

# Classical Test Theory Reliability and Item Analysis, Part 1

- Today's topics:
  - Reviewing Basic Stats...
  - Fundamentals of Classical Test Theory
  - 3 Approaches to Reliability
    - Test-Retest
    - Alternative Forms
    - Internal Consistency (Alpha)
  - Wrapping Up...

# Back to Basic Statistics...

Given 2 continuous random variables, X & Y

- The expected **value** of each is its mean:
  - $E(X) = \mu_x$        $E(Y) = \mu_y$
- The expected **variance** of each is as follows:
  - $\text{Var}(X) = E[(X - \mu_x)^2]$        $\text{Var}(Y) = E[(Y - \mu_y)^2]$
- Their expected **covariance** is as follows:
  - $\text{Cov}(X, Y) = E[(X - \mu_x)(Y - \mu_y)]$

## Correlation and Covariance

TABLE 5-1 Computation of Pearson Product-Moment Correlation Coefficient

Pupil	Mathematics X	Reading Y	x	y	x <sup>2</sup>	y <sup>2</sup>	xy
Bill	41	17	+1	-4	1	16	-4
Carol	38	28	-2	+7	4	49	-14
Geoffrey	48	22	+8	+1	64	1	8
Ann	32	16	-8	-5	64	25	40
Bob	34	18	-6	-3	36	9	18
Jane	36	15	-4	-6	16	36	24
Ellen	41	24	+1	+3	1	9	3
Ruth	43	20	+3	-1	9	1	-3
Dick	47	23	+7	+2	49	4	14
Mary	40	27	0	+6	0	36	0
$\Sigma$	400	210	0	0	244	186	86
M	40	21					

$$SD_x = \sqrt{\frac{244}{10}} = \sqrt{24.40} = 4.94 \quad SD_y = \sqrt{\frac{186}{10}} = \sqrt{18.60} = 4.31$$

$$r_{xy} = \frac{\Sigma xy}{(N)(SD_x)(SD_y)} = \frac{86}{(10)(4.94)(4.31)} = \frac{86}{212.91} = .40$$

## Rules about Expected Values...

- Expected value of a constant is just the constant:
  - $E(c) = c$
- Expected value of a sum of a constant and a random variable is the mean plus the constant:
  - $E(X + c) = \mu_x + c$
- The variance of a sum of a constant and a random variable is just the variance of the random variable:
  - $\text{Var}(X + c) = \sigma_x^2 \rightarrow$  Adding a constant does not change the X variance
- Multiplication of a random variable by a constant:
  - $E(cX) = c\mu_x$
  - $\text{Var}(cX) = c^2 \text{Var}(X) = c^2\sigma_x^2 \rightarrow$  Pull out any constant and square it

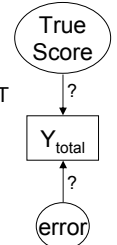
## More Relevant Rules...

- The expected value of a sum of random variables is the sum of their expected values:
  - $E(X+Y) = E(X) + E(Y) = \mu_x + \mu_y$
- The *variance of a sum* of random variables is given by the sum of ALL variances and covariances:
  - $Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$
  - Where does the '2' come from?
    - Covariance matrix is symmetric
    - Sum the whole thing to get to the *variance of the sum* of the items

	X	Y
X	$\sigma_x^2$	$\sigma_{xy}$
Y	$\sigma_{xy}$	$\sigma_y^2$

## Now, back to your regularly scheduled measurement class...

- In CTT, the **TEST** is the unit of analysis:  $Y_{total} = T + e$ 
  - True score T:**
    - Best estimate of 'latent trait': Mean over infinite replications
  - Error e:**
    - Expected value (mean) of 0, expected to be uncorrelated with T
    - e's are supposed to wash out over repeated observations
  - So the expected value of T is  $Y_{total}$**
  - In terms of observed test score variance:
    - Observed variance = true variance + error variance
- Goal is to quantify **reliability**
  - Reliability = true variance / (true variance + error variance)
  - Reliability calculation is conducted on sums across items (so type of item is not relevant), but will require assumptions about the items...



## Conceptualizing Reliability:

$$Y_{total} = \text{True Score} + \text{error}$$

- Wait a minute... if  $E(Y) = T$ ...
  - This idea refers to a single person's data... if a test is reliable, then a given person should get pretty much the same score over repeated replications...(except for random processes)
  - But we can't measure everybody a gazillion times...
  - So, we can conceptualize reliability as something that pertains to a sample of persons instead... by writing it in terms of variances
- $Var(Y) = Var(T) + Var(e)$ 
  - $= Var(T) + Var(e) + 2Cov(T, e)$
  - $= Var(T) + Var(e)$
- Reliability =  $Var(T) / Var(Y)$ 
  - Proportion of variance due to 'true score' out of total variance

## How do we get $Var(e)$ ?

### 3 main ways of quantifying reliability:

- Consistency of same test over time
  - Test-retest reliability
- Consistency over alternative test forms
  - Alternative forms reliability
  - Split-half reliability
- Consistency across items within a test
  - Internal consistency (alpha)

\*\* FYI: Some would say we have violated 'ergodicity' by quantifying reliability in this sample-based way...

