Classical Test Theory (CTT) for Assessing Reliability (and Validity)

- Topics:
  - Review of concepts and summary statistics
  - Characterizing differences between indicators
  - CTT-based assessments of reliability
    - Why alpha doesn’t really matter
    - Why standard errors of measurement should matter
  - Hand-waving at validity

Review: What are we trying to do?

- Measure a **latent trait**: unobservable ability, characteristic, attitude, or other type of individually-varying construct
  - "**Latent**" = Not directly observable
  - "**Trait**" = true score, factor score, or theta as predictor(s) in measurement models; *aka*, latent construct, variable, or factor
  - The LTMMs we will cover are for **continuous latent traits**

- How to measure a latent trait? Collect **observed responses from indicators** chosen to measure the latent trait
  - "**Indicator**" = item, trial, or other response-specific outcome
  - Indicators can be any kind of variable (categorical or quantitative)

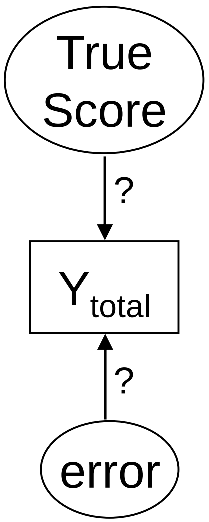
- How do we know we've done good job measuring the trait? **Collect evidence using the indicator responses**…
  - Two distinct ways such evidence gets used to represent a trait:
    - Build a **composite** (sum or average across indicator responses) → CTT
    - Use **all indicator responses** as outcomes of latent trait predictor instead: this is what happens in latent trait measurement models (LTMMs)

Big Picture of Instrument Development

- Primary concerns about the use of an instrument to measure one or more latent traits have a **hierarchical structure**:
  - **Validity**: Extent to which an instrument measures what it is supposed to
    - Validity is always a matter of degree and depends critically on how it is used
    - Almost always demonstrated by **external evidence**: relationships to measures of other constructs in expected directions (e.g., discriminant and convergent validity)
  - An essential **precursor** to validity is **reliability**: Extent to which an instrument measures a latent trait with **sufficient consistency** (i.e., extent to which the same result would be obtained repeatedly)
    - "Validity is measuring the right thing; reliability is measuring the thing right"
    - Reliability indices will be provided differently across CTT and LTMMs (stay tuned)
  - An important **precursor** to reliability is **dimensionality**: Accuracy of the mapping of the observed indicators to the latent traits they measure
    - Reliability is per trait! Most reliability indices assume **unidimensional traits**
    - **What follows in this lecture presupposes that dimensionality is KNOWN!**

Classical Test Theory (CTT)

- The **TOTAL** is the unit of analysis: $\boldsymbol{Y_{total} = True + Error}$
  - **True score $\boldsymbol{T}$:**
    - Best estimate of latent trait is **mean over infinite replications**
  - **Error $\boldsymbol{e}$:**
    - Expected value (mean) of 0; theoretically uncorrelated with T
    - Errors are supposed to wash out over repeated observations
  - **So the expected value of $\boldsymbol{T}$ is $\boldsymbol{Y_{total}}$**
    - This translates into $\boldsymbol{Y_{total}}$ = $\boldsymbol{T}$ true-score in practice
    - $\boldsymbol{Y_{total}}$ is referred to as a **total-score**, test-score, or scale-score



- Provides a framework with which to quantify ***reliability***
  - What proportion of **total-score** variance is due to **true-score** variance?
  - Understanding parts of CTT logic for quantifying reliability relies on traditional univariate and bivariate **summary statistics** for indicators…

Means, Variances, Covariances, and Correlations

Using population notation: N = # subjects, s = subject, i = item for y_{is}

(Arithmetic) Mean (μ):
Central tendency of y_{is}

$$\mu_i = \frac{\sum_{s=1}^{N} y_{is}}{N}$$

Variance (Var):
Dispersion of y_{is}
in squared units

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

Covariance (Cov):
How outcomes (e.g., y_{1s} and y_{2s}) go together in original metrics (unstandardized)

$$Cov(y_1, y_2) = \sigma_{y_1,y_2} = \frac{\sum_{i=1}^{N} [(y_{1s} - \bar{y}_1)(y_{2s} - \bar{y}_2)]}{N}$$

Pearson Correlation (r):
Covariance that has been standardized: -1 to 1

$$r(y_1, y_2) = \frac{Cov(y_1, y_2)}{\sqrt{Var(y_1)}\sqrt{Var(y_2)}}$$

What about Categorical Indicators?

- Computing means, variances, covariances, and correlations is standard and intuitive for **quantitative indicators**
  - When the numbers are actually numbers (interval measurement)
  - e.g., magnitude estimation slider bars, response times
- But observed indicators are **more often categorical**:
  - Binary (i.e., dichotomous) → 2 options
  - Ordinal (i.e., "Likert scale") → 3+ ordered options
  - Nominal (i.e., multinomial) → 3+ unordered options
- For **nominal indicators**, means and variances make no sense…
  - Frequency of each category is needed instead (stay tuned)
  - But what about summarizing binary or ordinal indicators?

