Introduction to Psychometrics

- Psychometrics & Measurement Validity
- Some important language
- Properties of a "good measure"
  - Standardization
  - Reliability
  - Validity
- Common Item types
- Reverse Keying

Psychometrics

(Psychological measurement)

The process of assigning values to represent the amounts and kinds of specified attributes, to describe (usually) persons.

- We do not "measure people"
- We measure specific attributes or characteristics of a person

Psychometrics is the "centerpiece" of empirical psychological research and practice.

- All data result from some form of "measurement"
- For those data to be useful we need "Measurement Validity"
- The better the measurement, the better the data, the better the conclusions of the psychological research or application

Most of what we try to measure in Psychology are **constructs**

They're called because most of what we care about as psychologists are **not** physical measurements, such as height, weight, pressure & velocity…

…rather the "stuff of psychology" → learning, motivation, anxiety, social skills, depression, wellness, etc. are things that "don't really exist".

They are attributes and characteristics that we've constructed to give organization and structure to behavior. Essentially all of the things we psychologists research, both as causes and effects, are Attributive Hypotheses with different levels of support and acceptance!!!!

Measurement of constructs is more difficult than of physical properties!

We can't just walk up to someone with a scale, ruler, graduated cylinder or velocimeter and measure how depressed they are.

We have to figure out some way to turn their behavior, self-reports or traces of their behavior into variables that give values for the constructs we want to measure.

So, measurement is, much like the rest of research that we've learned about so far, all about representation !!!

Measurement Validity is the extent to which the data (variable values) we have represent the behaviors (constructs) we want to study.

What are the different types of constructs we measure ???

The most commonly discussed types are ...

- Achievement -- "performance" broadly defined (judgements)
  - e.g., scholastic skills, job-related skills, research DVs, etc.
- Attitude/Opinion -- "how things should be" (sentiments)
  - polls, product evaluations, etc.
- Personality -- "characterological attributes" (keyed sentiments)
  - anxiety, psychoses, assertiveness, etc.

There are other types of measures that are often used…

- Social Skills -- achievement or personality ??
- Aptitude -- "how well some will perform after then are trained and experiences" but measures before the training & experience"
  - some combo of achievement, personality and "likes"
- IQ -- is it achievement (things learned) or is it "aptitude for academics, career and life" ??

Each question/behavior is called an → item

Kinds of items ➔ objective items vs. subject items

- "objective" does not mean "true" "real" or "accurate"
- "subject" does not mean "made up" or "inaccurate"

Items are names for "how the observer/interviewer/coder transforms participant's responses into data"

Objective Items - no evaluation, judgement or decision is needed

- either "response = data" or a "mathematical transformation"
- e.g., multiple choice, T&F, matching, fill-in-the-blanks

Subjective Items – response must be evaluated and a decision or judgment made what should be the data value

- content coding, diagnostic systems, behavioral taxonomies
- e.g., essays, interview answers, drawings, facial expressions

Some more language …

A collection of items is called many things…

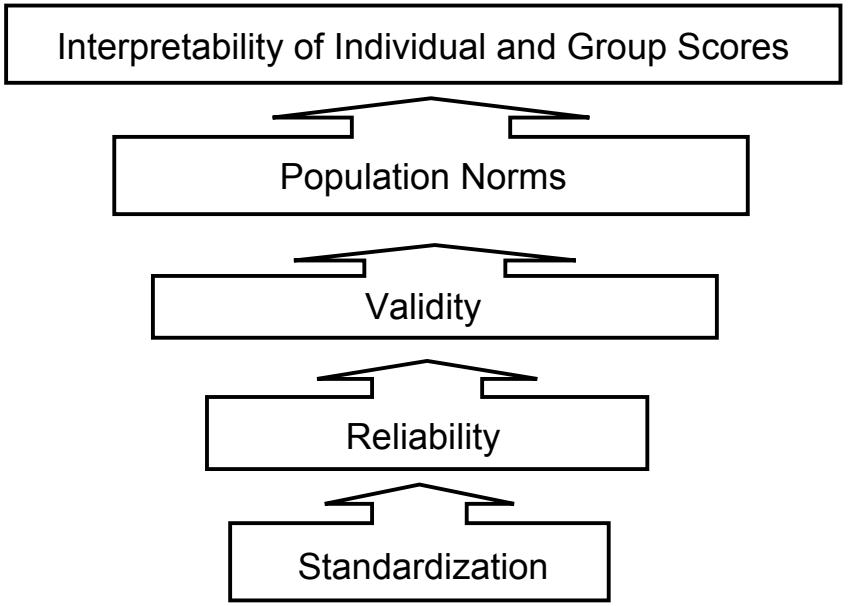
• e.g., survey, questionnaire, instrument, measure, test, or scale

Three "kinds" of item collections you should know ..

