3-2015

Introduction to Educational Measurement: Cramming a semesterlong course into a one-hour presentation

Tony Albano
University of Nebraska-Lincoln, albano@unl.edu

Introduction to Educational Measurement

Cramming a semester-long course into a one-hour presentation

Tony Albano

University of Nebraska-Lincoln
albano@unl.edu

Outline

1. Test design
2. Item writing
3. Scale development

Test Design

1. **Construct** – the unobservable trait or attribute we want to measure
2. **Operationalizing** – translating the construct into something observable
3. **Measurement** – using scores to represent amounts of the construct via operations
4. **Scale** – a set of operations (items) used to create composite scores

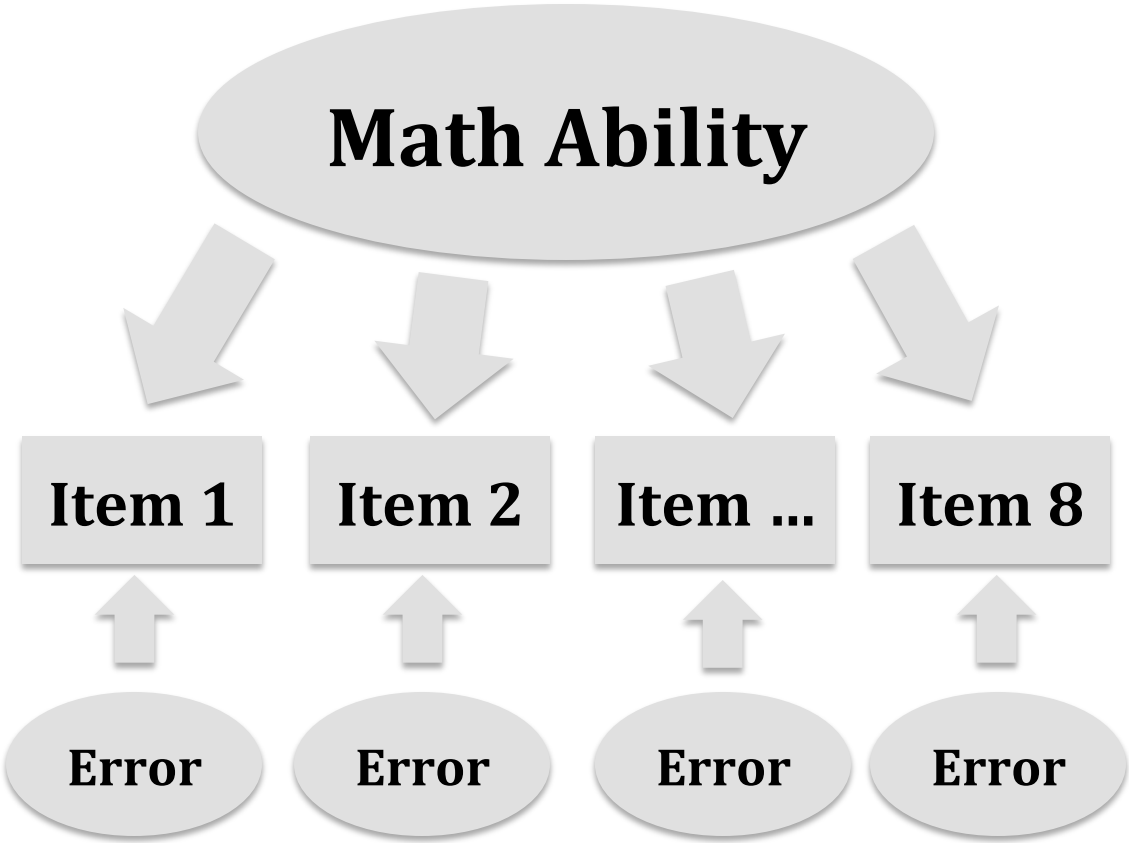
Test Design

5. **Inference** – assuming our scores describe some change in the construct
6. **Reliability** – extent to which inferences are consistent
7. **Validity** – extent to which inferences are accurate
8. **Purpose** – the who, what, and why