Binary and Ordinal Indicators

- For **binary indicators (y_{is} coded 0 or 1)**, variance is not a separate property (as it is in quantitative indicators)
  - If $\boldsymbol{p_i}$ = proportion of 1 values, and $\boldsymbol{q_i}$ = proportion of 0 values:
  - Mean $\boldsymbol{\mu_i = p_i}$, $Var(\boldsymbol{y_i}) = \boldsymbol{p_i * q_i}$ (same result even if computed as usual)

Mean and Variance of a Binary Variable

Mean ($\boldsymbol{p_i}$)	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

- For **ordinal indicators**, you may see means and variances calculated as usual, but they should give you pause…
  - e.g., 1=Strongly Disagree, 2=Disagree, 3=Neutral, 4=Agree, 5=Strongly Agree…. could equally be 1, 20, 300, 4000, 50000
  - Maximum variance is limited by k = # of response options used

$$Var_{max}(y_i) = \frac{(k-1)^2}{2}$$

Differences Between Indicators

- All indicators can be characterized by **two properties** with respect to how they map onto the latent trait that they measure: **item difficulty** and **item discrimination**
  - Item = indicator, but the term "item" always used in this context
  - Properties will be indexed differently across CTT and LTMMs

- **Item difficulty** is the indicator's **location** on the metric of the latent trait; also known as item "**severity**" for non-ability traits
  - i.e., an item of difficulty level X measures people at trait level X well
  - So to measure people with a range of trait levels accurately, you need to include indicators that have a corresponding range of item difficulty

- **Item discrimination** is how **strongly the indicator relates** to the trait ("discrimination" is used for ability or non-ability traits)
  - Is the degree to which the **indicator differentiates among persons** in their latent traits (should be positive, and stronger is always better)

Difficulty and Discrimination in CTT

- Under the belief that the best estimate of the **latent trait is the total-score** across indicators (i.e., sum or average) in CTT:

- **Item difficulty/severity** (location on the latent trait) is the **indicator's mean** across respondents
  - Only applicable to binary or quantitative items; also ordinal if you believe in the numbers (which is usually what people do in CTT)
  - Note that the difficulty terminology is conceptually backwards: An item with a higher mean is labeled as "higher difficulty" even though more people did well than not (so items with higher means are actually easier)
  - In LTMMs, difficulty/severity will become some kind of model intercept (which will break the problematic tie of respondents to indicators)

- Item difficulty/severity is often ignored in evaluating items in CTT, except when it causes problems with discrimination…

Difficulty and Discrimination in CTT

- Under the belief that the best estimate of the **latent trait is the total-score** across indicators (i.e., sum or average) in CTT:

- **Item discrimination** (relationship to the latent trait) is the **indicator's Pearson correlation** with the total-score
  - Called "**item–total**" correlation; often replaced with "**item–remainder**" correlation (i.e., total without that item) so the correlation isn't inflated
  - Only applicable to binary or quantitative items; also ordinal if you believe in the numbers (which is usually what people do in CTT)
  - In LTMMs, discrimination will become some kind of model slope

- Items of **extreme difficulty/severity** have a restricted range, which may result in **smaller item–total correlations**
  - Following common advice to remove extreme items will reduce your ability to measure respondents of corresponding extreme trait levels!

Reliability of CTT Total-Scores

- Before and after screening/selecting items (i.e., an iterative process), a **total-score** is created: a sum or mean across indicator responses
  - The **total-score** is now the unit of analysis: $\boldsymbol{Y_{total} = True + Error}$
  - Even though the total-score doesn't know what kind of indicators were used to create it, the **total-score is always treated as a quantitative variable** (i.e., "ordinal-treated-as-interval")
- Then need to quantify **reliability**: the **consistency** with which Y_{total} measures $True$ for a given respondent (i.e., subject)
  - Best index of T for each subject is supposed to be the mean Y_{total} over infinite replications… but that's not the kind of data usually collected!
  - Instead of *multiple replications* of total-score for a *single respondent*, more often collected are *single total-scores* for *multiple respondents*!
  - So reliability is instead defined using **between-subject sources** of respondent variance: $Reliability = Var(True) \,/\, Var(Y_{total})$
    - But to quantify reliability, we need more than one Y_{total} per subject…

How Only **Two Total-Scores** Can Yield a Reliability Coefficient in CTT

- $y_{1s} = T_s + e_{1s}$
- $y_{2s} = T_s + e_{2s}$

CTT assumptions to calculate reliability:

- Errors e_{1s} and e_{2s} have equal variance
- Total-scores y_{1s} and y_{2s} have equal variance
- Same subject-specific true score (T_s) at both times
- e_{1s} and e_{2s} are uncorrelated with each other and T_s

- Pearson Correlation between total-scores:
  - $r(y_1, y_2) = \frac{\sigma_{y_1,y_2}}{\sigma_{y_1}\sigma_{y_2}} = \frac{\sigma_{T+e_1,T+e_2}}{\sigma_{y_1}\sigma_{y_2}} = \frac{\sigma_{T,T} + \sigma_{T,e_1} + \sigma_{T,e_2} + \sigma_{e_1,e_2}}{\sigma_{y_1}\sigma_{y_2}} = \frac{\sigma_T^2}{\sigma_y^2}$

- In other words: $r(y_1, y_2) = Reliability = Var(True) \,/\, Var(Y_{total})$
  - So the Pearson correlation of two total-scores indexes how much of the observed total-score variance is due to "true" between-subject differences (if we believe all these untested assumptions)

3 Ways of Quantifying Reliability

- After measuring variance across subjects* two ways:
  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 between-subjects way:
  - What factors cause differences between respondents is not the same as what factors causes differences within a respondent over occasions...