- Scale (Test) - all items are "put together" to get a single score
- Subscale (Subtest) – item sets "put together" to get multiple separate scores
- Surveys – each item gives a specific piece of information

Most "questionnaires," "surveys" or "interviews" are a combination of all three.

Desirable Properties of Psychological Measures



Standardization

Administration – test is "given" the same way every time

- who administers the instrument
- specific instructions, order of items, timing, etc.
- Varies greatly - multiple-choice classroom test → hand it out)
  - WAIS -- 100+ page administration manual

Scoring – test is "scored" the same way every time

- who scores the instrument
- correct, "partial" and incorrect answers, points awarded, etc.
- Varies greatly -- multiple choice test (fill in the sheet)
  -- WAIS – 200+ page scoring manual

Reliability (Agreement or Consistency)

Inter-rater or Inter-observers reliability

- do multiple observers/coders score an item the same way ?
- important whenever using subjective items

Internal reliability -- do the items measure a central "thing"

- Cronbach's alpha → α = .00 – 1.00 ← higher values mean stronger internal consistency/reliability

External Reliability -- consistency of scale/test scores

- test-retest reliability – correlate scores from same test given 3-18 weeks apart
- alternate forms reliability – correlate scores from two "versions" of the test

Validity (Consistent Accuracy)

Face Validity -- do the items come from "domain of interest" ?
non-statistical -- decision of "target population"

Content Validity -- do the items come from "domain of interest"?
non-statistical -- decision of "expert in the field"

Criterion-related Validity -- does test correlate with "criterion"?

- statistical -- requires a criterion that you "believe in"
- predictive, concurrent, postdictive validity

Construct Validity -- does test relate to other measures it should?

- Statistical -- Discriminant validity
  - convergent validity -- correlates with selected tests
  - divergent validity -- doesn't correlate with others

"Is the test valid?"

Jum Nunnally (one of the founders of modern psychometrics) claimed this was "silly question"! The point wasn't that tests shouldn't be "valid" but that a test's validity must be assessed relative to…

- the construct it is intended to measure
- the population for which it is intended (e.g., age, level)
- the application for which it is intended (e.g., for classifying folks into categories vs. assigning them quantitative values)

So, the real question is, "Is this test a valid measure of this construct for this population in this application?" That question can be answered!

Face Validity

Does the test "look like" a measure of the construct of interest?

- "looks like" a measure of the desired construct to a member of the target population
- will someone recognize the type of information they are responding to?

- Possible advantage of face validity ..
  - If the respondent knows what information we are looking for, they can use that "context" to help interpret the questions and provide more useful, accurate answers
- Possible limitation of face validity …
  - if the respondent knows what information we are looking for, they might try to "bend & shape" their answers to what they think we want -- "fake good" or "fake bad"

Content Validity

Does the test contain items from the desired "content domain"?

- Based on assessment by "subject matter experts" (SMEs) in that content domain

- Is especially important when a test is designed to have low face validity
  - e.g., tests of "honesty" used for hiring decisions
- Is generally simpler for "achievement tests" than for "psychological constructs" (or other "less concrete" ideas)
  - e.g., it is a lot easier for "math experts" to agree whether or not an item should be on an algebra test than it is for "psychological experts" to agree whether or not an items should be on a measure of depression.
- Content validity is not "tested for". Rather it is "assured" by the informed item selections made by experts in the domain.

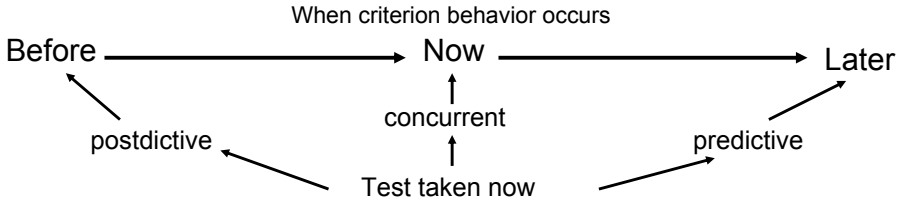
Criterion-related Validity

Do the test scores correlate with criterion behavior scores??

