Classical Test Theory (CTT) for Assessing Reliability and Validity

- Today's Class:
  - Hand-waving at CTT-based assessments of validity
  - CTT-based assessments of reliability
    - Why alpha doesn't really matter

2 Big Concerns about Scale Scores

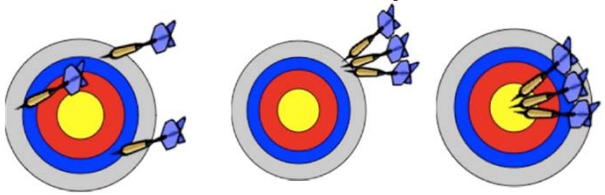
- **Reliability:**
  - "Extent to which the instrument does what it is supposed to with sufficient consistency for its intended usage"
  - "Extent to which same results would be obtained from the instrument after repeated trials"
  - Operationalized in several ways (stay tuned)…

- **Validity:**
  - "Extent to which the instrument measures what it is supposed to (i.e., it does what it is intended to do)" or "Validity for WHAT?"
  - Is measure of degree, and depends on USAGE or INFERENCES
    - Scales are not "valid" or "invalid" – validity is NOT a scale property
    - e.g., Test of intelligence: Measure IQ? Predict future income?

Another Way to Think About Reliability and Validity

Observed score = true score + error (Y = T + e)

- Error can be 'random'
  - Random error can be due to many sources (internal, external, instrument-specific issues, rater issues)
  - **Random error compromises reliability**
- Error can also be 'non-random'
  - Non-random error is due to constant source of variation that get measured consistently along with the construct (e.g., acquiescence)
  - **Non-random error compromises validity**
- In other words… reliability concerns how well you can hit the bulls-eye of the target…Validity concerns whether you hit the right target!



More about Validity

- The process of 'establishing' validity should be seen as building an argument:
  - To what extent can we use this instrument for its intended purpose (i.e., as a measure of construct X in this context)?

- Validity evidence can be gathered in two main ways:
  - Internal evidence
    - From construct map—does the empirical order of the items along the construct map match your expectations of their order?
    - From 'explanatory' item response models… if time permits
  - External evidence
    - Most of CTT is focused on this kind of evidence
    - This will be our focus for now…

Historical Classification of Types of Validity

- In 1954, the American Psychological Association (APA) issued a set of standards for validity, defining 4 types
  - Predictive, Concurrent, Content, Construct

- Cronbach and Meehl (1955) then expanded (admittedly unofficially) on the logic of construct validity
  - Predictive and concurrent → criterion-related (external)
  - Construct validity

- More recent versions (e.g., Messick, 1989)
  - Content, substantive, structural, generalizability, consequential, external…

Predictive and Concurrent Validity

- **Predictive** and **concurrent** validity are often categorized under 'criterion-related validity' (which makes it 3 kinds)
  - Predictive validity/utility: New scale relates to future criterion
  - Concurrent validity: New scale relates to simultaneous criterion

- **Criterion-related validity** implies there is some known comparison (e.g., scale, performance, behavior, group membership) that is immediately and undeniably relevant
  - e.g., Does newer, shorter test 'work as well' as older, longer test?
  - e.g., Do SAT scores predict college success?
  - This requirement limits the usefulness of this kind of evidence, however… why make a new scale if you already have one?

Content and Construct Validity

- **Content validity** concerns how well a scale covers the plausible universe of the construct…
  - e.g., Spelling ability of 4th graders—Are the words on this test representative of **all** the words they should know how to spell?
  - 'Face validity' is sometimes mentioned in this context (does the scale 'look like' it measures what it is supposed to?)

