Test Reliability

Reliability

- What is reliability?
- Types of reliability
  - Test-Retest reliability
  - Parallel forms reliability
  - Internal Consistency measures
  - Inter-rater reliability
- Standard measure of error

What is reliability?

- When we say a car or our best friend is 'reliable' what do we mean?

What is reliability?

- The reliability of a test is the extent to which can be relied upon to produce 'true' scores = a measure of its consistency/ variability across occasions or with different sets of equivalent items

Reliability = True variance / Observed variance

$= S^2_{true} / S^2_{observed}$

Reminder: What is variance?

s^2 = *variance* = the average squared difference from the mean.

- It is the square of standard deviation

Reminder: what is a coefficient?

a. A number or quantity placed (usually) before and multiplying another quantity known or unknown.
   Thus in $4x^2 + 2ax$, 4 is the coefficient of x^2, 2 of ax, and 2a of x

b. A multiplier that measures some property of a particular substance, for which it is constant, while differing for different substances.
   e.g. coefficient of friction, expansion, torsion, etc.

Classical reliability theory

Test reliability = $S^2_{true} / S^2_{observed}$

- This ratio will never be greater than 1: Why?
- This ratio will usually be quite a bit lower than 1: Why?

Classical reliability theory

Test reliability = $S^2_{true} / S^2_{observed}$

- Observed variance in scores includes an error (unsystematic) component:

$$S^2_{observed} = S^2_{true} + \mathbf{S^2_{error}}$$

- This error variance $\mathbf{S^2_{error}}$ is (by definition):

$$S^2_{error} / S^2_{true} = 1 - \text{reliability}$$
$$= 1 - (S^2_{true} / S^2_{observed})$$

So: How can we get S^2_{true}?

Alas…you can't!

- S^2_{true} cannot be directly computed
- For this reason, *we must estimate reliability* by indirect means:
  - look at the effects of variation in test administration conditions
  - look at the effect of variations in test content
- And 'look at' here means 'compute correlations'

What is a correlation?

- In a correlation, we want to find the equation for the (one and only) line (the line of regression) which describes the relation between variables with the least error.
  - the idea is simply that we draw a line such that the squared distances on two (or more) dimensions of points from the line would not be less for any other line

What co-relates?

- r = The covariance of x and y / the product of the SDs of X and Y
- *Covariance* is related to variance
  - Variance = the average squared difference from the mean
  - Covariance = the average value of all the pairs of differences from the mean for X multiplied by the differences from the mean for Y (the average product of differences from the two group means)

Why does it work?

- r = The covariance of x and y / the product of the SDs of X and Y
- When X and Y are related, large numbers will be systematically multiplied by large numbers with the same sign (for differences on both sides of the mean) = covariance will be large & close to the product of the SDs of X and Y, so r will be close to 1.
- The root of a correlation is the amount of variance explained.

Test-retest Reliability

- Correlate scores of the same people with two different administrations
  - The r is called the *test-retest coefficient* or *coefficient of stability*
- There is no variance due to item differences or conditions of administration
- Shorter inter-test intervals give larger r

Parallel-forms reliability

- One factor that does impact on test-retest reliability is individual differences in memory
- Solution is to give two or more forms of the test, to get a *parallel forms coefficient* or *coefficient of equivalence*

Parallel-forms reliability

- How can we deal with error from two sources: error due to different test times and error due to different forms?
  - Use Form A with half the sample, and Form B with the other half at T1; then switch at T2
  - The correlation between scores on both forms is the *coefficient of stability and equivalence*, taking into account errors due to both time of administration and due to different test items on the two forms

Internal consistency: Split-half method

- We can treat a single form as two forms: split it into two arbitrary halves and correlate scores on each half (*Split half reliability*)
- To get the reliability of the test as a whole (assuming equal means and variance), use *Spearman-Brown prophecy formula*:

$$r_{whole} = 2r_{half}/(1 + r_{half})$$

Ramping up the split-half method

- The split half method takes arbitrary halves
- However, different arbitrary halves might give different r values
- A better method might be to take all possible split halves, and average their values
- Luckily, there is a (fairly) easy way to do this...

Internal consistency: *Cronbach's alpha*

- Cronbach's (1951) alpha (*coefficient alpha*) is a widely used and widely reported measure of the extent to which item responses obtained at the same time correlate highly with each other.
  - Note: This is not the same as being a measure of unidimensionality, though it is sometimes reported as being so
  - You can get a high alpha coefficient with distinct, but highly-intercorrelated, dimensions in a test.
- Cronbach's alpha is mathematically equivalent to taking an average of all split halves

Cronbach's alpha

Alpha $= (k/(k-1)) * [1- \{SUM (s^2_i)\} / s^2_{total}]$

k = the number of items

s^2_i = the variances of scores for item I

$\{SUM (s^2_i)\}$ = the sum of all item variances

s^2_{total} = the total variance for all items.

How much reliability is enough?

- As usual, in this uncertain world there is no hard answer to this question
- Alphas for personality tests (0.46 - 0.96) tend to be lower than alphas for achievement and aptitude tests (0.66 - 0.98)
- If you are comparing groups means, modest alphas of 0.6 to 0.7 are sufficient
- If you want to make claims about differences between single individuals, you need to have more reliable scores; alphas of 0.85 or better

How can we increase reliability?

- Analyze your items
  - Bad items decrease reliability
- Increase the number of items
  - Longer tests are generally more reliable (Why?)
- Factor analyze
  - Unidimensional tests are more reliable (Why?)
  - Factor analyze to find if you are looking for 'spurious reliability'

How can we increase reliability?

- You can figure out how many items you need to get a given reliability using a generalization of Spearman's prophecy formula

Inter-rater reliability

- On tests requiring evaluative judgments (projective tests; personality ratings), different scorers may give different scores
- *Inter-rater reliability* is the correlation between their scores
- Generalized you get an *intraclass coefficient* (or *coefficient of concordance*) as the average correlation between many raters

What is error?

- **Error** is *the amount of uncertainty you have in a measurement*
  - **By definition**, it is random
  - If it is not, then it is not error
- Why can we be extremely thrilled about this fact?
  - You know why: because randomly distributed things that can vary in two directions have certain beautiful properties (Such as?)

Why is this thrilling?

- Because error is normally distributed, we can quantify it in the same ways we can quantify any normally distributed measure
- In particular, we can give the average and standard deviation of any error measure, and thereby compute the probability that any given error is likely- or we can quantify confidence bounds on any measure
  - eg. There is a ~95% chance that true score falls with two SDs of the obtained score

Standard error of measurement

- Reliability allows us to estimate standard error

$$S_{mr} = S * (1 - r)^{0.5}$$

- S = population SD of test scores
  - Note that the lower the reliability r, the higher the error
- s_{err} estimates the SD a person would obtain if he took the test infinitely many time

Standard error of measurement

- IQ tests have a mean of 100 and a SD of 15. One test has a reliability of 0.89. You score 130 on that IQ test. What is the standard error of measurement? What is the 95% [1.96 SD] confidence interval on your IQ?

$S_{mr} = S * (1 - r)^{0.5}$
$= 15 * \text{sqrt}(1- 0.89)$
$= 4.97$

We can be 95% sure that your true IQ is between 120.3 & 139.7.

Validity and reliability

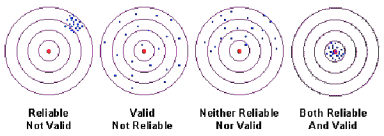

Image from: http://trochim.human.cornell.edu/kb

I highly recommend this site.