Introduction to Survey Statistics – Day 3
Measurement in Surveys

Federico Vegetti
Central European University

University of Heidelberg

Sources of error in surveys

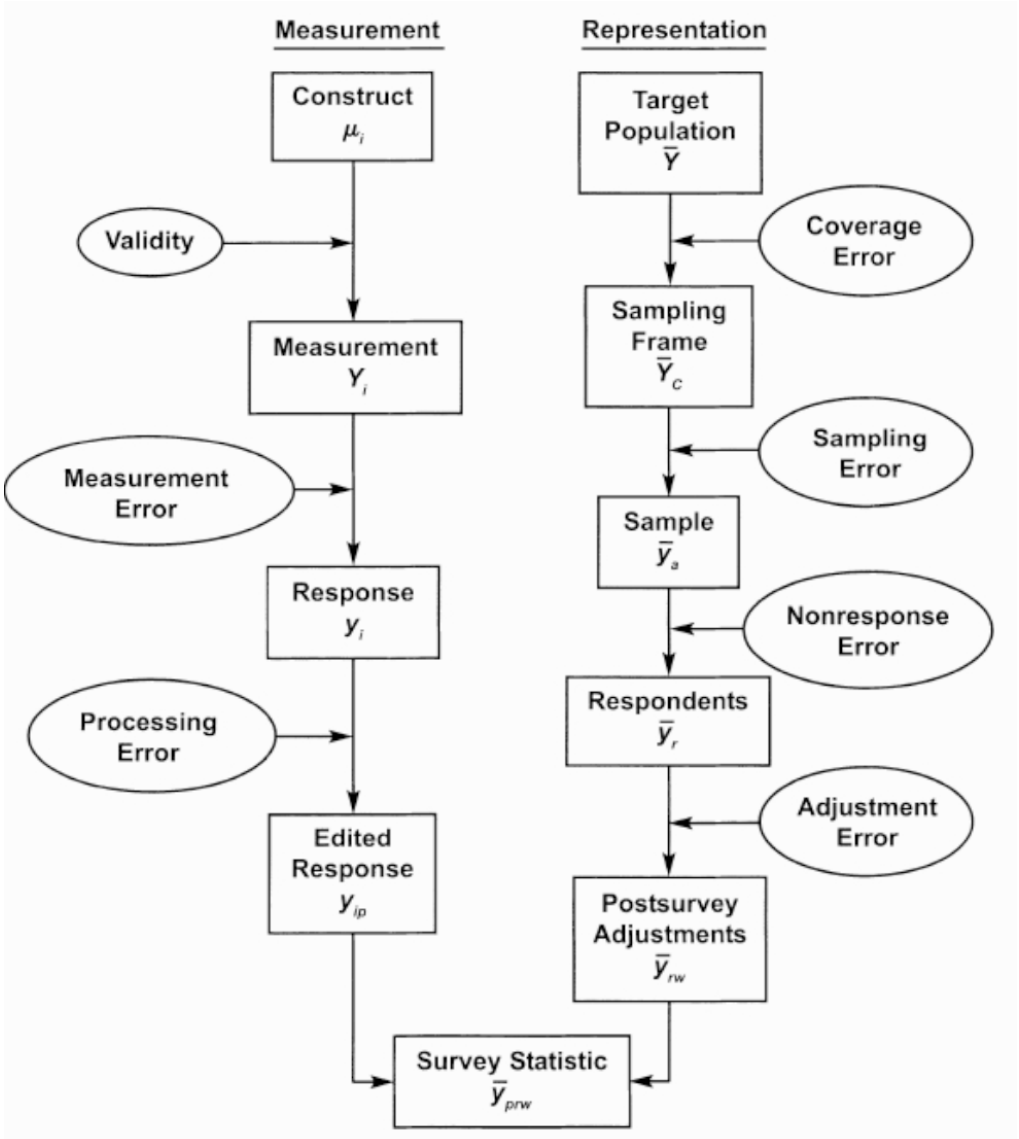

Figure 1: From Groves et al. (2009)

Measurement in surveys

- **Measurement error** is the difference between the values that we observe for a given observation, and the true values for that observation
- This can be reduced in the design stage of the survey by enforcing some good practices
- Even if we use secondary data, it is good to know what are the sources of it

Measurement in surveys (2)

- The quality of measurement in surveys also depends to a great extent on the **validity** of the questions we ask
- This is also reduced in the design stage, but has roots even earlier, during theory development
- With secondary data, we often use proxy measures for the concepts that we talk about. This may pose a validity problem
- There are diagnostics that can be performed once the data is collected to prove (or disprove) the quality of measurement

Measurement error

- When we interpret the data, we abstract from an observed quantity to a concept
  - E.g. 65% of respondents to a survey in Germay said that they either "*Agree*" or "*Strongly Agree*" with the statement "*Income and wealth should be redistributed towards ordinary people*"
  - This implies that the content of the statement resonates with their own attitude
  - We write that "*two thirds of German citizens are in favor of economic redistribution*"
- To understand measurement, we need to take this path backwards
- How does an attitude convert into a survey response?