- **Construct validity** concerns the extent to which the scale score can be interpreted as a measure of the latent construct (and for that context, too)
  - Involved whenever construct is not easily operationally defined…
  - Required whenever a ready comparison criterion is lacking…
  - Requires a 'theoretical framework' to derive expectations from…

Construct Validity: 3 Steps for Inference

1. **Predict** relationships with related constructs
   - Convergent validity
     - Shows expected relationship (+/-) with other related constructs
     - Indicates "what it IS" (i.e., similar to, the opposite of…)
   - Divergent validity
     - Shows expected lack of relationship (0) with other constructs
     - Indicates "what it is NOT" (unrelated to…)
2. **Find** those relationships in your sample
   - No small task… especially if your sample is deliberately different
3. **Explain** why finding that relationship means you have shown something useful
   - Must argue based on 'theoretical framework'

3 Ways to Mess Up a Construct Validity Study…

1. Is your instrument broken?
   - Did you do your homework, pilot testing, etc?
   - Did you measure something reliably in the first place? Reliability precedes validity, or at least examination of it does
   - Is that something the right something (evidence for validity)?

2. Wrong theoretical framework or statistical approach?
   - Relationships really wouldn't be there in a perfect world
   - Or you have the wrong kind of sample given your measures
   - Or you lack statistical power or proper statistical analysis
     - Watch out for "discrepant" EFA-based studies…

The 3rd Way to Mess Up a Construct Validity Study…

3. Did you fool yourself into thinking that once the study (or studies) are over, that your scale "has validity"?

- **MEASURES ARE NEVER "VALIDATED"! Say "evidence for validity"**
- Are the items still temporally or culturally relevant?
- It is being used in the way that's intended, and is it working like it was supposed to in those cases?
- Has the theory of your construct evolved, such that you need to reconsider the dimensionality of your construct?
- Do the response anchors still apply?
- Can you make it shorter or adaptive to improve efficiency?

Summary: CTT Validity

- Reliability is a precursor to validity… coming up next

- CTT approaches to validity are largely external…
  - Depend on detecting expected relationships with other constructs, which can be found or not found for many other reasons besides problems with validity
  - This kind of externally-oriented validity is sometimes called "nomological span"
  - There is an alternative, more internal approach, too …
    - "Construct representation" via explanatory IRT models…

Review: Variances and Covariances

Variance:
Dispersion of y

$$\text{Variance}(y_i) = \frac{\sum_{s=1}^{N}(y_{is} - \bar{y}_i)^2}{N-1}$$

Covariance:
How y's go together, unstandardized

$$\text{Covariance}(y_1, y_2) = \frac{\sum_{s=1}^{N}(y_{1s} - \bar{y}_1)(y_{2s} - \bar{y}_2)}{N-2}$$

Correlation:
How y's go together, standardized (–1 to 1)

$$\text{Correlation}(y_1, y_2) = \frac{\text{Covariance}(y_1, y_2)}{\sqrt{\text{Variance}(y_1)} * \sqrt{\text{Variance}(y_2)}}$$

N = # people, s = subject, i = item

Means and Variances in **Binary** Items

- Binary item mean = number correct / # items = p
- Binary item variance = $p * (1 - p)$
  - **Note that the variance is dependent on the mean**

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

  - **This means that residual variance will not an estimated parameter when we analyze binary responses**

What Goes into the Sum of Items…

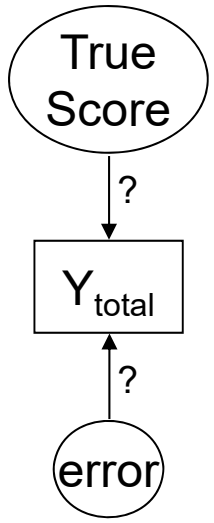
- The expected value of a sum of items is the sum of their expected values (which are just their means):
  - $E(y_1 + y_2) \rightarrow E(y_1) + E(y_2) \rightarrow \mu_{y1} + \mu_{y2}$

