Interpretation and Follow Up of Assessment Results

Undergraduate Medical Education Course Directors' Retreat

Don A. Klinger
Associate Professor
Assessment and Evaluation
Faculty of Education

Overview

- Norm Referenced and Criterion Referenced Assessment
- Reliability and Validity
- Test and Item Analyses

The Purpose of Testing

- Measure student achievement of learning expectations.

The determination of achievement is relative.

Norm-referenced Assessment

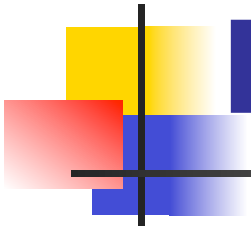
- Achievement is measured relative to the performance of others.
- Historically, this is the most common form of assessment
  - Bell curve
  - Tests are designed to maximise the spread between students

Criterion-referenced Assessment

- Achievement is measured relative to a identified standard.
- Theoretically, this is now the most common form of assessment
  - Tests are designed to maximise our confidence that the student has mastered the intended learning.

Our Previous/Present Models

- The use of easy and difficult test items to better identify the stronger students.
- Testing a broad domain of skills.
- Pass scores based on pre-determined percentages.



Reliability

Reliability is all about consistency.

- If a student wrote parallel forms of the same test, how consistent would their scores be?

Reliability

Reliability is a technical procedure

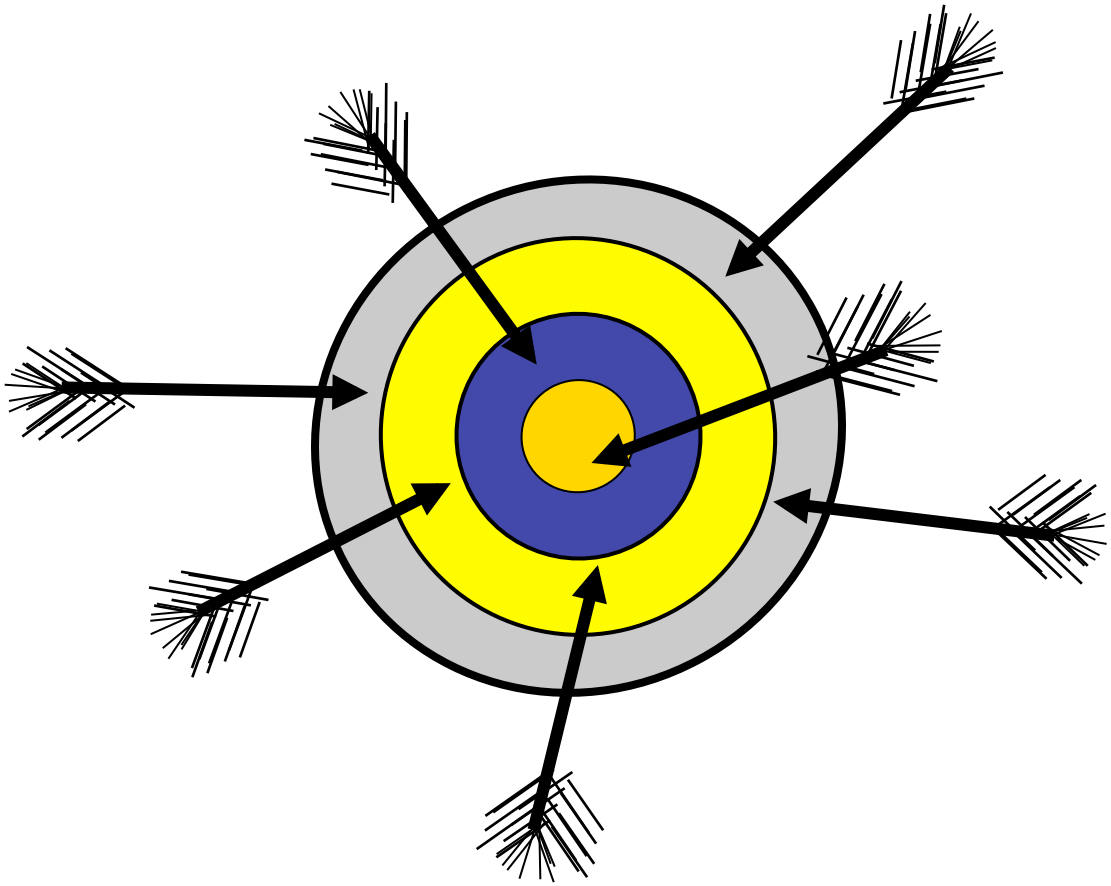
- There are different techniques to conduct reliability analyses
  - Parallel forms (Gold Standard)
  - Cronbach's Alpha (the most commonly used estimate of reliability)
- The result is a value between 0 and 1
  - For high-stakes decisions, >0.85

