PPOL 6818 - Spring 2025 - Week 6

Béatrice Leydier

Data Sources and Measurement Error

Georgetown University Initiative

gui^{2}de

on Innovation, Development and Evaluation

PPOL 6818 - Spring 2025

Data sources

GEORGETOWN UNIVERSITY

Examples of data?

Data sources

	Collected by	Challenges	Strengths	Examples
Administrative or Observational				
Survey or Behavioral				

Data sources

	Collected by	Challenges	Strengths	Examples
Administrative or Observational	Public entity or implementation partner	Obtaining, Matching, Processing	Cheap or free, Large quantities, Exhaustive	Census data, school records, police files, financial transactions, satellite images
Survey or Behavioral				

Data sources

	Collected by	Challenges	Strengths	Examples
Administrative or Observational	Public entity or implementation partner	Obtaining, Matching, Processing	Cheap or free, Large quantities, Exhaustive	Census data, school records, police files, financial transactions, satellite images
Survey or Behavioral	Researcher or subject	Collecting, Targeting, Making unbiased	Direct, Custom, Quality Controlled	Household surveys, experimental surveys, opinion polls

PPOL 6818 - Spring 2025

Administrative data

Characteristics of admin data

GEORGETOWN UNIVERSITY

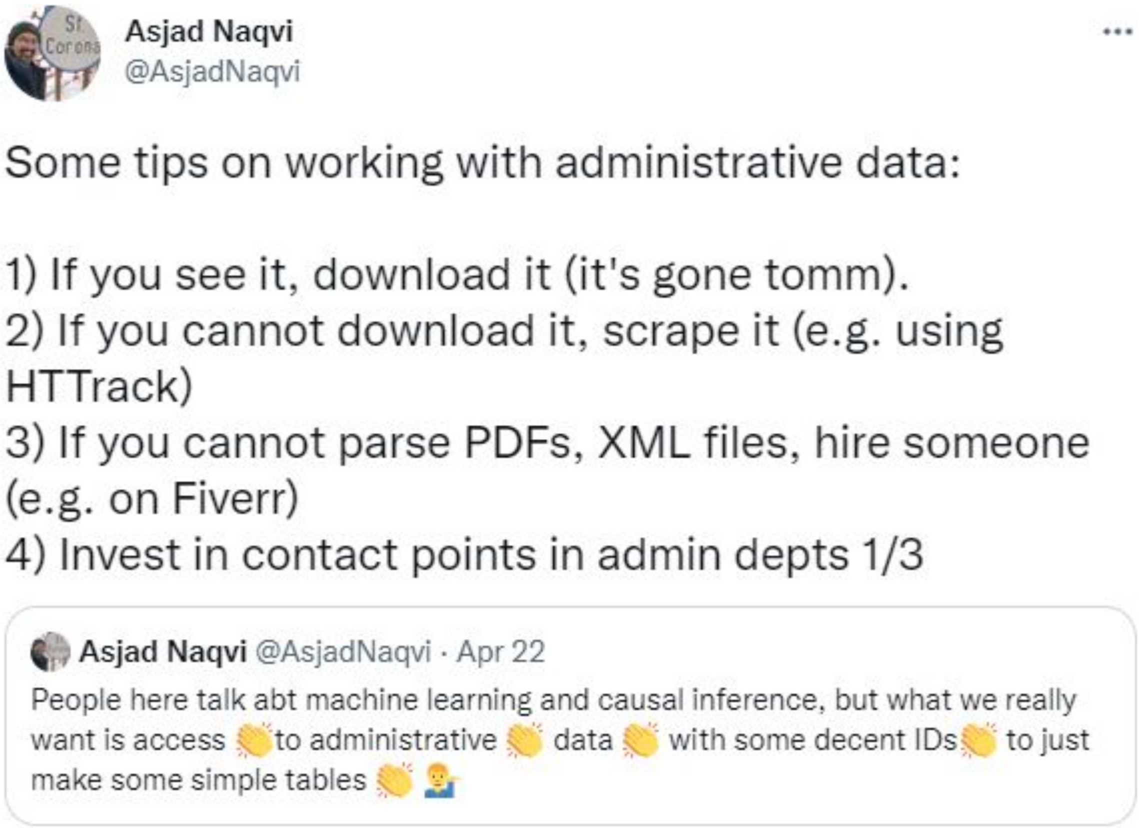
Admin data characteristics

- For the **universe of observations**

- Data is **generated for functional purposes**
  - usually by the **entity** managing that universe
  - can be human or machine generated
  - can be collected or observed

