ISS 305:0012
Evaluating Evidence:
Becoming a Smart Research Consumer

5. Problems of measurement

Reminder: Turn on your I<CLICKER

Introduction

- We've stressed need for operational definitions to
  -
  -
- But, there might be many ways of operationalizing a concept or variable, like…
- Take the variable from clinical psychology of "depression".
  - Q: How might we operationalize it?
- How do we choose among them--what makes one any better than another?

5. Problems of measurement

I. Basic concepts:
   A. Variable
   B. Measurement
      1. Levels of measurement

II. Errors of measurement
   A. Random Error (Noise)
   B. Systematic Error (Bias)

III. Evaluating measures
   A. Lowering random error (Reliability)
   B. Lowering systematic error (Validity)
   C. Ways of establishing reliability and validity

Introduction:

- Basic concepts
  - **Variable** - Any **attribute** which can assume different values among the members of a class of participants/subjects or events, but which has only one value for any given member of that class at any time.
  - Variables are the way that things can differ, and for which we can observe differences
- Examples?
  - Physical variables:
  - Psychology variables:
    - Usually called **constructs or concepts** – A hypothetical factor that is not observed directly rather its existence is inferred from certain observations and assumed to follow from certain situations. Examples?

Conceptual/Construct variables to operational variables

- Conceptual/Construct variable
  - Theoretical constructs (e.g. shyness, sleep quality, intelligence)
- Operational variables
  - Turn a conceptual variable into a variable that can be measured or manipulated
  - Connect unobservable traits, experiences, or qualities into things that can be observed and measured

Conceptual variable	Operational variable	Measured or manipulated
Marijuana usage	Inhaling 0.3 grams of marijuana	Manipulated
Math ability	Number of correct items on math test	Measured

Desirable qualities: Observations ➔ Numbers

- Observations should be recorded so they
  -
  -
- Words are often inadequate
  - very smart = quite smart = above average?
  - Think about beer
- Numbers are usually better
  - Meaning is usually more precise (IQ of 120 vs. "very smart")
  - Meaning is usually shared (120 means the same thing to different people)
- Numerical summaries can be meaningfully compared and can be combined using simple math (e.g., compute mean values)
- Measurement is the assignment of numbers to particular observations so as to reflect the variations in those observations

Levels of measurement:

- Some numbers carry more information than others
- There are 4 *levels/scales of measurement* (see NOIR on the next slide)
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- So, all else being equal, better operational definitions are those which
  - **Results in numeric values (permit measurement)**
  - **Result in higher levels of measurement**

Scales of Measurement: NOIR

Categorical Variables

Nominal:

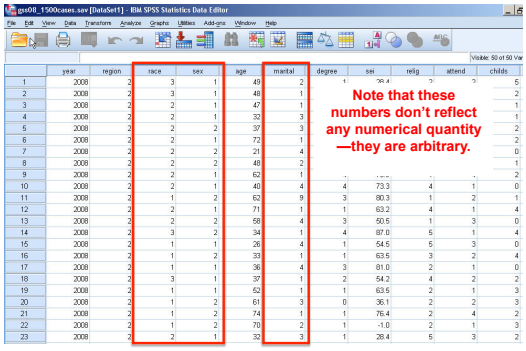
- Numbers are names only. Has two or more categories, but no intrinsic order to the categories. **Discrete variable** – no values in between
- Examples:
  - Marital status (married, unmarried, separated, etc.)
  - HIV-status (positive or negative)
  - Pet ownership (yes or no)
- Use Frequencies and Chi-Square

Numeric Variables

Ordinal:

- Numbers indicate order but distance between numbers not equal (**rank ordered**, but don't know how far apart score are)
- **Discrete variable**
- Example – rating of how much someone likes a painting (1 through 4)
- Use nonparametric statistics

Nominal variable



Scales of Measurement: NOIR

Numeric Variables

Interval:

- Distance between numbers equally spaced, but there is not true 0 or no meaning to 0 (**no absolute zero point**)
- **Continuous variable** – infinite number of values in between two values
- Example – temperature in Fahrenheit
- Can use ANOVA

Ratio:

- Like interval, but with a true 0 point (**has an absolute zero point** = indicates absence of quality)
- **Continuous variable**
- Examples – height and money
- Can use ANOVA

Desirable qualities: Measurements that have little error

- Useful to distinguish 2 kinds of error:
  - Random error (or noise) – Sources of error which bounce randomly around the underlying "true" value on the variable
    - the "flutter" or "noise" of a speaker
  - Systematic error (or bias) – Error which distorts measurements consistently **in a single direction** from the underlying "true" value

Desirable qualities: Low Random Error

- Random error – **Sources of error which bounce randomly around the underlying "true" value on the variable**
  - the "flutter" or "noise" of a speaker
- Across repeated measurements, they would tend to cancel out, so that the mean random error is zero
  - Specifically, they are statistical fluctuations in either direction around the mean
  - For example, there is random error for each choir member trying to hit a note, but the combined choir is closer to the right note

Examples of sources of random error when measuring...