Reliability

Random measurement error lowers reliability

- Test item quality
- Testing conditions
- Preceptors (markers)
- The test taker

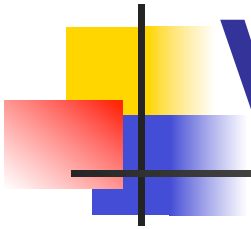
Random Error



Validity

Validity is about Truthfulness.

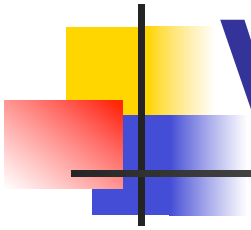
- Are the interpretations we make about a student's achievement correct?



Validity

Validity is an ongoing judgmental procedure.

- Does the test measure the intended construct?
  - Relevance and representativeness
- Does the test result in correct decisions?

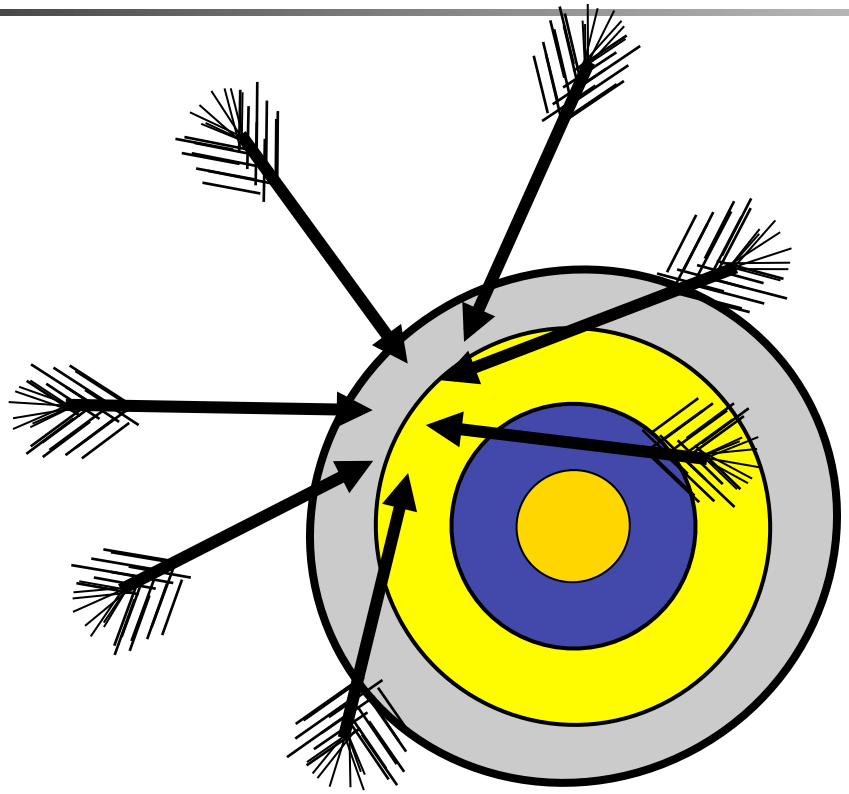


Validity

Systematic measurement error lowers validity

- Poorly defined constructs
- Test/construct misfit
- Biases
- Construct irrelevant variance

Systematic Error



Our Reality

- Errors are built into the **instruments** we construct to measure what students know and are able to do.

- Errors are built into our **interpretation** of the information students provide us on these instruments.

