UTTyler™

THE UNIVERSITY OF TEXAS AT TYLER

UTTyler
THE UNIVERSITY OF TEXAS AT TYLER

Validity, Reliability, and the Science of Measurement

Brandon L. Bretl, PhD
Assistant Professor
School of Education
University of Texas at Tyler

Sessions

- Foundations of High-Quality Assessment: Purpose, Clarity, and Fairness
- **Validity, Reliability, and the Science of Measurement (October 22)**
- Item Writing Mastery: From Multiple-Choice to Open-Ended Excellence (Nov. 19)
- Beyond the Basics: Introduction to Advanced Psychometrics
- Language Matters: Neurolinguistic and Cognitive Considerations in Assessment
- Applying Theory to Non-Traditional Assessment and Research Applications

Validity, Reliability, and the Science of Measurement

- Definitions and practical applications of key psychometric concepts
- Tools for identifying common sources of error in measurement
- Understanding trade-offs in assessment design

Introductions

Dr. Brandon Bretl

Assistant Professor, School of Education

PhD in Human Development and Learning

University of Kansas

- Math and science teacher
- Researcher on state standardized science tests
- Cognitive and social psych research

Introductions

Why attending?

What assessments/surveys are you using, creating, or plan on creating?

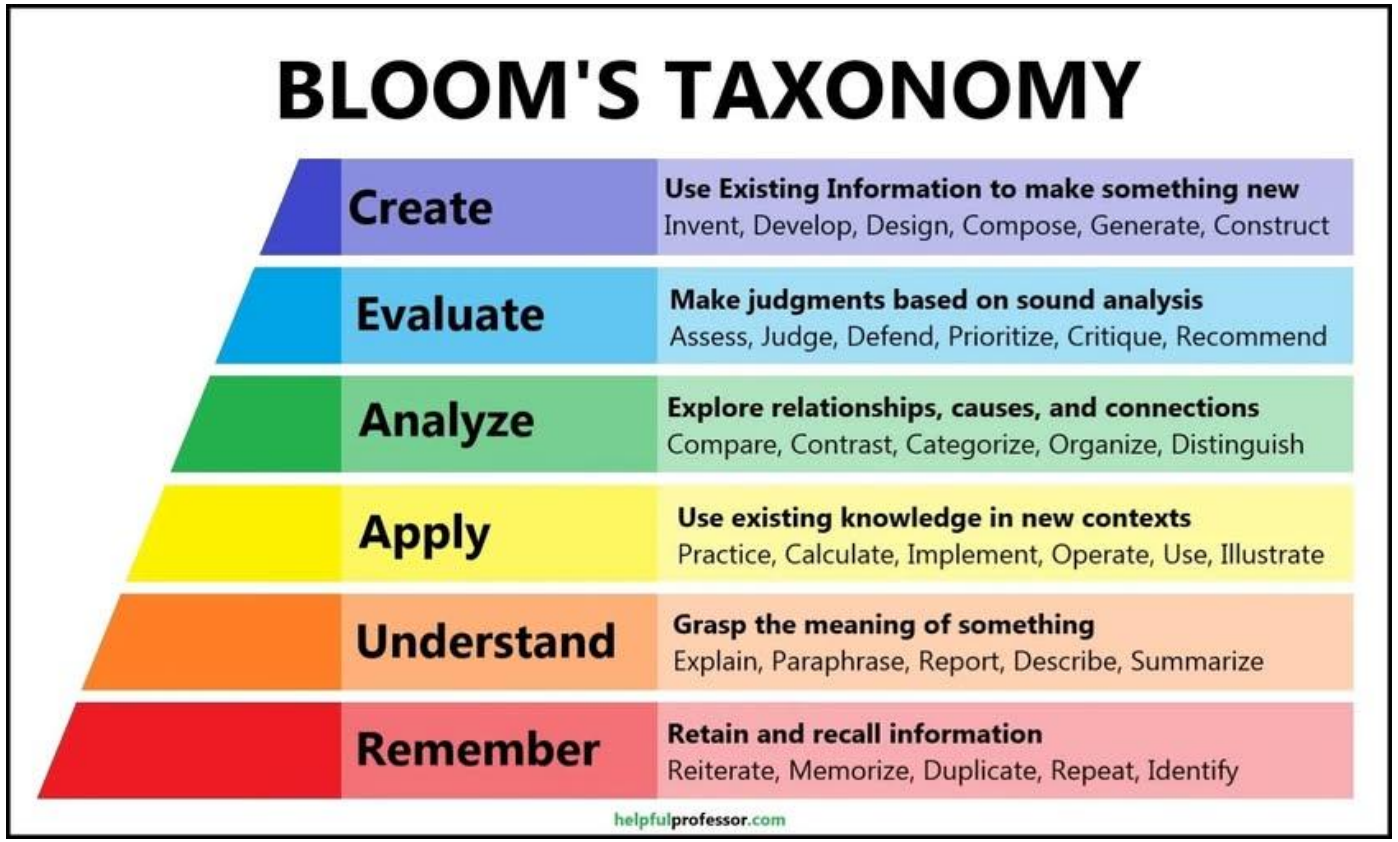
Quick Review

What makes an assessment good?

An assessment is an evidence system to generate information that reduces uncertainty.

A ***good*** assessment provides high quality evidence and the greatest reduction in uncertainty.

Epistemological



Real measurements

Latent Constructs

Latent Factor Analysis

PATIENT HEALTH QUESTIONNAIRE (PHQ-9)

NAME:_____ **DATE:**_____

Over the last *2 weeks*, how often have you been bothered by any of the following problems?
(use "✓" to indicate your answer)

	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3

Validity

Is an argument for how your assessment is measuring what you are claiming it is measuring.

Never valid or not valid…

Only better and worse validity arguments.

Clarity and Reducing Bias

A bias in assessment is a feature of the assessment that causes a decrease in the reliability and/or validity of the information obtained.

