

How to make your thesis, dissertation, or scientific paper transparent and reproducible

Florio Arguillas, Ph.D.
Daniel Alexander
CISER



Cornell Institute for Social and Economic Research

- CISER was founded in 1981 to support the evolving computational and data needs of social scientists and economists throughout the entire research lifecycle
- The CISER Data Archive provides access to approximately 27,000 social and economic dataset files
- CISER staff offers appraisal, curation, and replication services to researchers preparing for manuscript submission to scholarly journals



Goals of workshop

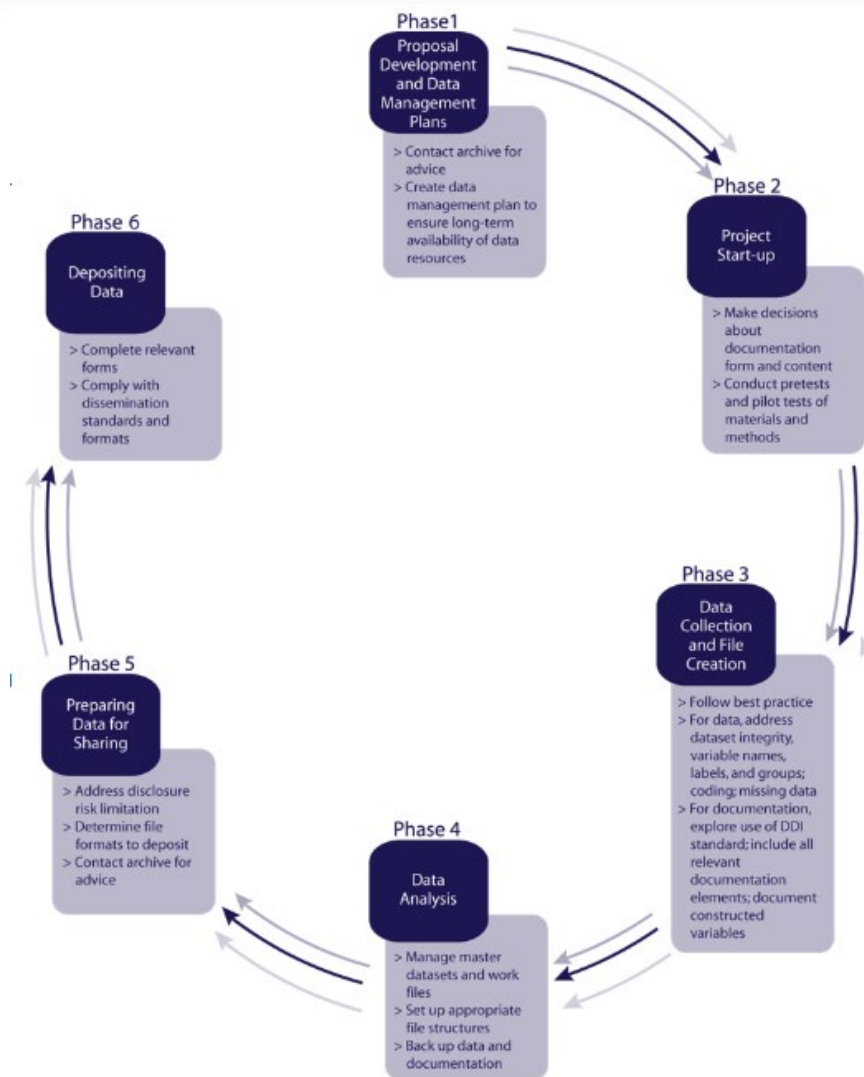
- Walk through the steps of creating a Reproduction Materials package
- What to include in your RM
 - File, documentation, data, code, output quality reviews
- Importance of Independently Understandable Materials
 - Only need to run the command file(s) to get identical results
- Make it easily findable/discoverable and accessible



Changing Mindset of Purpose of Researching

What is your end-goal when writing a thesis, dissertation, or scientific paper?

- ★ Plan ahead for sharing and archiving of supplementary materials for your scientific paper for discovery, access, and reuse



Outline

WHY

- What is Transparency? Reproducibility?
- Why make research transparent and reproducible

HOW

- R Squared and other tools
- Hands on
 - Quality Review Framework
 - Code
- Q and A



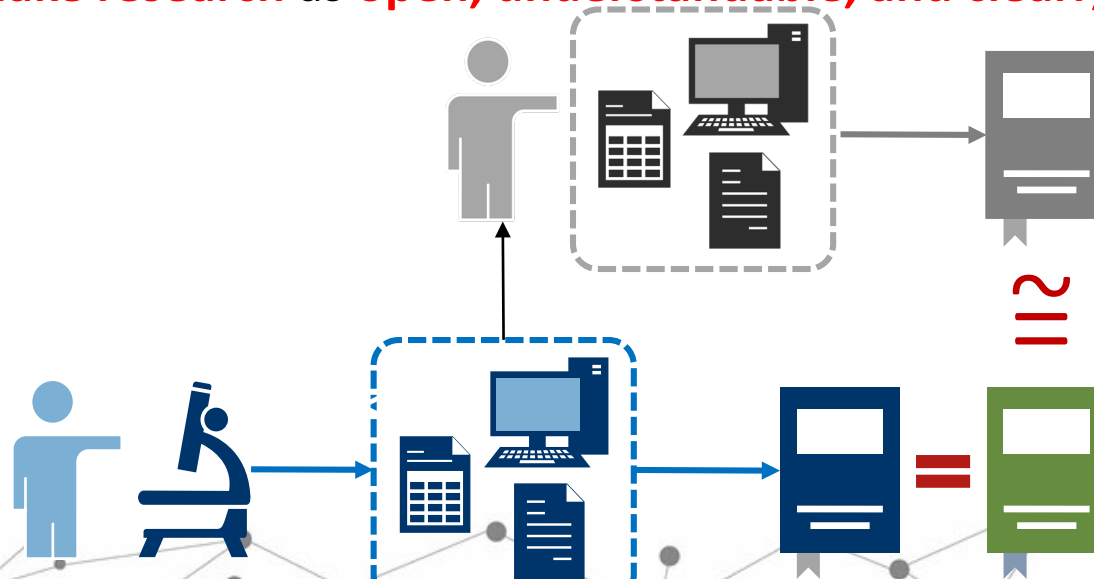
Transparency

Definition:

to make the process of research – including **data collection, coding, and analysis** – clearly visible to all readers

provides readers full access to the data, **methodology** (including coding and analysis scripts), and, whenever possible, **tools** used in data analysis

The end goal is **to make research** as **open, understandable, and clearly replicable** as possible



Pre-Registration



<https://cos.io/prereg/>

“When you preregister your research, you're simply committing to your plan in advance, before you gather data. Preregistration separates *hypothesis-generating* (exploratory) from *hypothesis-testing* (confirmatory) research. Both are important, but the same data cannot be used to generate and test a hypothesis, which can happen unintentionally and reduce the clarity and quality of your results. Removing these potential conflicts through planning improves the quality and transparency of your research, helping others who may wish to build on it.”