- The ***variance of a sum of items*** is given by the sum of all the item variances **AND the covariances** among them:
  - $Var(y_1 + y_2) = Var(y_1) + Var(y_2) + \mathbf{2}Cov(y_1, y_2)$
  - 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

	y_1	y_2
y_1	σ_{y12}	$\sigma_{y1,\,y2}$
y_2	$\sigma_{y1,\,y2}$	σ_{y22}

Now, back to your regularly scheduled measurement class…

- In CTT, the **TEST** is the unit of analysis: $\mathbf{Y_{total} = T + e}$
  - **True score T:**
    - Best estimate of 'latent trait': Mean over infinite replications
  - **Error e:**
    - Expected value (mean) of 0, by definition is uncorrelated with T
    - e's are supposed to wash out over repeated observations
  - **So the expected value of T is $\mathbf{Y_{total}}$**
  - 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} = True Score + 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 "error" 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

$$\begin{aligned} Var(Y) &= Var(T) + Var(e) \\ &= Var(T) + Var(e) + 2Cov(T, e) \\ &= Var(T) + Var(e) \end{aligned}$$

$$Reliability = Var(T) \,/\, Var(Y)$$

  - Proportion of variance due to "true score" out of total variance

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

$$r_{y1,y2} = \frac{\sigma_{y1,y2}}{\sigma_{y1}\sigma_{y2}} = \frac{\sigma_{t+e1,t+e2}}{\sigma_{y1}\sigma_{y2}} = \frac{\sigma_{t,t} + \sigma_{t,e1} + \sigma_{t,e2} + \sigma_{e1,e2}}{\sigma_{y1}\sigma_{y2}} = \frac{\sigma_t^2}{\sigma_y^2}$$

- Same as: $Reliability\ of\ Y = Var(True) / Var(Y)$
- 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 test variance is "true" (if we believe all the assumptions)

$Y = T + e$, so how do we get $Var(e)$?

3 main ways of quantifying **reliability**:

1. Consistency of same test over time
   - Test-retest reliability
2. Consistency over alternative test forms
   - Alternative forms reliability
   - Split-half reliability
3. Consistency across items within a test
   - Internal consistency (alpha or KR-20)

** FYI: Some would say we have violated "ergodicity" by quantifying reliability in this sample-based way…

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… ("close" in time)
  - 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)

- Same exact logic… correlation between two forms is an index of reliability → or Var(True) / 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)
  - $Reliability_{new} = 2 * Reliability_{old} \,/\, 1 + Reliability_{old}$
  - Example: $Reliability_{old} = .75?$ $Reliability_{new} = 2 * .75 \,/\, 1.75 = .86$

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
  - 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 equivalent form of alpha for binary items: KR 20

- Alpha is described in multiple ways:
  - Is the mean of all possible split-half correlations
  - As an index of "internal consistency"
    - Although Rod McDonald dislikes this term… everyone else uses it
  - Is lower-bound estimate of reliability under assumptions that:
    - All items are unidimensional → measure a single latent trait
    - All items are **tau-equivalent** → equally related to the true score
    - Item **errors are uncorrelated** (can be biased low or high if correlated)

Where Alpha Comes From

- The **sum of the item variances** is given by:
  - $Var(I_1) + Var(I_2) + Var(I_3) \ldots . + Var(I_k)$ → just the item variances

- The **variance of the sum of the items (total score)** is given by the sum of ALL the item variances and covariances:
  - $Var(I_1 + I_2 + I_3) = Var(I_1) + Var(I_2) + Var(I_3) \ldots + 2Cov(I_1, I_2) + 2Cov(I_1, I_3) + 2Cov(I_2, I_3) \ldots$
  - 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	${\sigma_1}^2$	σ_{12}	σ_{13}
I_2	σ_{21}	${\sigma_2}^2$	σ_{23}
I_3	σ_{31}	σ_{32}	${\sigma_3}^2$

Cronbach's Alpha

Covariance Version:
k = # items

$$\text{alpha} = \frac{k}{k-1} \bullet \frac{\text{variance of total} - \text{sum of item variances}}{\text{variance of total}}$$

