

# Classical Test Theory Reliability and Item Analysis, Part 2

- Today's topics:
  - Binary and Quantitative Item Properties
  - Psychometric Properties of Items in CTT
  - Revisiting Reliability and Assumptions about Items
    - 2-Score Reliability
    - Alpha and KR20
    - Applications and Miscellany
  - Wrapping Up...
- On Friday: Example in SPSS and SAS

# Item Psychometric Properties: Means and Variances by Item Type

- Means:
  - Quantitative item mean = sum items / n =  $\mu_y \rightarrow \bar{y}$
  - Binary item mean = number correct / n =  $\pi_y \rightarrow p_y$
- Variances:
  - Quantitative item:  $\text{Var}(Y) = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$
  - Binary item:  $\text{Var}(Y) = p_y(1-p_y) = p_yq_y = \sigma_y^2 \rightarrow s_y^2$ 
    - **Note that the variance is dependent on the mean ( $p_y$ )**

TABLE 3.2  
Binary Item Variance and Difficulty

p	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

# Item Psychometric Properties: Discrimination

- “**Discrimination**” = how related item is to latent trait
  - In latent trait models, it becomes some kind of factor loading (slope)
  - Is degree to which the item differentiates among persons in the latent construct (should be positive, and stronger is better)
  - **In CTT → Is correlation of the item with the total score** (or with the total minus that item, the item-remainder correlation)
- Choosing between item-total and item-remainder correlations:
  - Item-total correlation will be larger than item-remainder, but is potentially inflated (because the item is included in it)...
  - Item-remainder correlation is less biased than item-total, but then your ‘total’ is different for every item...
  - With enough items, it doesn't really matter

# Item Psychometric Properties: Difficulty\*

- ‘**Difficulty**’ = location on latent trait metric
  - In latent trait models, difficulty becomes some kind of intercept
  - **CTT item difficulty for binary items is p → proportion passing**
    - Variance of binary item =  $p(1-p) \rightarrow$  Variance depends on the mean
      - Thus, items with  $p = .50$  have the chance to be most ‘sensitive’ → they can show the most variance (which also helps with discrimination)
  - **CTT item difficulty for quantitative items is the item mean**
    - If 3+ response options *are used*, variance is not determined by the mean, but maximum variance is limited by k (# of response options)
    - So, a 5-option item would have max variance = 4  $\sigma_{\text{max}}^2 = \left(\frac{k-1}{2}\right)^2$
- \* As you've probably guessed by now, ‘**difficulty**’ is backwards (higher scores go with easier items)

## Reliability in Classical Test Theory

- Reliability is supposed to be about the consistency of an individual's score over replications... but it's not, really
- Instead, we get 2 scores per person (test-retest; alternate forms) or  $k$  items for person (alpha), and conceptualize reliability as 'consistency' across those
- $Y_{\text{Total}} = T + E$  or  $\text{Var}(Y_{\text{Total}}) = \text{Var}(T) + \text{Var}(E)$ 
  - True score is an internal characteristic of the person
    - True score variance is assumed to differ across samples
  - Error is an external characteristic (test + environment)
    - Error variance is assumed to be the same across samples
- Reliability =  $\text{Var}(T) / \text{Var}(Y)$ 
  - **Reliability is a characteristic of a sample, not a test**

## 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)

## Using CTT Reliability Coefficients: Back to the People

- Reliability coefficients are useful for describing the behavior of the test in the overall sample...  $\text{Var}(Y) = \text{Var}(T) + \text{Var}(e)$
- But reliability is a means to an end in interpreting a score for a given individual – we use it to get the error variance
  - $\text{Var}(T) = \text{Var}(Y) \times \text{reliability}$ ; so  $\text{Var}(e) = \text{Var}(Y) - \text{Var}(T)$
  - **95% CI for individual score =  $Y \pm 1.96 \times \text{SD}(e)$**
  - Gives an indication of how precise the true score estimate is on the metric of the original variable
  - Example:  $Y = 100, \text{Var}(e) = 9 \rightarrow 95\% \text{ CI} \approx 94 \text{ to } 106$   
 $Y = 100, \text{Var}(e) = 25 \rightarrow 95\% \text{ CI} \approx 90 \text{ to } 110$
  - Note this assumes a symmetric distribution, and thus will go out of bounds of the scale for extreme scores
  - Note this assumes the  $\text{SD}(e)$  or the **SE for each person is the same**