- Data is **obtained for research** purposes
  - data sharing, cleaning, matching, storing issues

Admin data tips



Asjad Naqvi
@AsjadNaqvi

Some tips on working with administrative data:

1) If you see it, download it (it's gone tomm).
2) If you cannot download it, scrape it (e.g. using HTTrack)
3) If you cannot parse PDFs, XML files, hire someone (e.g. on Fiverr)
4) Invest in contact points in admin depts 1/3

Asjad Naqvi @AsjadNaqvi · Apr 22
People here talk abt machine learning and causal inference, but what we really want is access 👏to administrative 👏 data 👏 with some decent IDs👏 to just make some simple tables 👏 💁‍♂️

PPOL 6818 - Spring 2025

Survey data

Characteristics of survey data

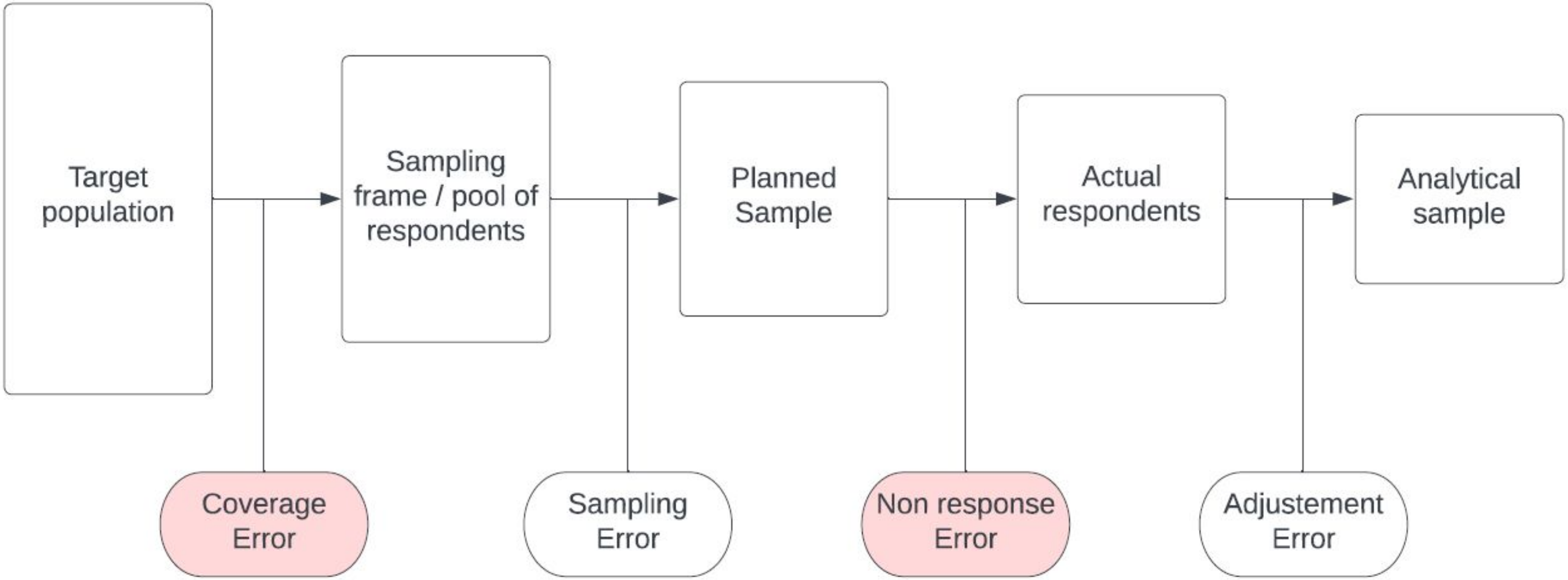
GEORGETOWN UNIVERSITY

Survey characteristics

- **Sample vs universe**
  - opinion polls : sample of the population
  - consumption modules : sample of behaviors (e.g. 7-day, 30-day recall period)

- **Data is generated for research purposes**
  - responses to direct questions
    - interpretation bias
    - desirability bias
  - data collection exercise biases
    - measurement error

