Getting Started in GER

Validity and Reliability in GER

Outline

- Introduction to Validity
- Introduction to Reliability
- Review of Instruments
- Revisiting the four studies

Validity

Tests and Instruments *DO NOT* have validity

- The ***interpretation*** of scores or measures have validity

Validity is a continuum, not an all-or-none concept

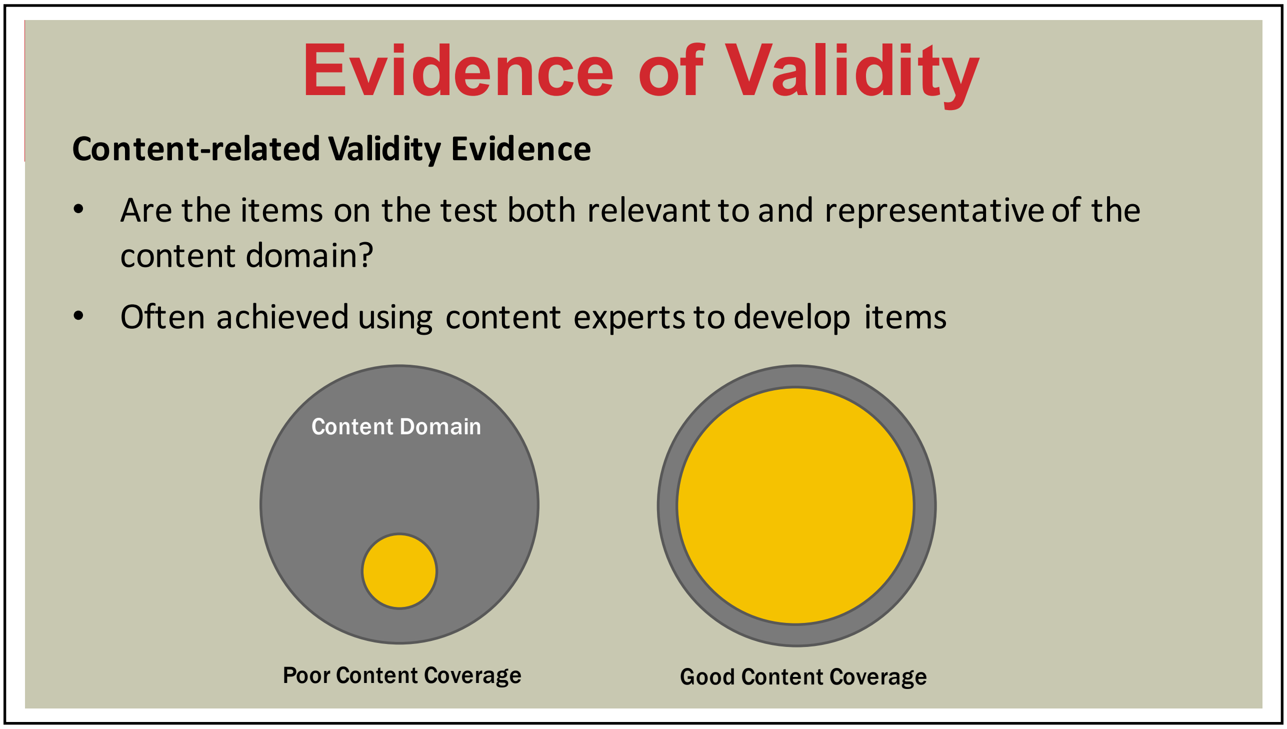
- There are numerous lines of evidence used to support the validity of interpretations

Imaginary Example: Plate Tectonics Concept Inventory (PTCI)