## Reliability of Difference Scores

- $Y_1 = T_1 + e_1$
  - $Y_2 = T_2 + e_2$
  - $D_T = T_2 - T_1$
  - $\text{Var}(D_T) = \text{Var}(T_2) + \text{Var}(T_1) - 2\text{cov}(T_2, T_1)$ 
    - $\text{Var}(D_T)$  is smallest when covariance is strong and positive
    - $\text{Var}(D_T)$  is largest when covariance is strong and negative
  - Reliability( $D_T$ ) =  $\text{Var}(D) - \text{Var}(E) / \text{Var}(D)$ 
    - Reliability is small when the covariance is positive
    - Reliability is large when the covariance is negative
  - Don't worry about this – just pick the most reliable  $Y_1$  and  $Y_2$  you can
- Assume errors are uncorrelated:  
 $\text{Var}(E_D) = \text{Var}(E_1) + \text{Var}(E_2)$
- $T_1$  and  $T_2$  are generally positively correlated

## Reliability in a Perfect World, Part 1

- Attenuation-corrected correlations
  - What would our correlation between two variables be if our measures were ‘perfectly reliable’?
  - $r_{\text{new}} = r_{\text{old}} * \text{SQRT}(\text{rel}_x * \text{rel}_y) \rightarrow$  all from same sample
  - For example:
    - Old x-y correlation = .38
    - Reliability<sub>x</sub> = .25
    - Reliability<sub>y</sub> = .55
    - New and “unattenuated” correlation = 1.03
  - Anyone see a problem here?

Psyc 948 Class 2b  
9 of 20

## Reliability in a Perfect World, Part 2

- What would my reliability be if I just added more items?
- Spearman-Brown Prophecy Formula
  - $\text{Reliability}_{\text{NEW}} = \text{ratio} * \text{rel}_{\text{old}} / [(\text{ratio}-1) * \text{rel}_{\text{old}} + 1]$ 
    - Ratio = ratio of new #items to old #items
  - For example:
    - Old reliability = .40
    - Ratio = 5 times as many items (had 10, what if we had 50)
    - New reliability = .77
- To use this formula, you must assume you have **PARALLEL** items
  - All discriminations equal, all error variances equal, all covariances and correlations among items equal, too

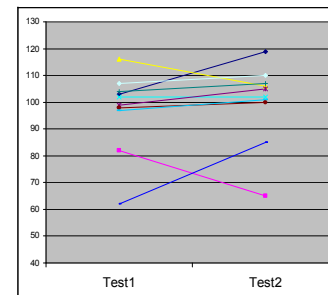
Psyc 948 Class 2b  
10 of 20

## Another Problem with Reliability

- Note that the formula for reliability is basically the Pearson correlation
  - Pearson r standardizes each variable, so that differences in mean and variance between variables don’t matter...
  - Pearson correlation indexes *relative*, not *absolute* agreement
- But the reliability formula assumes that the mean and variance of the true and observed scores are the same...
  - What if this is not the case?
  - Pearson correlation won’t pick this up!
  - A different kind of correlation is needed... **Intraclass correlation**
    - Note: There are LOTS of different versions of these... visit the McGraw & Wong (1996) paper for an overview

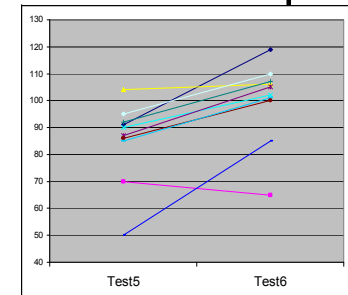
Psyc 948 Class 2b  
11 of 20

## Intraclass Correlation Example



M: 97            100  
SD: 15           15

Pearson  $r = .670$   
Intraclass (A,1)  $r = .679$



M: 85            100  
SD: 15           15

Pearson  $r = .670$   
Intraclass (A,1)  $r = .457$

Intraclass (A,1)  $r = \text{Var}(\text{people}) / [ \text{Var}(\text{people}) + \text{Var}(\text{tests}) + \text{Var}(\text{error}) ]$

Psyc 948 Class 2b  
12 of 20

# 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 = # 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'  $\rightarrow$  equally related to true score  $\rightarrow$  equal discrimination**

