

Reliability Handout

A. **Reliability:** The consistency of a measure. (does it always measure the same thing).

$$x_i = t_i + e_i \quad \text{Where} \quad \begin{array}{l} x_i = \text{subject score} \\ t_i = \text{true score} \\ e_i = \text{error} \end{array}$$

$$\sigma_x^2 = \sigma_t^2 + \sigma_e^2$$

Reliability Coefficient = ratio of true score variance to observed (total variance)

$$\text{Reliability} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\sigma_x^2 - \sigma_e^2}{\sigma_x^2}$$

If $\sigma_x^2 = \sigma_e^2$ then reliability = 0 If $\sigma_e^2 = 0$ then reliability = 1

B. Multiple Presentations of a test.

Test-Retest: give test at two points in time and correlated the results. High positive correlation indicated good Test-Retest Reliability. (May suffer from practice effects). Minimum test retest correlation of .50 across 3 months.

Parallel Forms: Two test that are identical conceptually but the items are different. (Controls for practice effects).

$$r_{xx'} = \frac{\sigma_t^2}{\sigma_x^2} \quad r_{xx'} = \frac{\Sigma (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{\left(\sqrt{\Sigma (X_1 - \bar{X}_1)^2}\right)\left(\sqrt{\Sigma (X_2 - \bar{X}_2)^2}\right)}$$

Where the numerator = covariance between time 1 and time 2 (or form 1 & form 2).

Where the denominator = pooled variance for t1 and t2

C. Internal Consistency

1. Estimating Reliability from Estimates of Error

Kuder-Richardson Formula 21: This formula is often useful for quick estimates of reliability given a limited amount of information for dichotomous data (correct vs. incorrect, yes vs. no, etc.)

$$KR21 = \left(\frac{K}{K-1}\right)\left(1 - \frac{\bar{X}(K - \bar{X})}{K(\sigma^2)}\right)$$

Where K = # of items

\bar{x} = the mean of the total test scores

σ^2 = the variance of the total test scores.

Kuder-Richardson Formula 20: This formula is used for items scored dichotomously (right vs. wrong)

$$r_{kr20} = \left(\frac{k}{k-1}\right)\left(1 - \frac{\Sigma pq}{\sigma^2}\right) \text{ where}$$

p = (# respondents getting it correct)/n

q = (# respondents getting it wrong)/n or (1-p)

Σpq = pq summed across all items

σ^2 = variance for the total test scores.

k = number of items.

2. Internal Consistency Based on Covariation

Split-Half: Split test items into two equivalent groups and correlate one half with the other.

Spearman Brown Prophecy formula: estimates the reliability for whole test based on the correlation for 1/2 the test. (e.g. what happens when you double the length of the halved test).

$$r_{xx'} = \frac{2r_{oe}}{1 + r_{oe}} \quad \text{Where } r_{oe} = \text{correlation between odd and even items.}$$

If used with $r_{xx'}$ from test retest, split half, or scale alpha, then Spearman Brown Prof. Form estimates the reliability you would expect if you doubled the length of the test.

In general increasing the number of items increases the reliability of a scale because it reduces the impact that bad items have on the overall score (this is true as long as you are not adding more bad items).

$$r_{xx'} = \frac{nr_{oe}}{1 + r_{oe}(n-1)} \quad \text{Estimates reliability for a test that is } n \text{ times longer.}$$

$$n = \frac{r_{xx'}(1 - r_{oe})}{r_{oe}(1 - r_{xx'})} \quad \text{Estimates the number of times longer (} n \text{) a test will need to be, to reach a desired level of reliability (} r_{xx'} \text{).}$$

Scale Alpha: (Chronbach's Alpha) the average intercorrelation between all of the items in a scale. All items should be responded to consistently. Minimum of .70 (Robinson, Shaver, & Wrightsman, 1991).

This averaging method also makes clear what happens when you increase the number of items. It reduces the impact that weakly correlated items have on the overall average.

- Item Intercorrelations

$$r_{ij} = \frac{\Sigma (X_i - \bar{X}_i)(X_j - \bar{X}_j)}{\sqrt{\Sigma (X_i - \bar{X}_i)^2} \sqrt{\Sigma (X_j - \bar{X}_j)^2}}$$

- Conceptually Alpha = $\alpha = \frac{\Sigma r_{ij}}{K(K-1)/2}$ Where K = # of items

$$\alpha = \frac{K(\bar{r}_{ij})}{1 + (K-1)\bar{r}_{ij}} = \text{is more accurate (spss uses this)}$$

- If Scale alpha is .70 then adequate internal consistency.

E. Estimating a True Score (Assessment Application)

Since every score is composed of error and true score, to estimate a true score then you need to develop a confidence interval based on the estimate of the error.

-**Standard Error of the Measure** does this.

Since $r_{xx'}$ is the ratio of true score variance to total variance, then $1 - r_{xx'}$ is the ratio of error variance to total variance. Thus

$$\text{Standard Error of Measure} = s\sqrt{1 - r_{xx'}}$$

This will give us a confidence interval within which we would expect a true score to fall. Since we use 1 standard deviation in the equation then we can be 68.26% confident (confidence level) that the true score will fall within the interval.

We can set our confidence level by multiplying the standard deviation by the appropriate Z score.

$$\text{Standard Error of Measure}_{\alpha/2} = Z_{\alpha/2} \left(s \sqrt{1 - r_{ss'}} \right)$$

95% confidence limit $Z = 1.96$

99% confidence limit $Z = 2.58$

Confidence Interval = $X \pm \text{SEM}$, where X = a given score

F. Observational Measures/Open-ended Measure Reliability

Interrater Reliability

– Continuous Data = Correlation

– Nominal Data = Cohen's Kappa, Minimum of .70 (Bakeman and Gottman).

$$K = \frac{\sum O - \sum E}{N - \sum E} \text{ But only for the diagonal of the matrix.}$$

Also E for each cell on diagonal = $(\text{ROW}_{\text{tot}} * \text{COLUMN}_{\text{tot}}) / N$

– Ordinal Data = Rho or Tao