- Gender
- Race/ethnicity
- Socioeconomic status
- Religion
- Etc.

Validity, Reliability, and the Science of Measurement

- Validity
  - Content
  - Construct
  - Criterion
- Reliability
- Fairness
- Trade-offs

Validity

- Content
- Construct
- Criterion

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Content Validity

Does the assessment fully represent the domain?

Grade 10 Biology Final Exam

If it's 90% genetics and nothing on ecology, it does not validly represent the content of a grade 10 biology curriculum.

Evidence:

- Expert judgments
- Comparing to curricular or other standards

Construct Validity

Does it actually measure the theoretical construct?

Example: Academic self-efficacy

Convergent validity: correlates positively with other self-efficacy measures.

Discriminant validity: does not correlate strongly with unrelated constructs, e.g., general optimism.

Additional evidence:

- Factor structure analyses
- Theoretical predictions, e.g., predicts persistance

Criterion Validity

Does it accurately predict relevant outcomes based on a specific criterion?

Predictive validity: a college entrance exam correlates strongly with first year GPAs or graduation rates.

Concurrent validity: clinician's diagnosis correlates with score on anxiety exam.

Evidence:

- Statistical correlation with outside benchmarks.
- Regression models showing predictive power.

Reliability

Is the assessment consistent and stable?

- Internal consistency
- Inter-rater
- Test re-test

Get ready for some math…

Scores

$$X = T + E$$

X = observed score

T = true score

E = error

Reliability

$$\text{Reliability} = \frac{\sigma^2_{\text{true}}}{\sigma^2_{\text{total}}}$$

Cronbach's Alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{V_t}\right)$$

Cronbach's Alpha

$$\frac{\sum V_i}{V_t}$$

Cronbach's Alpha

$$1 - \frac{\sum V_i}{V_t}$$

Cronbach's Alpha

$$\frac{k}{k-1}\left(1-\frac{\sum V_i}{V_t}\right.$$

Cronbach's Alpha

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{V_t}\right)$$

Cronbach's Alpha

Example data (4 items, 6 respondents; Likert-style 1–5)

Person	Item 1	Item 2	Item 3	Item 4	Total
A	2	2	2	2	8
B	3	3	3	2	11
C	4	3	4	3	14
D	5	4	5	4	18
E	4	4	4	5	17
F	3	2	3	4	12

