WS_Workflow_Presentation Outline Part 1
Krista K. Payne
July 24, 2017

Four steps in an effective workflow...

1. Cleaning data

   Things to do:

   - Verify your data are accurate
   - Variables should be well named
   - Variables should be properly labeled

   Ask yourself:

   - Do the variables have the correct values?
   - Are missing data coded appropriately?
   - Are the data internally consistent?
   - Is the sample size correct?
   - Do the variables have the distribution you'd expect?

   When developing my system and habits, I have kept the above questions in mind. Hopefully as we go through various examples, you will see this and understand why I do things the way I do.

2. Running analysis

   Typically, the easiest part of the workflow—HOWEVER, it is very easy to get lost when you are running multiple models or using more than one analytic sample.

3. Presenting results

   When moving results from Stata output to your paper or presentation, Long recommends:

   - Automation—not my strong suit
   - Document the provenance of ALL your findings (e.g. preserving the source of your results)
   - Make your presentation effective

4. Protecting files

   Be aware, backing up files and archiving files are two distinct things. The files saved on the server are backed up regularly by the University, that is why it is EXTREMELY important you save your files associated with your work at the Centers on the server.

   When your time at the Center is over, Hseuh-Sheng will archive your data and other items on to DVDs.

Tasks within each step of an effective workflow

1. Planning
2. Organization
3. Documentation
4. Execution (not going to focus on this here)

Planning

Ask yourself the following questions during your planning step:

- What types of analyses are needed?
- How will you handle missing data?
- What new variables must be constructed?

Next, draft a plan of what you need to do based on your answers to the above questions, then create a prioritized list. Some suggestions:

- Might not be a bad idea to place your plan and list at the start of your research log for the project (we'll get into what a research log is later).
- If projects are initiated via email, it's a good idea to save the original emails in the project's digital folder. I also print them out.

Organization

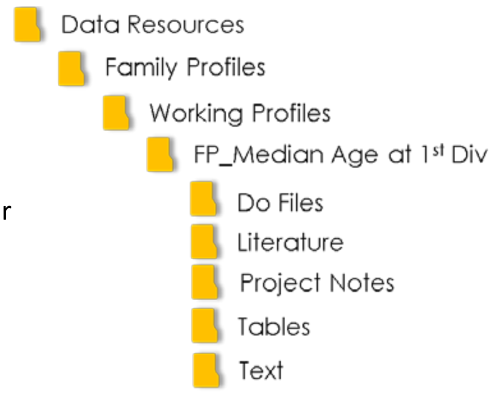
Requires you to think systematically about:

- How you name files and variables
- How you organize directories
- How you keep track of which computer has what information
- Where you store research materials

Some suggestions:

- Start early
- Simple, but not too simple
- Consistency
- Can you find it?
  - Place in the proper folder
    - Start with a carefully designed folder structure
    - When files are created, place them in the correct folder
  - Create and use project abbreviations/prefixes (mnemonics)
- Document your organization

Example: Carefully Designed Folder Structure



Documentation: Keeping track of what you have done and thought

Long's law of documentation: It is always faster to document it today than tomorrow.

Some suggestions:

- Include documentation as a regular part of your workflow.
- Long keeps up with documentation by linking it to the completion of key steps in the project
- Think of it as a public record that someone else could follow
  - "hit-by-a-bus" test…if you were hit by a bus, would a colleague be able to reconstruct what you were doing and keep the project moving forward?

What should you document?

- *Data sources*
  - If using secondary sources keep track of where you got the data and which release you are using
  - Why? Data updates
- *Data decisions*
  - How were variables created and cases selected?
  - Who did the work?
  - When was it done?
  - What coding decisions were made and why?
  - How did you scale the data and what alternatives were considered?
  - For critical decisions, also document why you decided *not* to do something.
- *Statistical analysis*
  - What steps were taken in the statistical analysis, in what order, and what guided those analyses?
  - If you explored an approach to modeling but decided not to use it, keep a record of that as well.
- *Software*
  - What version of Stata was used for coding and analyses?
- *Storage*
  - Where are results archived?
- *Ideas and plans*
  - Ideas for future research and lists of tasks to be completed should be included in the documentation.

