Reliability AND Validity

Fact checking your instrument

General Principles

- Clearly Identify the Construct of Interest
- Use Multiple Items
- Use One or More Reverse Scored Items
- Use a Consistent Response Format That Matches the Items

Not Anything Goes

- Not Necessarily *Easy*
- Requires Careful Creation and Empirical Evaluation

Evaluation

- The empirical focus of science must apply to the measures used as well as the study topics.
- Verifying the accuracy and consistency of tools that measure variables is important.

Evaluation

- Most commonly used measurements have gone through rigorous verification
  - Beck Depression Inventory
  - Recovery Assessment Scale
- New measures also need verification, this means that new scales have to be assessed when used
- Measures are assessed on two primary attributes

Dimensions of Evaluation

- Reliability – Are observations answering consistently?
- Validity – Is the instrument measuring what it says it is measuring?

True Score Theory

True score theory - statistical concept: there is a true value for a measurable characteristics, but error occurs

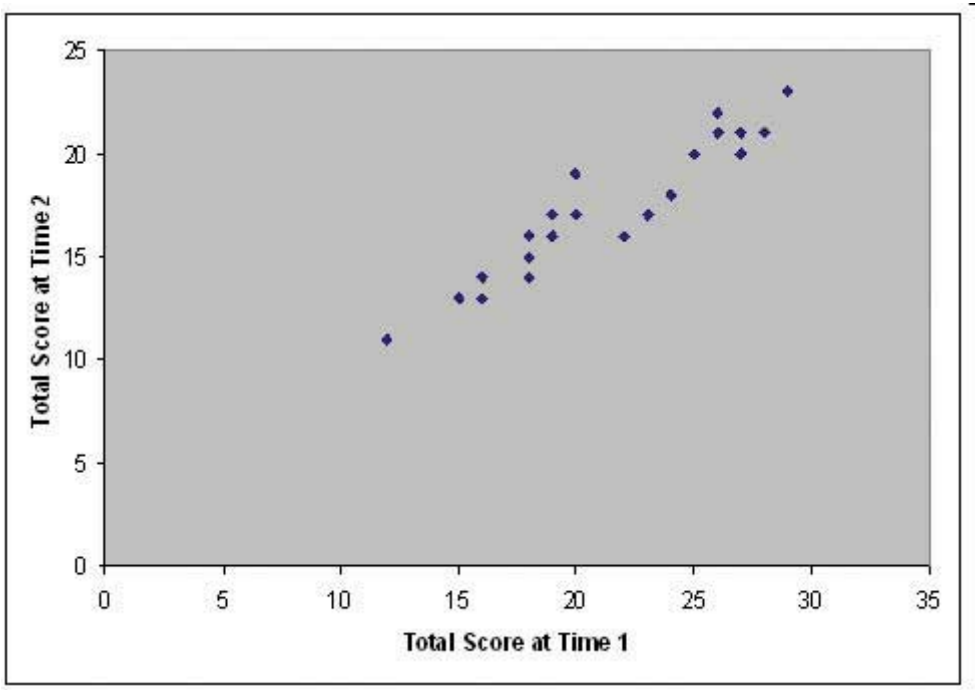
- True score theory of intelligence
  - One true ~~ring~~ value for intelligence
  - Extraneous variables cause errors
- Anything beyond the main characteristic of interest could be thought of as error
- Testing multiple times can give support to the reliability of a measure and account for error (though it may introduce other problems)

Reliability

- Are measurements consistent
- There are several measures of reliability but here are some common ones
  - Test-retest
  - Parallel forms
  - Internal consistency
  - Inter-rater

Test-Retest Reliability

- Test 1 vs Test 2
- Consistency Across Time
- Assessed by Test-Retest Correlation