- Length?
- Scholastic aptitude (e.g., with the SAT)?
- How to minimize random error?

Desirable qualities: Low Random Error

Sources of random error:

- Randomly varying aspects of the things/persons being studied
- Fuzzy or inconsistently applied measurement criteria
  - Child aggressiveness on the playground
  - Another reason for precise language in empirical statements
- Noise introduced by the observer(s)
  - By settling for a sample instead of all people (sampling error);
    - We'll discuss this more later (under *statistics* and *random assignment*).
  - By random factors affecting the observer (e.g. noise, distraction)
  - By using several observers rather than just one

Desirable qualities:

Low Random Error = High Reliability

- Reliability of a measure is an index of how well random error/noise has been controlled
  - How consistent/repeatable is the assessment? How precise is the measure? How much measurement error is involved?
  - Perfect reliability =
    -
  - A reliable measure measures (something) precisely
  - A measure with A LOT of random noise is what?
    -
  - What's the minimum level of reliability?
  - What would an IQ measure with "no reliability" look like?

Desirable qualities: Low systematic error

- Systematic error (or bias) – Error which distorts measurements consistently **in a single direction** from the underlying "true" value
- Across different measurements, a systematic error or bias does not cancel out
- bias in a speaker (so that the pitch is always too high or too low)
- the mean of a systematic error is not zero.
- Examples:
- Length?
- SAT?
- Child aggressiveness in the playground?

Desirable qualities: Low systematic error

There are many sources of systematic error or bias, including:

- Experimenter expectancy effects, selection biases, testing effects, demand characteristics
  - We have discussed some of these already and some we will cover later in the semester
- Systematic response biases
  - Moderation response biases
  - Acquiescence response bias – tendency to check the same response category in a series of similar questions.
    - "yea-sayers" and "nay-sayers"
    - May be due to social norms to be polite
    - Motivation of the responder
    - Skill of the responder

Desirable qualities: Low systematic error

There are many sources of systematic error or bias, including:

- Systematic response biases
  - Social desirability response bias – responding to look good, not to respond accurately
    - Solutions?
      -
      -
      -
      -
      -

» Note that as a skeptical consumer, you need to check if such solutions have been used when this bias is likely

Desirable qualities: Low systematic error

- Solutions?
  -
  -
- The "Bogus Pipeline"
  - –Reduce false answers by
    -

Desirable qualities: Low systematic error Self-reports of sensitive behaviors

	% Answering Yes	
	Control	Bogus Pipeline
Drink more than average?		
Ever drink and drive?		
Often have oral sex?		
Ever smoke pot?		
Do you smoke?		
Exercise 4 or more times a week?		

(Touraneau et al, 1997)

Desirable qualities: Low systematic error

- Solutions?
  -
  -
  -
  -
  -
  -
  -

Desirable qualities: Low systematic error = High Validity

- Validity of a measure is an index of how well systematic error has been controlled, or
- Validity of a measure is **how well you're measuring what you want to measure** and not something else (a source of bias)
  - **Perfect validity =**
  - A measure with **a lot** of systematic error is…
- Q: What's the minimum level?
- What would a measure with "no validity" look like?
- A:

Roughly 3 Kinds of Reliability in Measurement

1. Test-retest (Time)
   - There are random errors that perturb the measurement from one occasion to the next.
2. Inter-Rater (Observer)
   - Raters (observers) differ in how they classify a phenomenon.
3. Internal Consistency (Content)
   - Random error affects responses to items on an assessment.

Ways of establishing Reliability of a Measure or Scale

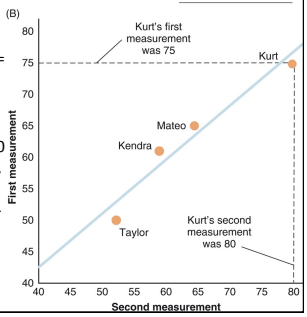
- For example, a scale to measure weight.
- How could I show that it has little random error?
  - Science assumes nature is consistent; under same conditions, nature behaves the same, and a "true" value of a variable (weight) is the same.
- **1. Temporal consistency of measurement (time)**
  - If you measure the same people at two points in time, the greater the reliability, the more similar the measurements should be.
  - Could look at some measure of the variability of repeated observations
    - What % of IQ tests vary by more than 15 points from one to next?
  - Usually assessed by computing a ***test-retest correlation***

