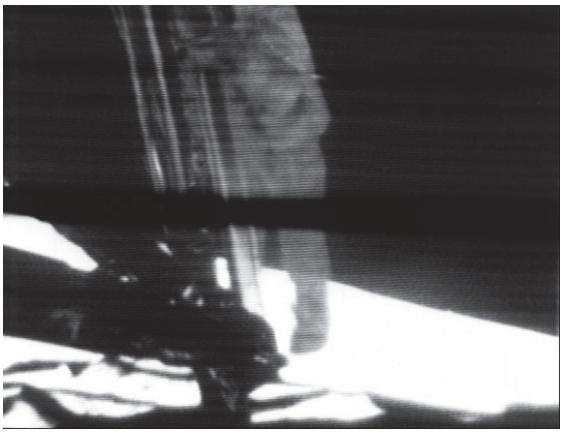


Examples of data loss

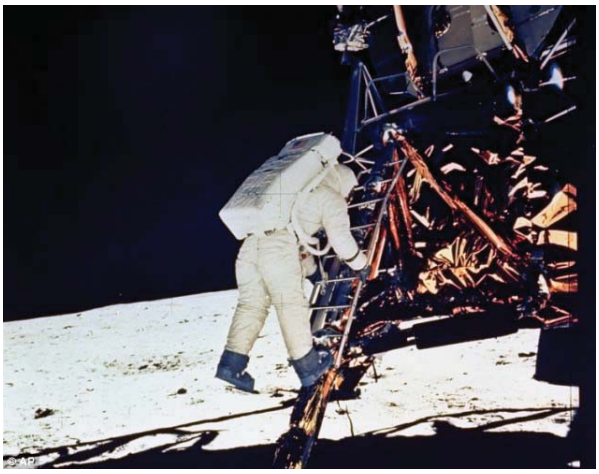
1. Water leaked above ICPSR server room.
2. 508K volumes are in obsolete formats at British Museum.
3. Data for Wolfgang's *Delinquency in a Birth Cohort* burned.
4. NASA lost 1000s of moon tapes.

Neil Armstrong's walk on the moon

In 1969, America saw "a *fuzzy* gray *blob* wading through an inkwell".



NASA had video that was too good to show on TV, but lost the tapes.



The produce for Pink Floyd rock video archived two moon tapes!

Dark Side of the Moon

Preserving files does *not* preserve content

"These files were saved six years ago as Gauss FMT files. We need to revise a paper and need the data in these files, but I can't open them. We have an old version of Gauss that doesn't run anymore. Any ideas?"

Conclusions

Replicability and reproducibility

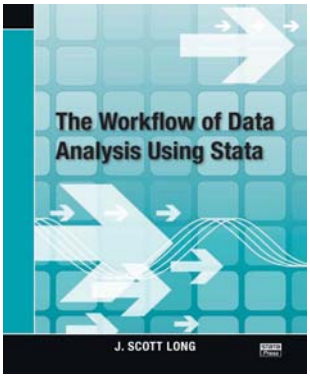
- Expectations for replicability and reproducability are growing.
- These *positive* developments increase demands for researchers.
- An effective workflow is essential.

Changing your workflow

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

Whose workflow

- There are *many* viable workflows, but it is nice to have a place to start.



Thank you!