Introduction to Psychological Measurement

- Psychometrics & some vocabulary
- Kinds of scale construction
- Desirable properties of a psychometric instrument
- Psychometric sampling
- Kinds of items
- Scaling models

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 of a person

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

- All data result from some form of "measurement"
- The better the measurement, the better the data, the better the conclusions of the psychological research or application

A bit of vocabulary to get us started…

Kinds of Items:

- survey item – an individual item that will measure the target construct
- scale item – one of a set of items that, when combined, will measure the target construct

e.g., age vs. emotional maturity

Scales

We'll use the terms "scale" to mean "a multi-item instrument designed to represent the amount or kind of a specific attribute for a specific individual"

A bit of vocabulary to get us started…

Population:

The group of individuals to which we want to apply our scale. We sample participants to represent the population.

Domain:

The type of information we want to measure with our scale. We sample items to represent the domain.

Construct value:

The individual's amount or kind of the attribute or characteristic we are trying to measure.

Variable:

The individual's measured score or code

Kinds of Scale Construction

We'll use the terms "scale" to mean "a multi-item instrument designed to represent the amount or kind of a specific attribute for a specific individual".

So… all kinds of things are "scales" !!!

The construction and validation of most scales follows a pretty similar overall logic and process. However, there are differences depending upon:

- what content/construct we're trying to measure
  - intellectual assessment, mental health evaluation, achievement testing, employment selection, research IVs/DVs, etc.
- to what legal and organizational auspices we're beholden
  - EEOC, APsychologicalA, APsyciarticA, EMAA, SIOP, etc.

Kinds of Scale Construction, cont.

Research scales – the emphasis in this class

- sometimes you can't find or can't afford a measure of some construct or behavior you'd like to research (new scale)
- sometimes you'll want to capture a "different version" of a construct than do existing instruments (new scale)
- sometimes you want to measure more things that you have time to measure using their current version (short form)
- sometimes you want an alternative version of a scale for pre-post or other repeated measures (alternate form)
- sometimes you'll want to "take apart" some construct or proxy variable, e.g., motivation (multiple subscales)

Kinds of Scale Construction, cont.

Classroom Tests – important for many of you and a useful counterpoint to research instruments

- writing tests to maximize assessment accuracy, educational benefit, both??

Selection Research – a "bonus" in this class

- Whether for education, employment or recreation, there are usually more "applicants" than "opportunities"
- "Fair" means of making these selections are increasingly required by organizations making the selections, by professional organizations and by the law

Another "kinds of Scales" – type of resulting information

Measurement or Equidiscriminating scales

- scales designed to give a score that represents a person's position along the contruct continuum
- e.g. % correct score to represent "how much they know"

Classificatory or Ordered Category scales

- scales designed to divide people into those that have "enough" vs. "not enough" of a construct
- e.g., pass – fail score

Many scales are a designed to provide a "blend" of these….

- to divide the construct continuum into multiple ordered categories
- e.g. "A" "B" "C" "D" "F" grade scale

Sometimes it is very hard to label the "type" of scale & measurement involved – take classroom grades…

- start with # or % correct on tests, homework, etc – looks equidiscriminating
- get a total % for the course – still looks equidiscriminating
- convert to letter grade (A, B, C, D, F – ordered categories) or pass/no pass (classification)
- letter grade is converted to number (A = 4.0, B = 3.0, -- back to equidiscriminating, but with fewer different values)
- averaged across classes to get GPA – looks equidiscriminating
- divide people up based on GPA – summa cum laude, magna cum lade, cum laude… (back to ordered categories)

The "type" of scale or measurement you are trying to make is important, because it will influence the items you want for it …