The response process

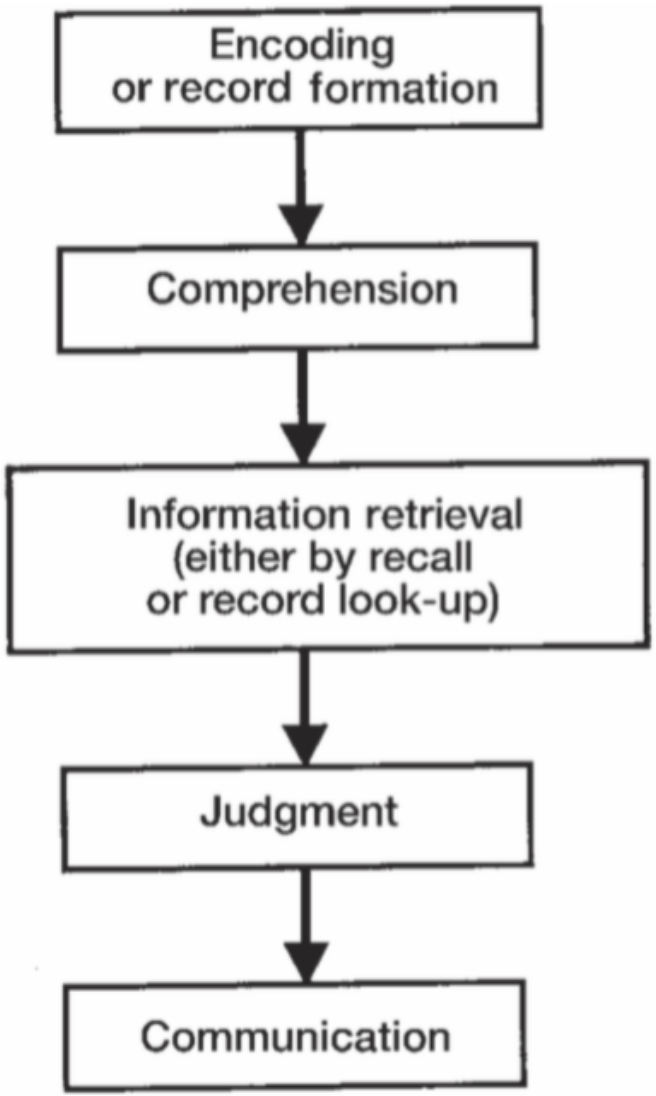

Figure 2: From Biemer and Lyberg (2003)

The response process (2)

- **Encoding**: the process of forming memories of events, recording information
- **Comprenehsion**: interpreting the question
- **Retrieval**: looking for the information necessary to answer the question
- **Judgment**: combining the information in a way that is useful to answer the question
- **Communication**: formulating the response (in the right format required by the survey)

Encoding

- This process happens before the survey takes place, possibly long before, when information is acquired and stored in memory
- However, it is very important: people can't provide information that they don't have!
- If some information has not been encoded, no matter how good the question, the response we get will not be accurate
- This is relevant for us with respect to what we can reasonably expect that respondents know
    - E.g. If we ask a young adult if her family had issues paying the bills when she was 14, we should keep in mind that the parents might have not talked about this topic with her

Comprehension

- The respondent reads the question and tries to understand what information is required
- This includes the instructions that come with the question
  - E.g. "*Now think of a family member*"
- The biggest problem arising at this step is when respondents **misinterpret** the question
- What we can do at this step is to try to avoid some known traps to comprehension

Misinterpreting survey questions

Some sources of comprehension problems (see Groves et al. 2009)

1. **Grammatical ambiguity**
2. **Excessive complexity**
3. **Faulty presupposition**
   - "*Immigrants commit more crimes than locals because they are being discriminated against*"
4. **Vague concepts**
5. **Vague quantifiers**
   - E.g. *Often*, *Somewhat*
6. **Unfamiliar terms**
7. **False inferences**
   - "*Under certain circumstances, it is acceptable that a policeman strikes a male citizen*"

Misinterpreting survey questions (2)

- There is a trade-off between vagueness of the concepts and complexity of the question
- We want to be as clear as possible, but at the same time avoid putting too much cognitive burden on the respondents
- It helps to think of how we would posit the question in a conversation