Test Design


0
NO HURT
2
HURTS
LITTLE BIT
4
HURTS
LITTLE MORE
6
HURTS
EVEN MORE
8
HURTS
WHOLE LOT
10
HURTS
WORST

Test Design


Math Ability
Item 1
Item 2
Item ...
Item 8
Error
Error
Error
Error

Test Design

- Test outline – structure the test around key features
  - Content areas
  - Learning standards or objectives
  - Types of items
  - Numbers of items
  - Depth of knowledge

Test Design

- Discussion
  - Construct and how it's operationalized
  - Scoring rules
  - The purpose of your test
  - Potential sources of unreliability and invalidity

Item Writing

- Standards or learning objectives – define what students should know or be able to do
- From the Nebraska state science test
  - Design and conduct investigations that lead to the use of logic and evidence in the formulation of scientific explanations and mode
  - Formulate a testable hypothesis supported by prior knowledge to guide an investigation.
  - Design and conduct logical and sequential scientific investigations with repeated trials and apply findings to new investigations.

Item Writing

- Depth of knowledge – level of complexity, based on required knowledge and skills
- Bloom's framework simplified by Webb
- Four levels of DOK:
  1. recall and reproduction
  2. skills and concepts
  3. strategic thinking
  4. extended thinking

Item Writing

- Item writing guidelines
  - Start with standard or learning objective
  - Keep it simple, while maximizing DOK
  - Minimize construct irrelevant variance
  - Utilize only as many distractors as you can write
  - Be cautious with “innovative” item types
- Planning guidelines
  - Follow your outline
  - Create 1.5 to 2 times as many items as you actually need
  - Get feedback before piloting (go to proola.org)

Item Writing

- Discussion
  - Define a learning objective for your test
  - Consider the appropriate DOK
  - Write an item for it

Scale Development

- Item analysis
  - Difficulty
  - Discrimination
  - Internal consistency
  - Bias

Scale Development

- Difficulty indexes average performance

Item	p-value	Corrected item total r	Alpha if item deleted
Item1	0.61	0.20	0.59
Item2	0.73	0.18	0.59
Item3	0.38	0.18	0.63
Item4	0.67	0.38	0.56
Item5	0.66	0.26	0.59
Item6	0.78	0.24	0.58
Item7	0.72	0.06	0.66
Item8	0.33	0.25	0.60

Scale Development

- Discrimination indexes relationship between single item and total score

Item	p-value	Corrected item total r	Alpha if item deleted
Item1	0.61	0.20	0.59
Item2	0.73	0.18	0.59
Item3	0.38	0.18	0.63
Item4	0.67	0.38	0.56
Item5	0.66	0.26	0.59
Item6	0.78	0.24	0.58
Item7	0.72	0.06	0.66
Item8	0.33	0.25	0.60

Scale Development

- Internal consistency indexes the shared relationship among all items

Item	p-value	Corrected item total r	Alpha if item deleted
Item1	0.61	0.20	0.59
Item2	0.73	0.18	0.59
Item3	0.38	0.18	0.63
Item4	0.67	0.38	0.56
Item5	0.66	0.26	0.59
Item6	0.78	0.24	0.58
Item7	0.72	0.06	0.66
Item8	0.33	0.25	0.60

Scale Development

- Bias tells us that performance differs for student with certain demographic or background characteristics

Item	p-value Women	p-value Men
Item1	0.56	0.74
Item2	0.69	0.83
Item3	0.31	0.53
Item4	0.61	0.78
Item5	0.63	0.74
Item6	0.74	0.85
Item7	0.71	0.73
Item8	0.21	0.57

Scale Development

- Bias tells us that performance differs for student with certain demographic or background characteristics

```
Mantel-Haenszel Chi-square
statistic:

      Stat.   P-value
Item1  0.1656  0.6841
Item2  0.0289  0.8651
Item3  0.2218  0.6377
Item4  1.2334  0.2667
Item5  0.2651  0.6066
Item6  0.4039  0.5251
Item7  0.0009  0.9765
Item8 14.3052  0.0002 ***
```