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
  - 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 subject's true-score, although:
  - 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)
  - Long enough to limit memory, but short enough to avoid real change… how long is that, exactly????

2. Alternative Forms or Split-Half Reliability

- **Two forms of same test** administered "close" in time
  - Different indicators on each, but still measuring same construct
  - Forms need to be "**parallel**" – this means no systematic differences between in the summary statistics of the total-scores
    - Responses should differ ONLY because of random fluctuation (error)

- OR just take one test and **split it in half**! → Ta-da, two forms!
  - e.g., odd indicators = y_{1s}, even indicators = y_{2s}
  - BUT reliability is now based on half as many indicators!
  - What if we could **extrapolate** what reliability would be with twice as many indicators… Can do so using a reduced form of the "Spearman Brown Prophecy Formula" (assuming parallel indicators; stay tuned)
    - $Reliability_{new} = 2 * Reliability_{old} \, / \, (1 \, + Reliability_{old})$
    - e.g., $Reliability_{old} = .75? \quad Reliability_{new} = 2 * .75 \, / \, 1.75 = .86$

More about Two Total-Score Reliability… What could go wrong?

Alternative Forms Reliability:

- In a word, **PARALLEL**:
  - Have to believe forms are sufficiently parallel: both total-scores have same mean, same variance, same true-scores and true-score variance, same error variance…
  - AND by extrapolation, all indicators within each test and across tests have equivalent psychometric properties and same correlations among them
  - Otherwise, indicator differences could create total-score differences
  - Still susceptible to problems caused by reactivity (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… (e.g.,125 splits for 10 indicators)

3. Internal Consistency Reliability

- For quantitative indicators, this is usually **Cronbach's Alpha**…
  - Or 'Guttman-Cronbach alpha' (Guttman 1945 < Cronbach 1951)
  - Equivalent form of alpha for binary items is named "KR 20"
- Alpha has been described in multiple ways:
  - Is the mean of all possible split-half correlations
  - As an index of "internal consistency"
    - Although Rod McDonald disliked this term… everyone else uses it
- Alpha is a lower-bound estimate of reliability under assumptions that all indicators:
  - Are **unidimensional** → measure a single latent trait
  - Are **tau-equivalent → "true-score equivalent"**→ equal item discrimination → equally related to the true score
  - Have **uncorrelated errors** (can be biased low or high if not)

Where Cronbach's Alpha comes from…

- The **sum of the I indicator variances** (e.g., $I = 3$ here):
  - $\sum_{i=1}^{I} Var(y_i) = Var(y_1) + Var(y_2) + Var(y_3)$ → only the variances
  - Will become a baseline for expected amount of total-score variation

- ***Variance of the I indicators' total-score*** is given by the sum the indicators' variances **PLUS their covariances**:
  - $$Var(Y_{total}) = Var(y_1) + Var(y_2) + Var(y_3) + \mathbf{2}Cov(y_1, y_2) + \mathbf{2}Cov(y_1, y_3) + \mathbf{2}Cov(y_2, y_3)$$
  - Where does the **2** come from?
    - Covariance matrix is symmetric
    - Sum the whole thing to get to the *variance of the sum* of the indicators
  - So should be greater than sum of indicator variances above if they have something in common → covariance

	y_1	y_2	y_3
y_1	$\sigma_{y_1}^2$	σ_{y_1,y_2}	σ_{y_1,y_3}
y_2	σ_{y_1,y_2}	$\sigma_{y_2}^2$	σ_{y_2,y_3}
y_3	σ_{y_1,y_3}	σ_{y_2,y_3}	$\sigma_{y_3}^2$

Cronbach's Alpha: It's not what you think.

- $\boldsymbol{alpha}\ (\alpha) = \frac{I}{I-1} * \frac{Var(Y_{total}) - \sum_{i=1}^{I} Var(y_i)}{Var(Y_{total})}$

  I = # items

  - Numerator reduces to the indicator covariances → if the indicators are related, the variance of the indicators' total-score, $Var(y_{total})$, should be bigger than the sum of the indicator variances, $\sum_{i=1}^{I} Var(y_i)$

- Easier way: $\boldsymbol{alpha}(\alpha) = \frac{I\bar{r}}{1 + [\bar{r}(I-1)]}$

  $\bar{r}$ = average inter-indicator Pearson correlation

  - Two ways to make alpha bigger: (1) Get more items, (2) increase the average inter-indicator correlation (but its's hard to do both at once)

- **Alpha** reliability assumes that all items are **unidimensional**
  - Formula does not take into account the spread of the inter-indicator correlations → **so alpha does NOT assess indicator dimensionality!**
- **Alpha** reliability assumes items have **equal discrimination** (tau-equivalent; equal relation to latent trait) with **uncorrelated errors**
  - Item properties are not included in the formula → exchangeable

Alpha: What could go wrong?

