Psychometrics, Measurement Validity & Data Collection

• Psychometrics & Measurement Validity
  – Measurement & Constructs
  – Kinds of items and their combinations
  – Properties of a “good measure

• Data Collection
  – Observational -- Self-report -- Trace Data Collection
  – Primary vs. Archival Data
  – Data Collection Settings

Psychometrics (Psychological measurement)
The process of assigning values to represent the amounts and kinds of specified behaviors or attributes, to describe participants.

• We do not “measure participants”
• We measure specific behaviors or attributes of a participant

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 validity, the better the data, the better the conclusions of the psychological research or application

Most of the behaviors and characteristics we want to study in are constructs

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

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

We have constructed them to give organization and structure to our study of behaviors and characteristics. Essentially all of the things we psychologists research and apply, both as causes and effects, are Attributive Hypotheses with different levels of support and acceptance!!!!
What are the different types of constructs we use???

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 someone will perform after then are trained . . . and experiences” but measures before the training & experience”
  - some combo of achievement, personality and preferences
- 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”
- “subjective” 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, judgment 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.
Reverse Keying

We want the respondents to carefully read 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 them 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)

Desirable Properties of Psychological Measures

Interpretability of Individual and Group Scores

Population Norms

Validity

Reliability

Standardization

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?
- especially important whenever using subjective items

Internal reliability
- do the items measure a central “thing”
- will the items add up to a meaningful score?

Test-retest reliability
- if the test is repeated, does the test give the same score each time (when the characteristic/behavior hasn’t changed) ?

Validity → Non-statistical types of Validity

Face Validity
- does the respondent know what is being measured?
- can be good or bad – depends on construct and population
- target population members are asked “what is being tested?”

Content Validity
- Do the items cover the desired “content domain”?  
- especially important when a test is designed to have low face validity e.g., tests of “honesty” used for hiring decisions
- simpler for more concrete constructs” ideas)
  - e.g., easier for math experts to agree about an algebra item than for psychological experts” to agree about a depression item

Content validity is not “tested for” rather it is “assured”
- having the domain and population expertise to write “good items”
- having other content and population experts evaluate the items

Validity → Statistical types of Validity

Criterion-related Validity
- do test scores correlate with the “criterion” behavior or attribute they are trying to estimate e.g., ACT → GPA
- types are labeled for when “the test” and “the criteria” msrd

Before → Now → Later
- postdictive
- concurrent
- predictive

concurrent -- test taken now “replaces” criterion measured now
- eg, written drivers test instead of road test

predictive -- test taken now predicts criterion measured later
- eg, GRE taken in college predicts grades in grad school

postdictive -- test taken now predicts criterion measured earlier
- eg, almost every self-report measure (ever)
Validity → **Statistical** types of Validity

Construct Validity
- refers to whether a test/scale measures the theorized construct that it purports to measure
- attention to construct validity reminds us that our many of the characteristics and behaviors we study are “constructs”
- Construct validity is assessed as the extent to which a measure correlates “as it should” with other measures of similar and different constructs

Statistically, construct validity 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”

Evaluate this measure of depression….

<table>
<thead>
<tr>
<th></th>
<th>New Dep</th>
<th>Dep1</th>
<th>Dep2</th>
<th>Anx</th>
<th>Happy</th>
<th>PhyHlth</th>
<th>FakBad</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Dep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old Dep1</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old Dep2</td>
<td>.59</td>
<td>.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anx</td>
<td>.25</td>
<td>.30</td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>-.59</td>
<td>-.61</td>
<td>-.56</td>
<td>-.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyHlth</td>
<td>.16</td>
<td>.18</td>
<td>.22</td>
<td>.45</td>
<td>-.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FakBad</td>
<td>.15</td>
<td>.14</td>
<td>.21</td>
<td>.10</td>
<td>-.21</td>
<td>.31</td>
<td></td>
</tr>
</tbody>
</table>

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, researcher-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 individual scores as “normal,” “high” & “low”
- lets us identify “cutoff scores” to put individual scores into importantly different populations and subpopulations
Desirable Properties of Psychological Measures

- Interpretability of Individual and Group Scores
- Population Norms
- Scoring Distribution & Cutoffs
- Validity
- Face, content, Criterion-Related & Construct
- Reliability
- Inter-rater, Internal Consistency & Test-Retest
- Standardization
- Administration & Scoring

All data are collected using one of three major methods...