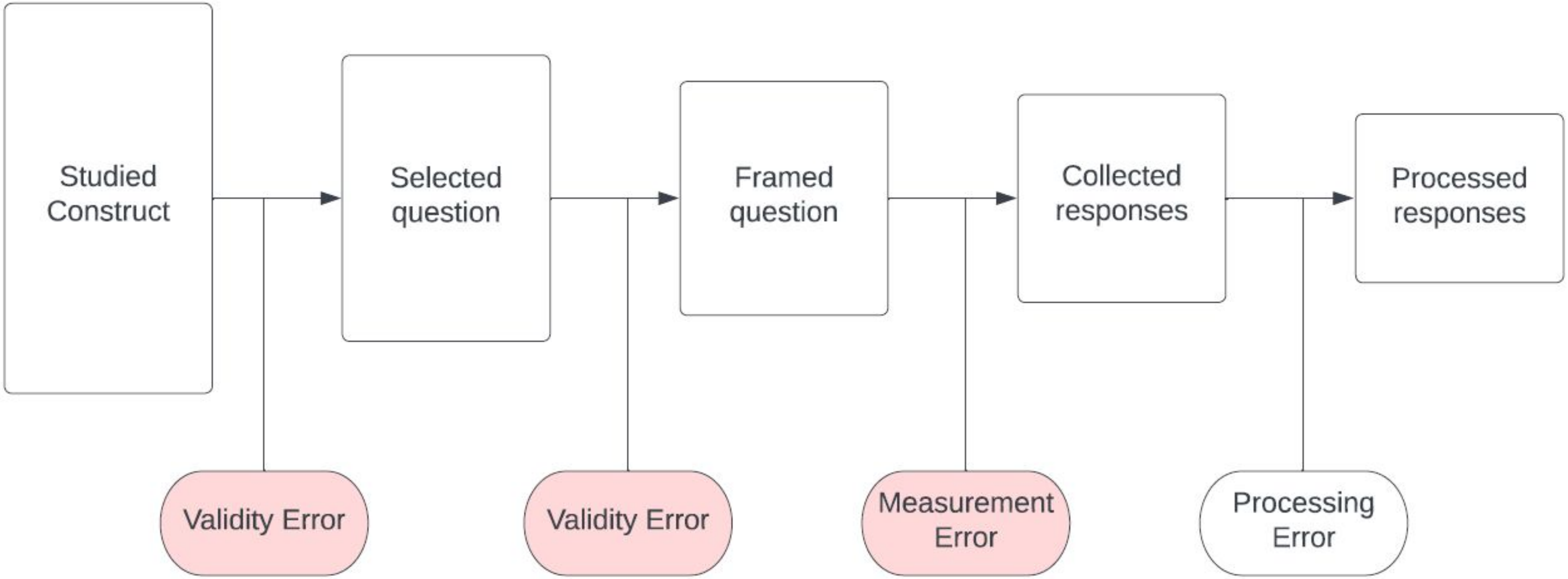
Total Survey Error : target (who)



Is our list of possible subjects correctly mapping the study subjects?

Item non response (missings) or survey non response (attrition)

Total Survey Error : construct (what)



Are the questions correctly covering the concept studied?

Incorrect responses (lie, memory, ...)

Measurement Error

How bad can it really get?

17 89

GEORGETOWN UNIVERSITY

TSE : empirical measures

- Compare survey data to admin data
  - overall difference = total survey error
  - individual differences = item non response, measurement errors