Numerator reduces to just the covariance among items

Sum of the item variances…

$Var(I_1) + Var(I_2) \ldots = Var(I_1) + Var(I_2)$ → just the item variances

Variance of the sum of the items (total score)…

$Var(I_1 + I_2 \ldots) = Var(I_1) + Var(I_2) + \mathbf{2Cov(I_1, I_2)}$

PLUS covariances

- So, if the items are related to each other, the variance of the sum of the items 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** reliability assumes that all items are **<u>unidimensional</u>, <u>tau-equivalent,</u> and have <u>uncorrelated errors</u>**

Correlation Version:
k = # items

$$\text{alpha} = \frac{k\bar{r}}{1 + \bar{r}\,(k-1)}$$

Where $\bar{r}$ is mean inter-item correlation

- You'll note alpha depends on two things (k and r), and thus there are 2 potential ways to make alpha bigger…
  - (1) Get more items, (2) increase the 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 the spread of the inter-item correlations, and thus **alpha does NOT assess dimensionality of the items**

How to Get Alpha UP

$$\text{alpha} = \frac{k\bar{r}}{1 + \bar{r}\,(k-1)}$$

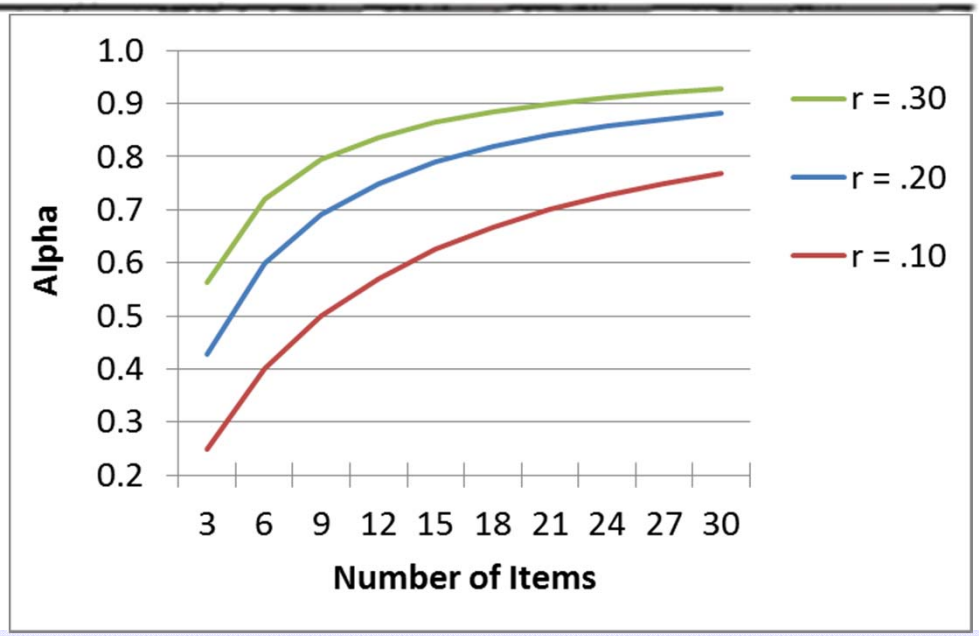
TABLE 1

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

$$\bar{r} = \frac{\text{alpha}}{k - (\text{alpha}*k) + \text{alpha}}$$

For the 2016 GRE psychology subject test,
alpha = .96…
for about 205 items,
which means **$\bar{r}$= .10**



Alpha: What could go wrong?