<http://help.osf.io/m/registrations/l/524205-register-your-project>



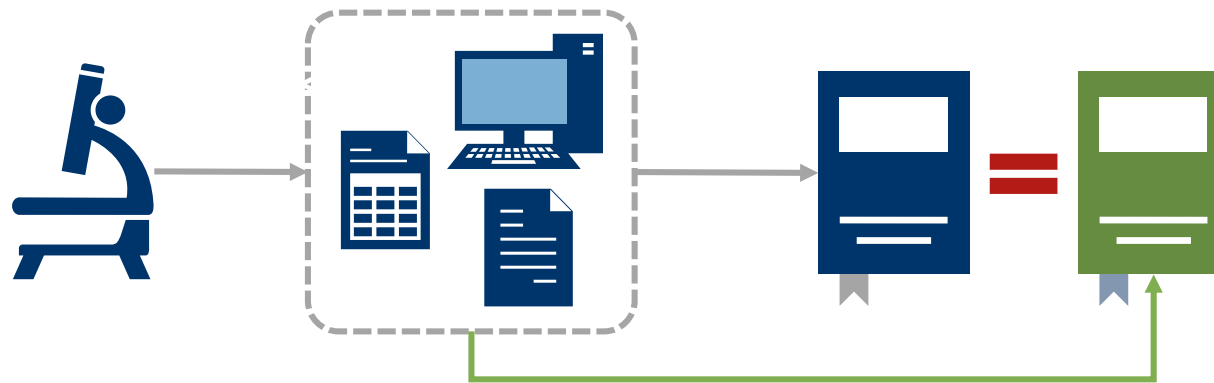
Citation Transparency

Making sure that everything cited is in your references, cite the specific pages used if possible especially in thesis dissertation. For articles it could be supplementary material.

Reproducibility (Computational Reproducibility)

Definition:

Changes in scientific practice and reporting standards to accommodate the use of computational technology...in particular whether the **same results** can be obtained from the **data and code used in the original study**.



Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2(1), 1–19. <https://doi.org/10.1146/annurev-statistics-010814-020127>

Independently Understandable RM

- Guarantees that the reproduction materials contain sufficient information for researchers to reproduce the results exactly.
- Decreases (or eliminates) the possibility of being contacted for questions about your reproduction materials



WHY PREPARE? FOR YOU

❖ Good Practice:

❖ Ability to document your own code (Vilhuber, 2017)

- ❖ For yourself in [t+n]
- ❖ For others on the team
- ❖ For human kind

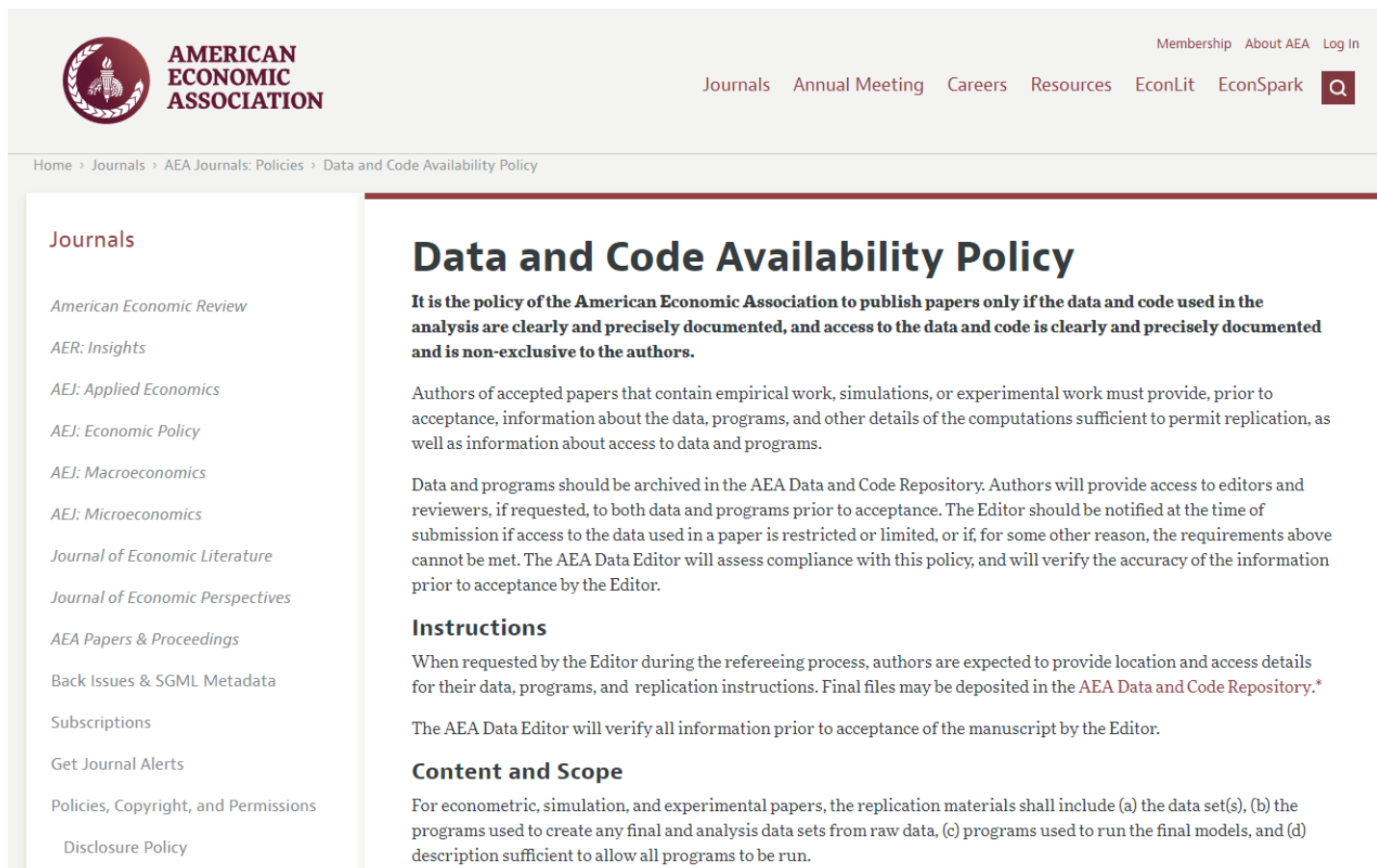
❖ Robustness to errors

❖ Discovery of error in an early step at a late stage of the research is easier to fix (Vilhuber, 2017)

❖ Early archiving may enable a researcher to enhance the impact (and certainly the visibility) of a project

❖ Studies with links to code and data gets cited more (Piwowar et., al, 2007)

WHY PREPARE? FOR YOU



The screenshot shows the American Economic Association (AEA) website. The header includes the AEA logo, navigation links (Membership, About AEA, Log In), and a search bar. The main content area is titled "Data and Code Availability Policy" and contains the following text:

It is the policy of the American Economic Association to publish papers only if the data and code used in the analysis are clearly and precisely documented, and access to the data and code is clearly and precisely documented and is non-exclusive to the authors.

Authors of accepted papers that contain empirical work, simulations, or experimental work must provide, prior to acceptance, information about the data, programs, and other details of the computations sufficient to permit replication, as well as information about access to data and programs.

Data and programs should be archived in the AEA Data and Code Repository. Authors will provide access to editors and reviewers, if requested, to both data and programs prior to acceptance. The Editor should be notified at the time of submission if access to the data used in a paper is restricted or limited, or if, for some other reason, the requirements above cannot be met. The AEA Data Editor will assess compliance with this policy, and will verify the accuracy of the information prior to acceptance by the Editor.