Retrieval

- The process of searching one's own memory for the information that is needed to answer the question
- In general, our ability to retrieve an object from memory depends largely on the characteristics of the object itself
- Events are easier to recall when they are
  - **Recent**
  - More **distinctive** (e.g. if you watch a lot of movies it is always difficult to answer the "favorite movie" question)
  - Close to **temporal boundaries** or other easily recalled events (e.g. you may find it easy to remember what you did on September 11, 2001)
  - **Important**, or otherwise emotionally involving for you

Retrieval error

- There are also other sources of retrieval error
- **Distortion in the representation of events**
  - It can be very difficult to distinguish in our memory what experiences we had first hand and what comes from other sources of information regarding the event
- **Reconstruction errors**
  - We tend to "connect the dots" and fill missing pieces of incomplete memories
  - E.g. we tend to remember the past by examining the present backwards, ignoring the change
  - However, when we think there has been a change, we tend to exaggerate it

Telescoping

- Telescoping is a kind of retrieval error where past events are remembered either closer (forward) or further away (backward) from the present moment
- Some survey questions ask whether a certain event occurred to the respondents, or how many times it occurred, within a certain reference period
  - E.g. How many jobs did you have in the past 5 years?
- Telescoping may lead to a bias when events that happened outside of the reference period are thought to have happened inside of the reference period

Judgment

- At this stage, the information retrieved is evaluated and a response that fits the format required is formulated
- It is useful here to distinguish between two kinds of question

1. **Behavior reports**
   - "*How many times did you watch TV for 1 hour or longer in the last seven days?*"
   - "*Did you vote at the European Parliament elections of 2014?*"
2. **Attitudes** or **judgments**
   - "*Do you think that the economic situation in Germany is now better or worse than 12 months ago?*"
   - "*Generally speaking, do you think that Germany's presence in the EU is a good or a bad thing?*"

Behavioral questions

- It is reasonable to expect that most people do not keep a running tally of how many times a specific event in their lives has occurred
- Hence, most respondents will have to make an estimation
- Several strategies
  - **Recall-and-count**
    - They remember specific events and sum them up
  - **Rate-based estimation**
    - They recall the rate by which events typically occur and make an inference to the reference period
  - **Impression-based estimation**
    - They guess a number from a vague impression

Behavioral questions (2)

- Generally, the more events are to be recalled, the larger is the error
- This is why response options tend not to be too precise when the task is to count the number of events happened in a given period
  - E.g. *5 or more*, etc.
- Two biases:

1. **Overreporting**
2. **Underreporting**

Attitude questions

- When asked about an attitude or a judgment we may have already a clearly defined view on the topic
- However, in most cases, respondents need to form the opinion right on the spot
- Two different strategies
  1. **Top-down**: deriving an opinion from a general value
     - E.g. When asked an opinion about immigration policy, a respondent recalls her own views on tolerance and generosity
  2. **Bottom-up**: deriving an opinion from some specific beliefs about the topic
     - E.g. When asked an opinion about immigration policy, a respondent recalls how she felt when she dealt with immigrants

Attitude questions (2)

- When respondents need to form an opinion on the spot, they tend to be more susceptible to the effect of
- Question wording
  - If the wording contains labels that refer to more general values (e.g. if a position is presented as “left” or “right”) or some other information (e.g. endorsements) that may lead the response
- The context where the question is presented
  - Previous questions in the survey may prime the respondents into thinking of some specific aspects of the question

Communicating the answer

- The last step is to map the answer into the right format and to communicate it
- Here the respondent can choose whether to provide the most accurate response or to answer strategically
- What is a **strategic** survey response?
  - A respondent may be not motivated to be accurate, but rather to give a certain impression to the interviewer
  - In this case, the survey response will not be genuine

Social desirability bias

- When we ask questions about sensitive topics, respondents may not want to admit an opinion or a behavior that they perceive as unpopular
  - E.g. questions about use of illegal drugs, vote for some specific parties, etc.
- This produces a systematic bias in the data
- It is more common in face-to-face interviews than in self-administered surveys (e.g. online surveys)

Fear of disclosure

- This is a special type of social desirability bias, when respondents are afraid of possible negative consequences that they will incur if they admit to have done something or to have a certain opinion
  - E.g. a wealthy person who cheated on the income tax form may underreport her income for fear to be caught
- The only way to limit this is to assure and possible prove that the data will be kept confidential

Acquiescence

- This is yet another special type of social desirability bias, when the respondent thinks that the interviewer (or the survey commissioner) wants to hear some specific responses
  - E.g. market surveys where people are asked to evaluate a product or a service
- Sometimes it is enough that the question has an implied direction to cause aquiescence
- To limit this bias, the question should be formulated as neutral as possible
- Multi-item batteries should contain a balanced number of positive and negative items with respect to the trait that they aim to measure

Formatting the answer

- The way the response has to be formatted to fit with the requirement of the survey may affect the response itself
- Some very common item types in standardized surveys are **ordinal** and **categorical**
- They both present some specific biases that are driven by the format

Ordinal items

To what extent do you agree or disagree with the following statement: In times of crisis, it is desirable for Germany to give financial help to another EU Member State facing severe economic and financial difficulties.