concurrent -- test taken now "replaces" criterion measured now
- often the goal is to substitute a "shorter" or "cheaper" test
- e.g., the written drivers test replaces road test

predictive -- test taken now predicts criterion measured later
- want to estimate what will happen before it does
- e.g., your GRE score (taken now) predicts grad school (later)

postdictive – test taken now captures behavior & affect of before
- most of the behavior we study "has already happened"
- e.g., adult memories of childhood feelings or medical history



Construct Validity

- Does the test interrelate with other tests as a measure of this construct should ?
- We use the term construct to remind ourselves that many of the terms we use do not have an objective, concrete reality.
  - Rather they are "made up" or "constructed" by us in our attempts to organize and make sense of behavior and other psychological processes
- attention to construct validity reminds us that our defense of the constructs we create is really based on the "whole package" of how the measures of different constructs relate to each other
- So, construct validity "begins" with content validity (are these the right types of items) and then adds the question, "does this test relate as it should to other tests of similar and different constructs?

The statistical assessment of Construct Validity …

Discriminant Validity

- Does the test show the "right" pattern of interrelationships with other variables? -- has two parts
  - Convergent Validity -- test correlates with other measures of similar constructs
  - Divergent Validity -- test isn't correlated with measures of "other, different constructs"
  - e.g., a new measure of depression should …
    - have "strong" correlations with other measures of "depression"
    - have negative correlations with measures of "happiness"
    - have "substantial" correlation with measures of "anxiety"
    - have "minimal" correlations with tests of "physical health", "faking bad", "self-evaluation", etc.

Population Norms

In order to interpret a score from an individual or group, you must know what scores are typical for that population

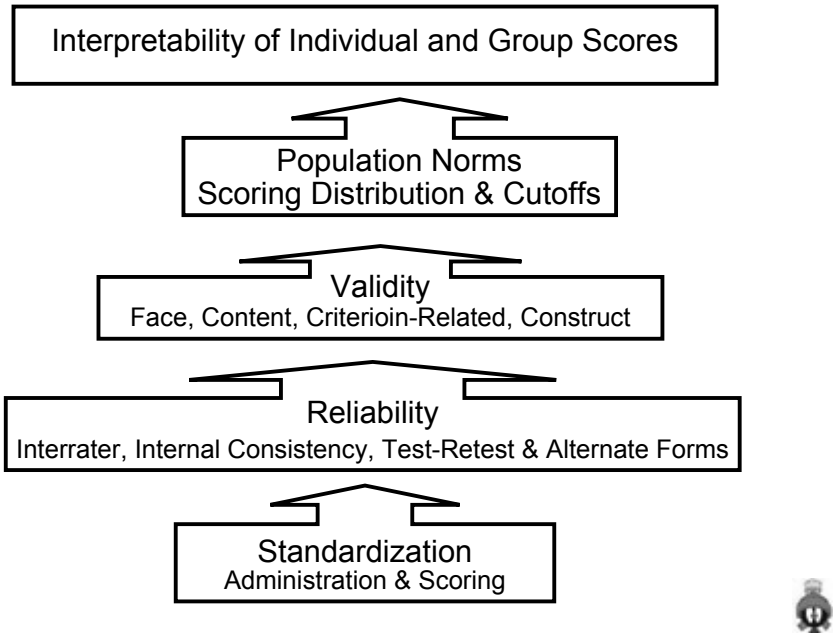
- Requires a large representative sample of the target population
  - preferably → random, research-selected & stratified
- Requires solid standardization → both administrative & scoring
- Requires great inter-rater reliability if subjective

The Result ??

A scoring distribution of the population.

- lets us identify "normal," "high" and "low" scores
- lets us identify "cutoff scores" to define important populations and subpopulations

Desirable Properties of Psychological Measures



Reverse Keying

We want the respondents to carefully read an separately respond to each item of our scale/test. One thing we do is to write the items so that some of them are "backwards" or "reversed" …

Consider these items from a depression measure…

1. It is tough to get out of bed some mornings.	disagree	1	2	3	4	5 agree
2. I'm generally happy about my life.		1	2	3	4	5
3. I sometimes just want to sit and cry.		1	2	3	4	5
4. Most of the time I have a smile on my face.		1	2	3	4	5

If the person is "depressed", we would expect then to give a fairly high rating for questions 1 & 3, but a low rating on 2 & 4.

Before aggregating these items into a composite scale or test score, we would "reverse key" items 2 & 4 (1=5, 2=4, 4=2, 5=1)