Example data (4 items, 6 respondents; Likert-style 1–5)

Person	Item 1	Item 2	Item 3	Item 4	Total
A	2	2	2	2	8
B	3	3	3	2	11
C	4	3	4	3	14
D	5	4	5	4	18
E	4	4	4	5	17
F	3	2	3	4	12

$$V_1 + V_2 + V_3 + V_3$$

Sum of item variances = 4.4667

Example data (4 items, 6 respondents; Likert-style 1–5)

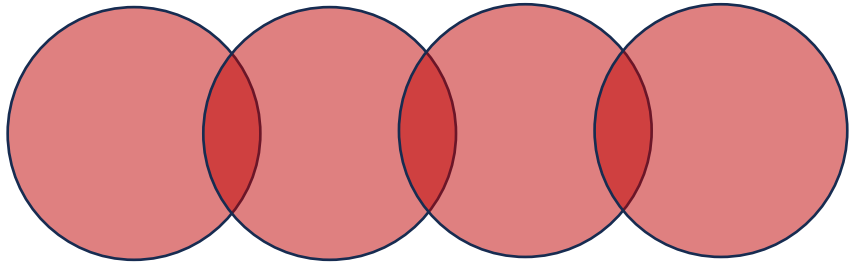
Person	Item 1	Item 2	Item 3	Item 4	Total
A	2	2	2	2	8
B	3	3	3	2	11
C	4	3	4	3	14
D	5	4	5	4	18
E	4	4	4	5	17
F	3	2	3	4	12

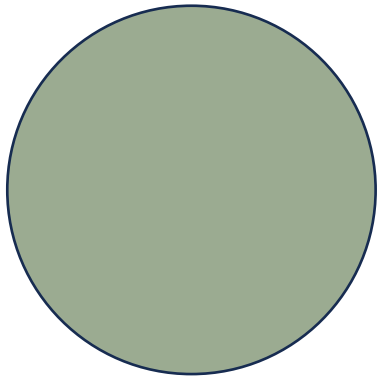
$$V_1 + V_2 + V_3 + V_3 \qquad V_t$$

Total score variance = 14.2667

Sum of item variances = 4.4667

Item total variance





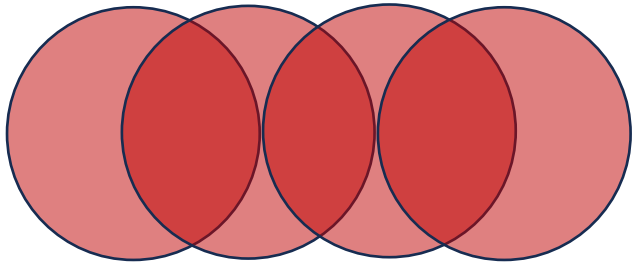
Total score variance

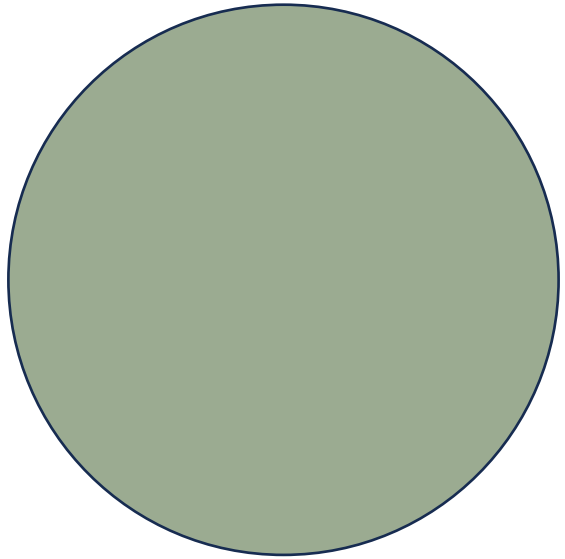
$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum V_i}{V_t}\right)$$

$$1 - \frac{\text{Item total variance (stays same)}}{\text{Total score variance is high}} = \textbf{Higher proportion, so lower Cronbach's alpha}$$

If there is low correlation in items, the differences in total scores will be less because the total score will be composed of less covariance between items.