Instructions

When requested by the Editor during the refereeing process, authors are expected to provide location and access details for their data, programs, and replication instructions. Final files may be deposited in the [AEA Data and Code Repository](#).*

The AEA Data Editor will verify all information prior to acceptance of the manuscript by the Editor.

Content and Scope

For econometric, simulation, and experimental papers, the replication materials shall include (a) the data set(s), (b) the programs used to create any final and analysis data sets from raw data, (c) programs used to run the final models, and (d) description sufficient to allow all programs to be run.

The left sidebar of the website lists various journals and resources, including the American Economic Review, AER: Insights, AEJ: Applied Economics, AEJ: Economic Policy, AEJ: Macroeconomics, AEJ: Microeconomics, Journal of Economic Literature, Journal of Economic Perspectives, AEA Papers & Proceedings, Back Issues & SGML Metadata, Subscriptions, Get Journal Alerts, Policies, Copyright, and Permissions, and Disclosure Policy.

WHY PREPARE? FOR BETTER SCIENCE

- ❖ When reproduction materials are available:
 - ❖ Allows scientists to test and replicate each others' findings
 - ❖ Reinforces open scientific inquiry. The self-correcting features of science work most effectively.
 - ❖ Encourages diversity of analysis and opinions. Researchers having access to the same materials can challenge each other's analyses and conclusions.
 - ❖ Promotes new research and allows for the testing of new or alternative methods.
 - ❖ Improves methods of data collection and measurement through the scrutiny of others by allowing the scientific community to reach consensus on methods.
 - ❖ Reduces costs by avoiding duplicate data collection efforts.
 - ❖ Provides an important resource for training in research. Secondary data are extremely valuable to students, who then have access to high-quality data as a model for their own work.

Reproduction of Results Service (R Squared)



SPECIFICS OF CISER R SQUARED

Do

- Ensure:
 - Transparency and Reproducibility
 - Ease of access to data and code
 - Independently Understandable
- Reproduce Output
- Archival packaging
- Suggestions for glaring code inefficiencies
- Data + code = results

Don't Do

- Methodology Check
- Share anything before article is published
- Challenge or check the client's conclusions or theories
- Change client's code or paper



Reasons for the service

1. We want researchers to have publication-ready reproduction materials prior to publishing their manuscripts.
2. To provide researchers confidence that their reproduction materials will produce results identical to the computational results written in their manuscripts.
3. To encourage researchers to share their code and data
4. To relieve researchers of stress when someone requests for a copy of their reproduction materials in the future. They can **PROMPTLY respond** by handing them a copy of their reproduction materials or by pointing to a trusted digital repository hosting these materials.



ARTICLE

- ➔ Highlight all sections (e.g., paragraphs, sentences, tables, charts) that reference output derived from your data.

tion of sampled activities in which the parent is alone with children—solo parenting. Mothers engaged in activities with their children are significantly more likely to be solo parenting: 49 percent of mothers' activities with children do not include another adult present compared to 32 percent of fathers' activities. This is true, in part, because women are much more likely than men to be parenting without a spouse or partner in the household. In our sample,

```
acts6|.0393|.0265|.0493|.034|.0291|
acts7|.1149|.1869|.1131|.232|.1168|.
acts8|.1145|.1537|.1081|.1697|.121|.
acts9|.0296|.0346|.0194|.0363|.0397|.
acts10|.0792|.1537|.0748|.1666|.0835|
wykidonly|0|.4299|0|.3249|0|.4874|.
.      } //end T2
.
```

```
acts7|.1149|.1869|.1131|.232|.1168|.
acts8|.1145|.1537|.1081|.1697|.121|.1
acts9|.0296|.0346|.0194|.0363|.0397|.
acts10|.0792|.1537|.0748|.1666|.0835|
wykidonly|0|.4299|0|.3249|0|.4874|.
.      } //end T2
```

COMMAND FILE

```
*****  
*****CODE TITLE*****  
*****  
clear  
log using "..\demoCodeLog.smcl", replace text  
import delimited "<path>\HelloWorld.txt"  
  
//Labeling the variables and values  
label variable treatment "The manipulation group"  
label define treatment1 1 "Well Documented" 2 "Poorly  
Documented"  
label value treatment treatment1  
label variable pieces "Good labels are descriptive but not  
too long"  
label variable gender "Gender"  
label define gender1 1 "Male" 2 "Female"  
  
***** Table 1 - Descriptive statistics of the sample  
tab mmff  
ttest age if mmff ==1 | mmff == 2, by(mmff) unequal  
ttest age if mmff ==3 | mmff == 4, by(mmff) unequal
```

- ✓ Label all variables and values
- ✓ Comment code to describe processes and map to paper sections
- ✓ Order code outputs in the same order as they appear in paper
- ✓ Think about variable names














CODE

- ➔ Specify the sequence of execution if it consists of multiple files. Prefix the filename with Step #.
- ➔ Add comments that map sections of code to results in the manuscript. Make sure commands that generate results are preceded by comments that indicate which result the command generates.

For example:

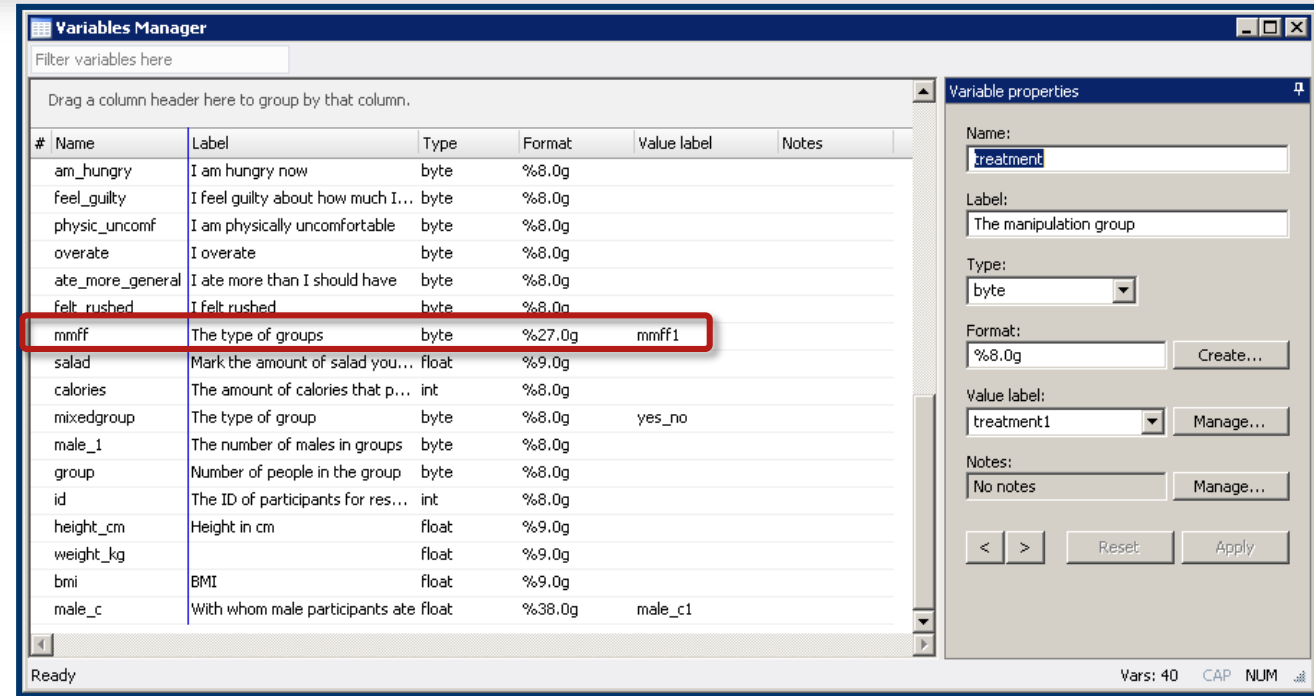
*The following command generates column 1 of Table 1

