Measurement

Fill In Your Name

15 February, 2021

What to measure

How to measure

What to measure

Measurement

- Measurement is an essential part of your research design.
  - Measurement follows from your theory of the way you think the world works and how you think your treatment manipulates that world. That is, measurement is a theoretical exercise as much as it is a technical and logistical one.
  - When we record a number, a word, or a letter, in a dataset to **represent** something more or less abstract in the world (like "financial stability," "hunger," or "math ability"), we are **measuring**.
  - Problems with measurement can lead you to draw incorrect (causal) inferences from your study (systematic error).
  - Noisy measurement reduces power (random error).
  - Data collection often takes up a very large portion of the time and financial resources available in the project budget.
  - New data can be a useful research output in its own right and an important foundation for future research. Data is a public good!

From Concept to Data or Scores (1)

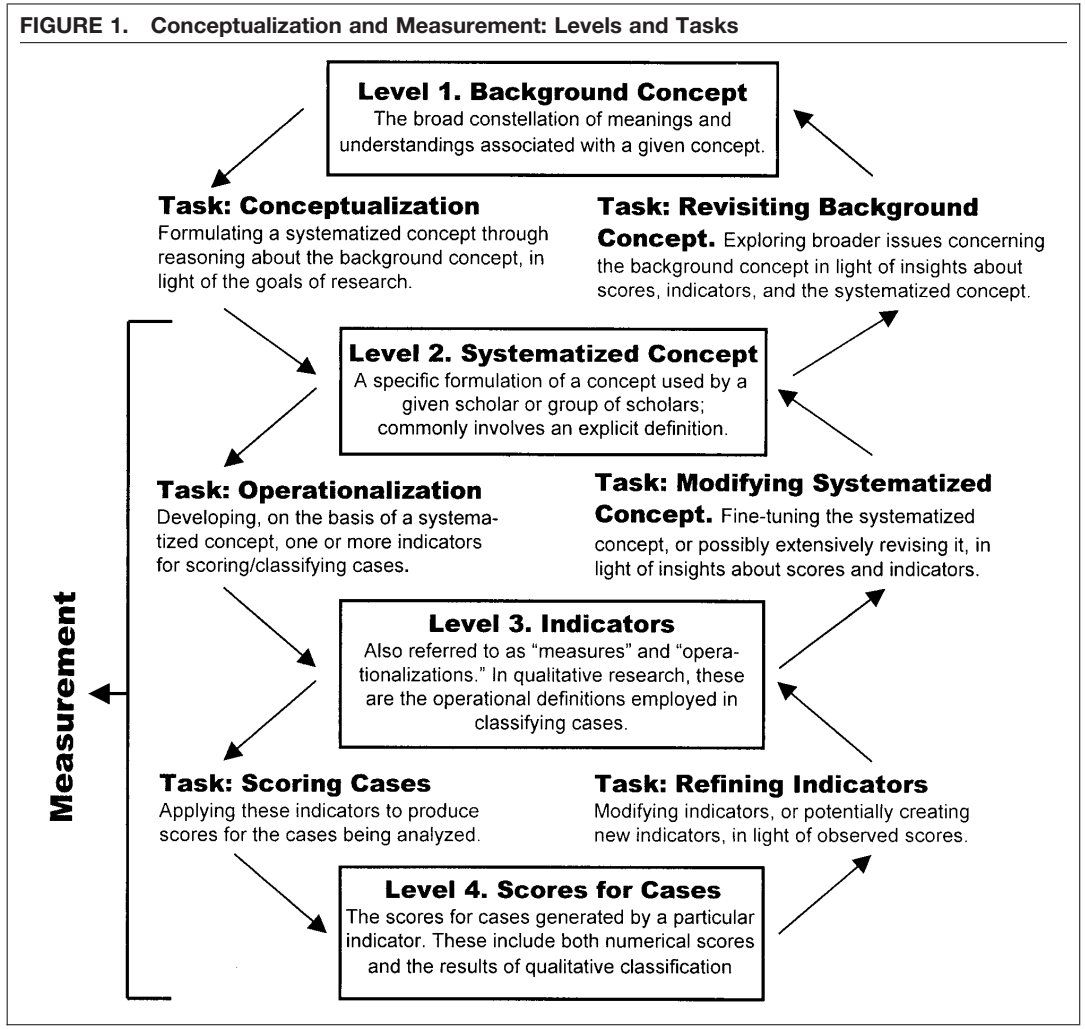

Figure 1: Measurement from concept to scores (or recorded data). From Adcock and Collier (2001).