Notes:

Levels of documentation

- *The research log –*
    - Cornerstone of your documentation
    - Chronicles
        - The ideas underlying the project
        - The work you have done
        - Decisions made
        - Reasoning behind each step in data construction and analysis
    - Includes
        - Dates when work was completed
        - Who did the work
        - What files were used
        - Where the materials are located
    - Also indicates
        - Other documentation available and where it is located

Mine is in paper form and more organic…not as organized as I'd like it to be. It also tends to become quite cumbersome. I'd like to move to something more digital.

- At the very least, your research log should include in a header:
    - Name of the file (the research log or project notes or whatever you are going to call it)
    - Your name
    - Date the project was initiated
- I also suggest the following:
    - Page numbers
    - Change margins to ½ inch
    - I also tend to change my font to Courier to match the Stata output—also helps with everything lining up, because it is a true type font
        - In the Appendix, I've included an example from Long's book

Notes:

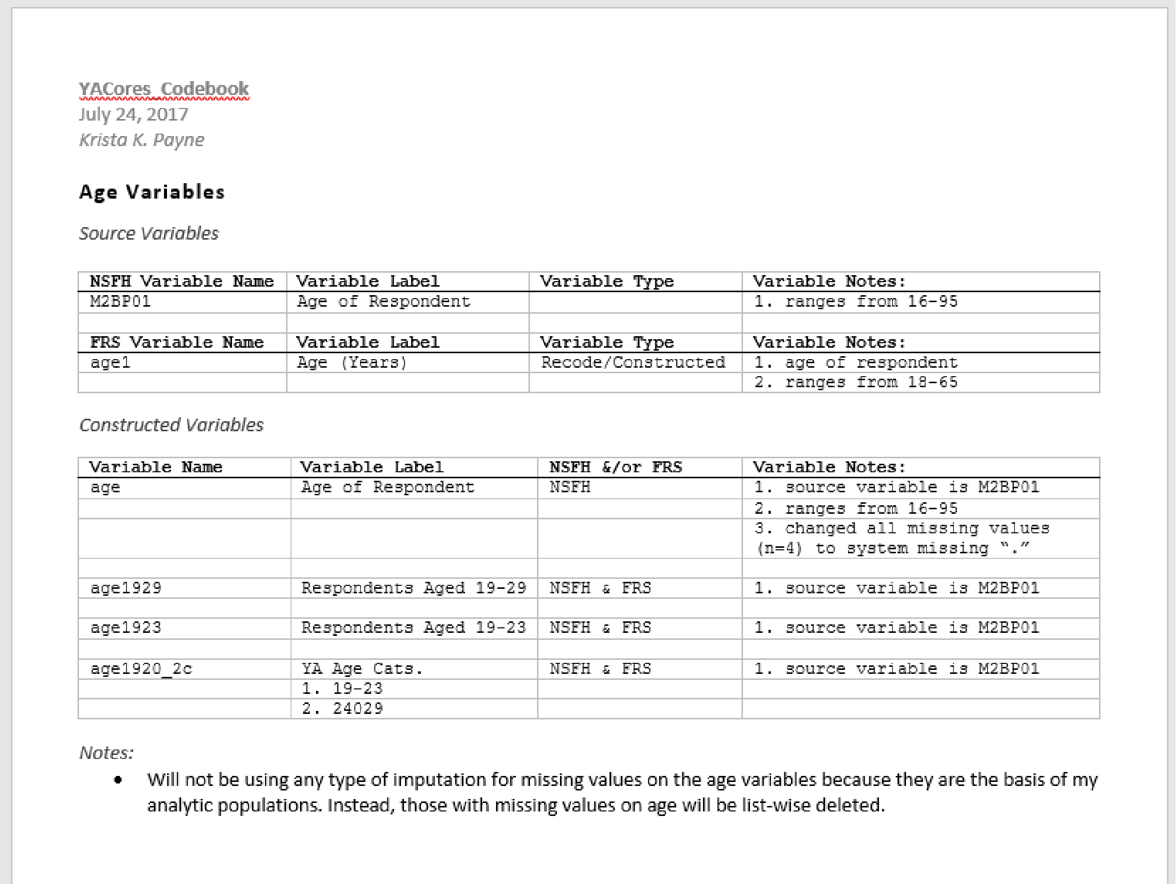
Levels of documentation, cont.