Test-Retest Reliability

- How to do this?
- Measure behavior at time 1 and again at time 2. Do this for all the participants and calculate a correlation
- correlation coefficient = Pearson *r* = a statistic to show how closely two variables are related to one another
- perfect test-retest reliability: *r* = 1.0 (knowing first measurement result, you could predict the 2nd perfectly)
- complete test-retest unreliability: *r* = 0.0 (knowing the first measurement result gives you absolutely no useful information about what the 2nd is)

(A)

	Measurement 1	2
Taylor	50	52
Kendra	61	59
Mateo	65	64
Kurt	75	80

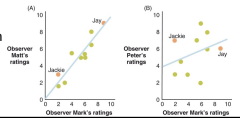


Reliability as temporal consistency of measurement

- Assumes that people (i.e., true value) does not change from first to second measure.
- When is this assumption likely to be a problem?
  - long intervals for most variables (lots of time for change)
    - Testing weight before and after the winter break
  - even very short intervals for unstable traits (e.g., moods; skills)
  - if there are carry-over or testing effects: simply being tested at Time 1 affects your true score at (immediate) Time 2
    - Testing effect
      - using the same exam two days apart to test the reliability of the test as a measure of knowledge of ISS305 material.
    - Fatigue effect
      - a marathon in the morning and the afternoon to test the reliability of marathons as measures of long distance running ability.

Ways of establishing Reliability of a Measure or Scale using Human Judges

- For example
  - judging sobriety by police officers
  - judging gymnastic performance
  - judging boxing
    - http://espn.go.com/boxing/story/_/id/9688631/cj-ross-steps-scoring-mayweather-alvarez-fight-draws-controversy
- How could we show that random error was low, that the observations were reliable?
- 2. **Interjudge (Inter-rater) agreement of measurement (Observer)**
  - If two human judges are measuring the same thing, the greater the reliability of their judgments, the more similar their two measurements should be.
  - Assessed with various statistics
    - interjudge reliability correlation
    - coefficient of concordance



Ways of establishing Reliability of a Multi-item/Summative Measure or Scale

- For example
  - SAT as a measure of scholastic aptitude
  - Final exam as measure of knowledge of course material
- **3. Internal consistency (Content)**
  - If a multi-item measure (a summative scale) is reliable, the measures obtained with different items ought to be similar
  - **Most common way of thinking about reliability**
    - BUT only one of several ways of doing so
  - Assessed using various statistics
    - alternate forms correlation (the bicycle questionnaire)
    - part-whole (split-half reliability) correlation
    - Cronbach's alpha (average all possible split-half correlations)

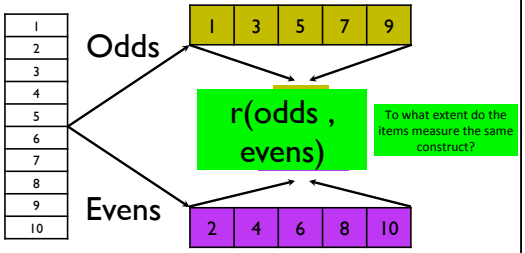
Internal Consistency/Reliability (Content)

- Assessed using various statistics
  - Alternate forms correlation
    - **"I like to ride Bicycles" and "I've enjoyed riding bicycles in the past" should have a strong agreement**
    - **"I hate bicycles" if disagreement with the above statements, good internal consistency of the test**
  - Split-Half
    - How would you measure the internal reliability of the extraversion scale?

1	I seek danger.
2	I dislike loud music.
3	I am willing to try anything once.
4	I enjoy being part of a loud crowd.
5	I would never go hang gliding or bungee jumping.
6	I seek adventure.
7	I enjoy being reckless.
8	I love action.
9	I love excitement.
10	I act wild and crazy.

average → Score

Split-Half Reliability



Cronbach's Alpha (α)

- Imagine computing the split-half reliability for
  - odds and evens → r_1,
  - 1st half vs 2nd half → r_2
  - etc.
- If we found every possible split-half correlation and calculated the average split-half correlation over all of these then that would be Cronbach's Alpha.

Assessing reliability:
What's a research consumer to do?

- Best:
  -
- Next best:
  - Look for indirect evidence of the measures' reliability.
    - Who uses the measure and where?
      - more likely to be reliable if used by scientists and reported in peer reviewed scientific journals
    - Is this a standard method of measurement?
      - widely and repeatedly used measures are more likely to have had their reliability assessed

Assessing reliability:
What's a research consumer to do?

- Least best:
  - Assess the "face reliability" of the measure. Is there any obvious source of random error?
    - highly subjective judgments;
    - judges poorly trained;
    - observations made carelessly

Assessing validity:
What's a research consumer to do?