The Fallibility of Assessment

Assessment is a human activity

1. What a student really knows and is able to do

+

2. What assessment instrument you choose

+

3. How difficult you make the tasks

+

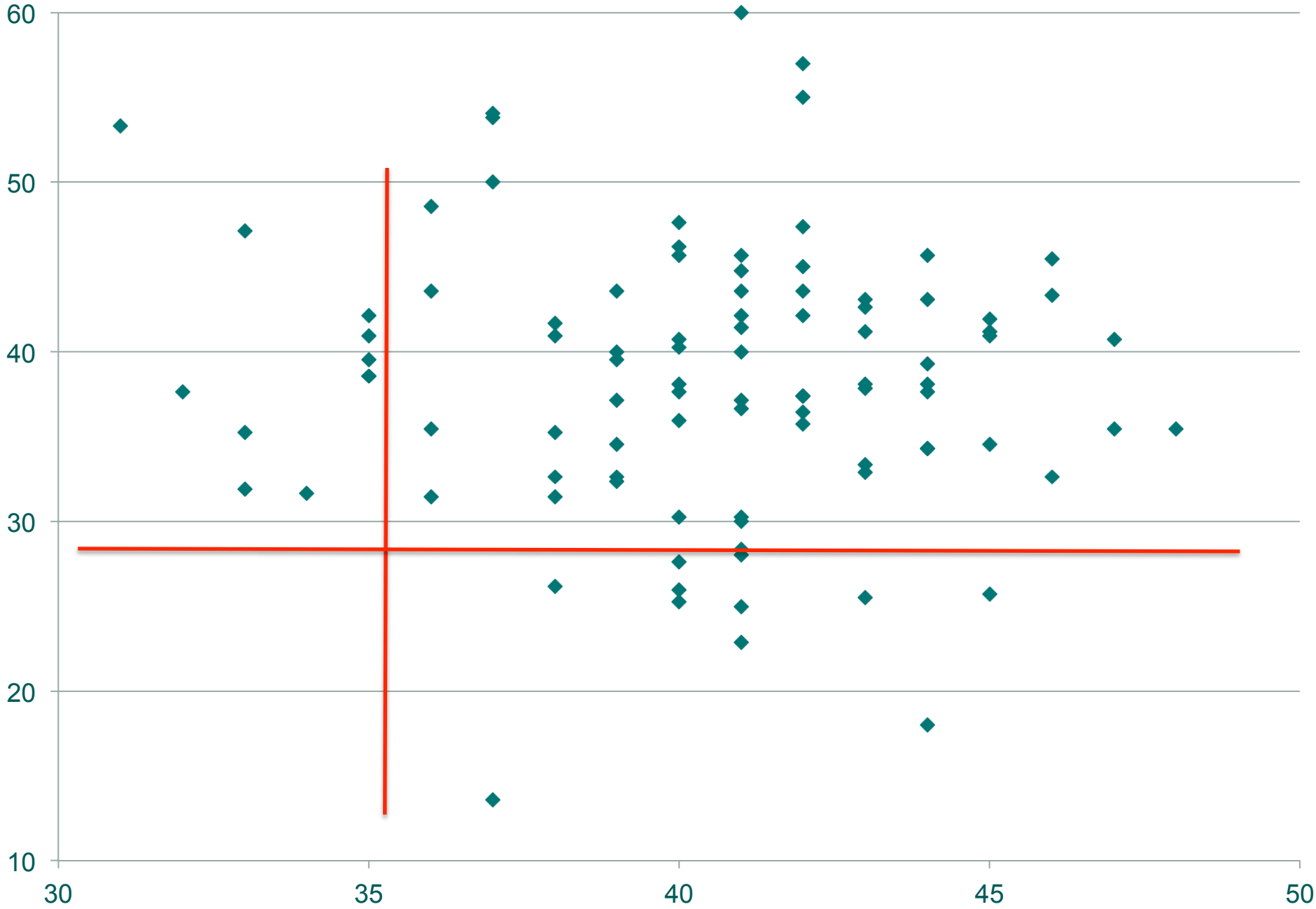
4. How you interpret what a student demonstrates

=

A Student's Score (mark)

Theories, policies, principles, standards and practice help us to reduce error in assessment.