- *Codebooks*

  I've found to be EXTREMELY useful when I'm using more than one dataset in a project

  - I generally include printouts from the codebooks of the datasets I'm using.
  - I'll also include my new variables and their new values
  - Long's suggestions of items/info to include:
    - The variable name and question number if the variables came from a survey
    - Text of the original question
    - Include information on how the branching was determined (who answered the question, if not everyone)
    - Descriptive statistics including value labels for categorical variables (use numlabel, add)
    - Descriptions of how missing data can occur along with codes for each type of missing data
    - If there was recoding or imputation, include details. If a variable was constructed from other variables in the survey (e.g., a scale), provide details, including how missing data were handled.

Example: Codebook



YACores_Codebook
July 24, 2017
Krista K. Payne

Age Variables

Source Variables

NSFH Variable Name	**Variable Label**	**Variable Type**	**Variable Notes:**
M2BP01	Age of Respondent		1. ranges from 16-95
FRS Variable Name	**Variable Label**	**Variable Type**	**Variable Notes:**
age1	Age (Years)	Recode/Constructed	1. age of respondent
			2. ranges from 18-65

Constructed Variables

Variable Name	**Variable Label**	**NSFH &/or FRS**	**Variable Notes:**
age	Age of Respondent	NSFH	1. source variable is M2BP01
			2. ranges from 16-95
			3. changed all missing values (n=4) to system missing "."
age1929	Respondents Aged 19-29	NSFH & FRS	1. source variable is M2BP01
age1923	Respondents Aged 19-23	NSFH & FRS	1. source variable is M2BP01
age1920_2c	YA Age Cats.	NSFH & FRS	1. source variable is M2BP01
	1. 19-23		
	2. 24029		

Notes:

- Will not be using any type of imputation for missing values on the age variables because they are the basis of my analytic populations. Instead, those with missing values on age will be list-wise deleted.

- *Dataset documentation*
  - I tend to document this within my Master Do-file
  - I just started using Stata's label and notes commands to add metadata to my datasets. I provide an example below (p. 8).

Levels of documentation, cont.

- *Documenting do-files*
  - Do-files should include detailed comments
  - You should have two goals when writing do-files:
    1. **Robust** do-files—write do-files that produce the same result when run at a later time or on another computer
    2. **Legible** do-files—are documented and formatted so that it is easy to understand what is being done.

Writing robust do-files

*For us, a robust location to save do-files is the CFDR server. It is **expected** that your do-files for any given project are saved on the server.*

Utilize a dual workflow—distinguish between (a) do-files for ***data management*** (e.g., reading in data, merging, coding, etc.) and (b) do-files for ***statistical analyses*** (e.g., descriptives, bivariate analyses, regression, etc.).

- Do-files for analyses never change the dataset
- Do-files for analyses depend on the datasets created by the data management do-files
- If done correctly, a dual workflow will make it easier to correct errors
- This means, you do **not** create and save new variables in your analyses do-files. If, in the course of running analyses, you realize you need a new variable you go back to the appropriate data management do-file and generate (and document) it there and resave your data file.

Naming do-files—A single project can require many do-files. Naming them carefully can make it easier to:

- Find results
- Document work
- Fix errors
- Revise analyses
- Replicate your work

The run order rule: Do-files should be named so that when run alphabetical order they exactly re-create your datasets and replicate your statistical analyses (run order, AKA, the order in which a group of do-files needs to be run).

Naming do-files to re-create datasets—use a prefix of `_data0n_`*`Purpose`*`.` Purposes related to data management might be *coding*, or handling *missing data*. Given he is a strong advocate for adding detailed comments to do-files, Long recommends breaking up your coding into multiple files. This helps files from getting too burdensome to de-bug.

Example: Grouping of file names

```
YACores_data01a_NSFH_Coding-agevars_06-21-17_kkp
YACores_data01a_NSFH_Coding-agevars_06-24-17_kkp
YACores_data01b_NSFH_Coding-relstvars_06-24-17_kkp
YACores_data01c_NSFH_Coding-coresvars_06-24-17_kkp
YACores_data02_NSFH_Missing Data_06-25-17_kkp
```

Naming do-files to reproduce statistical analyses—use of prefix of `_stat0n_Purpose`. Purposes related to data analysis might be *descriptives*, *univariate*, *bivariate*, *logistic*, etc.