From Concept to Scores (2)

- What is an example of a concept, like an outcome, you'd like to change with an experimental treatment?
- How would you know that a unit (like a person or a village) displayed high or low "ability" or "rank" on that concept? (Was "hungrier" or "happier" or "more peaceful?") What might you observe that would make you feel more or less confident that this person was hungrier or happier or less supportive of violence than this other person?
- What criticisms might you face if you said, "I think that this unit differs from this other unit on my conceptual outcome?"
- Note: we are talking about "measurement **validity**" here.
  - A valid measure is one that we can persuasively argue represents what we say it represents, and it doesn't represent other aspects of the unit.) (see Shadish et al. (2002) for more on validity and reliability of measurement).

From Concept to Scores (3)

- Does this measure "math ability?" Or knowledge of boating? *"A boat that can make forty miles an hour in still water makes a trip of one hundred miles down a certain stream. If this trip takes two hours, how long will the return trip take?" (from the 1926 SAT in the USA)*
- Does this measure "verbal ability?" Or knowledge of boating?

From Concept to Scores (4)

Choose the pair of words most like the top pair:

Runner : Marathon a) envoy : embassy b) martyr : massacre c) oarsman : regatta d) horse : stable

Correct answer is: c) oarsman : regatta. (from the 1980s SAT in the USA)

- Does this measure knowledge of elite college sports? Or "verbal ability?"

What should you measure? (1)

- You should measure *everything*:
  - Outcome for all units in your study (for each wave, and any attrition)
  - Treatment (assignment and compliance)
  - Indications that your treatment is received and interpreted as you expect (manipulation check)
  - Indications of possible harm being done by your research
  - Covariates, including context that might (a) affect how your intervention could work or (b) influence the variability in your outcome.

What should you measure? (2)

- We often have multiple theories for *how* an intervention might affect an outcome (different mechanisms).
- Measure indicators that are unique to each mechanism and that can help differentiate between them.
- Such indicators might include intermediate outcomes that are realized *before* the final outcome.
- They may include secondary outcomes such as
  - outcomes for which we expect effects only under some theories
  - placebo outcomes for which we expect no effects.

What should you measure? (3)

- The ideal case is direct measurement of the concept or phenomenon of interest with no error (rarely possible).
  - Treatment assignment under your control is the rare exception and something crucial to record.
- We are often only able to measure **indicators** connected to but not entirely determinative of the underlying concept or phenomenon of interest
  - Correct answers to specific problems (indicators) for underlying mathematical aptitude (the actual phenomenon).
  - Days without food (indicators) for hunger (the actual phenomenon).
- Reasonable people may disagree on the conceptualization. What do we really mean by mathematical aptitude or hunger? What else might our scores tell us about our people, villages, etc. in addition to what we hope to represent?

What should you measure? (4)

- Select indicators that closely connect to the what we **mean** when we talk about the phenomenon of interest. Select **valid** indicators. Ex. Math measures should not also measure boating knowledge.
- Select indicators that are **reliable**. Ex. A meter stick should always measure a meter no matter the temperature. A rubber meter stick can produce **unreliable** measurements of the abstract concept of length, especially if used by a 5 year old. A laser might produce more reliable measurements than even a wooden meter stick in the hands of a skilled user of wooden meter sticks.

What should you measure? (5)

- If you have *multiple indicators for the same phenomenon*, you will need to determine how to aggregate these indicators. Often, multiple indicators can improve **reliability** of measurement and perhaps bolster arguments for **validity**. Combining a wooden stick and a laser might be ideal to convince us that we have measured length in a way that we understand as **length** (rather than, as, say, temperature) *and* that the particular score does not depend crucially on the context of the measurement.