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

TABLE 18.2. Interitem Correlation Matrices for Two Hypothetical Tests with the Same Coefficient Alpha Reliability of .81

Test A with 10 items

Variable	1	2	3	4	5	6	7	8	9	10
1	–									
2	.3	–								
3	.3	.3	–							
4	.3	.3	.3	–						
5	.3	.3	.3	.3	–					
6	.3	.3	.3	.3	.3	–				
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	–

Test B with 6 items

Variable	1	2	3	4	5	6
1	–					
2	.6	–				
3	.6	.6	–			
4	.3	.3	.3	–		
5	.3	.3	.3	.6	–	
6	.3	.3	.3	.6	.6	–

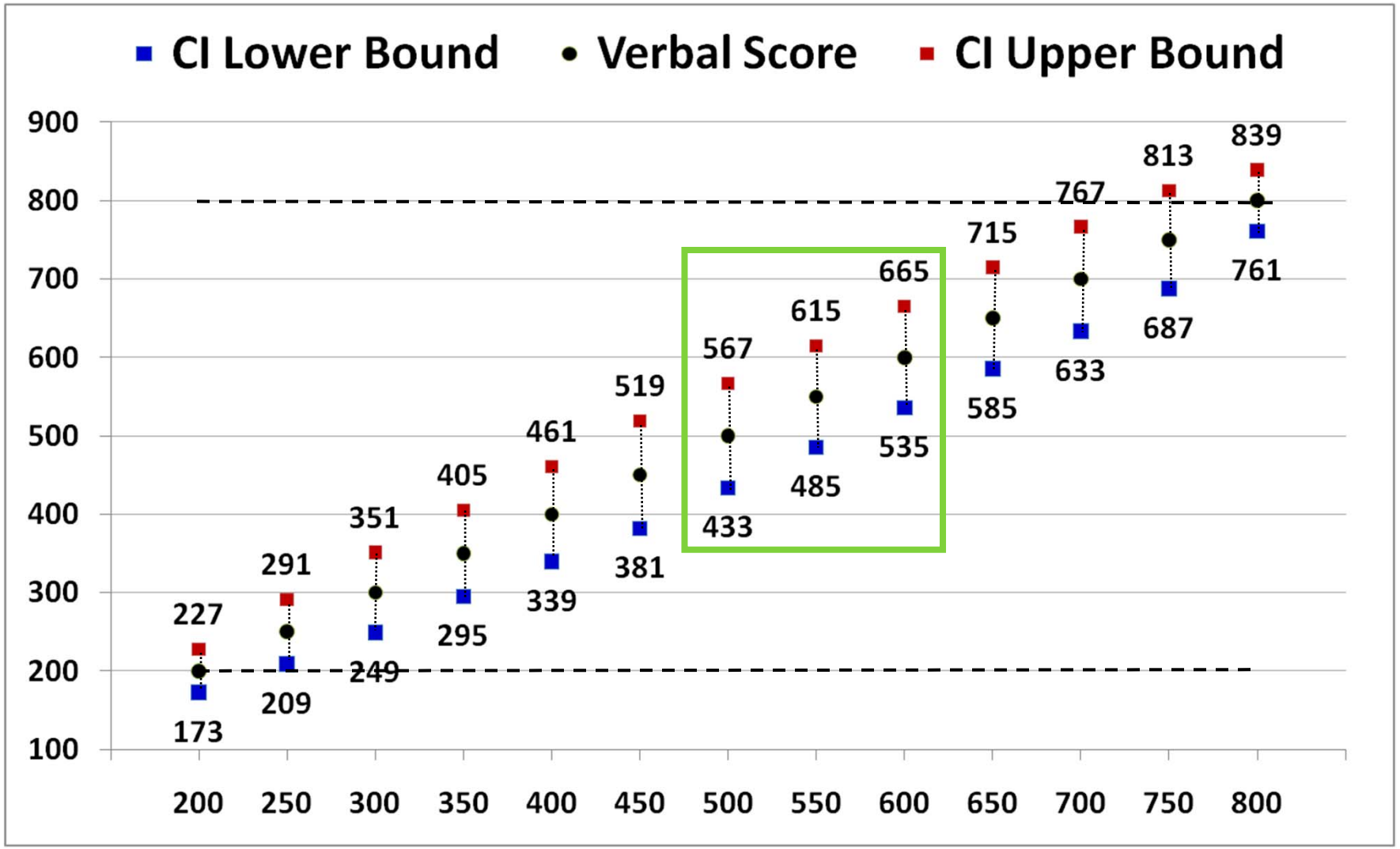
- The *variability* across the inter-item correlations matters, too!
- We will use item-based models (CFA, IRT) to examine dimensionality