*The following command generates the mean age mentioned on page 3, paragraph 3

| Name | Date modified | Type | Size |
|---|--------------------|----------------------|------------|
|  demog-tables-r&r-092214 | 9/22/2014 9:20 AM | Microsoft Excel W... | 167 KB |
|  m-m-092214 | 9/22/2014 9:20 AM | Microsoft Word D... | 105 KB |
|  1995Preg | 5/26/2010 11:39 AM | Stata Dataset | 10,494 KB |
|  1995Resp | 5/26/2010 11:35 AM | Stata Dataset | 124,745 KB |
|  200610FemResp | 4/11/2012 9:18 AM | Stata Dataset | 52,339 KB |
|  200610MaleResp | 4/11/2012 9:17 AM | Stata Dataset | 34,826 KB |
|  200610Preg | 4/11/2012 9:18 AM | Stata Dataset | 7,286 KB |
|  unfile-062214 | 9/5/2014 9:48 AM | Stata Dataset | 17,799 KB |
|  unfile-062214-month | 7/15/2014 2:07 PM | Stata Dataset | 65,657 KB |
|  1995_HARMONIZATION_kmsupp | 4/4/2012 8:31 AM | Stata Do-file | 25 KB |
|  2007_HARMONIZATION_kmsupp_new | 1/12/2014 2:21 PM | Stata Do-file | 32 KB |
|  Demography_R&R_6_22_14 | 7/15/2014 2:07 PM | Stata Do-file | 9 KB |
|  Demography_R&R_analysis_5_14_14 | 9/22/2014 8:53 AM | Stata Do-file | 18 KB |

DATA

- ➔ All variables and values labeled
- ➔ Free of errors and inconsistencies
- ➔ Data anonymized? Depends!



| # | Name | Label | Type | Format | Value label | Notes |
|---|------------------|-----------------------------------|-------|--------|-------------|-------|
| | am_hungry | I am hungry now | byte | %8.0g | | |
| | feel_guilty | I feel guilty about how much I... | byte | %8.0g | | |
| | physic_uncomf | I am physically uncomfortable | byte | %8.0g | | |
| | overate | I overate | byte | %8.0g | | |
| | ate_more_general | I ate more than I should have | byte | %8.0g | | |
| | felt_rushed | I felt rushed | byte | %8.0g | | |
| | mmff | The type of groups | byte | %27.0g | mmff1 | |
| | salad | Mark the amount of salad you... | float | %9.0g | | |
| | calories | The amount of calories that p... | int | %8.0g | | |
| | mixedgroup | The type of group | byte | %8.0g | yes_no | |
| | male_1 | The number of males in groups | byte | %8.0g | | |
| | group | Number of people in the group | byte | %8.0g | | |
| | id | The ID of participants for res... | int | %8.0g | | |
| | height_cm | Height in cm | float | %9.0g | | |
| | weight_kg | | float | %9.0g | | |
| | bmi | BMI | float | %9.0g | | |
| | male_c | With whom male participants ate | float | %38.0g | male_c1 | |

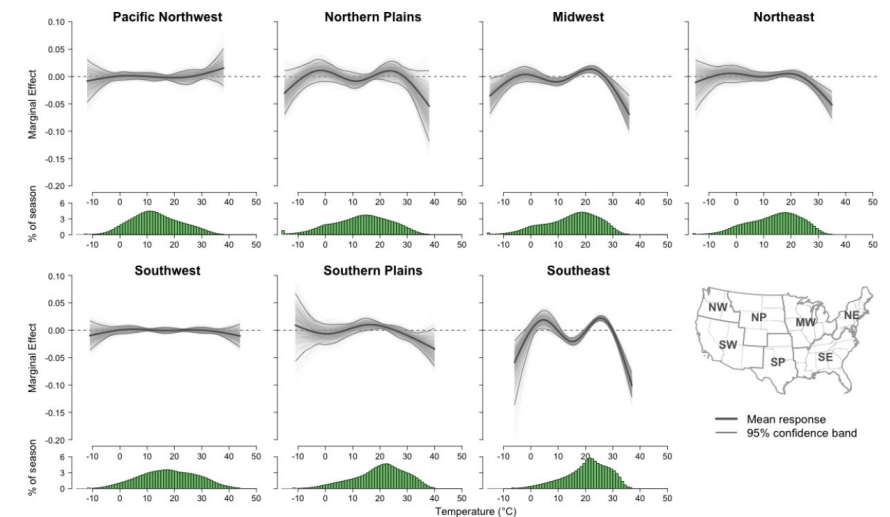
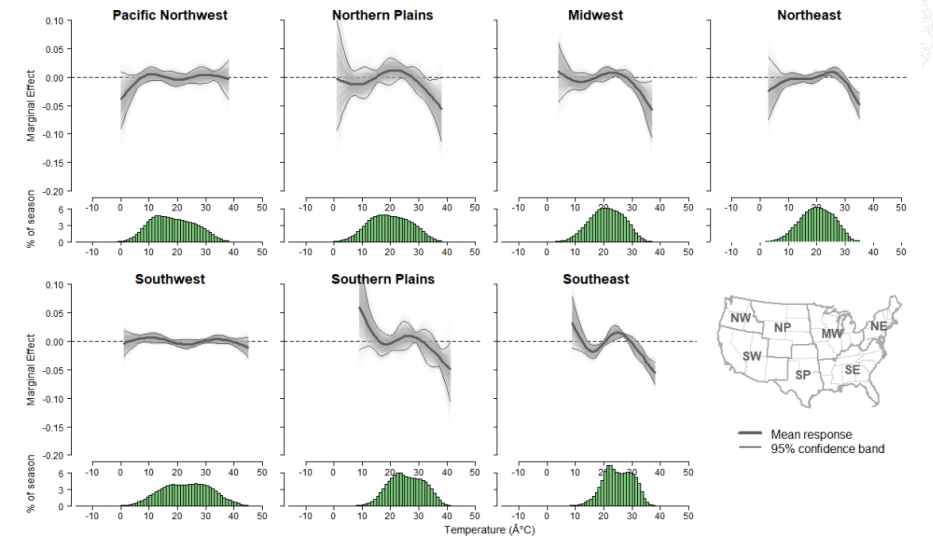
| Variable properties | |
|---|-----------------------------------|
| Name: | treatment |
| Label: | The manipulation group |
| Type: | byte |
| Format: | %8.0g Create... |
| Value label: | treatment1 Manage... |
| Notes: | No notes Manage... |
| < > Reset Apply | |