~~The PTCI is a valid instrument~~

Students who score high on the PTCI better understand geology

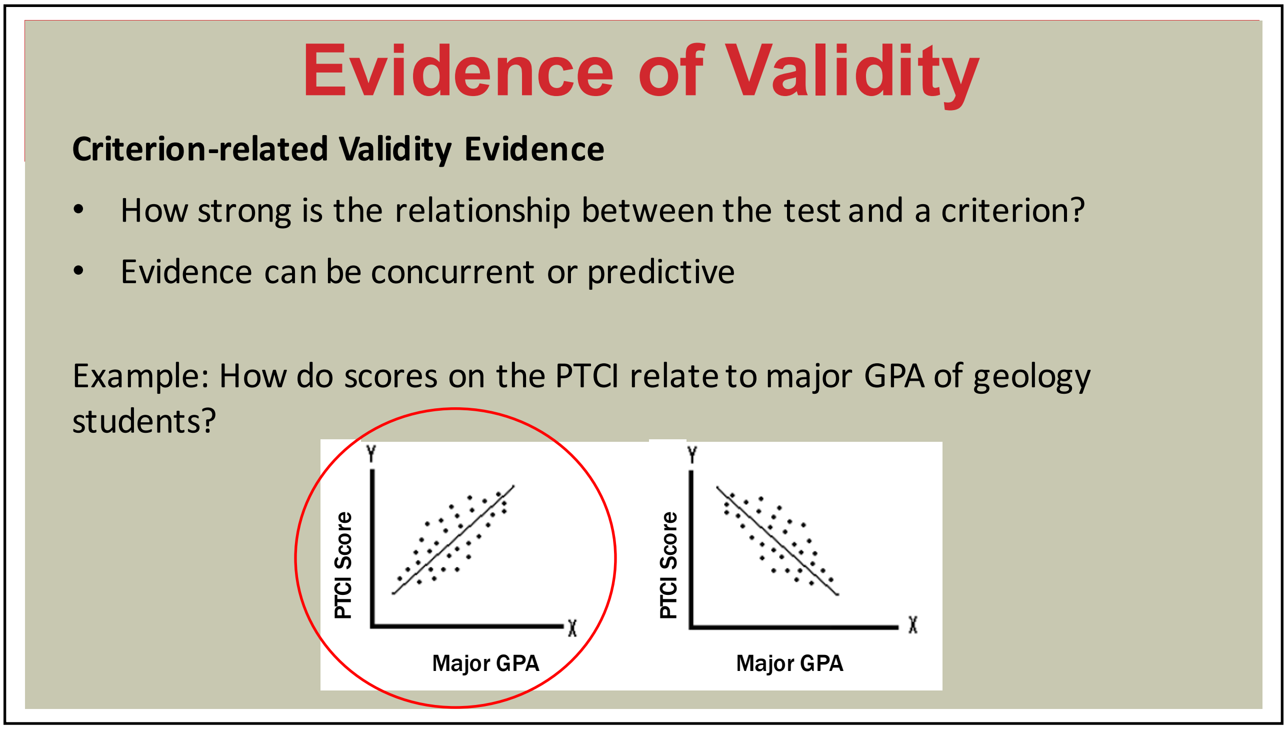
Students who score high on the PTCI better understand plate tectonics



Evidence of Validity

Content-related Validity Evidence

- Are the items on the test both relevant to and representative of the content domain?
- Often achieved using content experts to develop items



Evidence of Validity

Criterion-related Validity Evidence

- How strong is the relationship between the test and a criterion?
- Evidence can be concurrent or predictive

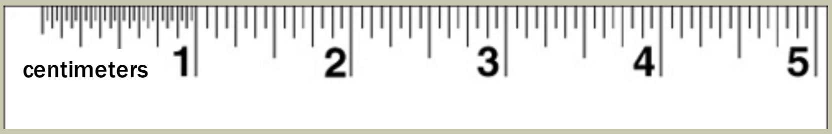
Example: How do scores on the PTCI relate to major GPA of geology students?

Evidence of Validity

Construct-related Validity Evidence

Does the test or instrument measure what it claims?

- Convergent or discriminant evidence
- Internal structure of the test (factor analysis, item homogeneity, etc.)