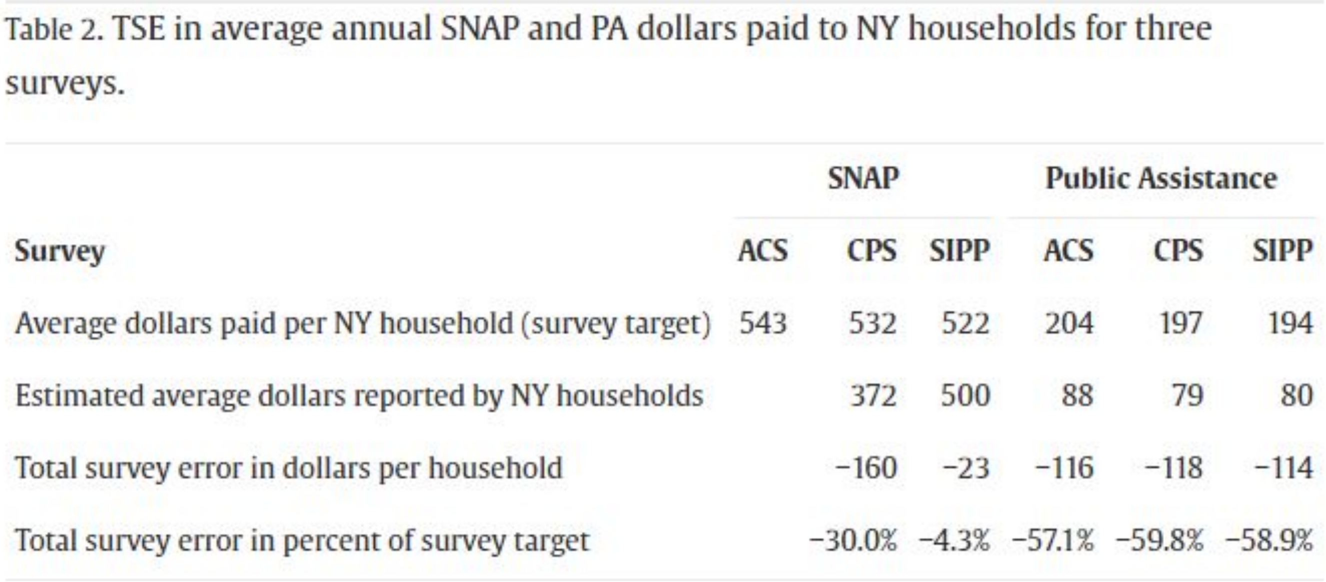


Table 2. TSE in average annual SNAP and PA dollars paid to NY households for three surveys.

	SNAP			Public Assistance		
Survey	**ACS**	**CPS**	**SIPP**	**ACS**	**CPS**	**SIPP**
Average dollars paid per NY household (survey target)	543	532	522	204	197	194
Estimated average dollars reported by NY households		372	500	88	79	80
Total survey error in dollars per household		−160	−23	−116	−118	−114
Total survey error in percent of survey target		−30.0%	−4.3%	−57.1%	−59.8%	−58.9%

TSE : empirical decomposition

- Compare survey data to admin data
  - overall difference = total survey error
  - individual differences = item non response, measurement errors

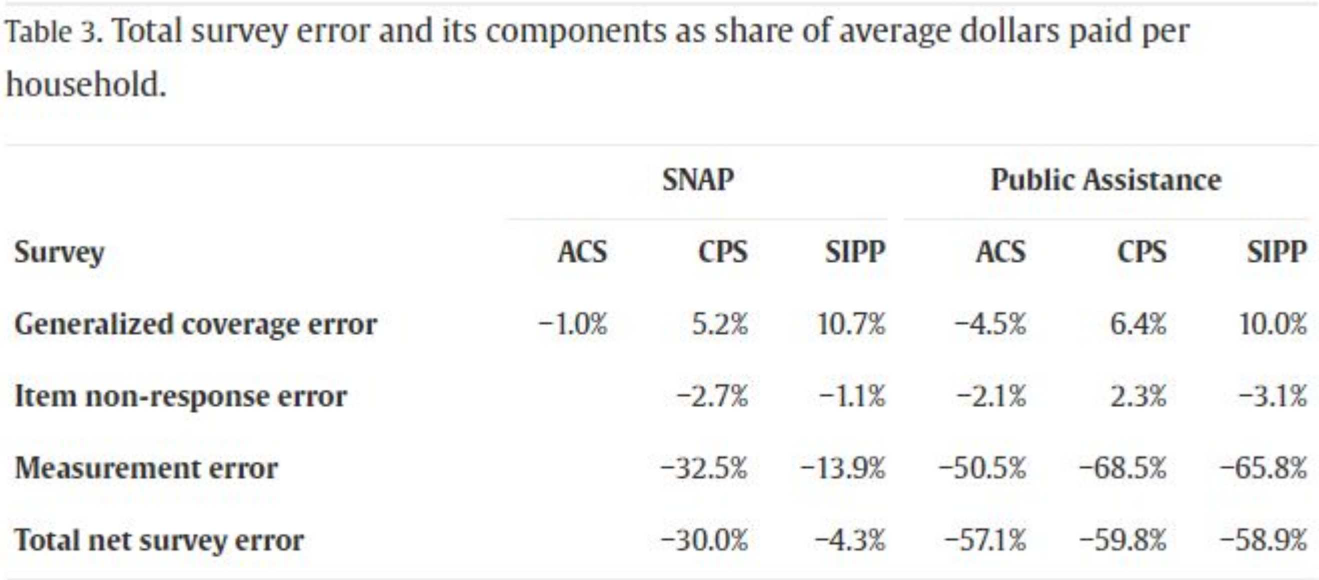


Table 3. Total survey error and its components as share of average dollars paid per household.