Ready Vars: 40 CAP NUM

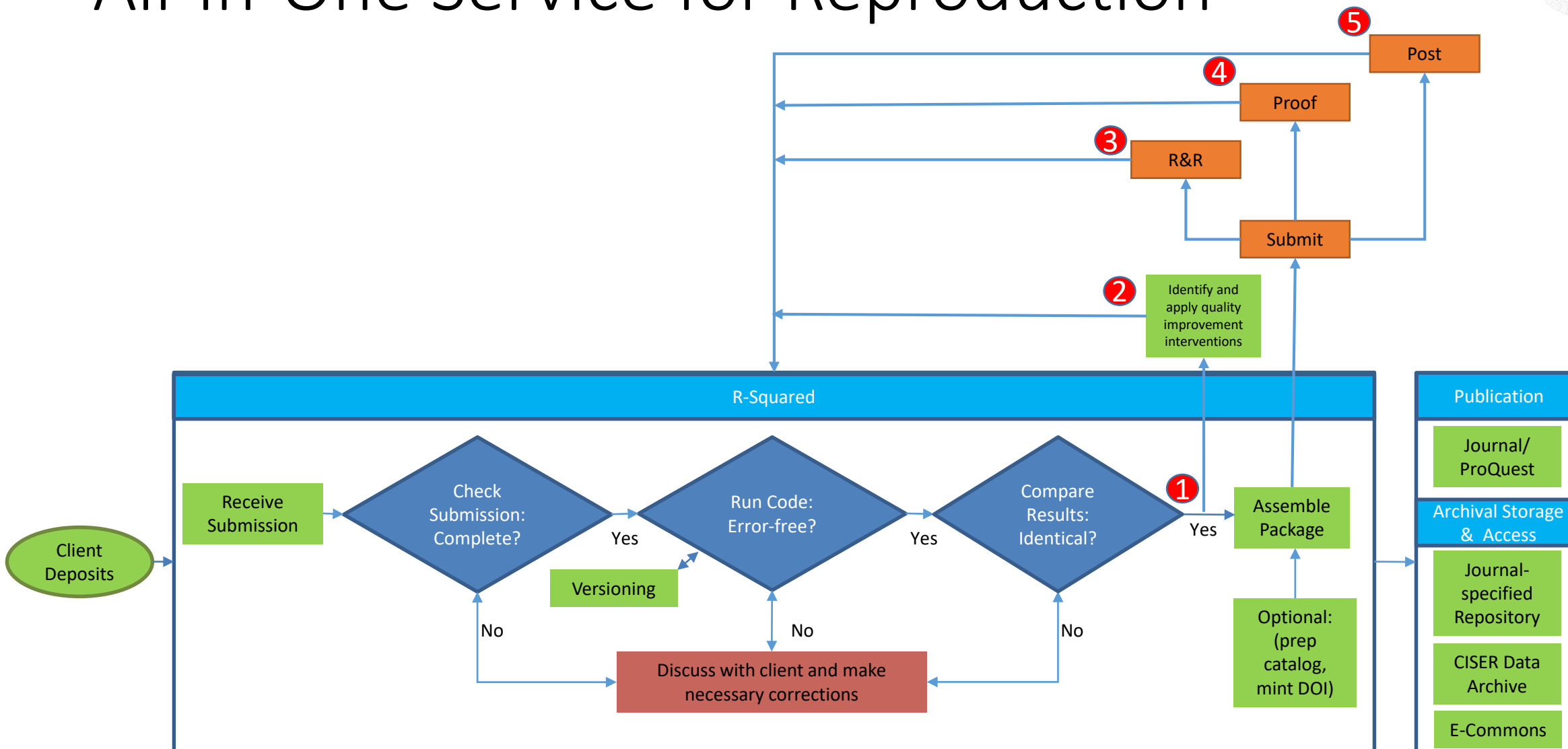
Lack of Seed

Any random generation requires a seed
or else it can have different results and
will not be reproduced.

Example on the top is without the seed
and bottom is with it.



All-In-One Service for Reproduction



Examples of good reproduction packages

- Kim
 - <https://doi.org/10.6077/80ph-y162>
- Ortiz-Bobea:
 - <https://doi.org/10.6077/f26v-xz15>





Most Common Error Types present in Manuscripts

- Rounding
- Typos
- Misidentification
- Out of-Date Tables
- Missing from Package
- Incorrectly Marking p-values



Rounding

- Extremely common and easy to miss
- Different software/people round in different ways, compounding if it is an error or difference of style
 - However inconsistent rounding is always a problem
- Double rounding can cause errors as well



Typos

- Sometimes obvious, such as missing a decimal place when all the others have it
- Can also just be a mistyped number or two





Misidentification (looking at wrong cell in output file)

- Very easy when there aren't output management software/practices
- Difficult to detect from repeated checks, as the number is present on the output page, but simply is not the result that should be included





Out of Date Tables

- Descriptive Tables are most common culprit
 - Researchers will generate their descriptive tables and then further clean data
 - Can be calculated inconsistently as well, such as valid vs total percentage
- Plenty of other tables can have this happen as well, as researchers often don't re-run their entire code (especially for long execute times)



Missing From Package of Scripts

- Can be similar to “hard coded variables” in that the calculation just done by hand/excel and put into the manuscript
- Very common with regards to figures as they are often edited or done in non or different statistical package programs (such as excel) and then edited to look better





Incorrectly Marking P-Values

- Often happens with regards to rounding
 - 0.045 rounds to 0.05 but should be marked as ≤ 0.05 , but can mistakenly be marked as $=0.05$
 - 0.101 shouldn't be marked as ≤ 0.10 , as it is >0.10
- Can come from inconsistency of having multiple people doing the tables
- Easy to make one of the other errors

Most Common Error Types present in Scripts



- Not calling the data initially
- Hard Coded Variables
- Missing Files/Commands
- No (or poor) Commenting
- Uncertain Order of Execution
- Does not follow the order of the Output
- Prerequisites
- Extraneous output/lines of code



Not calling the data initially

- Many clients simply open the data with the GUI instead of via code
- Can cause confusion when there are multiple data files present (including raw data, cleaned data, intermediate data, data from different sources, etc.)





Hard Coded Variables

- Generally done either by hand calculator or Excel and results entered into program file
- Can cause errors when data is further refined or if mistake was made when calculating it
- Causes confusion as to where the number came from and what it represents





Missing Files/Commands

- Impossible to fully reproduce results with needed files/commands missing
- Often because researchers will work on multiple computers or utilize poor version control and delete things
- Can also be caused by hard coding a variable and then discarding the code or by using the GUI and not pasting the commands



Poor or No Commenting and Poor Variable Names

- People will write the code for themselves, and thus commenting can tend to fall by the wayside instead of commenting as they go
- What makes sense to someone who has worked with the code/data for four years may not make sense to someone else and may need further explanation





Uncertain or No Order Specified in Title of Scripts

- Researcher knows the steps in running the program files so doesn't order them
 - Alternatively, they know the step, yet use confusing names such as "first-first..." vs "actually-first..." vs "initialize..."
 - Should have numerical component e.g. Step1... Or 1_1-cleaning...
- If not run on a 'clean install' then can have previously generated results affect outcomes



Doesn't Follow the Order of the Manuscript

- Things are rarely coded in the exact order they appear in the manuscript and can be impossible to do so. However, having the results generally follow the manuscript will dramatically increase transparency
- Often calculated on other computers and then the code thrown together to put everything in the same script file





Prerequisites (ado, library, extension, etc.)