Interpreting Test-Retest Correlations

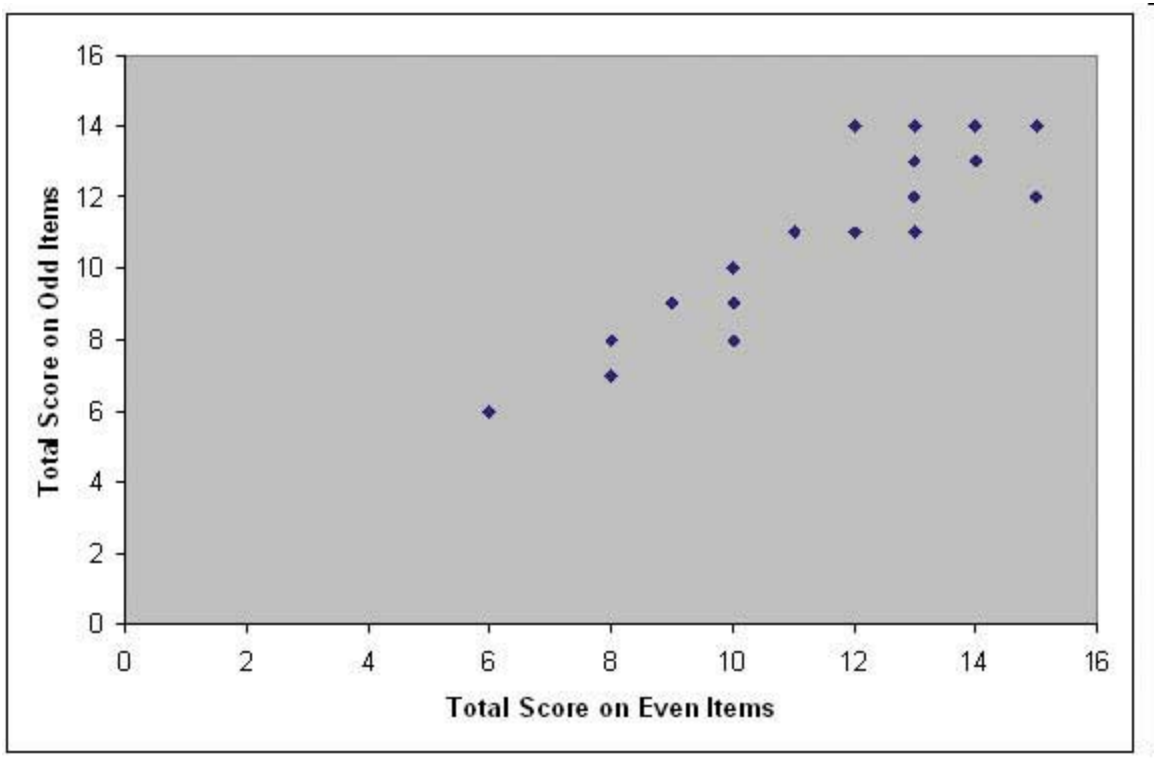
- In General, r = .8 is “Good”
- However, One Must Take Into Account …
  - Expected Consistency of Construct
  - Time Between Tests

Parallel Forms Reliability

- I create two forms with items measuring the same construct and see if they correlate with each other
- Testing the reliability of a measure of depression
  - I create one ten item form (form A)
  - I create a second ten item form with similar (but still distinct) questions (form B)
- I then administer both forms to a group of people and see how responses on the two forms correlate
- One drawback is that you need to create a lot of items

Internal Consistency

- Consistency Across Items
- Assessed by Split-Half Correlation



Internal Consistency

- A measure of internal consistency that is far superior to split-half because it essentially measures the mean of all possible split-halves
  - There are 252 ways to split a set of 10 items into two sets of five.
  - Cronbach's α would be the mean of the 252 split-half correlations
- Created by a Fresno State Alumus

Cronbach's alpha	Internal consistency
$0.9 \leq \alpha$	Excellent
$0.8 \leq \alpha < 0.9$	Good
$0.7 \leq \alpha < 0.8$	Acceptable
$0.6 \leq \alpha < 0.7$	Questionable
$0.5 \leq \alpha < 0.6$	Poor
$\alpha < 0.5$	Unacceptable

Inter-Rater Reliability

- A behavioral measure
- Scores are correlated between different raters on a behavioral measure

Validity

- Extent to Which Scores Represent the Construct They Are Meant To
- Dimensions
  - Face – Measure Appears Valid
  - Content – Measure "Covers" Construct
  - Criterion – Scores Correlate with Related Variables
  - Discriminant – Scores Do Not Correlate with Non-Related Variables

Face Validity

- Does a measure look like it is measuring what it says it is measuring?
- Sadism measurement on a scale from 1 (never) to 5 (always)
  - 1. I often get angry
  - 2. I get pleasure in hurting small animals
  - 3. I don't have pets for long
  - 4. etc…
- Face validity??
  - Good?
  - Bad?