## How Only Two Scores Give Us a Reliability Coefficient in CTT

- $Y_1 = T + e_1$
  - $Y_2 = T + e_2$
- CTT assumptions to calculate reliability:**
- Same true score (T) observed at both times
  - $e_1$  and  $e_2$  are uncorrelated with each other and T
  - $e_1$  and  $e_2$  have same variance
  - $Y_1$  and  $Y_2$  have same variance

$$r_{y_1, y_2} = \frac{\sigma_{y_1, y_2}}{\sigma_{y_1} \sigma_{y_2}} = \frac{\sigma_{t+e_1, t+e_2}}{\sigma_{y_1} \sigma_{y_2}} = \frac{\sigma_{t,t} + \sigma_{t,e_1} + \sigma_{t,e_2} + \sigma_{e_1, e_2}}{\sigma_{y_1} \sigma_{y_2}} = \frac{\sigma_t^2}{\sigma_y^2}$$

- Same as: Reliability of Y =  $\text{Var}(T) / \text{Var}(Y)$
- We express unobservable true score variance in terms of the correlation between the two total scores and the variance of the total scores (assumed to be the same across tests)
- We now have an index of how much of the observed variance is “true” (if we believe all the assumptions)

## 1. Test-Retest Reliability... What could go wrong?

- In a word, **CHANGE**: Test-retest reliability assumes that any difference in true score is due to measurement error
  - A characteristic of the test
  - It could be due to a characteristic of the person
- In a word, **MEMORY**: Assumes that testing procedure has no impact on a given person’s true score
  - Reactivity can lead to higher scores: learning, familiarity, memory...
  - Reactivity can lead to lower scores: fatigue, boredom...
- In a word (or two), **TEMPORAL INTERVAL**
  - Which test-retest correlation is the ‘right’ one?
  - Should vary as a function of time (longer intervals → smaller correlation)
  - Want enough distance to as to limit memory; not enough so as to observe change... how long is that, exactly?

## 2a. Alternative Forms Reliability

- Two forms of same test administered...
  - Different items on each, but still measuring same construct
  - Forms need to be ‘parallel’ – more about this later, but basically means no systematic differences between in the summary properties of the scales (means, variances, covariances, etc)
    - Responses should differ ONLY because of random fluctuation (e)
  - “Close” in time...
- Same exact logic... correlation between two forms is an index of reliability → or  $\text{Var}(T) / \text{Var}(Y)$

## 2b. Split-Half Reliability

- Don’t have two separate forms? No problem!
- Just take one test and split it in half! → Two ‘forms’
  - e.g., odd items =  $Y_1$ , even items =  $Y_2$
  - No problems with change or retest...
    - ...BUT – reliability is based on half as many items
- So let’s extrapolate what reliability would be with twice as many items... Use a reduced form of the Spearman Brown Prophecy Formula (more on this later)
  - $\text{Reliability}_{\text{new}} = 2 * \text{reliability}_{\text{old}} / 1 + \text{reliability}_{\text{old}}$
  - Example:  $r = .75$ ?  $\text{Reliability}_{\text{new}} = 2 * .75 / 1.75 = .86$

# Ta-da! More Reliability... What could go wrong?

## Alternative Forms Reliability:

- In a word, **PARALLEL**:
  - Have to believe forms are sufficiently parallel: both tests have same mean, same variance, same true scores and true score variance, same error variance... AND by extrapolation (more on this later), all items within each test and across tests have equivalent psychometric properties and same covariances and correlations between them
- In a word (or four), **SS,DD**:
  - Still susceptible to problems regarding change or retest effects

## Split-Half Reliability:

- In a word (or two), **WHICH HALF**: There are many possible splits that would yield different reliability estimates... (125 for 10 items)

# 3. Internal Consistency

- For quantitative items, this is Cronbach's Alpha...
  - Or 'Guttman-Cronbach alpha' (Guttman 1945 > Cronbach 1951)
  - Another version for binary items: KR 20 (more later on this)
- Alpha is described in multiple ways:
  - Is the mean of all possible split-half correlations
  - Is expected correlation with hypothetical alternative form of the same length
  - Is lower-bound estimate of reliability under assumption that all items are tau-equivalent (more about that later)
  - As an index of 'internal consistency'
    - Although Rod dislikes this term... everyone else uses it

