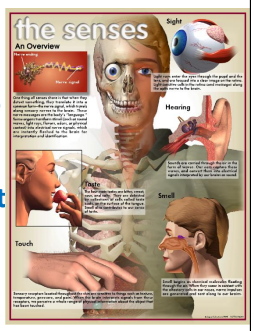
ISS 305:002
Evaluating Evidence:
Becoming a Smart Research Consumer

5. Problems of measurement

Reminder: Turn on your I<CLICKER

Introduction

- We've stressed need for operational definitions to
  - **satisfy the empirical requirement of testable and falsifiable sensory experience, and**
  - **to do so in a public (that means replicable) way**



Introduction

- 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?
  - **A1: Self report,**
  - **A2: Behavioral symptomotology,**
  - **A3: Physiologically,**
  - **A4: Peer reports, etc.**
- How do we choose among them--what makes one any better than another?

5. Problems of measurement

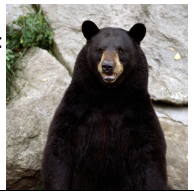
I. Basic concepts:
   A. Variable
   B. Measurement
      1. Levels of measurement

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:
    - **height**
    - **weight**
    - **eye color**
    - **mass**



Introduction:

- Psychological 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.
  - **depression**
  - **intelligence**
  - **need for achievement**
  - **aggression**

Desirable qualities: Observations ➔ Numbers

- Observations should be recorded so they
  - **have a precise meaning**
  - **have the same meaning for all**

Scales of Measurement: NOIR

Continuous 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

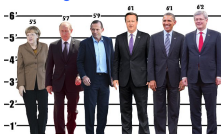
The difference in 20 degrees and 30 degrees is the same as between 30 and 40 degrees—but zero degrees does not mean that there is no temperature

Scales of Measurement: NOIR

Continuous Variables

- **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

The difference in 3 feet and 4 feet is the same as 5 feet and 6 feet and a height of zero is meaningful



Having $0 means something. X can have twice as much money as Y.

Examples of sources of random error when measuring...

- Length?
  - **lining up object carefully**
  - **angle of eye to instrument**
  - **rounding errors**
  - **mistakes in recording**
- Scholastic aptitude (e.g., with the SAT)?
  - **mood**
  - **guessing**
  - **errors filling out the answer sheet**
  - **mistiming by proctors (call STOP a bit too early or too late)**

Examples of sources of random error when measuring...

- How to minimize random error?
- **Take more data. Random errors can be evaluated through statistical analysis and can be reduced by averaging over a large number of observations.**

Desirable qualities:

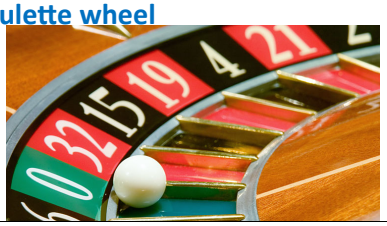
Low Random Error = High Reliability

- Reliability of a measure is an index of how well random error/noise has been controlled
  - Perfect reliability = **no noise/random error**
    - **rarely, if ever, achieved**
  - A reliable measure measures (something) precisely
  - **A measure with A LOT of random noise is unreliable**

Desirable qualities:

Low Random Error = High Reliability

- What's the minimum level of reliability?
- What would an IQ measure with "no reliability" look like?
  - **36-point IQ scale, measured using a roulette wheel**



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?
      - **increase anonymity (e.g., no names; randomized response techniques)**
      - **include filler items to mask/obscure the true purpose of the survey**
      - **add inducements for accurate reporting (e.g., bogus pipeline)**
      - **include catch items**
      - **cross check self reports (e.g., behaviorally)**
      - **avoid self reports altogether**
        - » **physiological measures (e.g., guilty knowledge test);**
        - » **unobtrusive measures (Milgram; Cialdini)**

» 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?
  - **Give participants anonymity**
  - **Include filler items to mask/ obscure the true purpose of the survey**
    - **Even Price = $5.00**
    - **Odd Price = $4.95**
    - **Filler Price = $4.75 or $5.25**

Desirable qualities: Low systematic error

- Solutions?
- The "Bogus Pipeline"
  - Reduce false answers by
    - **tricking participants into believing that the researchers can read their true feelings**

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

	% Answering Yes	
	Control	Bogus Pipeline
Drink more than average?	3.4%	21.0%
Ever drink and drive?	**17.2%**	**30.6%**
Often have oral sex?	32.1%	51.7%
Ever smoke pot?	**56.9%**	**71.0%**
Do you smoke?	20.7%	33.9%
Exercise 4 or more times a week?	**44.8%**	**22.6%**

(Touraneau et al, 1997)

Desirable qualities: Low systematic error

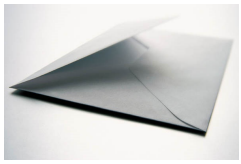
- **Solutions?**
- **Include catch items**
- "I have never lied."

Desirable qualities: Low systematic error

- **Solutions?**
- **Cross check self reports (e.g., behaviorally)**

Desirable qualities: Low systematic error

- **Solutions?**
- **Avoid self reports altogether**
- **Covert measures – the measurement of attitudes using unobtrusive techniques**
  - **behavioral observations**



  - **physiological measures**
    - **Facial Electromyograph (EMG)**

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 = **no systematic error**
- A measure with a lot systematic error is **invalid**

Desirable qualities:
Low systematic error = High Validity

- Q: What's the minimum level?
- What would a measure with "no validity" look like?
- A: **One which was all bias, totally unrelated to measured variable**
  - **IQ with a bathroom scale.**
  - **Ability to perform as a soldier based on sexual orientation.**

- Soon we'll note ways of determining and expressing the validity of measures.

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

- Best:
  - **Look for direct empirical evidence of the reliability of the measures as used in the project in question**
    - **test-retest; internal consistency correlation**
- 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

When is low face validity good?

Sometimes people don't want to acknowledge certain things about themselves

Number of stomach aches per week as a measure of anxiety in children

Pretty low face validity…

…but kids with anxiety do experience a lot of stomach aches

This could be a good measure of anxiety to use in kids that don't want to tell the experimenter (or maybe don't know) that they are anxious

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
      - **Fallible**
      - **Risky**
        - **1: use symptoms of mental illness as way to measure demonic possession**
        - **2: use number of interviews as measure of quality of an investigation**
        - **3: use tourist numbers as a measure of city safety**
        - **4: use parent satisfaction as measure of effectiveness of Head Start**
      - **a last resort, although better than nothing.**

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?
  - **That the variable actually does predict the behavior in question.**

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?
    - **the validity of the "well validated" measure**
  - Examples?
    - IQ?
    - Depression?

Convergent and Discriminant validity

Relationship Between Validity and Reliability

We need/want our measures to be both valid and reliable.

- **Reliability tells us if we are pretty good at measuring things.**
- **Validity tells us our measure measures what we want it to measure. No single test, but looking at the problem from different perspectives, with different methods.**
- **Reliability is a necessary, but not sufficient condition for validity.**
- **A reliable measure is not necessarily a valid measure.**
  - **Shoe size and IQ (High reliability and low validity)**
- **Low reliability and low validity = useless**
- **Low reliability and high validity = impossible**