

## Large Scale Assessment 101: The Design, Construction, and Evaluation of Large Scale Assessments

Mark D. Reckase  
Michigan State University

### Goals for this Workshop

- Gain an understanding of how large scale assessment work.
- Gain an understanding of how large scale assessments are designed.
- Consider issues about how results are reported.
- Become an informed consumer of test results and be able to critically evaluate results.

### A Little History

- Prior to about 1960, the focus of all test design and analysis was on a summed-score for tests. Test items were only seen as a means to get to the summed score.
- Around 1960, test analysts discovered that tests are composed of test items.
- This caused a major shift in the thinking about tests and resulted in the development of item response theory (IRT).

### The Test Item

- Before going forward, we need to gain an understanding of test items.
- Do the following with the first test item in your packet.
- Answer the item yourself.
- Think about what **lack** in skills and knowledge would result in a wrong answer.
- What profile of skills and knowledge would result in a correct response.
- What is the major distinction between these two sets of skills.

### The Test Item

- What does the score on the test item mean?
  - What does a score of 1 mean?
  - What does a score of 0 mean?

### The Test Item

- Do the following with the second test item in your packet.
- Answer the item yourself.
- Think about what **lack** in skills and knowledge would result in a wrong answer.
- What profile of skills and knowledge would result in a correct response.
- What is the major distinction between these two sets of skills.

### Interpreting the Two-Item Test

- We now have a two item test.
- What does the response pattern of 00 mean?
- What does the response pattern of 11 mean?
- Which question is easier?
- What does the response pattern of 10 mean?
- Does a response pattern of 01 mean the same thing as a pattern of 10?

### The Achievement Continuum

- Suppose that we think of an achievement scale as something like a temperature scale.
  - The scale goes from negative numbers (very cold) to big positive numbers (very hot).
  - The location of 0 is some place in the middle and its location depends on who developed the scale.
- If persons answers the first item correctly, where are they on the continuum?
- If persons answers the second item incorrectly, where are they on the continuum?

### The Achievement Continuum

- Where do you think persons are located with the response pattern 10?
- Where do you think persons are located with the response pattern 01?
- The important idea is that rather than thinking about persons as located on a scale from 0 to 2, they are now located on a general continuum and the responses to the items give us information about their location.

### The Achievement Continuum

- The concept of the achievement continuum used to be called domain sampling.
  - Define the domain of skills and knowledge that are the target of the curriculum.
  - Consider a large number of test questions that cover everything in the curriculum.
  - Figure out how to randomly sample from that set of test questions.
  - The proportion correct on the sample gives an estimate of what is known about the whole domain.
- But, no one knows how to actually do this.

### The Achievement Continuum

- An alternative idea of the achievement continuum.
  - Each test item measures a complicated combination of skills and knowledge.
  - The scale produced by a test (a set of test items) is the average over items of these skills and knowledge.
  - It is important for comparability over years that the same average of content and skills is produced for a test form so the scale has the same meaning.

### Achievement Continuum

- Which of the approaches makes sense to you?

### Item Classifications

- Whichever model for the achievement continuum that you like, we need to have a good sense of what each test item measures.
- Need to classify them according to the curriculum.
- A summary of that is given in the Grade Level Content Expectations
- Please classify the ten items in the packet into the GLCE categories -- there is a recording sheet for this process.

### Item Classifications

- How much agreement there is in the group?
- How well do the group judgments compare to the official classifications?

| Item Number | Content | Level |
|-------------|---------|-------|
| 41          |         |       |
| 51          |         |       |
| 52          |         |       |
| 50          |         |       |
| 33          |         |       |
| 26          |         |       |
| 38          |         |       |
| 31          |         |       |
| 46          |         |       |
| 44          |         |       |

### Alignment

- A big issue related to NCLB is alignment.
- Tests should measure what is specified in the curriculum.
- Typically, alignment studies ask experts to make judgments about the skills and knowledge that students apply to test question

### Creating the Test Form

- Large scale testing programs do not use a single form of a test.
  - Because of security concerns and the need to span the full range of the curriculum, new test forms are produced each year.
  - The reported results are expected to be comparable
- Need to produce test forms that yield results on the same achievement continuum.