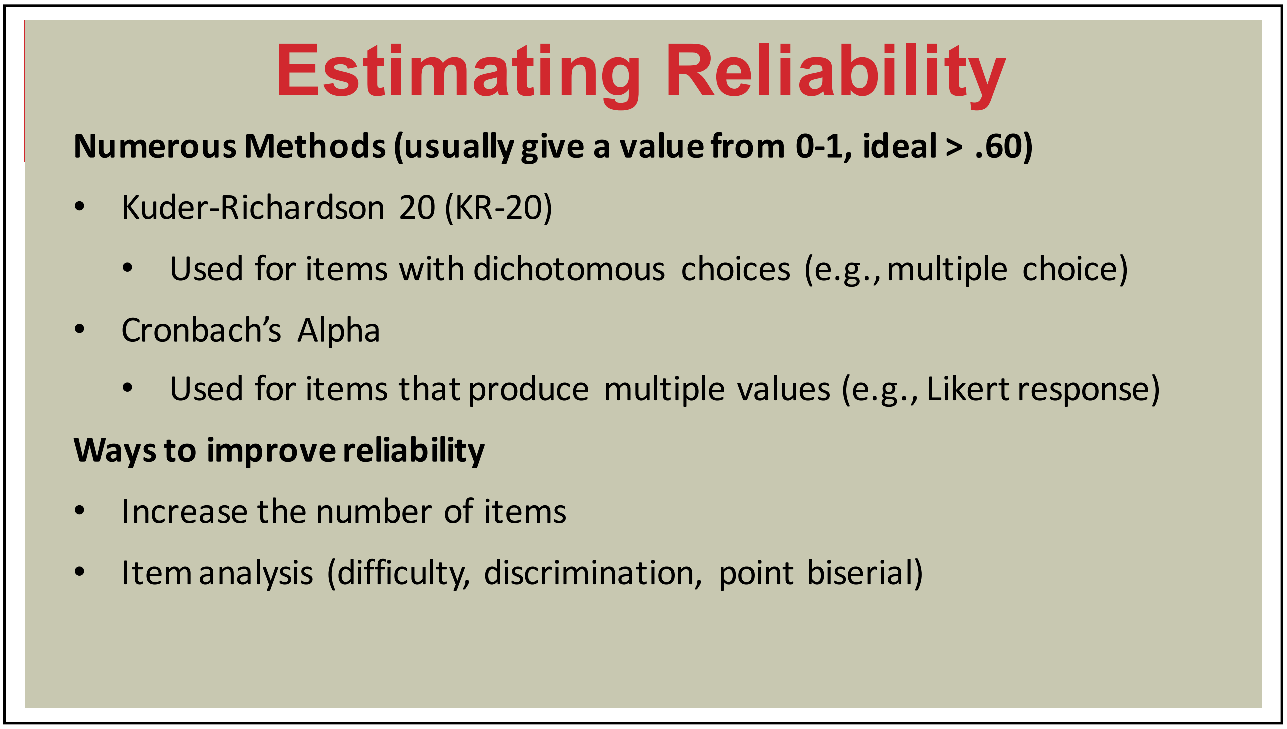
Would length measurements taken with this ruler be valid?

Reliability

Reliability is the consistency of test or measurement results

- True Score Theory (Obtained Score = True Score + Error)
  - Can never know an individual’s True Score with absolute confidence
- Sources of error include:
  - Content sampling error
  - Time sampling error (temporal instability)
  - Other Sources

Reliability is a necessary condition of validity!