Example: Grouping of file names

```
YACores_stat01a_NSFH_Descriptives_06-21-17_kkp
YACores_stat01a_NSFH_Descriptives_06-24-17_kkp
YACores_stat02a_NSFH_Bivariate-pcores_06-24-17_kkp
YACores_stat02b_NSFH_Bivariate-coresat_06-24-17_kkp
YACores_stat03a_NSFH_Regression-pcores_06-25-17_kkp
YACores_stat03b_NSFH_Regression-coresat_06-25-17_kkp
YACores_stat03b_NSFH_Regression-coresat_06-26-17_kkp
```

Using master do-files. Naming your do-files in this way requires the use of master do-files. The master do-file contains, in order and with comments/descriptions all the do-files related to creating your datasets and all the do-files related to analyzing your data sets. Some projects are so big, they may benefit from a separate master do-file for creating the dataset from analyzing it. Generally, for Profiles or OPNs, one master do-file should be sufficient.

Naming and internally documenting datasets

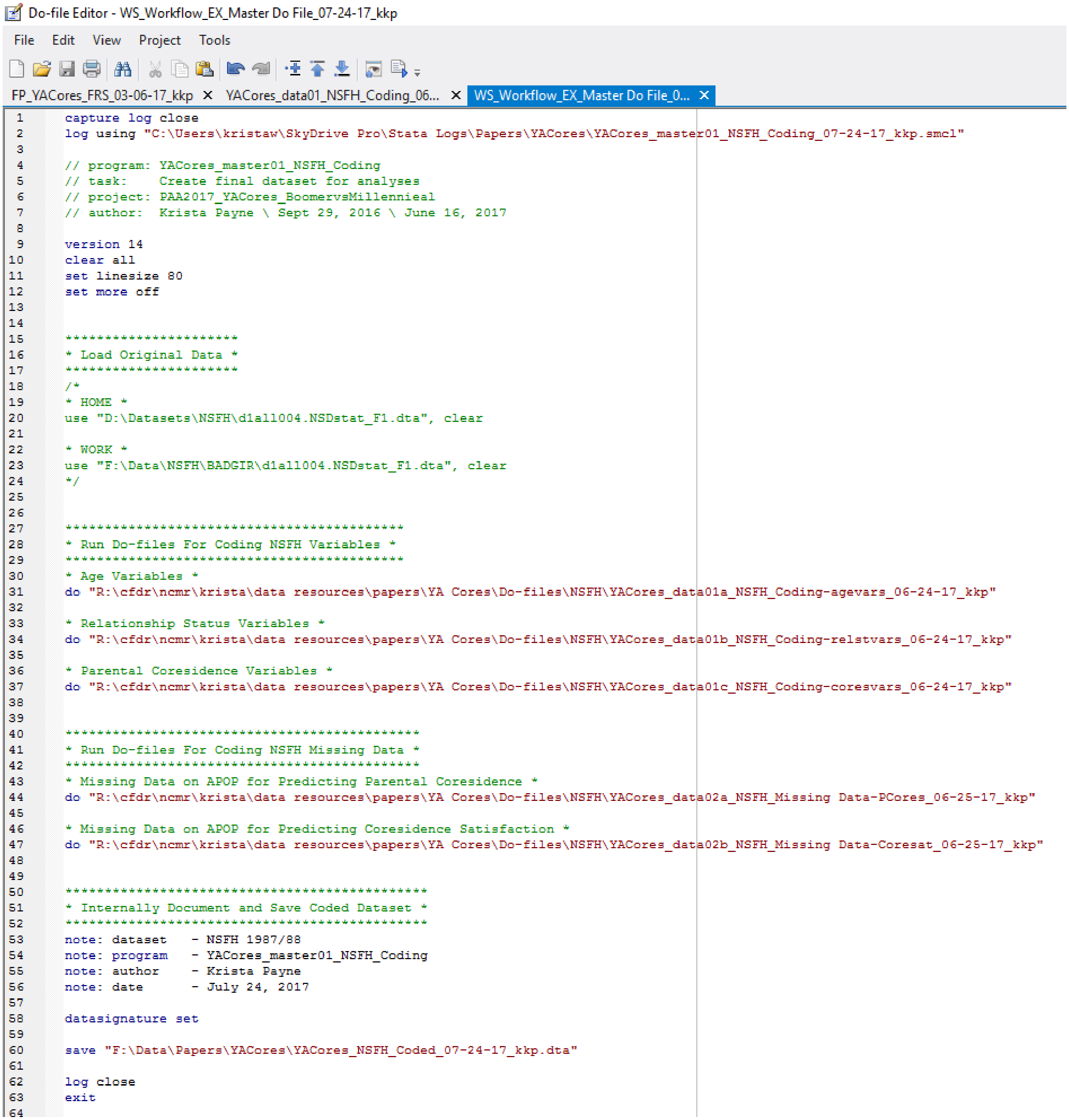
NEVER name it final!