- The idea of reproducibility is it will work on any computer and without the list of prerequisites, that will not work
- Very easy to forget as they don't need to be reinstalled from project to project
- Prone to breaking things when they, or the base software gets updated, so having unneeded ones can be problematic



Extraneous outputs/lines of code

- Makes the code and output files longer and unnecessarily confusing
- Not very transparent to have outputs that are not used/needed
- Can cause mis-identification or other issues detailed in the common errors in manuscript.
- Time consuming on both the re-user and checker's.



Hands-on: Data & code review

Distribute the handouts and introduce/begin first exercise

Quality Review for Curation



**FILE
REVIEW**



**DATA
REVIEW**



**CODE
REVIEW**



**DOC
REVIEW**



**OUTPUT
REVIEW**

curating for reproducibility

FILE QUALITY REVIEW



- ✓ Check for presence of all files
- ✓ Verify content of files matches expected format
- ✓ Assign persistent identifier
- ✓ Create study citation and study-level metadata record
- ✓ Record file size details
- ✓ Create non-proprietary versions of files
- ✓ Implement migration strategy for file formats

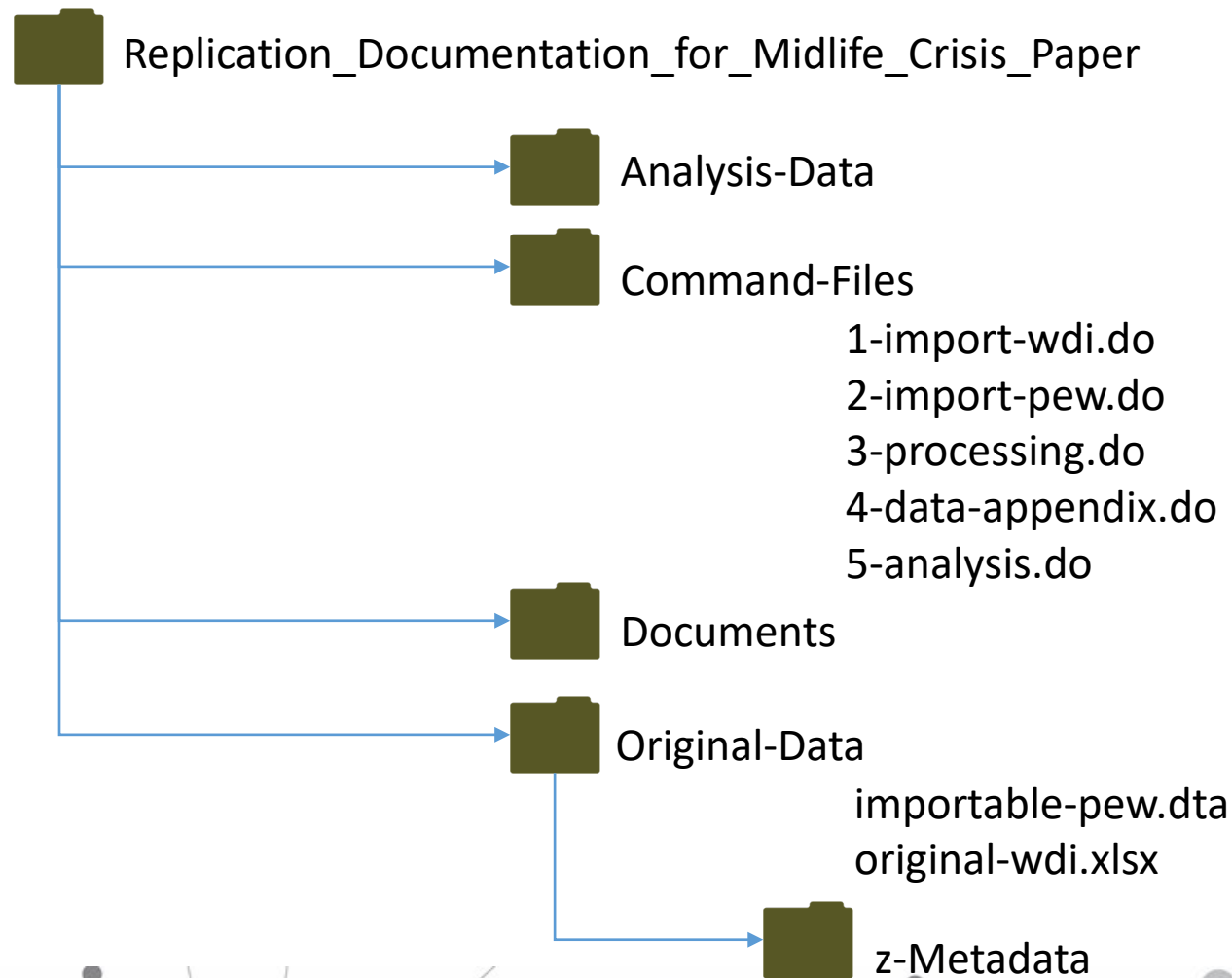


Hands-on: Data and Code Review

Hands on with File Review

- ➡ Look through the files and compare it to the manifest
- ➡ The next slide contains the necessary files, not the complete list. The program will run as long as you have those





curating for reproducibility

DATA QUALITY REVIEW



- ✓ Restricted Access Data
 - Provide detailed instructions on how to gain access to the data in a readme file or data documentation.
- ✓ Public-use Data
 - Provide detailed instructions on how to gain access to the data in a readme file or data documentation.
 - Make the re-user download the data and execute your code against it.
 - But, also provide the original data in case it can no longer be found online



curating for reproducibility

DATA QUALITY REVIEW



- ✓ Check for undocumented variable and value information
- ✓ Examine data for inconsistencies and errors
 - Discrepancies in number of observations
 - Out-of-range or wild codes
 - Undefined null values
- ✓ Review data for confidentiality issues
- ✓ Add question text to variables
- ✓ Identify and address foreign language characters
- ✓ Adjust format widths
- ✓ Optimize file size
- ✓ Standardize missing values
- ✓ Check for consistency and skip patterns

curating for reproducibility

DATA QUALITY REVIEW – Variables



**DATA
REVIEW**

- ❖ *The exact question wording or the exact meaning of the datum*
- ❖ *The text of the question integrated into the variable text*
- ❖ *Universe information, i.e., who was actually asked the question*
- ❖ *Exact meaning of codes*
- ❖ *Missing data codes*
- ❖ *Unweighted frequency distribution or summary statistics*
- ❖ *Imputation and editing information*
- ❖ *Details on constructed and weight variables.*
- ❖ *Location in the data file*
- ❖ *Variable groupings*