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

478 OLIVER P. JOHN AND VERONICA BENET-MARTÍNEZ

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-indicator correlations matters, too!
- We will use LTMMs predicting indicator responses to examine dimensionality

Example from: John, O. P., & Benet-Martinez, V. (2014). Measurement: Reliability, construct validation, and scale construction. In H. T. Reis & C. M. Judd (Eds.), *Handbook of research methods in social and personality psychology* (pp. 3473-503, 2nd ed.). New York, NY: Cambridge University Press.

How to Get Alpha UP: More Items!

Given indicator $\bar{r}$,

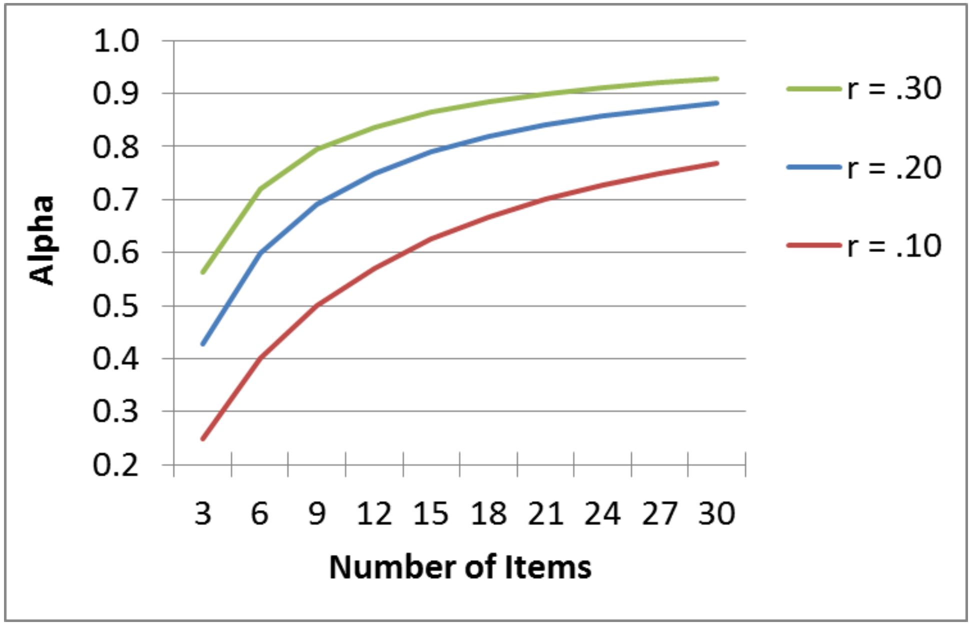
$$\boldsymbol{alpha} = \frac{I\bar{r}}{1 + [\bar{r}(I-1)]}$$

Given alpha (α),

$$\boldsymbol{\bar{r}} = \frac{\alpha}{I - (\alpha I) + \alpha}$$

Number of Indicators I	Average Indicator $\bar{r}$			
	.2	.4	.6	.8
2	.333	.572	.750	.889
4	.500	.727	.857	.941
6	.600	.800	.900	.960
8	.666	.842	.924	.970
10	.714	.879	.938	.976

Btw: For the 2020 GRE psychology subject test, (KR-20) **alpha = .95**… for about 205 items, this means $\boldsymbol{\bar{r} = .084!}$



Kuder Richardson (KR) 20: Alpha for Binary Items

- From 'Equation 20' in 1937 paper:

$$\mathrm{KR20} = \frac{k}{k-1}\left(\frac{\text{variance of total Y} - \text{sum of } pq \text{ over items}}{\text{variance of total Y}}\right)$$

k = # items (I before)
p = proportion of 1s
q = proportion of 0s

- Numerator again reduces to covariance among items…
  - **Sum of the indicator variances** (sum over indicators of pq) is just the variances
  - **Variance of the indicators' total-score** has their covariances in it, too
  - Numerator reduces to the indicator covariances → if the indicators are related, the variance of the sum of the indicators $Var(y_{total})$ should be bigger than the sum of the indicator variances $\sum_{i=1}^{I} Var(y_i)$
  - So KR20 is the same thing as alpha (it's just a computational shortcut)
  - Btw, this is how reliability is computed for the GRE subtests …

Kuder, G. F., & Richardson, M.W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Limited Reliability of Binary Indicators

