Standardized Measurement and Assessment
Measurement

Identify dimensions, quantity, capacity, or degree of something

Assign a symbol or **number** according to **rules**

(e.g., assign a **number** for height in inches according to the **rule** for a length of one inch on a ruler)
Nominal
   Name

Ordinal
   Rank Order

Interval
   Equal Distance Between Numbers
   Does Not Have Absolute Zero Point
   (e.g., Celsius, Fahrenheit)

Ratio
   Equal Intervals and an Absolute Zero Point
   (e.g., number correct, annual income)
Testing

Measure variables (e.g., technology skills, time spent on task)

Assessment

Gather and integrate data to make an educational evaluation (e.g., may use a collection of tests, interviews, and observations to make an evaluation)
Assume that psychological traits and states exist and can be measured in various ways.

**Trait**
- Distinguishable characteristics that vary among individuals
  - Characteristics stay that way over time
    (e.g., persistent anxiety)

**State**
- Temporary
  (e.g., short-term anxiety due to stressful event)
Assessment is influenced by other variables (e.g., taking test when sick can cause error in the assessment).

Interpret results with caution for weaknesses in assessment.

Testing and assessment can benefit society (e.g., performance standards).
Reliability

Consistency

When a test is reliable, we will get similar scores every time we use the test.
Test-Retest Reliability

Consistency of scores over time

Give same test to same group of people on two different occasions

High and high? Low and low?

Time interval can have an effect.
Too soon -- people may remember responses.
Too late -- other things influence learning.
Equivalent Forms Reliability

Consistency of scores on two equivalent tests

Tests are designed to measure the same thing.
Equivalent Forms Reliability (continued)

Specific items on test are different, but

• Same number of items
• Same difficulty level
• Administer, score, and interpret test in same way

Students take both tests. Scores are correlated.

Seldom used -- difficult to construct two tests
Split-Half Reliability

• Split one test into two equivalent halves

• Assess consistency of scores on each half

• Use when time is a factor
Split-Half Reliability (continued)

Ways to split test

Odd numbers and even numbers

Randomly assign items to one half or the other

Divide by content -- equal number of items in each half -- equivalent in content and difficulty
Steps to Follow

1. Score each half of the test for each person.

2. Compute correlation between the two scores.

3. Use Spearman-Brown formula to adjust the computed correlation coefficient.
Spearman-Brown Formula

Used to correct the split-half reliability coefficient because full-length test was split into two shortened tests.

Longer tests are usually more reliable.

Adjusted correlation must be high to indicate reliability.
Spearman-Brown Formula (continued)

Only use Spearman-Brown formula for tests measuring **only** one construct or characteristic.

Do not use when measuring more than one construct or characteristic.

Do not use for tests with a time limit.
Internal Consistency

Consistency in measuring a single construct (e.g., anxiety, self-esteem, learning disability)

“Inter-item consistency”

Are the test items measuring the same construct?
KR-20

Kuder-Richardson
Twentieth formula in the series

Used to measure internal consistency

Estimates reliability of a homogeneous test
(with test items on a single construct)
Coefficient alpha

Developed by Cronbach (1951)

Variant of KR-20 formula

Estimates reliability of a homogeneous test
Inter-Scorer Reliability

Also called Inter-Rater Reliability

Do two or more raters agree on a score?

Have each rater conduct an independent review and provide a score.

Compute correlation coefficient between the scores.
What evidence do we have to show that we have taken correct actions based on a test or assessment?
Gather evidence to support inference (for example, inference based on test scores)
Ways to Gather Evidence of Validity

Content-Related Evidence

Criterion-Related Validity
  Concurrent
  Predictive

Construct Validity

Factor Analysis
Content-Related Evidence

Determined by expert judgment

Do the test items sufficiently represent the area of interest?

(e.g., morale, shyness, computer skill)

(e.g., a test on Chapters 1 - 4)

Are items covered according to emphasis during instruction?
Criterion-Related Evidence

Can test scores be used to predict performance in some activity?

The activity is the criterion (the standard).

(e.g., design test that measures academic ability -- compare scores of that test to grade point averages)

Students with high test scores should also have high GPAs.
Criterion-Related Evidence

Concurrent Evidence

Gather test data and criterion data at nearly same time and compare results. (e.g., compare scores on new version of SAT with high school GPAs)
Criterion-Related Evidence

Predictive Evidence

Time interval between collection of test scores and criterion scores (e.g., correlate SAT scores with college GPAs)
Validity Coefficient

Correlation coefficient

e.g., compare test scores (SAT) and criterion scores (college GPAs)

-- looking for high, positive correlation between SAT scores and college GPAs
Factor Analysis

e.g., the Harter Self-Perception Profile for Children measures five dimensions of self-esteem (scholastic competence, social acceptance, athletic competence, physical appearance, and behavioral conduct)

Factor analysis is used to find evidence that each subtest is measuring a different construct.
Caution

Characteristics of people in your study must match the characteristics of people in the reliability and validity studies.