Labels and notes for datasets—Each time Long updates the data, he adds a note.

The datasignature command—protects the integrity of your data and should be used every time you save a dataset.

WS_Workflow_Presentation Outline Part 1
Krista K. Payne
July 24, 2017

Example: Master Do-file, Naming and Internally Documenting Datasets


Do-file Editor - WS_Workflow_EX_Master Do File_07-24-17_kkp
File
Edit
View
Project
Tools
FP_YACores_FRS_03-06-17_kkp
YACores_data01_NSFH_Coding_06...
WS_Workflow_EX_Master Do File_0...

```
capture log close
log using "C:\Users\kristaw\SkyDrive Pro\Stata Logs\Papers\YACores\YACores_master01_NSFH_Coding_07-24-17_kkp.smcl"

// program: YACores_master01_NSFH_Coding
// task:    Create final dataset for analyses
// project: PAA2017_YACores_BoomervsMillennieal
// author:  Krista Payne \ Sept 29, 2016 \ June 16, 2017

version 14
clear all
set linesize 80
set more off


**********************
* Load Original Data *
**********************
/*
* HOME *
use "D:\Datasets\NSFH\d1all004.NSDstat_F1.dta", clear

* WORK *
use "F:\Data\NSFH\BADGIR\d1all004.NSDstat_F1.dta", clear
*/


*******************************************
* Run Do-files For Coding NSFH Variables *
*******************************************
* Age Variables *
do "R:\cfdr\ncmr\krista\data resources\papers\YA Cores\Do-files\NSFH\YACores_data01a_NSFH_Coding-agevars_06-24-17_kkp"

* Relationship Status Variables *
do "R:\cfdr\ncmr\krista\data resources\papers\YA Cores\Do-files\NSFH\YACores_data01b_NSFH_Coding-relstvars_06-24-17_kkp"

* Parental Coresidence Variables *
do "R:\cfdr\ncmr\krista\data resources\papers\YA Cores\Do-files\NSFH\YACores_data01c_NSFH_Coding-coresvars_06-24-17_kkp"


*********************************************
* Run Do-files For Coding NSFH Missing Data *
*********************************************
* Missing Data on APOP for Predicting Parental Coresidence *
do "R:\cfdr\ncmr\krista\data resources\papers\YA Cores\Do-files\NSFH\YACores_data02a_NSFH_Missing Data-PCores_06-25-17_kkp"

* Missing Data on APOP for Predicting Coresidence Satisfaction *
do "R:\cfdr\ncmr\krista\data resources\papers\YA Cores\Do-files\NSFH\YACores_data02b_NSFH_Missing Data-Coresat_06-25-17_kkp"


**********************************************
* Internally Document and Save Coded Dataset *
**********************************************
note: dataset   - NSFH 1987/88
note: program   - YACores_master01_NSFH_Coding
note: author    - Krista Payne
note: date      - July 24, 2017

datasignature set

save "F:\Data\Papers\YACores\YACores_NSFH_Coded_07-24-17_kkp.dta"

log close
exit

```

Writing legible do-files

Use a header and a lot of comments

Ex: Proper Header Information for a Do-File

```
capture log close
log using "C:\Users\kristaw\SkyDrive Pro\Stata Logs\REQ-SLB\50pl Cohabs\REQ-
SLB_50pl Cohabs_CPS-2010-2012-2014-2016_06-26-17_kkp.smcl"

* File:          REQ_SLB_50pl Cohabs_CPS-2010-2012-2014-2016_
* Project:       LAT Paper for PAA
* Task:          Determine source of discrepancy in reports of # cohabiting
* Author:        Krista K. Payne
* Date:          April 18, 2017; June 26, 2017

version 14
clear all
set linesize 80
set more off
.
.
.
log close
exit
```

Use alignment and indentation

- Note how I used consistent alignment in the header information of the above do-file example?

Use short lines

- Long says using 82 will mean what you print will match what is on the screen. Generally, I copy and paste my output into either Word documents or Excel files.