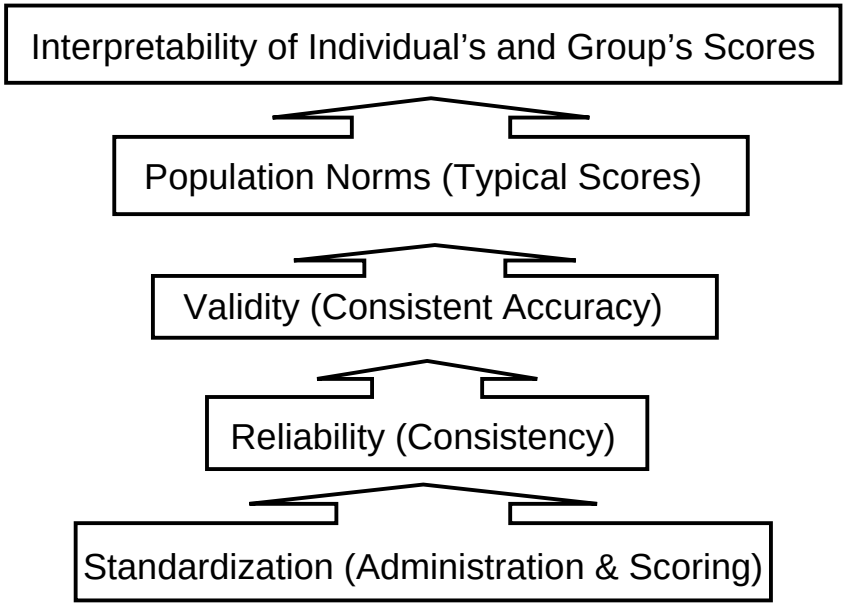
- a test to "identify remedial math students" will emphasize items that most students at that grade level can answer
- a test to "identify gifted math students" will emphasize items that very few students at that grade level can answer
- a test to "measure mathematical ability" will include items with a broad range of difficulty – so that we can place all students along the underlying continuum

these "depression scales" will all have different "levels" and "ranges" of items…

- identify clinically depressed individuals
- research outcome variable for treatment of clinically depressed individuals
- identify college students with little or no depression
- research measure to examine relationship between depression and school performance

You also have to pay attention to this when selecting scales for research – is the scale designed to give you the kind of measure you want (equidiscriminating vs. classificatory) for your target population????

Desirable Properties of Psychological Measures



Standardization

- Administration -- "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)
    -- Intelligence test -- 100+ pages of "how to" in manual -- about 1/2 semester in Psych 955
- Scoring -- "graded" 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)
    -- Exner System for the Rorschach -- 2 weeks of in depth training

Reliability (Consistency)

- Inter-rater or Inter-scorer reliability
  - can multiple raters produce the same score for a given test (assumes standardization)
- Internal reliability -- agreement among test items
  - split-half reliability -- randomly split into two tests & correlate
  - Chronbach's α -- tests "extent to which items measure a central theme"
- External Reliability -- consistency of scores from whole test
  - test-retest reliability -- give same test 3-12 weeks apart (r)
  - alternate forms reliability -- two "versions" of the test (r)

Validity (Consistent Accuracy)

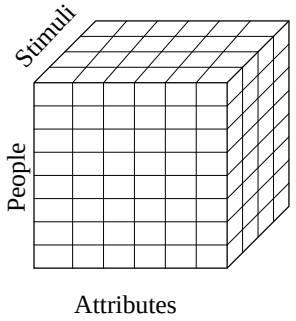
- Criterion-related Validity -- does test correlate with "criterion"?
  - statistical -- requires a criterion that you "believe in"
  - predictive, concurrent, postdictive validity
- Content Validity -- do the items come from "domain of interest"?
  - evaluated by "Subject Matter Experts"
- Face Validity -- do the items come from "domain of interest"
  - evaluated by "target population"
- 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

Psychometric Sampling

- from a statistical or research design perspective "sampling" usually refers to the selection of some set of people from which data will be collected, for the purposes of representing what the results would be if data were collected from the entire population of people in which the researcher is interested
- from a psychometric perspective "sampling" is a broader issue, with three dimensions
  - sampling participants (respondents) to represent the desired population of "individuals"
  - sampling attributes to represent some desired domain of "characteristics" or "behaviors"
  - sampling stimuli (may be other people or even themselves) to represent the desired domain of "objects"
- respondents give us values that represent attributes of stimuli

3-way sampling

Let's look at how "people", "attributes" and "stimuli" are used...



Examples

- 20 patients each rate the complexity, meaningfulness and pleasantness of the 10 Rorschach cards
- 3 co-managers judge the efficiency, effectiveness, efficacy and elegance of the 15 workers they share
- 10 psychologists rate each of 30 clients on their amenability to treatment, dangerousness and treatment progress

➔ 200 respondents complete a 50 item self-report personality measure

Reliability -- We have to consider two different kinds …

… of Statistical Decisions -- repeatability of our H0: testing result

• in general ... reliability increases with sample size (# respondents)

• this is the basis of "subject/variable ratios" & "power analysis"

• Basic tenet -- "When we want to decide whether or not there is a relationship between particular attributes/behaviors in a population, that population is best represented by a substantially sized sample of population members"