1. *Totally agree*
2. *Tend to agree*
3. *Tend to disagree*
4. *Totally disagree*

Issues with ordinal items

- Respondents tend to avoid the extreme categories of the scale
- The labels can affect the answer
  - Labeling all categories is generally better than labeling only the ending categories
- Fewer categories may reduce the variance, but too many categories may introduce useless measurement error
  - What is the difference between 7 and 8 on a 1-10 scale?

Unordered categorical items

- Here the response options have no natural order

Which party did you vote for in these recent European Parliament elections?

1. *CDU*
2. *SPD*
3. *Greens*
4. *FDP*
5. *AfD*
6. *Linke*

Issues with unordered items

- **Primacy effect**: options that come **first** have more chances to be picked
  - More common when respondents have to read through the options by themselves
- **Recency effect**: options that come **last** have more chances to be picked
  - More common when the interviewer reads the options to the respondents

Other sources of error

- Another source of error is the **interviewing** process
- There is quite some research on how to detect and reduct interviewer effects
- Ideally, our survey data would include also indicators about the interviewer, at the very least an ID that we could use to control for interviewer fixed effects

Reliability

- A useful concept in psychometrics is the one of reliability
- A reliable instrument measures the same construct consistently
- This implies that
  - If you repeat the same measure on the same individuals over time, you should get similar results
  - If you have a multi-item scale, the latent factor should be measured consistently by all items

Reliability (2)

- Several types of reliability
  - Inter-rater reliability: to what extent different raters produce the same outcome to a given task?
    - Very common in content analysis
  - Test-retest reliability: to what extent thr scores are similar when the measure is repeated?
  - Internal consistency: to what extent different items of a scale produce similar values?
    - This is an important source of validation for your result in case you usea multi-item index

Cronbach's alpha

- Used to measure internal consistency
- It is a function of the covariance between the items in a scale

$$\alpha = \frac{N\overline{c}}{\overline{v}(N-1)\overline{c}}$$

  - Where N is the number of items in the scale
  - $\overline{c}$ is the average covariance between all items
  - $\overline{v}$ is the average variance among all items
- A higher score means that the index has better internal consistency

Cronbach's alpha (2)

- It is criticized because it tends to grow as a function of the number of items
- However, it is one of the most common statistics that you calculate to check the goodness of your multi-item indicators
- As a rule of thumb, a value of 0.7 or higher indicates a good internal consistency

Validity

- Whereas measurement error has something to do with our ability to measure an already-defined construct, validity has more to do with the definition of the construct itself
- In many cases it is hard to tell whether a measurement problem is due to error or validity issues
- A measure can capture a concept very well, and still what we infer from it can be not valid at all
  - For instance, in case we are doing a bad conceptual stretch
- In the same way, a valid measurement can be poorly designed, and produce error for all the reasons we saw

Validity and theory

- The most important thing in order to have a valid measurement is to have a clear concept
- Example: populism
  - What are the distinctive features of a populist mindset?
  - How is it different from anti-immigrant attitude?
  - How is it different from authoritarianism?
- Looking at several media reports and op-eds about populism, the concept is not clear to most people

Validity and theory (2)

- One question to ask when you are writing your item is: who would answer with the highest value? And who would answer with the lowest?
- If you can not answer these questions (or if you can answer only for one direction), perhaps your item needs better calibration

Convergent validity

- There are some ways to provide evidence that your measurement is valid
- One is **convergence**
- Does it correlate with other variables that measure the same construct?
- This requires that there exist some measures of the same concept already

Discriminant validity

- By the same logic, you can look if your measure correlates with other indicators that are supposed to measure different things
- If that is the case, your measure can isolate the specific trait without conflating it with other concepts

Criterion validity

- Similar to convergent validity, but based on a concrete outcome
- Example:
  - Does your scale of populism predict the vote for a populist candidate?
  - Does it predict "liking" of populist posts on Facebook?

References

Biemer, Paul P., and Lars E. Lyberg. 2003. *Introduction to Survey Quality*. Hoboken, NJ: Wiley.

Groves, Robert M., Floyd J. Fowler Jr, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. 2009. *Survey Methodology*. 2 edition. Hoboken, N.J: Wiley.