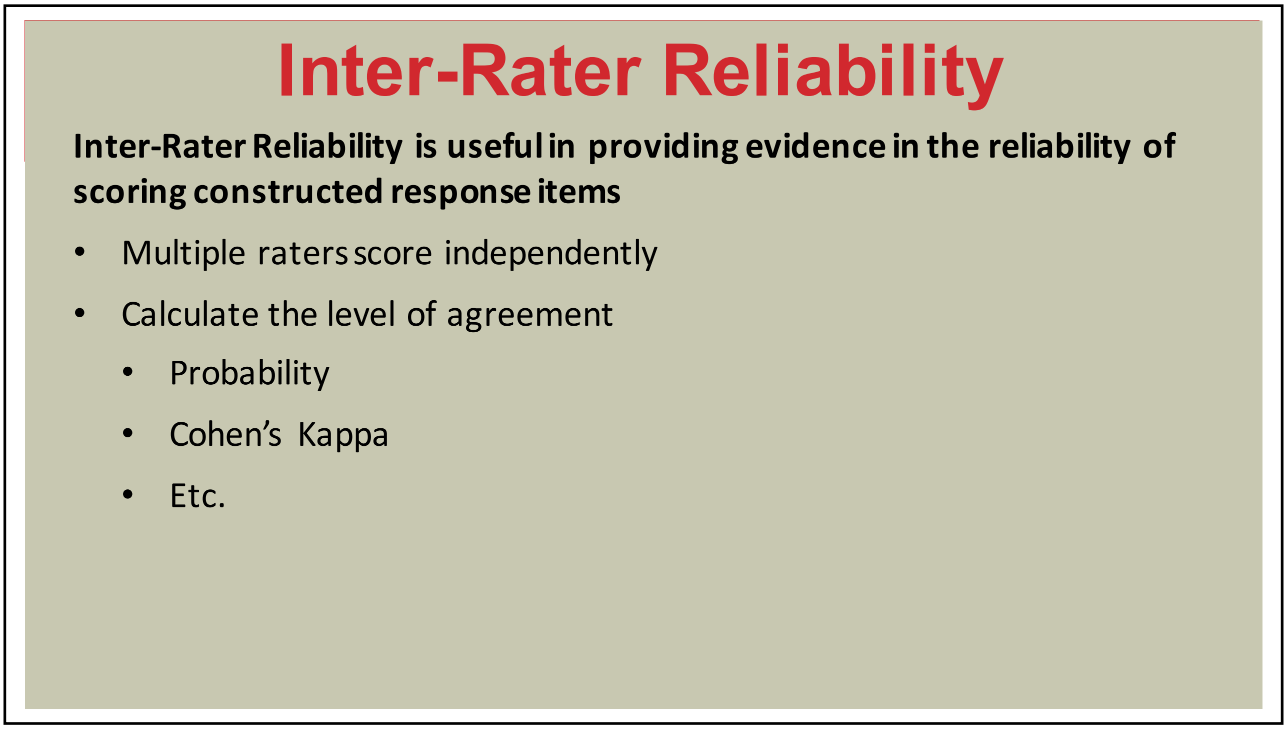
Estimating Reliability

Numerous Methods (usually give a value from 0-1, ideal > .60)

- Kuder-Richardson 20 (KR-20)
  - Used for items with dichotomous choices (e.g., multiple choice)
- Cronbach's Alpha
  - Used for items that produce multiple values (e.g., Likert response)

Ways to improve reliability

- Increase the number of items
- Item analysis (difficulty, discrimination, point biserial)



Inter-Rater Reliability

Inter-Rater Reliability is useful in providing evidence in the reliability of scoring constructed response items

- Multiple raters score independently
- Calculate the level of agreement
  - Probability
  - Cohen's Kappa
  - Etc.

Sample Instruments

Geoscience Content

- Geoscience Concept Inventory (GCI)
- Geoscience Literacy Exam (GLE)
- Landscape Identification and Formation Test (LIFT)

Attitudinal Instruments

- Scientific Attitude Inventory (SAI)
- Changes in Attitudes about the Relevance of Science (CARS)

Teaching Observational Instruments

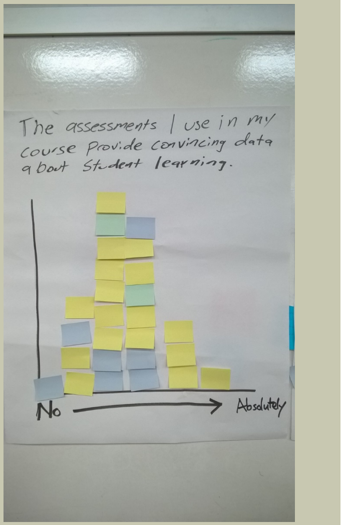
- Reformed Teaching Observation Protocol (RTOP)
- Classroom Observation Protocol for Undergraduate STEM (COPUS)

Your Turn!

Get with your research group and discuss the following prompts:

1. You decide to measure learning gains associated with a new teaching strategy in your classroom using your own exams. What sort of evidence can you report to demonstrate reliability and validity?

2. How will you provide evidence for validity and reliability in you own study?


The assessments I use in my course provide convincing data about student learning.
No
Absolutely