Reproducible Results and the Workflow of Data Analysis

Scott Long
Departments of Sociology and Statistics
www.indiana.edu/~jsloc/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

Assessing the fragility of published results
Measurement, methods, and divergent patterns: Reassessing the effects of same-sex parents

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**

![Diagram of Classical Model of Inference](https://example.com/classical_model_diagram)

Vic Barnett, Comparative Statistical Inference

---

**Multiple tests at 5% level**

![Diagram of Multiple Tests at 5% Level](https://example.com/multiple_tests_diagram)

---

**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

<table>
<thead>
<tr>
<th>Variable</th>
<th>explore1</th>
<th>verify1</th>
<th>explore2</th>
<th>verify2</th>
<th>explore3</th>
<th>verify3</th>
</tr>
</thead>
<tbody>
<tr>
<td>bmi</td>
<td>1.067***</td>
<td>1.066***</td>
<td>1.004</td>
<td>1.074***</td>
<td>1.101***</td>
<td>0.971</td>
</tr>
<tr>
<td>white</td>
<td>0.518***</td>
<td>0.547***</td>
<td>0.521***</td>
<td>0.543***</td>
<td>0.509***</td>
<td>0.562***</td>
</tr>
<tr>
<td>age</td>
<td>1.242***</td>
<td>1.251***</td>
<td>1.324***</td>
<td>1.288***</td>
<td>1.292***</td>
<td>1.341***</td>
</tr>
<tr>
<td>agesq</td>
<td>0.999***</td>
<td>0.998***</td>
<td>0.998***</td>
<td>0.998***</td>
<td>0.998***</td>
<td>0.998***</td>
</tr>
<tr>
<td>hsdegree</td>
<td>0.720***</td>
<td>0.680***</td>
<td>0.662***</td>
<td>0.749***</td>
<td>0.780***</td>
<td>0.650***</td>
</tr>
<tr>
<td>weight</td>
<td>1.006***</td>
<td>1.006***</td>
<td>1.016***</td>
<td>1.004**</td>
<td></td>
<td>1.043***</td>
</tr>
<tr>
<td>height</td>
<td></td>
<td>0.936**</td>
<td></td>
<td></td>
<td>0.909***</td>
<td></td>
</tr>
<tr>
<td>female</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
</tr>
<tr>
<td>_cons</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
<td>0.000***</td>
</tr>
</tbody>
</table>

N: 8036, 8035, 8036, 8035, 8036, 8035

---

**Model Robustness**

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

![Figure 1: Modeling distribution of union wage premium](https://example.com/model_robustness_graph)

*Figure 1.* Modeling distribution of union wage premium. Note: Kernel density graph of estimates from 1000 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

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

Why must the workflow be coordinated?

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 rejections

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

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. Mislabeled gene in a study of alcoholism
8. Collaborations that multiply the ways things go wrong
9. Misleading output such as...
**Definitel in a $3M study**

`tabulate female sdchild_v1`

<table>
<thead>
<tr>
<th>X is</th>
<th>Q15 Would let X care for children female?</th>
<th>Defintel</th>
<th>Probably</th>
<th>Definitel</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>41</td>
<td>99</td>
<td>155</td>
<td>197</td>
<td>492</td>
</tr>
<tr>
<td>Female</td>
<td>73</td>
<td>98</td>
<td>156</td>
<td>215</td>
<td>542</td>
</tr>
<tr>
<td>Total</td>
<td>114</td>
<td>197</td>
<td>311</td>
<td>412</td>
<td>1,034</td>
</tr>
</tbody>
</table>

**How important is it…**

`codebook tc1*, compact`

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Unique</th>
<th>Mean</th>
<th>Min</th>
<th>Max</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>tc1doc</td>
<td>1074</td>
<td>10</td>
<td>8.71</td>
<td>1</td>
<td>10</td>
<td>Q46 How important is it to go to …</td>
</tr>
<tr>
<td>tc1fam</td>
<td>1074</td>
<td>10</td>
<td>8.75</td>
<td>1</td>
<td>10</td>
<td>Q43 How important is it to turn t…</td>
</tr>
<tr>
<td>tc1friend</td>
<td>1073</td>
<td>10</td>
<td>7.79</td>
<td>1</td>
<td>10</td>
<td>Q44 How important is it to turn t…</td>
</tr>
<tr>
<td>tc1mhprof</td>
<td>1045</td>
<td>10</td>
<td>7.59</td>
<td>1</td>
<td>10</td>
<td>Q48 How important is it to go to …</td>
</tr>
<tr>
<td>tc1psy</td>
<td>1050</td>
<td>10</td>
<td>7.57</td>
<td>1</td>
<td>10</td>
<td>Q47 How important is it to go to …</td>
</tr>
<tr>
<td>tc1relig</td>
<td>1039</td>
<td>10</td>
<td>5.66</td>
<td>1</td>
<td>10</td>
<td>Q45 How important is it to turn t…</td>
</tr>
</tbody>
</table>

**Which number is which?**

`tab occ ed, row`

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Years of education</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>BlueColl</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Craft</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>WhiteColl</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

**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 enthographer 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
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?