- The possible **Pearson's r for binary variables will be limited** when they are not evenly split into 0/1 because their variance depends on their mean
  - Remember: Mean = p_i , Variance = $p_i(1-p_i) = p_i q_i$
- If two indicators (x and y) differ in p_i, such that $p_y > p_x$
  - Maximum covariance: $Cov(x, y) = p_x(1-p_y)$
  - This problem is known as **"range restriction"**
  - **Here this means the maximum Pearson's r will be smaller than ± 1 it should be:**

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

  - Some examples using this formula to predict maximum Pearson r values →
  - **So if indicator $\bar{r}$ is limited, so is reliability as measured by alpha (or KR-20)…**

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

Correlations for Binary or Ordinal Indicators

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

Line of Suspended Disbelief to Reduce Impact of Range Restriction

- **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 variables (not = Pearson); aka, between probits

- We will make use of **tetrachoric and polychoric correlations** in LTMMs predicting binary and ordinal indicator responses (limited-info estimation)

More Correlations: Pearson vs. Intraclass

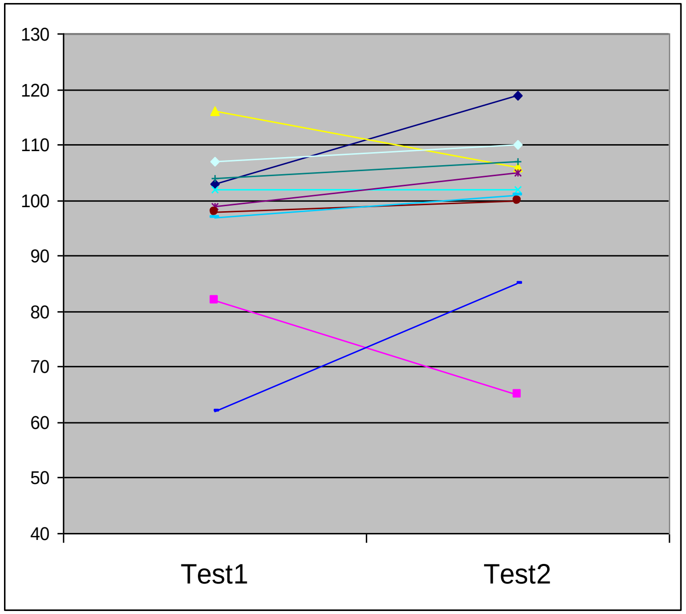
- **Pearson's r is problematic for assessing reliability across raters**, because it ignores relevant differences in mean and variance across raters by standardizing each variable separately
  - e.g., **multiple raters** (y_{1s}, y_{2s}) each provide scores for the same set of targets

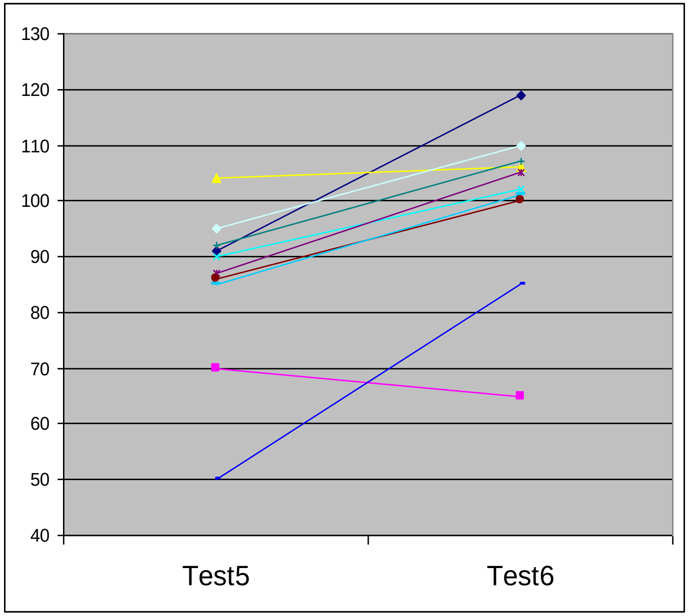
- Solution: use an "**Intraclass Correlation**" (ICC) instead, which standardizes across all raters using a **common mean and variance**
  - For example, for two raters: $\mathrm{ICC}(y_1, y_2) = \frac{\sum_{s=1}^{N}[(y_{1s}-\bar{y})(y_{2s}-\bar{y})]}{(N-1)*s^2}$

    where $\bar{y} = \frac{\sum_{s=1}^{N}[(y_{1s}+y_{2s})]}{2N}$ and $s_y^2 = \frac{\sum_{s=1}^{N}(y_{1s}-\bar{y})^2 + \sum_{s=1}^{N}(y_{2s}-\bar{y})^2}{2N-1}$
  - ICC is also a ratio of variances: $ICC = \frac{s^2_{Between-Targets}}{s^2_{Between-Targets} + s^2_{Between-Raters} + s^2_{within-both}}$

- **ICCs can readily be extended** to more than two raters, as well as to quantify the effect of multiple distinct sources of sampling variance
  - e.g., multiple raters of multiple targets across days—how much variance is due to each?
  - Btw, this is the basis of "Generalizability Theory" (or G-Theory) in measurement