Behavioral Observation Data
- Studies actual behavior of participants
- Can require elaborate data collection & coding techniques
- Quality of data can depend upon secrecy (naturalistic, disguised participant) or rapport (habituation or desensitization)

Self-Report Data
- Allows us to learn about non-public “behavior” – thoughts, feelings, intentions, personality, etc.
- Added structure/completeness of prepared set of ?s
- Participation & data quality/honesty dependent upon rapport

Trace Data
- Limited to studying behaviors that do leave a “trace”
- Least susceptible to participant dishonesty
- Can require elaborate data collection & coding techniques

Behavioral Observation Data Collection
It is useful to discriminate among different types of observation ...

Naturalistic Observation
- Participants don’t know that they are being observed
  - requires “camouflage” or “distance”
  - researchers can be VERY creative & committed !!!!

Participant Observation (which has two types)
- Participants know “someone” is there – researcher is a participant in the situation
  - Undisguised
    - the “someone” is an observer who is in plain view
    - Maybe the participant knows they’re collecting data…
  - Disguised
    - the observer looks like “someone who belongs there”

Observational data collection can be part of Experiments (w/ RA & IV manip) or of Non-experiments !!!!!
Self-Report Data Collection

We need to discriminate among various self-report data collection procedures...

- Mail Questionnaire
- Computerized Questionnaire
- Group-administered Questionnaire
- Personal Interview
- Phone Interview
- Group Interview (focus group)
- Journal/Diary

In each of these participants respond to a series of questions prepared by the researcher.

Self-report data collection can be part of Experiments (w/ RA & IV manip) or of Non-experiments !!!!!

Trace data are data collected from the “marks & remains left behind” by the behavior we are trying to measure.

There are two major types of trace data...

Accretion – when behavior “adds something” to the environment
  • trash, noseprints, graffiti
Deletion – when behaviors “wears away” the environment
  • wear of steps or walkways, “shiny places”

Garbageology – the scientific study of society based on what it discards – its garbage !!!

- Researchers looking at family eating habits collected data from several thousand families about eating take-out food
- Self-reports were that people ate take-out food about 1.3 times per week
- These data seemed “at odds” with economic data obtained from fast food restaurants, suggesting 3.2 times per week
- The Solution – they dug through the trash of several hundred families’ garbage cans before pick-up for 3 weeks – estimated about 2.8 take-out meals eaten each week

Data Sources …

It is useful to discriminate between two kinds of data sources...

Primary Data Sources
  - Sampling, questions and data collection completed for the purpose of this specific research
  - Researcher has maximal control of planning and completion of the study – substantial time and costs

Archival Data Sources (AKA secondary analysis)
  - Sampling, questions and data collection completed for some previous research, or as standard practice
  - Data that are later made available to the researcher for secondary analysis
  - Often quicker and less expensive, but not always the data you would have collected if you had greater control.
Is each primary or archival data?

- Collect data to compare the outcome of those patients I've treated using Behavior vs. using Cognitive interventions.  
- Go through past patient records to compare Behavior vs. Cognitive interventions.  
- Purchase copies of sales receipts from a store to explore shopping patterns.  
- Ask shoppers what they bought to explore shopping patterns.  
- Using the data from some else's research to conduct a pilot study for your own research.  
- Using a database available from the web to perform your own research analyses.  
- Collecting new survey data using the web.  

Data collection Settings

Same thing we discussed as an element of external validity…

Any time we collect data, we have to collect it somewhere – there are three general categories of settings:

Field

- Usually defined as “where the participants naturally behave”
- Helps external validity, but can make control (internal validity) more difficult (RA and Manip possible with some creativity)

Laboratory

- Helps with control (internal validity) but can make external validity more difficult (remember ecological validity?)

Structured Setting

- A “natural appearing” setting that promotes “natural behavior” while increasing opportunity for “control”
- An attempt to blend the best attributes of Field and Laboratory settings !!!

Data collection Settings identify each as laboratory, field or structured…

- Study of turtle food preference conducted in Salt Creek.  
- Study of turtle food preference conducted with turtles in 10 gallon tanks.  
- Study of turtle food preference conducted in a 13,000 gallon “cement pond” with natural plants, soil, rocks, etc.  
- Study of jury decision making conducted in 74 Burnett, having participants read a trial transcript.  
- Study of jury decision making with mock juries conducted in the mock trial room at the Law College.  
- Study of jury decision making conducted with real jurors at the Court Building.