- Remember, validity is the extent to which a measure of X truly measures X and NOT Y. Validity suggests "truthfulness", how well an idea fits reality. **How well you're measuring what you want to measure** and not something else (a source of bias).
- Not very good – **Subjective**:
  - **Face validity** – does it look like what it's supposed to measure?
  - **Content validity** – Are the actual items on a test consistent with the scientist's understanding of the construct the test measures?
  - **Validity by assertion or by authority**
  - **Consensual validity**
- Better – **Empirical**:
  - **Concurrent validity / Known-groups paradigm**
  - **Criterion validity** – Can the measure accurately forecast future behavior (Predictive validity)?
  - **Convergent and discriminant validity**- Is a measure (un)related to some other measure of behavior?

Assessing validity:
What's a research consumer to do?

- Not very good – **Subjective**:
  - **Face validity of the measure**
- Whether the measure seems to be valid to those who are taking it.
  - May want participants to take the inventory seriously.
- Face validity does not imply content validity.
  - Surveys in popular magazines have face validity but not necessarily content validity.


COSMO QUIZ

Are You Approachable?

Assessing validity:
What's a research consumer to do?

When is low face validity good?

Assessing validity: What's a research consumer to do?

- Not very good – **Subjective**:

  Content validity - rational, but still a non-empirical way of establishing validity.
  - the content of the measure is a fair, representative sample of all that should go into a measure of the variable?
    - Judge the content validity of these scales as measures of extraversion.

1	I seek danger.
2	I dislike loud music.
3	I am willing to try anything once.
4	I enjoy being part of a loud crowd.
5	I would never go hang gliding or bungee jumping

1	I often look up an unfamiliar word when I read it.
2	I often look through the book chapter that are not assigned in class.
3	I often read non-assigned books that pertain to subjects not learning in class.
4	I often research topics during my free time that are being discussed in class.
5	I often attend workshops to become a better student.

Assessing validity: What's a research consumer to do?

- Not very good – **Subjective**:
  - **Validity by assertion or by authority**
    - Someone who should know asserts that the measure is measuring what it is supposed to measure.
    - BUT

Assessing validity: What's a research consumer to do?

- Not very good – **Subjective**:

  Consensual validity - rational, but still a non-empirical way of establishing validity.
  - the agreement (consensus) of experts that the measure is valid
  - Problems?
    - Need to identify legitimate experts
    - Only useful when a consensus exists
    - "Best" content may not be the content which is easiest to obtain OR to reach agreement on
      - » teaching
      - » scholarly research

Assessing validity: What's a research consumer to do?

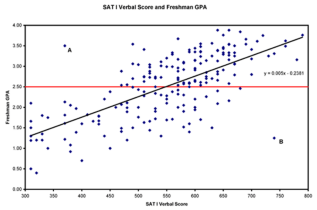
- Best: Empirical, criterion-related strategies. Does the measure do what it should do?

1. **Concurrent validity / Known-groups paradigm**
   - Researchers see whether scores on the measure can discriminate among a set of groups whose behavior is already well understood
     - Salivary cortisol as a measure of stress
     - Lie detectors
     - Subjective well-being scale
   - Problem--what if the groups "known" to differ actually do not?
   - This is a very general problem
     - You always make one or more assumptions when doing validity tests

Assessing validity: What's a research consumer to do?

- Best: Empirical, criterion-related strategies. Does the measure do what it should do?

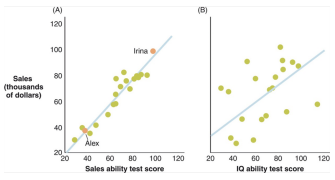
  2. **Criterion/Predictive validity** – A valid measure should predict future behavior that should be affected by the variable
- College GPA is a criterion that SAT/ACT scores should predict



Assessing validity: What's a research consumer to do?

- Best: Empirical, criterion-related strategies. Does the measure do what it should do?

  2. **Criterion/Predictive validity** – A valid measure should predict future behavior that should be affected by the variable
- Examples?
  - Racial prejudice
  - depression



- What's the hidden assumption in a predictive validity test?
  -

Assessing validity:
What's a research consumer to do?

Best: Empirical, criterion-related strategies. Does the measure do what it should do?

3. **Convergent validity** – A valid measure of a variable ought to produce similar scores (i.e., be correlated) with other (presumably) well validated measures
   - Such convergent validation depends on what assumption?
     •

Assessing validity:
What's a research consumer to do?

Best: Empirical, criterion-related strategies. Does the measure do what it should do?

4. **Discriminant validity** – A valid measure of a variable ought to produce dissimilar scores (i.e., be uncorrelated) with other (presumably) well validated measures of different/distinctive variables
   - Again, depends upon the assumption of independence of the criterion measure
   - A way of eliminating measures of too much
     - e.g., what if Exam 1 correlates strongly with a measure of verbal ability (e.g., Verbal SAT score)?
   - The most neglected form of validity.

Convergent and Discriminant validity

Relationship Between Reliability and Validity