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...

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?

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.

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

"I was fine, only blind for a few days."
Given reproducibility, criteria for choosing WF

1. **Accuracy**
   - If your program is not correct, then nothing else matters.  
   --Oliveira and Stewart

2. **Efficiency**
   - Complete work quickly
   - Working quickly competes with working accurately

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

4. **Automation**
   - Automated procedures prevent mistakes
   - Time invested learning automation can save time

5. **Simplicity**
   - Unnecessarily complicated procedures are abandoned

6. **Usability**
   - Your workflow should reflect the way you like to work
   - 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 starts here

Coordinating 30 pairs of workflows

Coordinating multiple workflows: Agree on a WF

Get an enforcer

Steps in your workflow

1. What motivates your research
2. 90% of the work unless you hurry
3. Often the simplest step
4. Maintaining provenance
Tasks within each step

- Plan
- Organize
- Document
- Compute

Tasks within each step

- Plan
- Organize
- Document
- Compute

Tasks within each step

- Plan
- Organize
- Document
- Compute

Tasks within each step

- Plan
- Organize
- Document
- Compute
**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.
   - 115,000 files on a research center’s LAN
   - 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

<table>
<thead>
<tr>
<th>Win</th>
<th>Mac</th>
</tr>
</thead>
<tbody>
<tr>
<td>Desktop</td>
<td>Desktop</td>
</tr>
<tr>
<td>Music</td>
<td>Music</td>
</tr>
<tr>
<td>Pictures</td>
<td>Pictures</td>
</tr>
<tr>
<td>Videos</td>
<td>Movies</td>
</tr>
<tr>
<td>Documents</td>
<td>Documents</td>
</tr>
</tbody>
</table>

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

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

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

\- 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...*

Execution and computing

1. Execution involves carrying out tasks within each step.
2. Effective execution requires *the right tools*.
   - **Software**
     - *File manager: Explorer and Finder do not work well*
     - *Macro program: don’t retry the same thing*
     - *Text editor: use just one for all programs*
     - *Word processor: page feed, outlines, headings*
     - *Statistical software*
   - **Hardware**: display, storage, memory, CPU
3. Planning is more important than computing power.
   - Consider the changes in computing...
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.
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

---

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 | 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

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>
| 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 | 19 |
| 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 | 94 | 7 | 84 | 0.00 | 3.57 | 2.38 | 2.38 | 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 | 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

<table>
<thead>
<tr>
<th>Years of education</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>
| 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 | 19 |
| 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 | 94 | 7 | 84 | 0.00 | 3.57 | 2.38 | 2.38 | 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 | 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

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>
| 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 | 19 |
| 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 | 94 | 7 | 84 | 0.00 | 3.57 | 2.38 | 2.38 | 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 | 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

<table>
<thead>
<tr>
<th>Occupation</th>
<th>3</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Menial</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>11</td>
</tr>
</tbody>
</table>
| 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 | 19 |
| 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 | 94 | 7 | 84 | 0.00 | 3.57 | 2.38 | 2.38 | 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 tcl*, 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...
   tc1prof 1045 10 7.58756 1 10 Q48 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 tcl*, compact
   Variable Obs Unique Mean Min Max Label
   tc2doc 1074 10 8.714153 1 10 Q46 How important is it to go to ...
   tc2fam 1074 10 8.755121 1 10 Q43 How important is it to turn t...
   tc2friend 1073 10 7.799627 1 10 Q44 How important is it to turn t...
   tc2prof 1045 10 7.58756 1 10 Q48 How important is it to go to ...
   tc2relig 1039 10 5.66025 1 10 Q45 How important is it to turn t...

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.

Planning labels
Even better labels
   * codebook tcl*, 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
   tc3prof 1045 10 7.58756 1 10 Q48 Help from mental health care
   tc3relig 1039 10 5.66025 1 10 Q45 Help from religious leader

Find errors with a graph
Data cleaning (prevents retraction)
**Remember coding decisions**

- Diagram showing the relationship between PhD prestige and first job prestige.

**Understanding the substantive process**

- Diagram showing the relationship between PhD prestige and first job prestige.

**Avoiding expensive mistakes**

- Diagram showing the relationship between years and enroll time.

**Analyze data**

1. Take classes in statistics (like the ICPSR Summer program)
2. Go to talks on data analysis
3. Find exemplars

**Presentations and provenance**

1. **Content and methods** are disciplinary decisions
2. **Presentations and provenance** are universal

**Tables too small**

- Table showing data with small font size.

**Colors that aren't**

- Chart showing data with male and female categories.

- Chart showing data that doesn't work for printing.
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 75, respectively (z=2.55, p<.01). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (75 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 75, respectively (z=2.55, p<.01). However, this gap has disappeared for the 1943-1947 cohort and, indeed, employed women have slightly more limitations (75 for non-

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...

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.
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?

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.

Questions?
Let me know what’s happening in your field.

Thank you!