Using CTT Reliability Coefficients: Back to the People

- Reliability coefficients like alpha are useful for describing the test behavior in the overall sample… Var(Y) = Var(T) + 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
  - Var(T) = Var(Y)*reliability; so Var(e) = Var(Y) – Var(T)
  - **95% CI for individual score = Y ± 1.96*SD(e)**
  - Precision of true score estimate in the metric of the original variable
  - Example: Y = 100, Var(e) = 9 → 95% CI ≈ 94 to 106
    Y = 100, Var(e) = 25 → 95% CI ≈ 90 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 SD(e) or the **SE for each person is the same**
  - *Cue mind-blowing GRE example*

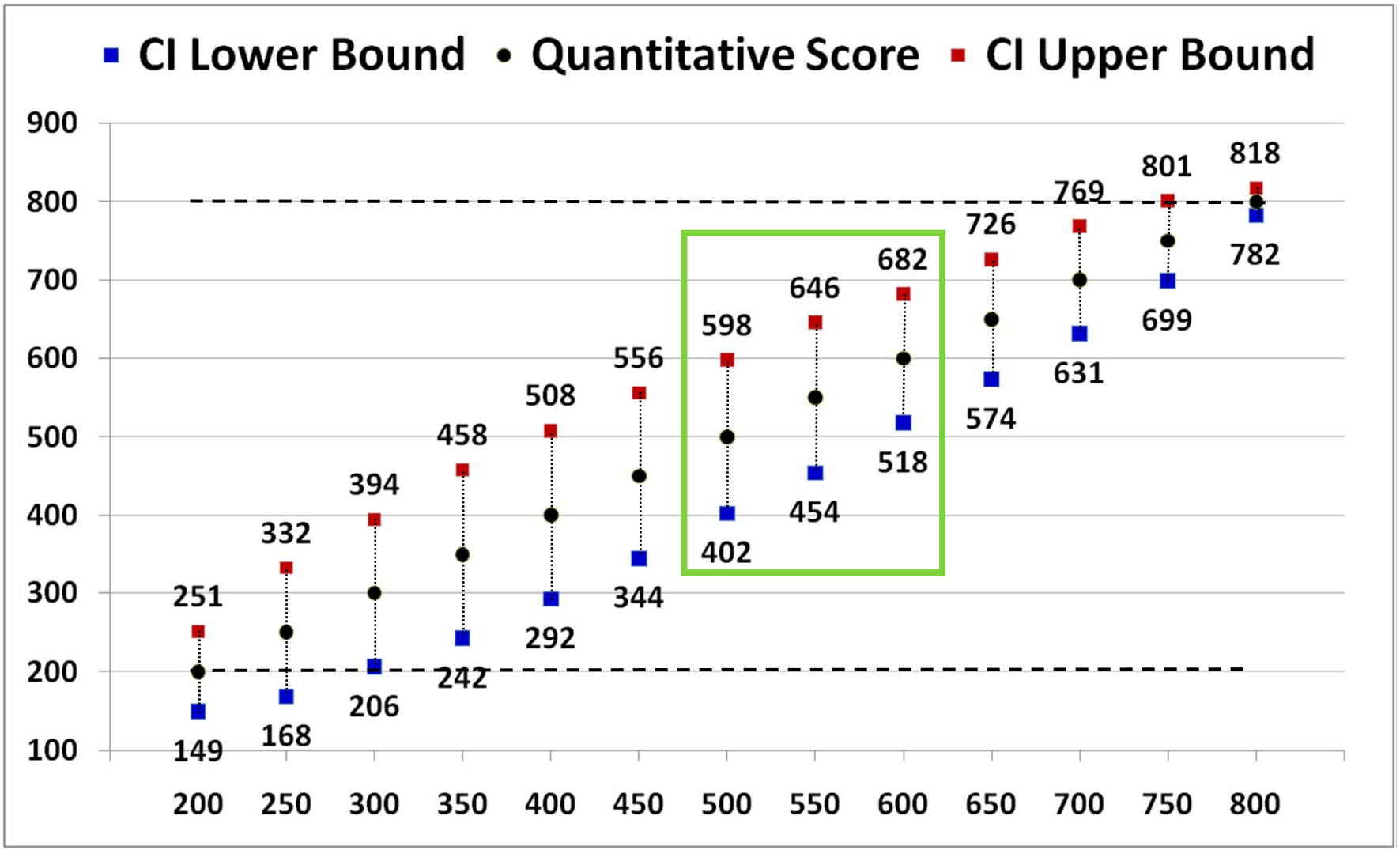
95% Confidence Intervals: Verbal

SEM ranges from 14 to 35



95% Confidence Intervals: Quantitative

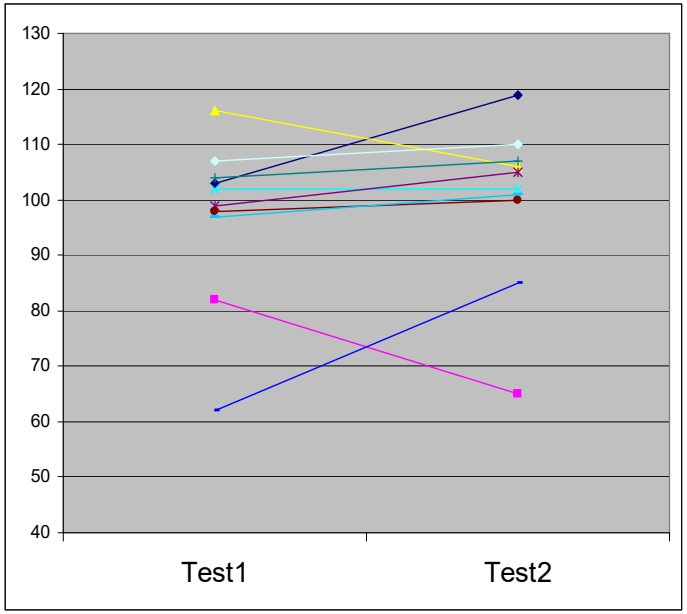
SEM ranges from 9 to 55

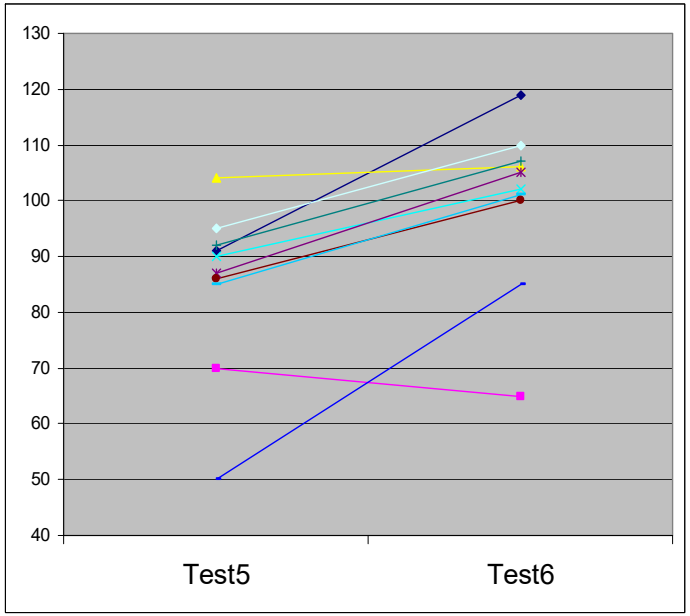


Another Problem with Reliability

- Note that the formula for reliability is the Pearson correlation
  - Pearson *r* standardizes each variable, so that differences in mean and variance between variables don't matter…
  - So 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

