Measurement Theory & Reliability

The Big Picture
- Measurement error makes it difficult to identify the true patterns of relationships between important variables
- Reliability indexes the amount of measurement error
- Relationships are attenuated by unreliability

\[ r_{XY} = \rho_{XY} \sqrt{r_{XX} r_{YY}} \]

Attenuation

\[ Y = .05 + .96(X) \]
\[ Y = .03 + .11(X) \]

R² = .49; r = 1.0
R² = .07; r = 0.5

Where does Error Come From?
- Reliability is the extent to which measurements are consistent, repeatable, dependable, or generalizable across measurement conditions
- Anything that results in inconsistencies across measurement conditions (e.g., items, raters, times, etc) is a source of measurement error
- Sources...

Error Sources
- Psychologically interesting variables that matter, but were not included in the model
  - Idiosyncratic Error
    - Mood, fatigue, boredom, language difficulties, attention
  - Generic Error
    - Poor instructions, confusing categories or anchors, setting, variations in administration
  - Additive Error
    - Acquiescent, leniency, severity response sets

The Big Picture
- The purpose of measurement theory is to provide reliability estimates for the measures
- To make good decisions, you must know how much error is in the data upon which the decisions are based
- The challenge is to identify/develop a measurement model that makes reasonable assumptions while still allowing a useful way to estimate reliability.
Error Sources
- Systematic error
- Social desirability
- Demand characteristics, experimenter expectancies
- Halo error
- Interviewer bias
- Rater dispersion bias, midpoint response sets

Classical Test Theory – Single Indicator
- Most common measurement model
- Based on the linear combination of true score and error and the assumption that error is random:
  \[ X = t + e \]

Classical Test Theory – Single Indicator
- Recall, that the variance of a linear composite is:
  \[ \sigma_X^2 = \sigma_t^2 + \sigma_e^2 + 2 \rho_{te} \]
- If you assume that the error and true score are independent, then
  \[ \sigma_X^2 = \sigma_t^2 + \sigma_e^2 \]

Classical Test Theory – Single Indicator
- Signal to Noise Ratio:
  \[ SNR = \frac{\sigma_t^2}{\sigma_e^2} \]
- Reliability:
  \[ \rho_{xx} = \frac{\sigma_{XX}}{\sigma_X^2} \]

CTT – Parallel Forms Reliability Estimation
- Original approach to estimating the reliability ratio was based on the logic of parallel forms
- Assume you have two exactly equal measures of the same latent variable (X and X')
- If two measures are parallel, then:
  * Same true scores \( \tau_X = \tau_X' = \tau \)
  * Therefore same true score variance \( \sigma_{XX}^2 = \sigma_{XX'}^2 = \sigma_{XX}^2 \)
  * Same error variance \( \sigma_{ee}^2 = \sigma_{ee}^2 = \sigma_{ee}^2 \)
  * Therefore same observed variance \( \sigma_{XX}^2 = \sigma_{XX'}^2 = \sigma_{XX}^2 \)
Given these results, the correlation between the two parallel measures is an estimate of the reliability
\[
\rho_{xx'} = \frac{COV(\tau_x + e_x, \tau_{x'} + e_{x'})}{\sigma_{x}\sigma_{x'}}
\]
\[
\rho_{xx'} = \frac{\sigma_{\tau_x e_x} + \sigma_{\tau_{x'} e_{x'}} + \sigma_{\tau_x e_{x'}} + \sigma_{\tau_{x'} e_{x}}}{{\sigma_{x}^2}} = \frac{\alpha^2}{\sigma_{x}^2}
\]

CTT – Multiple Indicators
- Theoretically, a person's true score may be defined as the mean response to an infinite set of observations – Domain sampling theory
- Therefore, one of the best ways to get closer to the true score (less error!) is to add measurements
- Now, reliability of a linear composite of indicators

A very simple measurement model . . .
\[
t = \sum_{i=1}^{k} \lambda_i \xi_i + \epsilon_t
\]
\[
\epsilon_t = 0 \quad r_{e_t} = 0 \quad r_{e_t e_{t'}} = 0
\]

CTT – Reliability of a linear composite
- By adding indicators (items, times, judges, etc) you get a better measure.
- The composite provides a more precise estimate of the true score because the random errors on each of the indicators cancel each other out
- Averages are better estimates than single observations