How to measure

Tools and Sources (1)

- After we determine what concept to measure and how we might know it when we see it, we need to figure out **how** to measure it.
- You need to maintain symmetry between treatment and control groups as you do the measurement.
  - How many times and for how long you interact with participants should be the same across all treatment arms.
  - The questions should be the same.
  - Be particularly careful on these points for indicators used for manipulation checks.
  - **Measurement should not be the experimental intervention.** We want the only difference between treatment arms to be the intervention, not the measurement.

Tools and Sources (2)

- We have various tools and sources:
  - Survey measures
  - Behavioral measures
  - Administrative data (tax records, election results, etc.)
  - Images/remote sensing
  - Text (speech transcripts, newspapers, etc.)
  - Sensors on wearables, phones, other devices (household items)
  - Others

Considerations in choosing tools and sources

- Different tools imply different trade-offs:
  - Validity (Does the score capture the concept and only the concept?)
  - Reliability (Would the same score used twice yield the same answer?)
  - Bias (systematic error)
  - Precision (random error)
  - Sample for which you can make measurements
  - Timing of the measurements
  - Cost

Measurement error (1)

- We want to avoid/minimize two types of measurement error:
  - Systematic error (bias)
  - Random error (lack of precision)

Measurement error (2)

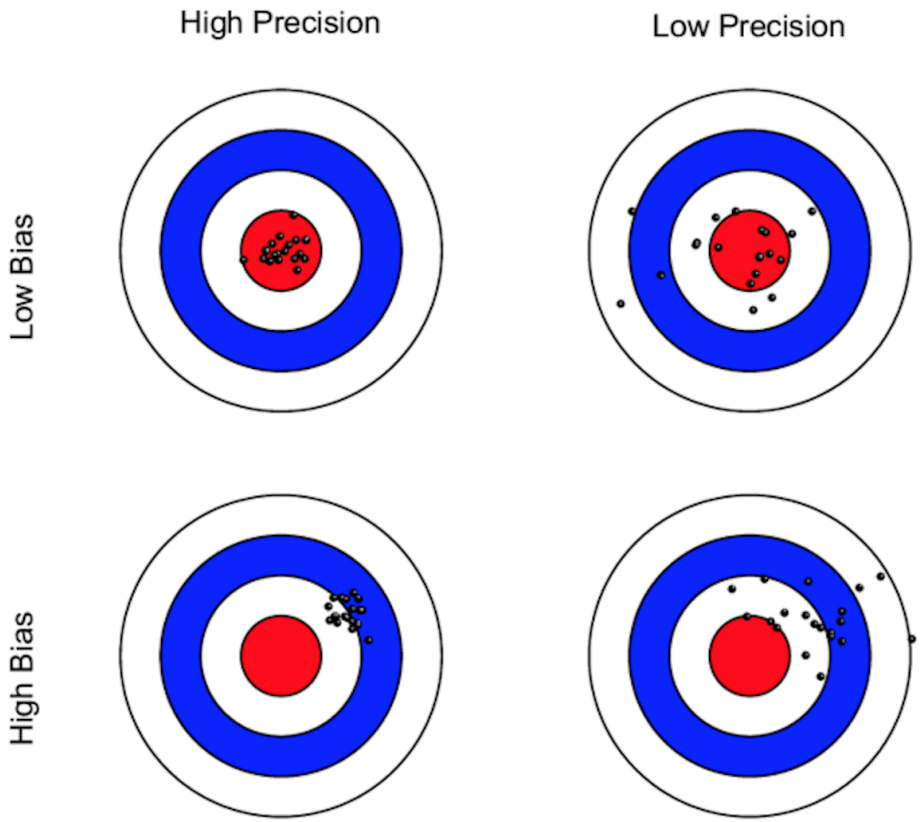

Figure 2: Bias and Precision

(Source: Precision and Bias of Masurement)

Measurement error examples

- Systematic error
  - A scale that is calibrated incorrectly, so everyone appears to be 2kg lighter than they actually are
  - A food diary that consistently under-reports snacks
  - Demand effects
  - Hawthorne effects
- Random error
  - A shaky hand while marking distance
- Notice that these are examples of **unreliable** measurement. They could also be **invalid**, but need not be.