Intraclass Correlation Example





M: 97 100

SD: 15 15

$Pearson\ r = .670$

$Intraclass\ r = .679$

M: 85 100

SD: 15 15

$Pearson\ r = .670$

$Intraclass\ r = .457$

$$ICC = \frac{s^2_{Between-Targets}}{s^2_{Between-Targets} + \boldsymbol{s^2_{Between-Raters}} + s^2_{within-both}}$$

Reliability in a Perfect World, Part 1

- What would my reliability be if I just added more indicators?
- **Spearman-Brown Prophesy Formula**
  - $Reliability_{NEW} = \frac{ratio * reliability_{old}}{1+[(ratio-1)*reliability_{old}]}$

$$ratio = \frac{\#\ new\ indicators}{\#\ old\ indicators}$$

  - For example:
    - Old reliability = .40
    - Ratio = 5 times as many indicators (had 10, what if we had 50)
    - New reliability = .77
- To use this formula, you must assume <u>**PARALLEL**</u> indicators
  - All item discriminations equal, all indicator error variances equal, all covariances and correlations among indicators are equal, too
  - (Unlikely) assumption of parallel items is testable in LTMMs

Assumptions about Indicators When Calculating Score Reliability in CTT

- Use of **alpha** as an index of reliability of total-scores requires an assumption of **tau-equivalent indicators**:
  - aka, "true-score equivalence" → equal item discrimination
  - Translates to **equal covariances** among indicators
    - But not necessarily equal correlations…(because different error variances)

- Use of **Spearman-Brown** Prophesy formula to predict new reliability requires an assumption of **parallel indicators**:
  - Tau-equivalent indicators PLUS equal error variances
  - This translates into equal correlations among indicators, too

- Btw, parallel indicators is also required to get a perfect correlation between latent trait estimates (of predictors as used in an LTMM) and total-scores as latent trait estimates in CTT (stay tuned)