CTT – Parallel Forms Reliability Estimation
- Parallel indicators
  - Equal lamdas (loadings)
  - Equal error variances
- Tau-equivalent indicators
  - Equal lamdas (loadings)
- Congeneric indicators
  - It's a free-for-all
CTT – Split Half Reliability

- **Split Half Reliability problems...**
  - How to split?
  - Spearman-Brown Prophecy formula
  - Reliability for a shorter or longer set of items

\[ r_{sb} = \frac{k \cdot r_{ij}}{k + (k-1) \cdot r_{ij}} \]

average item correlation

CTT – Reliability of a linear composite

- Cronbach's alpha \( (\alpha) \)
  - Generally used on items as an index of internal consistency, but applicable to any types of measures that meet the assumptions
  - Equals the average split half reliability corrected for length
  - Assumes tau-equivalence
  - Not a measure of uni-dimensionality
  - Equal to KR-20 when computed on dichotomously scored items

\[ \alpha = \frac{k}{k-1} \left[ 1 - \frac{\sum \sigma^2_i}{\sigma^2_{LC}} \right] \]

CTT – alpha

CTT – Reliability of a linear composite

- The denominator is simply the variance of the linear composite of all the indicators
- The variance of a linear combination is simply the sum of all elements in the variance-covariance matrix of the original components.
CTT – Reliability of a linear composite

- Standardized alpha
- Used when you will standardize the variance of each indicator before forming the composite

\[
\alpha_i = \frac{k \rho_{ij}}{1 + (k - 1) \rho_{ij}}
\]

CTT – Reliability of a linear composite

- Reliability of Congeneric test
- Estimated using confirmatory factor analysis

\[
\rho_{xx} = \frac{\sum \lambda_i^2 Var(\tau)}{\sum \lambda_i^2 Var(\tau) + \sum Var(\epsilon_i)}
\]

CTT – Alternative interpretation of reliability

- Correlation of observed test score with true score – Reliability index
- Estimated as the square root of the reliability coefficient

Some important points . . .

- The level of reliability does not increase with increases in sample size (N).
- Reliability depends on the conditions of measurement. There are as many reliabilities for a measure as there are different conditions of measurement.
- Application of the Spearman-Brown prophecy formula assumes that new items are the same as old items.

Uses

- Once you have a reliability estimate you can do many things
  - Dissattenuation
  - Standard Error of Measurement
  - Standard Error of the Difference
Uses

- Dissattenuation of an observed relationship for unreliability in either variable

\[ r_{xy}' = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \]

Standard Error of Measurement

- The standard error of measurement is similar to the standard error of estimate in regression (the standard deviation of the errors of prediction).

- Estimate of the average distance of observed test scores from an individual's true score.

\[ \text{SEM} = \sigma \sqrt{1 - r_{xx}} \]

Examples:

<table>
<thead>
<tr>
<th>( \sigma )</th>
<th>( r_{xx} )</th>
<th>SEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>.90</td>
<td>3.16</td>
</tr>
<tr>
<td>10</td>
<td>.60</td>
<td>6.32</td>
</tr>
</tbody>
</table>

Standard Error of Measurement

Confidence Interval: Use SEM to calculate a band or range of scores around the person's observed score that has a high probability of including the person's true score.

\[ \text{CI} = \text{Obtained score} \pm z \times (\text{SEM}) \]

- z = 1.0 for 68% level
- z = 1.44 for 95% level
- z = 1.65 for 99% level
- z = 2.58 for 99.9% level

Normal Curve

- 68% of scores fall within 1SEM
- 95% of scores fall within 2SEM
- 99% of scores fall within 3SEM

- Actual z-score = 1.96
- Actual z-score = 2.58

- 3.16 x 1.96 = 6.19 (95% confidence)
- 3.16 x 2.58 = 8.15 (99% confidence)

Standard Error of the Difference

When we want to compare two scores:

\[ \text{SED} = \sqrt{\text{SEM}_1^2 + \text{SEM}_2^2} \]

How reliable is reliable?

- Reliability estimates of .70–.80 are usually “good enough” for research.
- However, if important decisions are being made, the test should have reliabilities of at least .90 and .95 is the desirable standard.
- It’s usually not worthwhile to try to improve reliability beyond .90 in most cases. The exception is when individual decisions of great import are being made. In this case, it’s important to also consider the SEM.
- Nunnally is the standard source for these standards.