MEASUREMENT

Introduction

Most research questions can't be directly studied. We have to rely on indicators for the concepts in them.

Example: What is the influence of religiosity on health?
What influences attitudes and values about inter-racial marriage?

Measurement = operationalizing concepts

The goal with most social research is to capture our observations of social life (our concepts in our research question) with as much systemic variance as possible and with as little extraneous variance as possible. (Goes back to original purpose of research design.)

We can measure anything that physically exists. The problem is that in the social sciences we often want to measure things that don't physically exist. Rather, they symbolically exist, such as prejudice, anomie, religiosity, health, attitudes and values about inter-racial marriage, intelligence...

These are IDEAS rather than THINGS. We can measure them too, but it is harder and there is more measurement error.

Need to use multiple questions to measure "ideas"

Ex: How to measure social class: Marx, Weber, economists, feminists

Classical Measurement Theory: $X_{ij} = T + e$

x=observed variable value
T=true value (latent variable)
e=error

x-T=e

Even the most directly observable variables are abstractions to some extent, and will have some error in measurement. Examples: Height, weight, income

With increasing error it becomes more and more difficult to find relationships in analysis.

Criteria for Measurement Quality

Precision — fineness of distinctions in measure. Want to measure concept/phenomenon so as to represent the variation that exists empirically. But sometimes, categorizing that variation makes more real life sense because of how the concept impacts a person's behavior/life

Example. Do you measure education in categories or number of years of formal schooling? Age? Income?

Reliability — whether your measurement technique applied repeatedly to the same subject or across different populations yields the same result.

Example. Scale measuring weight. Question measuring household income.

Validity — extent to which the empirical measure you come up with reflects the *real meaning of the concept*. Do you measure what you intended to measure?

Example. Scale measuring weight. Question measuring race (w, b, o?).

To obtain accurate measurement, you must consider:

- What to measure/observe: theory/concepts
- How to measure/observe: set of operations/instruments/instructions

Your RQ and your theory(ies) will influence both, as well as your knowledge of the population/study participants, and common sense

Research question

↓

C (conceptual level/theoretical level) (conceptualization) (latent variable)

↓

X indicator/observed level (operationlization): exactly how you measured concept

↓

T (empirical level; true value)

Ideally C= X=T, but this doesn't happen in real life. We have **measurement error**.

T is generally unknown, or incompletely known; so can't really assess true relationship between measurement levels (exceptions: Hurricane Evacuation study)

Examples: attitudes or beliefs – slippery; behavior less slippery
Hypothetical behavior – slippery

When we suspect we know what T means in the real world it is appropriate to go through a conceptualization process and develop a standardized measurement strategy (Quantitative research only)

Examples: satisfaction with a product; drug use; SPOT

When we are unsure that T exists, or what its properties are, then we disregard the formal development of X; instead focusing on direct observations of C (Qualitative research)

Examples: quality of marriage or friendship, children's play

Quantitative observations methods are strong on reliability (standardized measurement strategy); but qualitative methods maximize validity

MEASUREMENT ERROR: b/c the link between C and X is poor (validity), or link between C and T is poor (validity), or X and T is poor (reliability)

$$X-T=e$$

Types of error: random and non-random (systematic)

Reliability issues center on random error
Validity issues center on systemic error

1. Random error: error due to chance (coding errors, ambiguous instructions, interviewer fatigue, slight changes in how different interviewers ask questions, fluctuations in data collection setting such as whether it rained the day the person answered)

   Examples: happiness, parenting

   the more reliable, the less random error
   high reliability in a measure = variable doesn't fluctuate due to random error

a. Always have some random error: the process of imposing an " artificial" measure causes some random error

b. Random error can make X over or underestimate T:

random error often cancels out and has no big impact on analyses
Example: "how many times did you eat out last month?"

2. Nonrandom error (systematic): systematic biases in measures: always over or underestimates T

Examples: health, income, children's play

validity issue: the more systematic error, the less valid

if outside factors systematically influence X, that suggests that X may measure something other than T, (it measure something else, perhaps a related concept)

Example: health (is it measure with questions on physical health? Emotional health? Spiritual health?

non-random error causes major problems for analyses. Data doesn't represent what it was supposed to.

Consequences of Measurement Error

1. Univariate: biased estimates. Means, variance, etc.. Don't represent true pop values

2. Bivariate: biased and inconsistent:

a. makes a correlation stronger or weaker than it is in pop

if random error in X: r is less than true (not affected by non-random in Y)

b. Doesn't get any better with larger sample sizes

3. Multivariate: biased, inconsistent and inefficient

a. Relationships could be too weak or too strong, or could appear to exist when they don't or vice versa (because of the cumulative effect of measurement error in all variables in analysis)

b. Significance tests are off (variances are off)

c. Standardized coefficients are off

What Influences How Well You Measure

1. Validity of theory. Is the concept accurately conceived? Does it adequately reflect reality?

2. Operational Validity. Does the measure measure what it is supposed to measure (i.e. the concept)? Types of validity: face, content, construct, convergent, criterion, discriminant

3. Reliability: is the measurement instrument (a question) stable? Does it produce the same set of observations through repeated applications or over different contexts.