Reliability in a Perfect World, Part 2

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

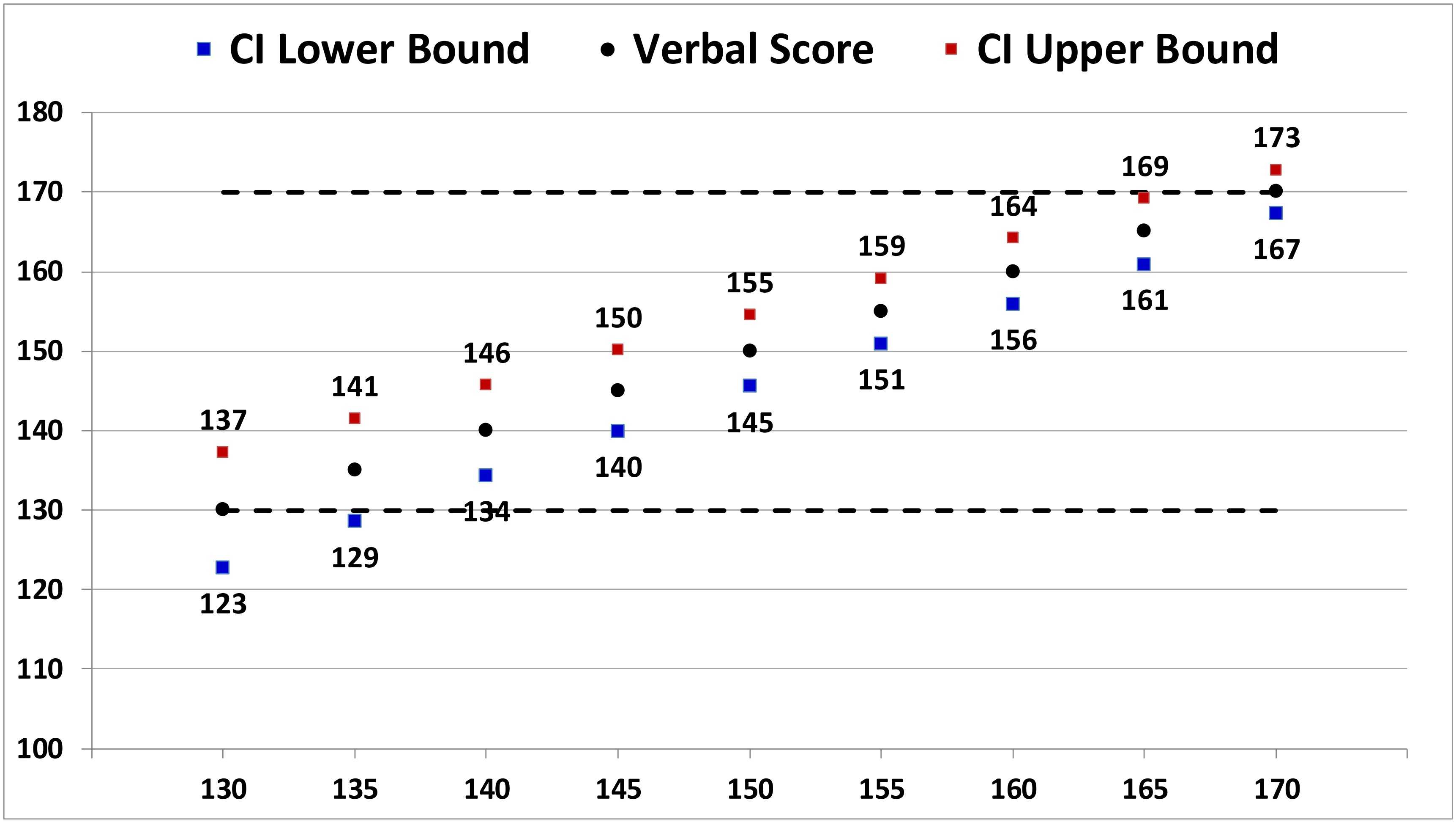
Using Reliability Coefficients → SE

- **Reliability coefficients** (Rel) are sample-level statistics…
  - But reliability is a means to an end in interpreting a score for a **given individual**—we use reliability to get the **error variance**
  - $Var(True) = Var(Y_{total}) * Rel$; so $Var(Error) = Var(Y_{total}) - Var(True)$
  - $\boldsymbol{SD(error)}$ **is individual standard error of measurement,** $\boldsymbol{SE}$
  - **95% CI for individual total-score =** $\boldsymbol{Y_{total} \pm (1.96 * SE)}$
    - Gives precision of true score estimate in the metric of the original total-score

- e.g., if $Var(Y_{total}) = 100$ and y_{total} for subject $s = 50$
  - $Rel = .91, Var(Error) = 9, SE = 3$ → $95\%\ CI \approx 44$ to 56
    $Rel = .75, Var(Error) = 25, SE = 5$ → $95\%\ CI \approx 40$ to 60
  - Note this assumes a symmetric distribution, and thus the limits of **CI can go out of bounds** of the scale for extreme scores
  - Note this also assumes the **SE for each person is the same!**
  - *Cue real-world example using the GRE…*