… of attribute/behavior measurements -- repeatability of scores

• in general … reliability increases with test size (# items)

• this is the basis of "internal reliability analysis" (more later)

• Basic tenet -- "When we want to measure the value of a particular attributes/behavior for a respondent, that attribute/behavior is best represented by a substantially sized sample of domain items

These two types of reliability interact in all our work…

Theoretical --

• we then concern ourselves with the "proper" sampling of individuals from the target population, so that our statistical results are reliable (either H0: decision or size of parameter estimate CI)

• empirical hypothesis/theory testing assumes that the attributes & behaviors are being measured reliably

• where "proper" means a sufficiently large sample of participants to adequately represents all the of the population (e.g., stratified sampling)

Psychometric --

• evaluation of the reliability of a measurement assumes that the population is being represented reliably

• when then concern ourselves with the "proper" sampling of items from the target domain

• where "proper" means a sufficiently large sample of items that adequately represents all the "niches" of the domain

Interrelationship --

• Evaluating one type reliability requires assuming the other type of reliability

Kinds of Items -- several distinctions

Survey Items

- individual items expected to "capture" the attribute of interest
- e.g., age, height, political registry

Scale Items

- items that are expected to "capture" the attribute of interest only when aggregated together to form a scale
- e.g., emotional maturity, body image, liberalism-conservatism
- each item has a combination of "specificity" and "error"
  - specificity -- only taps a portion of the target domain
  - error -- two kinds..
    - random -- unreliability/inconsistency
    - systematic -- intrusion of other domains

Psychometrics emphasizes the measurement of specific, relatively complex attributes, and so, emphasizes the use of multi-item scales.

another "kinds of items"

Any item can be defined by specifying three central attributes…

Judgment item vs. Sentiment item

- judgments -- have "correct answers"
- sentiments -- have no "correct answers"

Comparative item vs. Absolute item

- responding to two or more "stimuli"
- responding to a single "stimulus" (relative to "internal scale")

Preference item vs. Similarity item

- ranking or ordering items
- giving a value to define similarity
- can be comparative or absolute (vs. internal stim)

Most Common Types of Items ???

Personality, Attitude, Opinion ("Psychology") Items

1. How do you feel today ?

   Unhappy 1 2 3 4 5 happy

2. How interested are you in campus politics ?

   Interested 1 2 3 4 5 6 7 Uninterested

Absolute / Similarity / Sentiment -- most common

Pick a number from the scale (stimulus) that best depicts how you feel (no "correct answer)

Most Common Types of Items ???

Personality, Attitude, Opinion ("Psychology") Items

1. Which of these best describes you ?

   a. I am mostly interested in the "social side" of college.

   b. I am mostly interested in the "intellectual side" of college.

2. Would you rather spend time with a friend ...

   • at your favorite restaurant

   • watching a sporting event

Comparative / Preference / Sentiment -- less common

Pick from the two responses (stimuli) that which best depicts how you feel (no "correct answer)

Most Common Types of Items ???

"Test" Items

1. Which of these is one of the 7 dwarves ?

   a. Grungy b. Sleazy c. Kinky d. Doc e. Dorky

2. What should you do if the traffic light turns yellow as you approach an intersection ?

   a. Stop b. Speed up

   c. Check for Police and then choose "a" vs. "b"

Comparative / Preference / Judgment

Pick from the responses (stimuli) that which best depicts the correct answer

A bit more about judgment vs. sentiment items…

Most of the items used on psychological measures are "somewhere between" judgments and sentiments -- called "keyed sentiments".

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 sometimes just want to sit and cry.		1	2	3	4	5
3. I'm generally happy about my life.		1	2	3	4	5

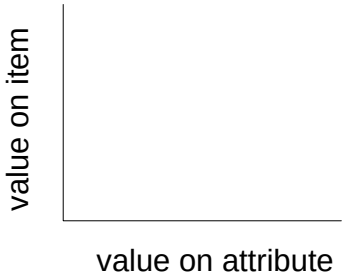
• If the person is "depressed", we would expect then to give a fairly high rating for questions 1 & 2, but a low rating on 3.