  Ex: `set lines 82`

Use command abbreviations sparingly:

- If you want your code to be completely legible to others, do not use too many abbreviations. Some that are considered universally acceptable:

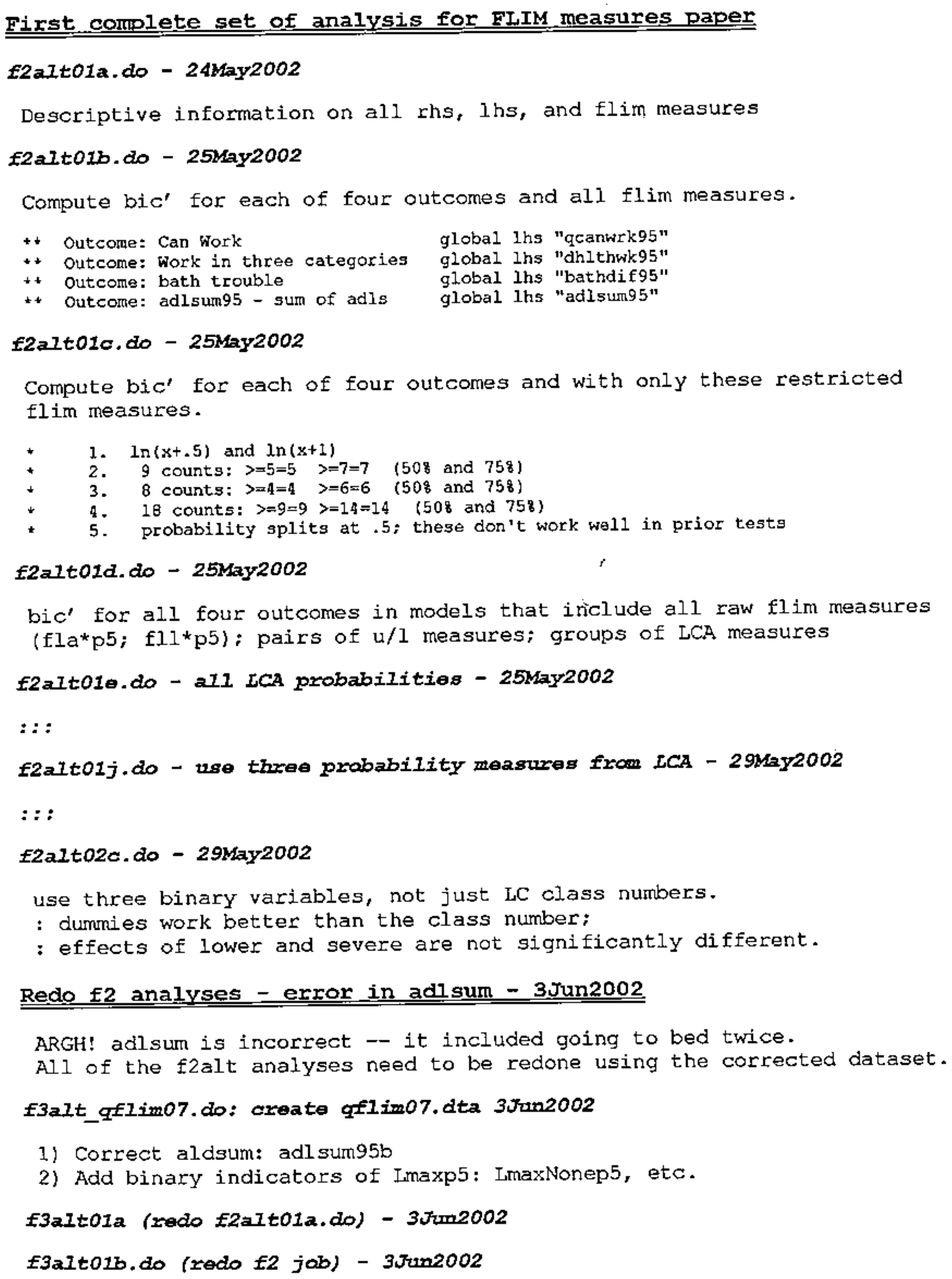
`generate`	→	`gen`
`summarize`	→	`sum`
`tabulate`	→	`tab`
`missing`	→	`mi`
`no label`	→	`nol`

Be consistent

APPENDIX

Example of a Research Log

(Taken from J. Long's book)



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 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_qflim07.do: create qflim07.dta 3Jun2002

1) Correct aldsum: adlsum95b
2) Add binary indicators of Lmaxp5: LmaxNonep5, etc.

f3alt01a (redo f2alt01a.do) - 3Jun2002

f3alt01b.do (redo f2 job) - 3Jun2002