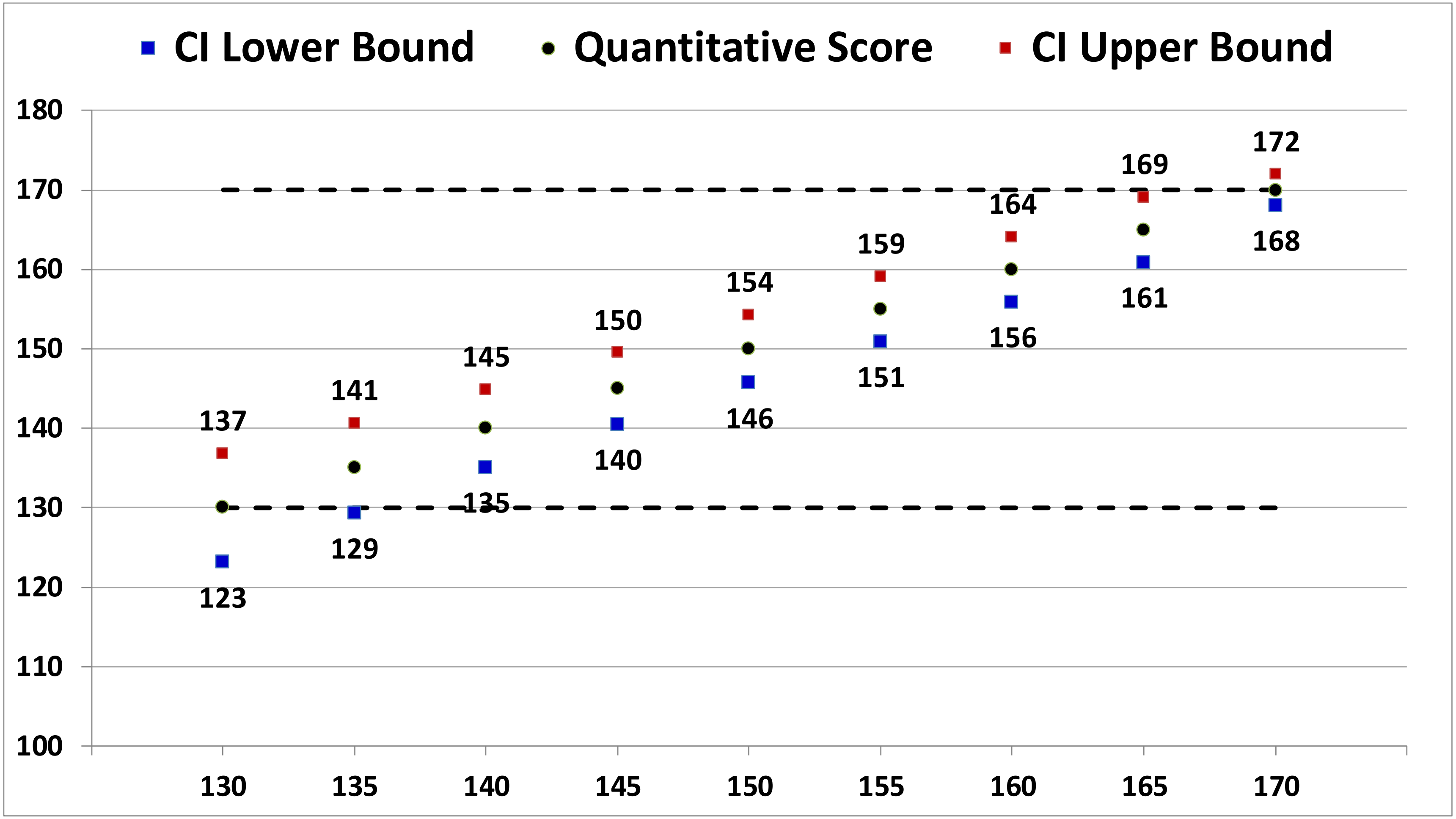
95% CIs for Individual Score: Verbal

M=150.4, SD=8.5, range=130 to 170; SE=1.4 to 3.7



95% CIs for Individual Score: Quantitative

M=153.4, SD=9.4, range=130 to 170; SE=1.0 to 3.5



Intermediate Summary: CTT Reliability

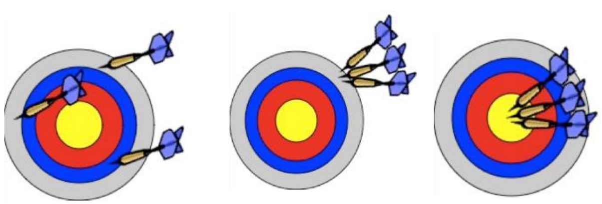
- **CTT unit of analysis is the TOTAL:** $\boldsymbol{Y_{total} = True + Error}$
  - Total-score is best estimate of True Score (i.e., the Latent Trait)
  - ASU measurement model (ASU = Add Stuff* Up)
    - ASU model assumes unidimensionality – the only thing that matters is the one $True$
  - Reliability of total-score cannot be quantified without assumptions that range from somewhat plausible to downright ridiculous (testable in item-level models)

- **Indicator 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**

From Reliability to Validity

- Given a belief of unidimensionality of the indicators measuring a latent trait, we can then assess reliability (of the total-scores in CTT)
  - Btw, both can be examined simultaneously in LTMMs (stay tuned)
  - **Random error limits reliability** (ability to hit a consistent target)

- **Reliability precedes validity**: the degree to which evidence and theory support interpretation of scores (i.e., hits the right target)
  - Validity is measure of degree, and **depends on USAGE or INFERENCES**
  - **Non-random error** limits validity: other sources of variation that get measured consistently along with the construct (e.g., acquiescence)
  - So how do you know **which source of variation** you've measured??? **Examine how latent trait estimates relate to those of other constructs!**



Left to right: Not reliable or valid, reliable but not valid, reliable and valid

Reliability vs. Validity "Paradox"

- Given the assumptions of CTT, it can be shown that the correlation between a total-score and an outside criterion (i.e., to demonstrate evidence for validity) 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 indicators with the strongest discriminations (or inter-correlations) can help to 'purify' or homogenize an instrument, but potentially at the expense of validity
  - Can end up with a "bloated specific" (i.e., too narrow, redundant items)
  - Items that are least inter-related may be most useful in keeping the construct well-defined conceptually and thus relatable to other things

Building an Argument for Validity

- Validity evidence can be gathered in two main ways:
  - Internal structure: preferable, but less frequent
    - From construct map—does the **empirical order** of the indicators along the construct map match your expectations of their order?
    - From 'explanatory' item response models that predict item difficulty or item discrimination (see Lecture 6 and Example 6 from this class)
  - External evidence: almost always used instead!
    - To what extent do the **latent trait estimates relate to those of other constructs** in theoretically or practically expected directions?

- Long history of debate about ways to classify or label "kinds" of validity available from external evidence
  - Want to know more about validity?
    Look for a new PSQF course in Spring 2021 (Dunbar)

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
  - For 2nd edition, APA was joined by AERA and NCME in 1966; subsequent editions published in 1974, 1985, 1999, and 2014: "The Standards for Educational and Psychological Testing"
- Individual authors have had their own takes:
  - Cronbach and Meehl (1955): Predictive and concurrent validity → criterion-related validity; construct validity still separate
  - Messick (1989; 1995): Content, substantive, structural, generalizability, consequential, external…
- I'll wave my hands at the most common, using "scale" as an abbreviation for "collection of indicators from which some kind of latent trait estimate is derived" (i.e., more general than CTT)

Predictive and Concurrent Validity

- **Predictive** and **concurrent** validity are often categorized under "criterion-related validity"
  - **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 shorter scale 'work as well' as original longer scale?
  - 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?
  - Can be compromised by differences in range of sensitivity

Content and Construct Validity

- **Content validity** concerns how well the indicators cover 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 intended latent trait (and for a given 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"?

   - Are the indicators 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? To all kinds of samples?
   - Can you make it shorter or adaptive to improve efficiency?

- **MEASURES ARE NEVER "VALIDATED" OR "RELIABLE"! Say "evidence for validity (or reliability)" instead!**