False Positive and False Negative Decisions



Contingency Tables
(Confusion Matrices)

	Pass	Fail
Competent	60	5
Not Competent	15	20

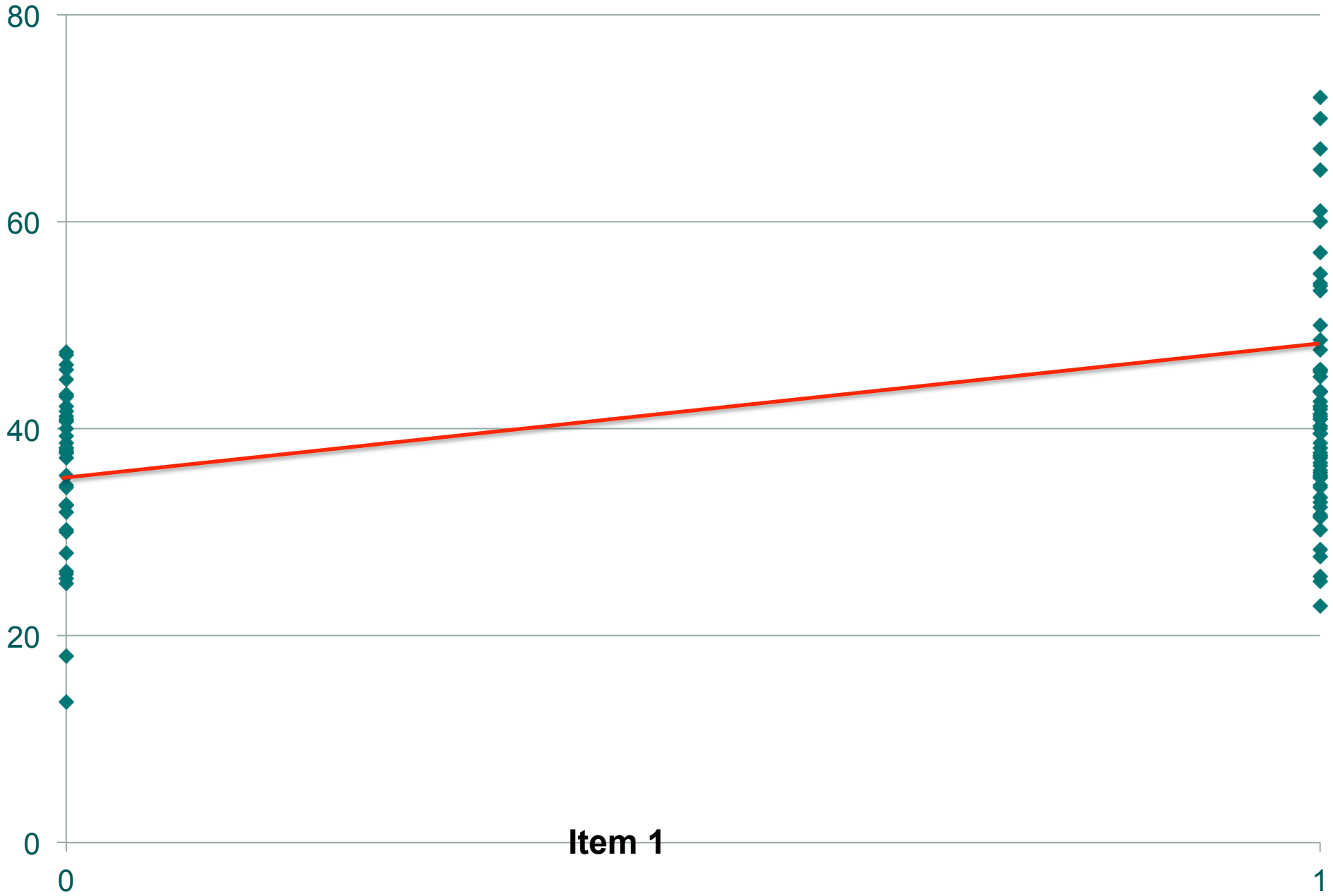
Test and Item Analyses

- Item Analyses
  - Difficulty
    - *p*-values
    - mean scores
  - Discrimination (Validity)
    - Point biserial
    - Other indices

Examples

Quest	# Cor.	A	B	C	D	% Corr.	Diff.	Validity
1	21	5	4	21	20	42	.42	0.18
2	50	0	50	0	0	100	1.0	0.00
3	40	2	0	40	8	80	.80	0.30
4	30	5	30	4	11	60	.60	0.06
5	50	1	25	18	6	50	.50	-0.10

Biserial Correlation



Formulas used by the Medicine program

$$\text{Difficulty} = \frac{R1 + R2}{N1 + N2}$$

$$\text{Validity} = \frac{R1 - R2}{N1}$$