• Before aggregating these items into a score, we would "reverse key" item #3 (1=5, 2=4, 4=2, 5=1)

• Summary: for keyed items we don't score answers as right or wrong, but "key" them, so they can be aggregated sensib

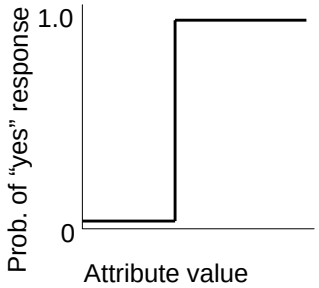
one more -- how items represent respondents "value" of the attribute

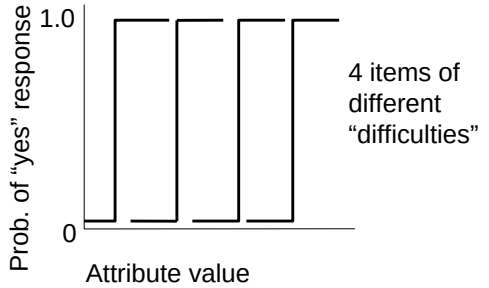
- this is sometimes called the "Scaling Model" being used
- relates to the "item trace" or "item characteristic curve"
- the item trace is the shape of the function of the relationship between individual items and the specific behavior/attribute being measured



Important terms
- linear
- monotonic vs. nonmonotonic
- deterministic vs. probabilistic
- specified vs. unspecified distribution form

Deterministic Model
- each item perfectly discriminates those who are above vs. below that "extent" of the target attribute
- so, a properly chosen item set will allow specific identification of "where" each respondent is along the attribute continuum
- a respondent answers "no" to all items "less difficult" than a specific value, and "yes" to all items "more difficult" than that value





This is also known as a "Guttman Scaling Model" (50-60's)

Best applied to sets of quantitative, one-dimensional constructs …

Are you taller than 6'6"

Are you taller than 6'3"

Are you taller than 6'0"

Are you taller than 5'9"

etc...

Work more poorly multidimensional or qualitative constructs …

The United Nation is the savior of all people

The UN is our best hope for peace

The UN is a constructive force in the world

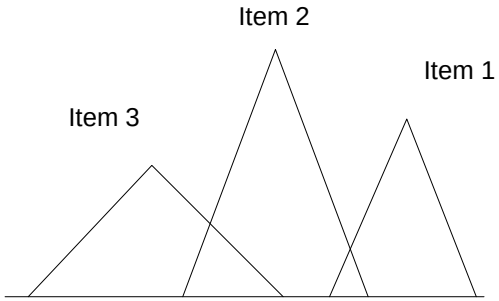
The US should participate in the UN

etc...

Problems with Deterministic Scaling Models
- assumes items have r = 1.00 with attributes (r > .40 is "huge")
- unless there is an underlying metric (like height) this type of item does little more than "order" people -- don't know "spacing"

Nonmonotone Model

- each item is maximally responded to by respondents at a certain position on the attribute continuum
- both those above and below that position respond with lower probability



Examples of Item 2

- I think all graduate students need 15 hours of stats courses
- Religion should play some role in peoples everyday decisions.
- Politically speaking, I am a moderate

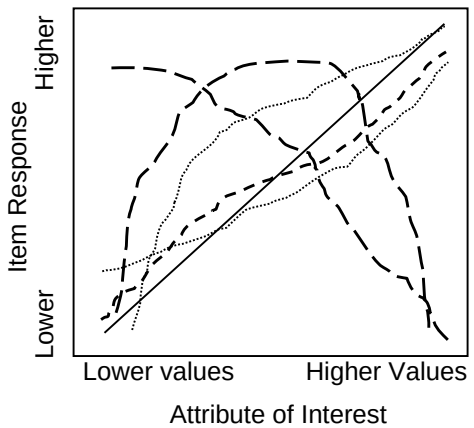
Very hard to write these items well for most constructs. Very repetitious to take surveys written this way.

Monotone Model with Specified Distribution Form

- Rasch models, Item Trace Theory, Item Trace Characterists
- each item has a trace with a specific mathematical form (usually logistic or normal ogive)
- increasingly popular for standardized tests (especially achievement tests)
  - allows application of strongest math/stat models
  - allows direct assessment of difficulty, variability, discrimination of the items
- hasn't "caught on" as well in behavioral measurement
  - requires large numbers of participants
  - often considered "overall kill" (more later)

Monotone with Unspecified Distribution Form

- most commonly used model in "scale construction"
- a "good item" is monotonically related to the attribute of interest
- sum of "good items" is nearly linearly related to attribute



Scatter plot of each persons score in the item and construct

great item --------

common items ....

bad items – – – – –

sum of good items _____