Item total variance





Total score variance

If they score high on one, they score high on others, i.e., high correlation, high covariance.

And if they score low on one, they score low on others.

So total score variance is larger.

$$1 - \frac{\text{Item total variance (stays same)}}{\text{Total score variance is } \textbf{LARGER}} = \textbf{Smaller proportion, so greater Cronbach's alpha}$$

Example data (4 items, 6 respondents; Likert-style 1–5)

Person	Item 1	Item 2	Item 3	Item 4	Total
A	2	2	2	2	8
B	3	3	3	2	11
C	4	3	4	3	14
D	5	4	5	4	18
E	4	4	4	5	17
F	3	2	3	4	12

$$V_1 + V_2 + V_3 + V_3 \qquad V_t$$

Total score variance = 14.2667

Sum of item variances = 4.4667

Cronbach's Alpha

For $k = 4$ items,

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_X^2}\right)$$

Plug in the numbers:

$$\alpha = \frac{4}{3}\left(1 - \frac{4.4667}{14.2667}\right) = \frac{4}{3}(1 - 0.3130) = \frac{4}{3} \times 0.6870 \approx 0.916$$

Result: $\boxed{\alpha \approx 0.916}$

Cronbach's Alpha Guidelines

Cronbach's alpha	Internal consistency
$\alpha \geq 0.9$	Excellent
$0.9 > \alpha \geq 0.8$	Good
$0.8 > \alpha \geq 0.7$	Acceptable
$0.7 > \alpha \geq 0.6$	Questionable
$0.6 > \alpha \geq 0.5$	Poor
$0.5 > \alpha$	Unacceptable

On a related note…

- Canvas's new quizzes provide Cronbach's alpha and additional validity, reliability, and discrimination statistics

Item Difficulty

$$D = \frac{C}{T}$$

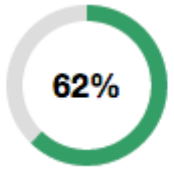
General guidelines:

- Under 0.30 is too difficult
- Above 0.85 is too easy

Item Discrimination

Attempts: 52 out of 53

During the recent Thor movie eagle eyed viewers got a glimpsed of the superhero Hawkeye. Which actor played him in Thor and will play him in the upcoming Avengers film?



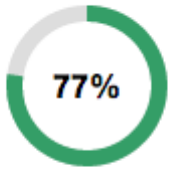
Correct answer

62% of your students correctly answered this question.

+0.18 Discrimination Index

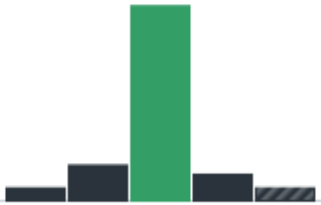
Attempts: 52 out of 53

Dozens of mutants have been members of the X-Men over the years. What is the better known alter ego of Remy Etienne LeBeau?



Correct answer

77% of your students correctly answered this question.



+0.26 Discrimination Index

Discrimination Index

Discrimination index	Interpretation
0.40 and above	Very good discrimination
0.30 - 0.39	Good discrimination
0.20 - 0.29	Fair discrimination
0.10 - 0.19	Not discriminating
Below 0.10	Poor item
Negative	Reversed relationship

Corrected Item-total Correlation Coefficient

- Available for new quizzes
- Better because it doesn't just consider discrimination between high, mid, and lowest scores
- -1 to +1
- Aim for above +0.20 or +0.30

Trade-offs

- Validity vs. reliability
  - Multiple choice more reliable
  - Open-ended may be more valid
- Validity vs. fairness
  - Context may be more authentic, but may disadvantage certain demographics
- Reliability vs. fairness
  - High standardization may increase reliability but stray from authentic assessment, privileging good test takers
- Precision vs. practicality
  - Highly valid, highly reliable instruments take a lot of time and resources to create

Need additional help?

Thank you!

BBRETL@uttyler.edu

Next session…

Item Writing Mastery: From Multiple-Choice to Open-Ended Excellence (Nov. 19)