Example: Social Class influences Health

Conceptualizations: Social Class = SES (household income, educational level, occupation)

Health=physical health (absence of disease, fitness, ability to perform daily functions, nutrition)

Operationalizations: What is your household income? What was the highest educational degree you obtained? Do you have any illnesses? Are you able to walk a flight of stairs? Are you able to dress yourself?

Empirical: Someone's true household income, their true educational level, their true health

Types of Validity: Classical Theory

Conceptual Validity

1. Face Validity: does it make common sense

Religiosity = # times pray a week

Assessment: does it make common sense (no quantitative method to assess)

2. Content Validity: does it reflect all dimensions of the concept

example: if you measured poverty with income alone you wouldn't get at the fact that a person is a student or whether she or he receives social support

*you would over-estimate poverty

need to specify all dimensions of a concept in order to measure it accurately

assessment: 1. Go to literature, identify all dimensions of a concept
2. Does measure reflect all dimensions (No quantitative way to assess)

*usually use content validity in developing a measure

Empirical Validity

3. Criterion Validity: does the measure (the proxy) accurately predict the behavior in question (the criterion)

use when you can measure True Score

use when you are trying to use a proxy

usually not for concepts, particularly abstract concepts (can use if you are doing attitude-behavior research and you have a way of measuring the behavior, such as with Hurricane Evacuation study)

Example: prejudice attitude as a proxy for prejudice behavior

test score as a proxy for knowledge vs. talking with a student to assess true knowledge
SAT as a proxy for academic performance (get this from later behavior)

Assessment: how strong is the correlation between the proxy and the criterion

*In social sciences we often don't have criterions

Theoretical Validity

4. Construct Validity: is the observed relationship what was theoretically expected? If so, than evidence for construct validity

Example: Does our observed poverty level influence health?

Assessment:

1. State theoretical hypotheses
2. Get data, examine relationship
3. Conclusion: does data support hypothesis?
4. If multiple indicators: relationships between all the measures and the concept should have similar strength and direction
5. Repeated assessment over multiple studies

*if results support hypotheses: interpretation= data suggest the measure has construct validity. Still need to do repeated assessments in other studies before you say measure has construct validity.

Ex: Contingent Valuation, influence of masculinity on criminal activity

*if results do not support hypotheses, the interpretations are:

a. Measure lacks construct validity: using that question to measure that concept needs adjustment.

b. Theory not right, not accurately specified: maybe more complicated relationship (need more independent variables, need to specify a process....)

c. Other variables in the analysis lack validity and reliability: if these aren't measured well it could dirty the measure in question

Using Multiple Indicators to Measure Complex Concepts (Scales)

Example: Attitudes about women

Indicators =
1. would you vote for a female presidential candidate
2. would you want your wife to work outside the home
3. how would you react if your wife made more money that you did
4. who should have primary responsibility for childcare

Example: Attitudes about Pornography
Example: Health

Reliability

Central to the notion of reliability is REPLICATION

1. across observations
2. repeated measurement of same observation
3. repeated measurement by different observers
4. repeated parallel measures of some underlying concept

Improve Reliability by:

1. asking people what they know, not what other people know
2. ask only questions that are relevant to them
3. simple questions, well worded, clear
4. train researchers/workers for consistency
5. pretests
6. use established measures

Assessing Reliability

many require parallel measures = two or more questions which are intended to measure the same concept

One Indicator

1. Test-retest

have same measure for same people at two points in time
do you get the same scores

example: support for Bill Clinton

Assessment: correlation between two variables

Problems:

a. Assumes True score doesn't change over time (attitudes about Clinton, and Lewinsky scandal or Rich pardon occurs between time 1 and time 2)

Result: underestimates reliability

b. Testing effects: inflates reliability

c. Errors are assumed to be uncorrelated. They probably are correlated. Overestimates reliability

d. History effects: changes in how well X measures T across time (reliability can be over or underestimated)

2. Alternative Forms

slightly different measures at Time 1 and Time 2

example: health, deviance

Assessment: correlation between two variables

Problems:

a. Assumes True score doesn't change over time (health didn't change, perceptions of deviance didn't change)

Result: underestimates reliability

b. Assumes parallel measures (= weights)

Advantages over test-retest:

1. Less testing effects

2. Errors less likely to be correlated, but still can be because if you answer 1 question inconsistently, good chance you will inconsistently answer a related question

Agreement Among Multiple Indicators At One Point In Time

3. Split Halves

many indicators of T have to be available
split indicators into two halves (comprise first and second half of survey)
if measures consistent, should have high correlation between halves

Assessment: compute correlation each half, then correlation between each of those

Advantages: No change in time, so less of a chance that T changes, less memory/testing effects, less correlated error (but still can happen)

Disadvantages:

1. Requires parallel measures: Are the measures really parallel?
2. How to split the halves
3. Assumes uncorrelated errors

VALIDITY AND RELIABILITY IN QUALITATIVE RESEARCH

Validity: accurate interpretation of the world
Reliability = consistent interpretation of the world

Effects on:

1. Completeness of observations/interview questions
2. Researcher bias in interpreting behaviors/words
3. Reactive effects of researcher's presence
4. inter-rater reliability