DATA README

- ➔ **Create one readme file for each data file, whenever possible.** It is also appropriate to describe a "dataset" that has multiple, related, identically formatted files, or files that are logically grouped together for use (e.g. a collection of Matlab scripts). When appropriate, also describe the file structure that holds the related data files.
- ➔ **Name the readme so that it is easily associated with the data file(s) it describes.**
- ➔ **Write your readme document as a plain text file**, avoiding proprietary formats such as MS Word whenever possible. Format the readme document so it is easy to understand (e.g. separate important pieces of information with blank lines, rather than having all the information in one long paragraph).
- ➔ **Format multiple readme files identically.** Present the information in the same order, using the same terminology.
- ➔ **Use standardized date formats.** Suggested format: [W3C/ISO 8601 date standard](#), which specifies the international standard notation of YYYY-MM-DD or YYYY-MM-DDThh:mm:ss.
- ➔ **Follow the scientific conventions for your discipline for taxonomic, geospatial and geologic names and keywords.**

curating for reproducibility

DOCUMENTATION QUALITY REVIEW



**DOC
REVIEW**

- Title of data collection
- Principal investigators
- Description or abstract
- Sponsor or funding agency
- Type of Data (Quanti, Geo, Quali, Mixed)
- Methodology
- Universe
- Units of analysis (Individual, Household, Metropolitan area, other)
- Mode of data collection
- Type of data collection (admin, census, clinical, observational data, etc)
- Study time period
- Data collection dates
- Geographic coverage areas
- Geographic unit
- Sampling information
- Response Rate
- Measurement tools/Scales
- Weights
- Keywords
- Data-Related publications
- Confidential and/or copyrighted
- Grant managers
- Original data source citation
- Analysis software version

curating for reproducibility

DOCUMENTATION REVIEW – METADATA



- **Related publications.** Citations to publications based on the data, by the principal investigators or others.
- **Technical information on files.** Information on file formats, file linking, and similar information.
- **Data collection instruments.** Copies of the original data collection forms and instruments. Other researchers often want to know the context in which a particular question was asked, and it is helpful to see the survey instrument as a whole. Copyrighted survey questions should be acknowledged with a citation so that users may access and give credit to the original survey and its author.
- **Flowchart of the data collection instrument.** A graphical guide to the data, showing which respondents were asked which questions and how various items link to each other. This is particularly useful for complex questionnaires or when no hardcopy questionnaire is available.
- **Index or table of contents.** A list of variables either in alphabetic order or organized into variable groups with corresponding page numbers or links to the variables in the technical documentation or codebook.
- **List of abbreviations and other conventions.** Variable names and variable labels often contain abbreviations. Ideally, these should be standardized and described.
- **Interviewer guide.** Details on how interviews were administered, including probes, interviewer specifications, use of visual aids such as hand cards, and the like.
- **Coding instrument.** A document that details the rules and definitions used for coding the data. This is particularly useful when open-ended responses are coded into quantitative data and the codes are not provided on the original data collection instrument.

curating for reproducibility

CODE QUALITY REVIEW



- ✓ Identify packages and version required to execute code
- ✓ Identify the execution order of codes (if multiple files)
- ✓ Set up working directories or convert absolute file paths in your code to relative file paths
- ✓ Check code is well-annotated i.e., there are non-executable comments that document analysis processes
- ✓ Identify the types of code included – e.g, variable transformation and data cleaning; statistical procedures; presentation (tables and graphs); saving/logging

curating for reproducibility

CODE QUALITY REVIEW



- ✓ Open/Print and identify the results in the paper
- ✓ Execute code to ensure code is error-free
- ✓ Compare code output to findings presented in the article



curating for reproducibility

CODE QUALITY REVIEW



- ✓ In case of discrepancies:
 - ✓ Mark/highlight the discrepancies in the manuscript
 - ✓ Write next to it the actual figures you derived from the output
 - ✓ If necessary, write notes at the margins of the printed manuscript
- ✓ Try to establish the reason for the discrepancies or errors
- ✓ Retrace your steps
- ✓ Scan and send to the client. Ideally, have the client sit next to you during the entire process so they can respond immediately.



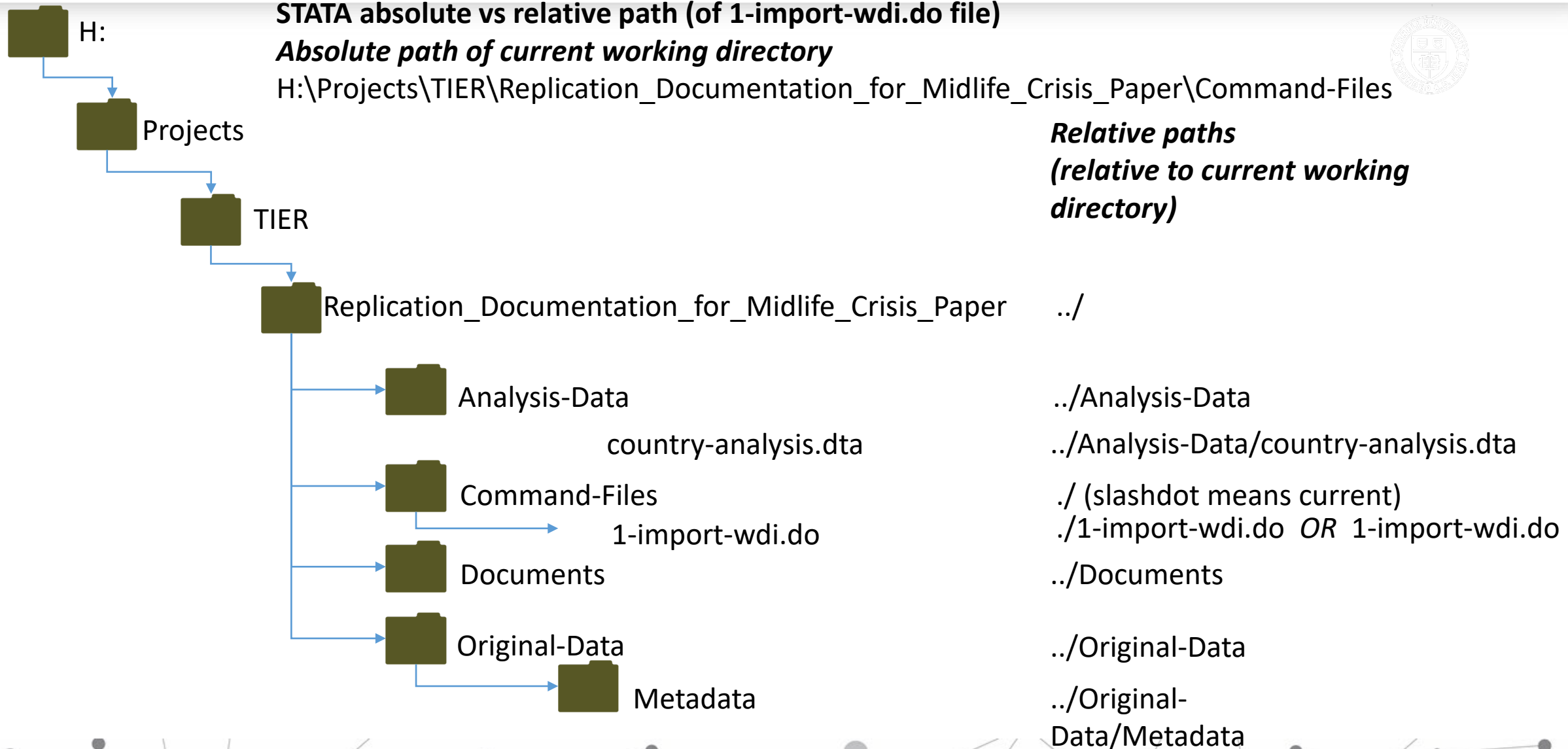
curating for reproducibility

CODE QUALITY REVIEW



- ✓ Convert absolute file paths to relative file paths
 - No need to recreate the absolute path of your working directory
 - The paths will be relative to the current working directory. Invoking the software by opening the program file makes the program's folder location the current working directory.
 - Portability of your Reproduction Materials Package
- SAS:
http://sascommunity.org/wiki/Assign_library_in_a_relative_path_in_Windows
- R: <https://www.youtube.com/watch?v=fe6GA200dks>





Hands-on: Data and Code Review

Hands on with Relative Paths

- ➡ Open up the workshop folder
 - ➡ Note the TIER protocol, we will be modifying this but it is a good starting point
- ➡ Take a min to familiarize yourself, see how each folder is clearly labeled and it is clear what goes where, with a readme in the top level
- ➡ Now navigate back up to the top level and open the “Command Files” folder
 - ➡ Note how the file order is designated by a number.
- ➡ Open up the first file titled “1-import-wdi”
- ➡ Try to run the code
 - ➡ Shouldn't work because of absolute paths



Hands-on: Data and Code Review

Hands on with Relative Paths (cont.)