### Creating the Test Plan

- For the table in your materials, divide the percent space into three columns.
- For the 12 cells in the table, put in the percent of the entire test that should be related to that cell.
- The full set of 12 numbers should add to 100.

| Strand                | Knowledge | Application | Higher Understanding |
|-----------------------|-----------|-------------|----------------------|
| Number and Operations |           |             |                      |
| Algebra               |           |             |                      |
| Geometry              |           |             |                      |
| Data and Probability  |           |             |                      |

## Forms Construction

- Need to select items to match content distribution.
  - How much can you trust the content classification?
- Need to select items to match cognitive skill distribution.
  - How much can you trust the cognitive skill classification?

## Forms Construction

- Have classifications from item writers.
- Have some data from tryout of items.
- What do you do when you don't have an item you need to fill out the plan?

## Comparability of Results from Different Test Forms

- Suppose we have constructed two test forms according to the processes just discussed.
- One test form is administered in 2009.
- One test form is administered in 2010.
- Do the number-correct scores on the 2010 form have the same meaning as the 2009 form?

## Equating

- It is virtually impossible to construct test forms that give test scores on the number-correct score scale that are comparable.
- However, if achievement continuum idea is used, the idea is to use the information from the item responses from the forms to find locations of students on the same continuum.
- For this to work, we need to have the locations of the test items from the two forms on the same continuum.

## Equating

- Remember that for temperature, the guys that developed the scale set the numerical values for various points on the scale.
- This is also true for the computer programmers who developed the analysis program for test data.
  - Some decided to set 0 at the average student performance.
  - Some decided to set 0 at the average location of test items.
- To get everything on the same scale, we need to find the equivalent of the conversion equation from Celsius to Fahrenheit.

### Equating

- Consider the graphs describing the functioning of two test items.
- The items on the two graphs are identical, but the scales have been shifted.
- How do you need to shift the scale to get the graphs to be identical?

### Equating

- The actual equating process is somewhat more difficult, but it has the same concept.
- Find the equation that links values from one scale to those on the other.
- Once the items are put on the same scale, the estimates of students' locations using the items are on the same scale.

### Setting the Units and Numerical Values for the Reported Score Scale

- The people who wrote the analysis programs have made some arbitrary decisions about the unit size and numbers for the score scale.
- We could go with these decisions.
  - 0 is someplace near the middle of the scale.
  - The size of unit is approximately 1/6 of the range of performance.
- Do you like this scale, or would you like to set the numerical values to something different and use different size units?

### Setting the Reporting Score Scale

- Need to set two values to set a scale.
  - One point and size of unit.
  - Two points.
- How would you set the scale for the achievement continuum?

### Setting the Reporting Score Scale

- Make the average student 500 and the range of scores 200 to 800.
- Set the location for a minimally qualified student, like proficient, at a particular numerical value.
- Consider the current MEAP scale.
- Issues
  - Size of units versus accuracy of the test.
  - Match of test accuracy to decision points.

### Summary – Where have we been and what should you do next?

- Tests are composed of test items. Poor test items result in poor tests.
  - Review test items to determine if they measure appropriate content using appropriate contexts.
- The first test form for a testing program defines the achievement continuum for the program.
  - Review the full set of test items to determine if the achievement continuum is properly defined.

### Summary – Where have we been and what should you do next?

- Test forms must be constructed to define the same achievement continuum if they are to yield comparable results.
  - Tests should have a clear and detailed development plan.
  - Each form should be developed with careful attention to the plan.
- Even with the most careful development, test forms will not be perfectly equivalent – there scales need to be transformed to have the same units and specified numerical values.
  - Determine how test forms are equated and learn how to tell when equating is working.

### Summary – Where have we been and what should you do next?

- The numerical values for the reported scores and the size of units are selected by the developer.
  - Do they seem like reasonable choices to you?
  - Do the size of the units match the accuracy of locations for students on the achievement continuum.
- One test item does not give much information about the location of students on the achievement continuum.
  - Subscores need about 12 test items to give reasonable accuracy for locations on an achievement continuum.
  - How much testing time are you willing to spend to get useful test scores?

### Final Thoughts

- Work to become an informed consumer of test information.
- Thank you for being here!