Consequences of mismeasuring the treatment

- In addition to generating an incorrect description of the level of some phenomenon, measurement error can affect our causal inferences.
- If the treatment variable is binary (a unit can be in treatment or control only), then measurement error is negatively correlated with the true variable. (A 1 is miscoded as 0, so the error is -1; a 0 is miscoded as 1, so the error is 1.)
- If you use OLS to calculate estimators of average treatment effects, this kind of error leads to smaller estimates of the causal effect (coefficient on the treatment variable).

Consequences of random measurement error

- In an OLS analysis, greater random error in the outcome variable leads to less precise estimates of the causal effect (coefficient on the treatment variable).
- Reducing random measurement error in the outcome can increase statistical power (because the outcome has less non-treatment related noise.)

Consequences of systematic measurement error (1)

- If all measurements are off by the same amount, like -2kg for everyone:
  - this makes no difference when treatment effects are defined as differences in potential outcomes,

$\tau_i = Y_i(1) - Y_i(0)$, $\tilde{\tau}_i = (Y_i(1) - 2) - (Y_i(0) - 2)$, so $\tau_i = \tilde{\tau}_i$

  - but this is problematic if the treatment effect is defined as the ratio of potential outcomes ($Y_i(1) > 0$ and $Y_i(0) > 0$).

$\tau_i = Y_i(1)/Y_i(0)$, $\tilde{\tau}_i = (Y_i(1) - 2)/(Y_i(0) - 2)$, so $\tau_i \neq \tilde{\tau}_i$ except when $Y_i(1) = Y_i(0)$.

- How far off $\tilde{\tau}_i$ is from τ_i depends on how large 2 (the error) is relative to the actual values $Y_i(0)$, $Y_i(1)$.
- Notice that logistic regression coefficients are ratios of potential outcomes.

Consequences of systematic measurement error (2)

- Measurement error may be correlated with the true value of Y.
- For example, people who engage in frowned-upon, embarrassing, or illegal behavior may under-report that behavior, while those who do not may report their level accurately. (This is known as social desirability bias.)
  - This under-reporting may happen with victims of this kind of behavior by others, who blame themselves or fear sanctions by others who would rather not know about the behavior (e.g., victims of intimate partner violence).
- This makes is more difficult to detect an effect of an intervention designed to reduce this behavior.

Consequences of systematic measurement error (3)

- Another form of social desirability bias may also lead to measurement error being correlated with the treatment.
- For example, your intervention might aim to reduce hostile attitudes towards members of other social groups. If participants can figure out the goals of your study, they may (subconsciously) try to please the researcher by telling him what he wants to see. Those in the treatment group may under-report their hostility towards other groups compared with the control group.
- This makes it difficult to know whether the difference in observed outcomes between treatment and control groups is due to the intervention actually reducing hostility or knowledge of the treatment changing reporting of hostility.

How can we limit measurement errors?

Some options:

- Self-reporting by a subject (on a survey) is more problematic than unobtrusive observation of the subject ("in the wild") by someone else.
- Behavioral measures are less subject to social desirability bias.
- Administrative records for which misreporting has legal penalties might be more accurate.
- Providing more privacy so that scoring can happen without observation by others or the experimenter.
- Keeping some hypotheses and aims of the study hidden from study participants.
- If you can't control measurement error, study it — figure out whether it's a problem and how large. Consider measurement oriented pilot studies.

Example - administrative records

- Attendance records for a meeting, instead of asking whether someone attended.
  - There may only be regular attendance records for meetings that are not of interest to your original target population.
  - You may need to plan ahead for data collection at the meeting instead.