- ➡ To change everything from absolute to relative paths, we want to do a find/replace
 - ➡ To do this, push 'ctrl' + 'h' and fill in the appropriate fields (U:/dra65/Documents/TIER Protocol/) and (../) respectively
- ➡ Now try to execute the code. It should work on everyone's machine despite every path being different. This is the power of relative paths.



Hands-on: Data and Code Review

Master Files

- ➡ Master Files can help with reproduction and transparency
- ➡ Show the exact order things get done, make it extremely easy to do, just execute one file and produce the results in your manuscript
- ➡ Relatively easy and recommended when dealing with just a single programming language, more difficult but still do-able with different languages



Hands-on: Data and Code Review

Master Files (cont.)

- ➡ Create a new do file “Master”
- ➡ The command to run other files in stata is “do”
- ➡ Type out
do 1-import-wdi.do
- ➡ Repeat for the rest of the files
- ➡ Can/should



Hands-on: Data and Code Review

Ados and Prerequisites

- ➡ I already changed the rest of the code to be relative paths for you, so the code for 2 through 4 should work.
- ➡ However 5 has an ado requirement and won't run properly. To get around this, most people just list out the prerequisites that are needed and rely on the reproducers to get the files themselves
 - ➡ However that can be difficult as software updates the addons get updated too, but your code won't. Is there a way to solve both the problem of requiring downloads and versioning??
- ➡ Yes!
- ➡ Prerequisites folder preparation



Hands-on: Data and Code Review

Ados and Prerequisites (cont.)

- ➔ The first step is to designate a new folder as your “PLUS” folder. This can be done in every statistics language but is called different things depending on what software it is. (R calls them libraries, etc.) However fundamentally they are the same thing.
- ➔ Make a new subfolder in the “WorkshopDemo” folder called “Prerequisites”
 - ➔ It should appear below the “Original-Data” folder
- ➔ Now type into stata the following set of commands:
`sysdir set PLUS ../Prerequisites`
`ssc install outreg2`
- ➔ This will install the missing ado and put it in the Prerequisites folder you just created, now all your code will work



Hands-on: Data and Code Review

Saving Outputs

- ➡ Save the figures generated by the code into the documents folder, so all outputs are generated by the code and don't need to keep the code open
- ➡ This can be done with the following line of code:
 - ➡ `graph export ../Documents/figure1.png`
 - ➡ Place this code after the graph is generated in figure 1
 - ➡ Modify this code to work for figure 2 (just change the file name to figure2.png)



curating for reproducibility

OUTPUT QUALITY REVIEW



- ✓ All outputs are in the manuscript
 - Highlight values in the output used in the manuscript
 - Preferably only one statistical software used
 - If output is not in manuscript, explain its purpose
- ✓ Output in same order as the flow of the manuscript



Hands-on: Data and Code Review

Output comparison

- ➡ Examine the log file handout and compare the un-highlighted output to the table in the manuscript



Hands-on: Data and Code Review

Output comparison (cont.)

- ➡ Examine the outreg2 output
 - ➡ Compare this with earlier log output. Automatically outputting it to csv with outreg2 allows for much more organized and neat outputs while preserving everything you do in the code
 - ➡ Automatically marks significance levels etc.
 - ➡ There are similar programs in most statistical software if you look for them, can be very helpful





Hands-on: Data and Code Review

Readme

- ➡ Good transparency requires a good readme file
 - ➡ However what should be in this readme is up for debate and depends on your field
- ➡ Generally speaking it should explain concisely and completely explain your file structure and how to get started should explain any additional files needed
- ➡ This is the first point of contact with users and thus needs to be both a guide and a place they can turn to for help



Hands-on: Data and Code Review

Readme (cont.)

- ➡ In the workshop demo folder there is an example of a readme called “Introduction to the Tier Protocol” which is now out of date as we changed some of the file structure as well as a one created for this exercise titled “readme”
- ➡ For an exercise, discuss or write up what a readme for this workshop should contain to briefly guide people through the transparency/reproduction process



Hands-on: Data and Code Review

Readme (cont.)

➡ <https://data.research.cornell.edu/content/readme>



Tools for Transparency and Reproducibility

- Versioning software (git)
- Versioning software websites (Github, bitbucket, etc.)
- Code Ocean - <https://codeocean.com/> (research collaboration platform)
- OSF - <https://osf.io/> (place to share your research – pre-register, store and share files, collaboration, etc.)



Dynamic Documentation

- R Notebook and R Markdown
- *Advantages*
 - R code can be embedded in the report, so it is not necessary to keep the report and R script separately. Including the R code directly in a report provides structure to analyses.
 - The report text is written as normal text
 - The output can be HTML, PDF, Word, OpenDocument, RTF and includes pictures, code blocks, R output, and text.
 - Update data, model, and results on the fly, if changes have to be made. But, changes may affect results and analysis text may have to be modified.
- *Disadvantages*
 - More difficult to edit. The “finished” markdown may look confusing before it is knit
 - May involve lots of experimentation to get the desired output (table, figures, etc).

Dynamic Documentation

- STATA ([new in STATA15](#))
- Dynamic documents – if data ever change we can run the original source document to create an update
- dyndoc - use markdown text-formatting language intermixed with Stata. Output produced is an HTML file.
sample command to execute: `dyndoc dyndoc_example1.do`
- putdocx - creates Word documents with formatted paragraphs (analysis text), tables, embedded output from Stata in paragraphs and tables, embedded Stata graphs.
sample command to execute: `do putdocx_example.do`
- putpdf - creates PDF files with formatted paragraphs (analysis text), tables, embedded output from Stata in paragraphs and tables, embedded Stata graphs.
sample command to execute: `do putpdf_example.do`



Conclusion

- Change your mindset: Graduating/Publishing AND reproduction materials sharing are your ultimate goals
- Include in your workflow time for Data Quality Review, such as code review, output review, adherence to TIER protocol in managing files for ease of reproduction
- Process is time-intensive, costly, and confusing if data and code quality is low
- Suggest code review, output review, adherence to TIER protocol in managing files for ease of reproduction
- CURATE AS YOU CODE AND CODE WITH REUSE IN MIND.



Discussion/Questions

THANK YOU

Feedback welcome and encouraged!

Florio Arguillas, Ph.D.

Daniel Alexander

CISER@Cornell.edu