	SNAP			Public Assistance		
Survey	**ACS**	**CPS**	**SIPP**	**ACS**	**CPS**	**SIPP**
Generalized coverage error	−1.0%	5.2%	10.7%	−4.5%	6.4%	10.0%
Item non-response error		−2.7%	−1.1%	−2.1%	2.3%	−3.1%
Measurement error		−32.5%	−13.9%	−50.5%	−68.5%	−65.8%
Total net survey error		−30.0%	−4.3%	−57.1%	−59.8%	−58.9%

Why so large?

- Measurement “errors” are **not random**
  - tend to be in the same direction (e.g. under-reporting of income) and correlated with other variables

- **Non sampling errors are rarely estimated**
  - weights and confidence intervals typically focus on measuring and reporting sampling errors
  - non-sampling measurement errors are harder to document and correct for

Why does the type of error matters

- Different errors lead to different **interpretations**
        - eg with reporting of cash transfer program
    - if it is a coverage (sampling) error
        - the result may not be generalizable to the population
    - if it is a measurement (non sampling) error
        - due to under-reporting recipient status, the survey may not accurately reflect take-up and targeting
        - due to under-reporting amounts, the survey may not accurately reflect the poverty impact of the program

Admin data has measurement error too

- Often subject to **data entry**
  - same response, coding and processing errors
  - same potential biases due to reporting incentives
    - examples? lack of reporting, under-reporting
- **Rarely well maintained**
  - a lot of non-response errors
- **Magnitude** (examples from the literature)
  - 24% of the variance in Dutch official hourly wage records was random measurement error
  - 20% - 30% of osteoarthritis cases are not registered in Quebec hospital administrative records
- **Tricky to estimate**
  - admin data used to estimate survey data error
  - survey data used to estimate admin data error

So what do we do?

- **Data exploring**
  - know your data and understand its sources of biases
  - know orders of magnitude
- **Data cleaning**
  - Manual vs automatic entries
  - Imputation
  - Weighting
  - Winsorizing (outliers)
- **Triangulation**
  - with other sources of data
- **Sensitivity analysis**

PPOL 6818 - Spring 2025

Surveying for Experiments

Process

GEORGETOWN UNIVERSITY

Computer assisted surveying

- **Minimize errors at data collection**
  - design for more intuitive survey flow
  - constrain responses
- **Minimize errors at data cleaning**
  - ready-to-use dataset
  - monitor survey responses live
- **Implement interactive surveys**
  - case management to track survey attempts
  - connectivity to pull data from other/previous surveys

Know your goal

In some circumstances these convenience features are not actually desirable.

In which cases may we prefer :

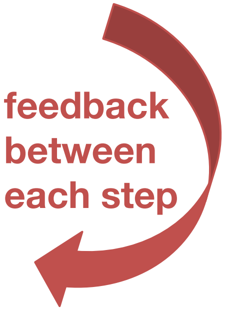
- **Paper based surveys**
- **No constraints to the data**
- **No input from previous datasets**

Major steps of good data collection

0. Outcomes definition
1. Stocktaking project codebook & previous surveys
2. Survey writing
3. Survey coding
4. Survey deployment
5. Data management
6. Data exporting
7. Data monitoring and quality checks

Major steps of good data collection

0. Outcomes definition
1. Stocktaking project codebook & previous surveys
2. Survey writing
3. Survey coding
4. Survey deployment
5. Data management
6. Data exporting
7. Data monitoring and quality checks



feedback
between
each step

Major steps of good data collection

0. Outcomes definition

week 4: research design

week 2 : project structure, documentation

1. Stocktaking project codebook & previous surveys
2. Survey writing
3. Survey coding

this week

4. Survey deployment
5. Data management
6. Data exporting
7. Data monitoring and quality checks

Major steps of good data collection

0. Outcomes definition

week 4: research design

week 2 : project structure, documentation

1. Stocktaking project codebook & previous surveys

2. Survey writing

week 12 : survey optimization

3. Survey coding

this week

4. Survey deployment

5. Data management

6. Data exporting

week 7 : survey deployment

7. Data monitoring and quality checks

week 11 : data quality

Georgetown University Initiative

gui2de

on Innovation, Development and Evaluation

www.gui2de.org