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 = Var(people) / [Var(people) + **Var(tests)** + Var(error)]

Kuder Richardson (KR) 20: Alpha for Binary Items

- KR20 is actually the more general form of alpha
- From 'Equation 20' in 1937 paper:

$$\mathrm{KR}20 = \frac{k}{k-1}\left(\frac{\text{variance of total Y} - \text{sum of } pq \text{ over items}}{\text{variance of total Y}}\right)$$

k = # items
p = prop. passing
q = prop. failing

- 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 sum of the items should be bigger than the sum of the item variances
    - How much bigger again depends on how much covariance among the items – the primary index of relationship

Problems with Reliability for 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: $Cov(X,Y) = p_x(1-p_y)$
  - **Maximum correlation will be smaller than -1 or 1:**

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

  - For Example:

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

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 (≈ Pearson)
- **Point-biserial correlation**: between one binary and one quantitative variable, still working with the observed distributions (and still ≈ Pearson)

Line of Suspended Disbelief

- **Tetrachoric correlation**: between 'underlying continuous' distributions of two actually binary variables (not ≈ Pearson); aka, between probits
- **Biserial correlation**: between 'underlying continuous' (but really binary) and observed quantitative variables (not ≈ Pearson); aka, between probits
- **Polychoric correlation**: between 'underlying continuous' distributions of two ordinal (quant-ish) variables (not ≈ Pearson); aka, between probits

Summary: Reliability in CTT

- 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 do:

- **$Y_{Total} = T + E$ or $Var(Y_{Total}) = Var(True) + Var(Error)$**
  - **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 is a characteristic of a sample, not of a test***

- Want to improve reliability? Examine the items…
  - Because individual items are not in the CTT measurement model, we have to make assumptions about them instead

Item 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 = $p(1-p)$ → 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

$$\sigma^2_{\max} = \left(\frac{k-1}{2}\right)^2$$

  - Difficulty is usually ignored in CTT (other than as a cause of range restriction, which then limits the item's relationship to the trait)

* ***Difficulty* is backwards** (higher scores go with easier items)

Item 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)
  - Discrimination is always considered in evaluating items across models

- 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

Reliability in a Perfect World, Part 1

- What would my reliability be if I just added more items?
- Spearman-Brown Prophesy Formula
  - $Reliability_{NEW} = ratio*reliability_{old} / [(ratio-1)*reliability_{old} + 1]$
    - 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 **PARALLEL** items
  - All discriminations equal, all error variances equal, all covariances and correlations among items equal, too

Assumptions about Items When Calculating Reliability 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…(because of different error variances)

- Use of the Spearman-Brown Prophesy formula requires an assumption of **parallel** items:
  - Tau-equivalence PLUS equal error variances
  - So translates into equal correlations among items, too

Reliability in a Perfect World, Part 2

- Attenuation-corrected correlations
  - What would our correlation between two variables be if our measures were 'perfectly reliable'?
  - $r_{new} = r_{old}$* SQRT(rel_x*rel_y) → all from same sample
  - For example:
    - Old x-y correlation = .38
    - $Reliability_x$ = .25
    - $Reliability_y$ = .55
    - New and "unattenuated" correlation = 1.03
  - Anyone see a problem here?
    - Btw—this logic forms the basis of SEM ☺

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 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 conceptually and thus relatable to other things

Wrapping Up…

- **CTT unit of analysis is the WHOLE TEST: Y_{total} = True + error**
  - Total score → 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, which means:**
  - No way 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**