Content Validity

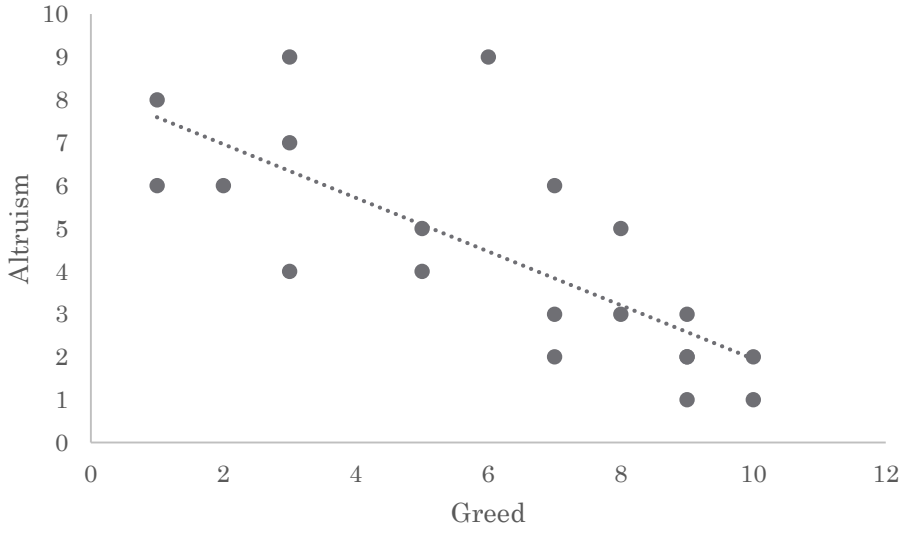
- Based on the conceptual definition of the construct
  - Experts are often the ones who determine content validity
  - Based on current theories of the construct

- Conceptual definition of trauma: "When a person experiences an event or events that drastically affect the person's ability to function on a day to day basis."
  - Items in a trauma instrument should reflect this definition of trauma

Criterion Validity

- Related variables should be related
  - Pearson's r value should be close to 1

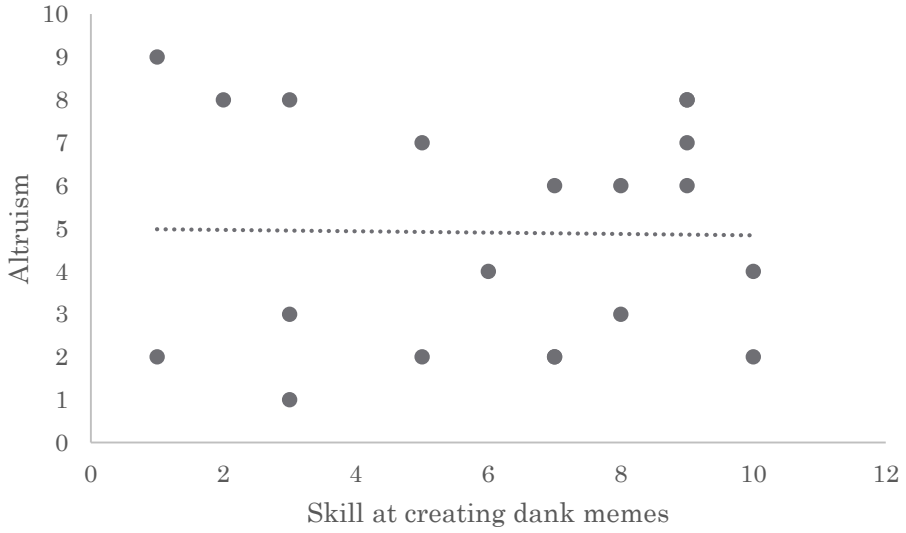
- r = –0.745



Discriminant validity

- Unrelated variables should be unrelated
  - Pearson's r value should be close to 0

- r = –0.018



- A strong measure needs to have both reliability and validity
  - If a measure does not measure what it claims (valid), then it is not very useful
  - Likewise, it is not useful if it not consistent (reliable)
- Body weight scale is accurate if it produces weight measurements that are representing the true weight of a person standing on it
  - It is reliable if it produces consistent results **under similar circumstances**
  - This means that a measure can be reliable, but not valid. But a score that is valid must be reliable.