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**

# Kuder Richardson (KR) 20: Alpha for Binary Items

- KR20 is actually the more general form of alpha
  - From 'Equation 20' in 1937 paper:
    - $k = \# \text{ items}$
    - $p = \text{prop. passing}$
    - $q = \text{prop. failing}$
- $$KR20 = \frac{k}{k-1} \left( \frac{\text{variance of total Y} - \text{sum of } pq \text{ over items}}{\text{variance of total Y}} \right)$$
- Numerator again reduces to covariance among items...
    - **Sum of the item variances** (sum of pq) is just the item variances
    - **Variance of the sum of the items** has the covariance in it, too
    - 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

# Problems with Correlations Among Binary Items...

- In binary items, the variance is dependent on the mean
- If two items (X and Y) differ in p, such that  $p_y > p_x$  :
  - Maximum covariance:  $\text{Cov}(X, Y) = p_x(1-p_y)$
  - Maximum correlation will be smaller than -1 or 1:

$$r_{x,y} = \frac{p_x(1-p_y)}{\sqrt{p_y(1-p_x)}}$$

px	py	max r
0.1	0.2	0.67
0.1	0.5	0.33
0.1	0.8	0.17
0.5	0.6	0.82
0.5	0.7	0.65
0.5	0.9	0.33
0.6	0.7	0.80
0.6	0.8	0.61
0.6	0.9	0.41
0.7	0.8	0.76
0.7	0.9	0.51
0.8	0.9	0.67

- For Example:

## Some other kinds of 'correlations' you may have heard of before:

- **Pearson correlation:** between two quantitative variables, working with the distributions as they actually are
- **Phi correlation:** between two binary variables, still working with the observed distributions ( $\approx$  Pearson)
- **Point-biserial correlation:** between one binary variable and one quantitative variable, still working with the observed distributions (and still  $\approx$  Pearson)

---

*Line of Suspended Disbelief*

---

- **Tetrachoric correlation:** between 'pretend continuous' distributions of two actually binary variables (not  $\approx$  Pearson)
- **Biserial correlation:** between 'pretend continuous' (but really binary) and observed quantitative variables (still not  $\approx$  Pearson)
- **Polychoric correlation:** between 'pretend more continuous' distributions of two ordinal (quant-ish) variables (still not  $\approx$  Pearson)

Psyc 948 Class 2b  
17 of 20

## Reliability vs. Validity 'Paradox'

- Given the assumptions of CTT, it can be shown that the correlation between a test and an outside criterion cannot exceed the reliability of the test (see Lord & Novick 1968)
  - Reliability of .81? No observed correlations possible  $>$  .9, because that's all the 'true' variance there to be relatable!
  - In practice, this may be false because it assumes that the errors are uncorrelated with the criterion (and they could be)
- Selecting items with the strongest discriminations (or the strongest inter-correlations) can help to 'purify' or homogenize a test, but potentially at the expense of construct validity
  - Can end up with a 'bloated specific'
  - Items that are least inter-related may be most useful in keeping the construct well-defined and thus relatable to other things

Psyc 948 Class 2b  
18 of 20

## Summary: Assumptions about Items in CTT

- Use of alpha as an index of reliability requires an assumption of **tau-equivalent** items:
  - "True-score equivalence", or
  - Equal discrimination, or
  - Equal covariances among items
    - But not necessarily equal correlation...
- Use of the Spearman-Brown Prophecy formula requires an assumption of **parallel** items:
  - Tau-equivalence PLUS equal error variances
  - Translates into equal correlations among items, too

Psyc 948 Class 2b  
19 of 20

## Wrapping Up...

- **CTT unit of analysis is the WHOLE TEST:**  $Y_{total} = T + e$ 
  - Total score  $\rightarrow$  True Score (Latent Trait)
  - ASU measurement model (Add Stuff Up)
    - ASU model assumes unidimensionality – the only thing that matters is T
  - Assumes linear relationship between total score and latent trait
  - Reliability cannot be quantified without assumptions that range from somewhat plausible to downright ridiculous (testable in item-level models)
- **Item responses are not included:**
  - No means of explicitly testing dimensionality
  - Assumes all items are equally discriminating ("true-score-equivalent")
    - All items are equally related to the latent trait (also called "tau-equivalent")
  - To make a test better, you need more items
    - **What kind of items? More.**
  - Measurement error is assumed constant across the latent trait
    - **People low-medium-high in True Score are measured equally well**

Psyc 948 Class 2b  
20 of 20