Scaling, Validity, and Reliability
Measuring a Construct

- Not every phenomena of interest can be measured through direct observation

- Constructs are phenomena that exist but cannot be directly observed
Measuring a Construct

- Instruments can be created to measure constructs.
- Researchers interested in a particular construct (e.g. substance abuse) or characteristic (e.g. sex) refer to these phenomena as *variables*.
Scaling

Nominal: classified and counted (ethnicity)

Ordinal: Ranked in order (as in a race)

Interval: Equal interval, no true zero

Ratio: Equal intervals, true zero, ratio relationship

Discrete (categorical)

Continuous
Nominal

- Method of categorization
- No label is quantitatively higher or lower than another label—it is simply a label
- For example:
  - African American
  - Caucasian
  - Asian
  - Latino/Latina
  - Native American
Ordinal

- Denotes an order or ranking
- You cannot make comparisons beyond the knowledge of the order
- Both nominal and ordinal variables are known as disperse variables
Interval

- Denotes equality between levels with no true zero
  - Fahrenheit temperature

- Social science often uses quasi-interval scales
  - mood
Ratio

- Denotes equality between levels with a true zero
  - Number of words spelled correctly
- Interval and ratio variables are sometimes referred to as continuous variables
Importance

- The ability to analyze validity and reliability is the cornerstone to identifying whether an experiment utilized
  - proper instrumentation
  - Proper procedure
  - Achieved meaningful results
So, who comes up with this stuff?

- American Educational Research Association (AERA)
- American Psychological Association (APA)
- National Council on Measurement in Education (NCME)
- Standards for Educational and Psychological Testing
1999 *Standards*

- “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations” (AERA, APA, & NCME, 1999, p. 9).
Evidence based on test content

- Logical analyses and experts’ evaluations of the content of the measure
  - Items
  - Tasks
  - Formats
  - Wording
  - Processes required of examinees
- Addresses whether the content of a measure represents a specified content domain
Evidence based on test content

- Based on
  - a review of the literature
  - Expert opinion
Evidence based on response processes

- Extent to which the tasks or types of responses required of the examinees fit the intended, defined construct
  - e.g. A self-report, paper-pencil inventory measures an affective construct
Evidence based on internal structure

- *Constructs* are traits that cannot be observed. They are hypothesized characteristics.
- Theories are developed related to specific constructs
- Tests items load on specifics aspects of the theory
- The extent to which the internal components of a test match the defined construct
  - Principle component analysis
  - Confirmatory factor analysis
Principle Component Analysis and Factor Analysis (Stat Soft, Inc)

- Both are data reduction procedures
- They combine 2 or more correlated variables in to a single factor
- But there are differences (we will not go into the computational aspects)
Evidence based on relations to other variables

- Utilizes correlation studies to investigate group comparisons
  - Convergent and discriminant validity
  - Concurrent and predictive validity
Criterion related evidence

- Evidence may be *predictive*
  - The test is designed and used to predict other variables (i.e. freshman gpa).
    - ACT, SAT, GRE
  - This can be demonstrated via correlation or regression analyses
Criterion related evidence

- Evidence may be *concurrent*
  - The test is related to other measure (i.e. other tests, grades, ratings, etc).
    - K-BIT is a shortened version of the WISC-III, but has strong criterion related evidence
  - Often demonstrated through correlational analyses
Criterion related evidence

- Evidence may be *concurrent*
  - When we demonstrate that a test measures *similar* constructs, we say the evidence is *convergent*
  - When we demonstrate that a test measures *different* constructs, we say the evidence is *discriminate*
The MTMM

- Assesses convergent and discriminate evidence

**Convergent evidence** is the degree to which concepts that should be related theoretically are interrelated in reality.

**Discriminant evidence** is the degree to which concepts that should *not* be related theoretically are, in fact, *not* interrelated in reality.
Multi-trait method

<table>
<thead>
<tr>
<th></th>
<th>SE₁</th>
<th>SE₂</th>
<th>SE₃</th>
<th>LOC₁</th>
<th>LOC₂</th>
<th>LOC₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE₁</td>
<td>1.00</td>
<td>.83</td>
<td>.89</td>
<td>.02</td>
<td>.12</td>
<td>.09</td>
</tr>
<tr>
<td>SE₂</td>
<td>.83</td>
<td>1.00</td>
<td>.85</td>
<td>.05</td>
<td>.11</td>
<td>.03</td>
</tr>
<tr>
<td>SE₃</td>
<td>.89</td>
<td>.85</td>
<td>1.00</td>
<td>.04</td>
<td>.00</td>
<td>.06</td>
</tr>
<tr>
<td>LOC₁</td>
<td>.02</td>
<td>.05</td>
<td>.04</td>
<td>1.00</td>
<td>.84</td>
<td>.93</td>
</tr>
<tr>
<td>LOC₂</td>
<td>.12</td>
<td>.11</td>
<td>.00</td>
<td>.84</td>
<td>1.00</td>
<td>.91</td>
</tr>
<tr>
<td>LOC₃</td>
<td>.09</td>
<td>.03</td>
<td>.06</td>
<td>.93</td>
<td>.91</td>
<td>1.00</td>
</tr>
</tbody>
</table>
Evidence based on consequences of testing

- Some benefit should be realized from the intended use of scores
  - Placement
  - Effectiveness
- Previously not considered as evidence for validity
- Few guidelines on how to measure
Reliability

- The extent to which a test measures a construct accurately and consistently
- Substantiated by high correlations and reduced error
Reliability

- Various methods may be utilized to produce strong correlations:
  - Stability over time (test-retest)
  - Parallel forms
  - Internal consistency
    - Split half
    - Cronbach’s Coefficient alpha
  - Inter-rater reliability
Computing Coefficient Alpha

\[
\frac{n}{n-1} \left( \frac{S_{tot}^2 - \sum S_i^2}{S_{tot}^2} \right)
\]

- \((\text{# of items}/(\text{# of items} - 1)) \ast ((\text{Variance of the total score} - \text{Sum of the Variances of the items})/\text{Variance of the total score})\)
Types of Instruments

- Achievement
- Aptitude
- Personality
- Projective
- Observations

- Questionnaire
- Interest, Attitude, Values
  - Likert Scale
- Interview