




THE PITFALLS OF DATA COLLECTION

Lessons Learned from a Multi-Year Data Project

Methods Workshop

10 April 2024

Megan Lloyd



Lessons learned the hard way

- Data collection always takes more time than you think it will or want it to.
- Concepts or events of interest are harder to code than to define.
- Micro-managing *how* coders enter data is a necessary evil.
- Fast, good, or cheap. You can have two, not all three.

Objectives of a gameplan

- Producing high-quality, replicable, and reliable data
- Avoiding the need for data cleaning and retroactive coding
- Safeguarding professional relationships with supervisors and colleagues
- Developing skills as a coder and a manager

Envision and organize

- Brainstorming: Who are the end users, and what types of questions will the data be used to answer?
 - Personal research interests and/or related literature
 - Gaps in extant research
 - Micro □ Meso □ Macro
- Prioritize and order steps
 - Stages of data collection
 - Short-term versus long-term data needs

Lay the groundwork for data collection

- Identify main data sources
- Draft codebook
 - Write **working** definitions of variables
 - Specify variable types: who, what, where, when, why?
 - Identify variables that require time- and context-specific knowledge
 - “Data link” variables
 - Establish clear coding rules
 - Character separators
 - Names of people, entities (lower, upper case)
 - Format of date/time variables

Pilot data collection method

- Build data frame
 - Unit of analysis
 - Data link variables

- Code sample of cases
 - Select small timeframe
 - Randomly select a case
 - Selectively choose easy cases and hard cases

dispute	gwcode	kgcid	start_date	end_date
India - Assamese	750	410	1/1/91	1/1/91
Cyprus - Turkish Cypriots	352	108	2/11/91	2/11/91
Ethiopia - Oromos	530	309	5/26/91	5/27/91
Angola - Cabindans	540	301	Nov-91	Nov-91

Review and revise

- Review the pilot coding process
 - How fast did coding go?
 - Did the codebook definitions make sense in the real world?
 - Do variables need to be added or removed?
 - Were similar results produced across RAs?
- Revise the coding process and codebook
 - Note all changes in the codebook
 - Perform another round of pilot coding

Working with a supervisor

Common mistakes	Solutions
Waiting to ask questions, provide updates	<ul style="list-style-type: none">• Schedule regular meetings with consistent structure<ul style="list-style-type: none">• Updates on the project• Questions/hurdles □ answers/solutions• Next steps and objectives to complete before next meeting• Take detailed notes on all decisions
Hesitating to suggest alternative methods	<ul style="list-style-type: none">• Provide reasoning• Provide possible solution with a plan to test
Overcommitting, then underperforming	Do what you say you're going to do.

Managing other research assistants

Common mistakes	Solutions
Playing the role of peer, not manager	Fake it 'till you make it
Failing to micro-manage the details of data entry	Check in proactively and regularly with RAs
Communicating infrequently or irregularly across RAs	Take detailed notes on RAs questions, communicate with supervisor, and follow up with everyone

Additional tips

- Data management and storage
 - Always work in the cloud
 - UMD Box has weird gimmicks
 - Understand how date and time variables are automated in software
- Dataset construction
 - Take clear notes in R script
 - Write code with end audience in mind
 - Plan out steps of data merging before assembly

Thank you!

■ Questions? Thoughts?