Example - behavioral measures (1)

- Ask subjects to induce effort like sign a petition, make a donation, or do some other task which has a small personal cost, instead of asking subjects whether they support a particular issue.
  - This may only capture subjects who care strongly about that issue.
  - Example: Pedro Vicente's work on vote-buying in Sao Tome and Principe uses surveys, administrative records, and whether respondents mailed a pre-paid postcard to measure outcomes. see Brief 20: Is Vote Buying Effective?

Example - behavioral measures (2)

- Play "lab games" to measure cooperation or generosity towards out-groups, instead of asking subjects whether they would cooperate with others.
  - Scacco and Warren's study of prejudice and discrimination uses variants of dictator games. see Can Social Contact Reduce Prejudice and Discrimination? Evidence from a Field Experiment in Nigeria

Example - cover or privacy (1)

- Provide "cover" so that respondents believe that their responses cannot be traced back to them.
  - Randomized response: chance determines whether the respondent should respond to a question truthfully or respond "yes" regardless of the truth. The enumerator does not know which condition chance put the respondent in.

Example - cover or privacy (2)

- List experiments: give respondents a list of items or statements and ask how many many are true for them. Respondents are randomized into seeing different lists, where one contains an additional sensitive item (ie “my husband beats me”). This allows the researcher to estimate the prevalence of a particular item. Note that this approach reduces power for a given sample size.
- Simple privacy: for questions such as vote choice, ask the respondent to fill out a mock ballot and place it in a locked box instead of responding directly to the enumerator.

Example - blinding respondents to hypotheses (3)

- See Scacco and Warren's study of social contact theory in Kaduna, Nigeria, where participants were recruiting into a computer skills program, not one advertised as a program to reduce prejudice and discrimination.

How much error is too much?

- Reducing measurement error is important, but can be quite costly. So how much do you need to do?
- Depends on the scale and your goals.
  - Compare size of errors to the size of the treatment effect. Changing attitudes is difficult. Social desirability bias may be large compared with small treatment effects on attitudes.
  - Compare size of errors to the possible range of that outcome. Being one cent off on your bank balance does not appreciably affect our measure of your overall wealth.

General advice for measurement (1)

- Start with standard practice for your indicators; indicators that the community of researchers has agreed represent the concept of interest. These will have been road-tested for you and comparability of measures across research studies and sites is a virtue.
- But be careful to consider whether standard indicators make sense. For example, how one measures income might differ between rich and poor areas. Similarly, how one measures political attitudes might differ in more or less democratic regimes.

General advice for measurement (2)

- Recall that the boating related measures of math and verbal ability were standard practice in the 1920s and 1980s in the USA. Remember that you are measuring social constructs and thus must pay attention, as much as you can, to what prior partially or unexamined beliefs you and the community of researchers bring to the study (for some more vivid examples of this issue see Gould ([1981] 1996)).
- Connect with other researchers in your subject area.
- Focus on finer measurement in the range of the variable where you expect change.

Resources for survey research and survey items

- Major data archives are searchable by topic. One major archive is ICPSR:

https://www.icpsr.umich.edu/web/pages/ICPSR/

- Pew Research Center on Questionnaire Design

https://www.pewresearch.org/methods/u-s-survey-research/questionnaire-design/

- Tools4Dev on Writing Questions

http://www.tools4dev.org/resources/how-to-write-awesome-survey-questions-part-1/

- University libraries will often have guides to data resources:

https://guides.libraries.emory.edu/c.php?g=944707&p=6810109

References

- 10 Things to Know About Measurement in Experiments

Adcock, Robert, and David Collier. 2001. "Measurement Validity: A Shared Standard for Qualitative and Measurement Validity: A Shared Standard for Qualitative and Quantitative Research." *American Political Science Review* 95 (3): 529–46.

Gould, Stephen Jay. [1981] 1996. *The Mismeasure of Man*. WW Norton & Company.

Shadish, William R, Thomas D Cook, Donald Thomas Campbell, and others. 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference/William r. Shedish, Thomas d. Cook, Donald t. Campbell.* Boston: Houghton Mifflin,.