# Prepping for Alpha...

- The **sum of the item variances** is given by:
  - $\text{Var}(I_1) + \text{Var}(I_2) + \text{Var}(I_3) \dots + \text{Var}(I_k) \rightarrow$  just the item variances
- The **variance of the sum of the items** is given by the sum of ALL variances and covariances:
  - $\text{Var}(I_1 + I_2 + I_3) = \text{Var}(I_1) + \text{Var}(I_2) + \text{Var}(I_3) \dots + 2\text{Cov}(I_1, I_2) + 2\text{Cov}(I_1, I_3) + 2\text{Cov}(I_2, I_3) \dots$
  - Where does the '2' come from?
    - Covariance matrix is symmetric
    - Sum the whole thing to get to the variance of the sum of the items

	$I_1$	$I_2$	$I_3$
$I_1$	$\sigma_1^2$	$\sigma_{12}$	$\sigma_{13}$
$I_2$	$\sigma_{21}$	$\sigma_2^2$	$\sigma_{23}$
$I_3$	$\sigma_{31}$	$\sigma_{32}$	$\sigma_3^2$

# Cronbach's Alpha

Covariance

Version: 
$$\alpha = \frac{k}{k-1} \cdot \frac{\text{variance of total Y} - \text{sum of item variances}}{\text{variance of total Y}}$$

$k = \# \text{ items}$

- Numerator reduces to just the covariance among items
  - **Sum of the item variances...**
    - $\text{Var}(X) + \text{Var}(Y) = \text{Var}(X) + \text{Var}(Y) \rightarrow$  just the item variances
  - **Variance of the sum of the items...**
    - $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y) \rightarrow$  **PLUS covariances**
  - So, if the items are related to each other, the variance of the total sum should be bigger than the sum of the item variances
    - How much bigger depends on how much covariance among the items – the primary index of relationship

# Cronbach's Alpha:

- **Alpha** is a lower-bound estimate of reliability under assumption that **all items are 'tau-equivalent'** → **equally related to true score**

Correlation Version: 
$$\alpha = \frac{k\bar{r}}{1 + \bar{r}(k-1)}$$
 Where  $\bar{r}$  is mean inter-item correlation  
 k = # items

- You'll note alpha depends on two things (k and r), and thus there are 2 potential ways to make alpha bigger...
  - Get more items and/or increase average inter-item correlation
- Potential problems:
  - But can you keep adding more items WITHOUT decreasing the average inter-item correlation???
  - Does not take into account spread of inter-item correlation, and thus **alpha does NOT assess dimensionality of the items**

# How to Get Alpha UP

TABLE 1  
Values of Cronbach's Alpha for Various Combinations of Different Number of Items and Different Average Interitem Correlations

Number of Items	Average Interitem Correlation					
	.0	.2	.4	.6	.8	1.0
2	.000	.333	.572	.750	.889	1.000
4	.000	.500	.727	.857	.941	1.000
6	.000	.600	.800	.900	.960	1.000
8	.000	.666	.842	.924	.970	1.000
10	.000	.714	.870	.938	.976	1.000

# Ta-da! Alpha as Reliability... What could go wrong?

- Alpha does not index **dimensionality** → it does not index the extent to which items measure the same construct

TABLE 13.2. Interitem Correlation Matrices for Two Hypothetical Tests with the Same Coefficient Alpha Reliability of .81

Test A with 10 items										Test B with 6 items							
Variable	1	2	3	4	5	6	7	8	9	10	Variable	1	2	3	4	5	6
1.	—										1.	—					
2.	.3	—									2.	.6	—				
3.	.3	.3	—								3.	.6	.6	—			
4.	.3	.3	.3	—							4.	.3	.3	.3	—		
5.	.3	.3	.3	.3	—						5.	.3	.3	.3	.6	—	
6.	.3	.3	.3	.3	.3	—					6.	.3	.3	.3	.6	.6	—
7.	.3	.3	.3	.3	.3	.3	—										
8.	.3	.3	.3	.3	.3	.3	.3	—									
9.	.3	.3	.3	.3	.3	.3	.3	.3	—								
10.	.3	.3	.3	.3	.3	.3	.3	.3	.3	—							

- The *variability* across the inter-item correlations matters, too!
- We will use item-based models (CFA, IRT) to examine dimensionality

# Wrapping Up...

- Reliability theoretically concerns the consistency of an individual's performance, but this gets operationalized in terms of variance decomposition for a sample
  - Proportion of true score (T) variance relative to total (Y) variance
    - Assuming true score measured exact same way over time with same test (retest) or with different sets of items (alternative forms)
  - Proportion of variance of sum of items relative to just the variance of the items (i.e., how much covariance there is)
    - Covariance will be an important part of all subsequent models
- As we will see in more detail later, these strategies of operationalizing reliability require assumptions that may not be plausible... but some of which will be testable