Grade 3 FCAT - Test Construction & Equating

June 1, 2007

Cornelia S. Orr, Assistant Deputy Commissioner of Accountability, Research, and Measurement (ARM)

Office of Assessment and School Performance

Florida Department of Education

"Experience teaches only the teachable."
Aldus Huxley (1894-1963)

Topics

- The Grade 3 Test in 2006
- Test Construction
  - Process and Product
  - Science and Art
- Psychometric Primer
- Test Calibration and Equating

The Grade 3 Test in 2006

- Passages – Questions – Forms
- Student scores based on 5 passages & 45 questions
- 30 different forms, each with 1 passage & 7-8 questions
- Forms are used for anchor and field test questions
- One of the 6 passage positions is used for anchor and field test questions

2006 Grade 3 FCAT Test Passages and Positions					
Day 1 – Session 1			Day 2 – Session 2		
1	2	3	4	5	6

The Grade 3 Test in 2006

2006 Grade 3 FCAT Test Passages and Positions			
Day/ Session	**Passage Position**	**Number of Questions**	**Passage Description**
1	1	8	Ladybird, Ladybird, Fly Away Home (Lit.)
	2	7 or 8	Anchor and Field Test Passages (Varies)
	3	10	A Gift of Trees (Inform.)
2	4	13	Swim, Baby, Swim (Lit.)
	5	8	Slip, Slop, Slap/Sunny Sidebar (Inform.)
	6	6	Making Spring (Lit.)
	TOTAL	52-53	

"Test Construction"

- Process of building the test
  - Occurs the summer before a test
  - Based on available passages, questions, and statistics
- Guidelines for building the test
  - *Test Construction Specifications*
- Building the test is an iterative process

Test Construction Cycles

1999	2000	2001	2002	2003	2004	2005	2006
						June Test Construction	2006 Test Admin.
				Oct/Nov Item Review Meetings	June Field Test Construction	March Field Test	
					June Test Construction	2005 Test Admin.	
			Oct/Nov Item Review Meetings	June Field Test Construction	March Field Test		
				June Test Construction	2004 Test Admin.		
		Oct/Nov Item Review Meetings	June Field Test Construction	March Field Test			
			June Test Construction	2003 Test Admin.			
	Oct/Nov Item Review Meetings	June Field Test Construction	March Field Test				
		June Test Construction	2002 Test Admin.				
Nov/Dec Item Review Meetings	June Field Test Construction	March Field Test					
	June Test Construction	2001 Test Admin.					
May/Aug Item Review Meetnigs	March Field Test						
Sept/Oct Field Test Construction							

Test Construction Specifications - 1

- Guidelines for building the test
  - Ranges for each category
  - Iterative process
- Content Guidelines
  - Reading Passages (type and word counts)
  - Benchmark Coverage
  - Reporting Category (Strand) Coverage
- Multicultural & Gender Representation
- Cognitive Level Guidelines

Test Construction Specifications - 2

Statistical Guidelines for Questions

- Classical Item Difficulty and Discrimination
- IRT Difficulty, Discrimination, and Guessing
- Differential Item Functioning (DIF)
- IRT Model Fit Statistics

Statistical Guidelines for Tests

- Test Characteristic Curves
- Test Information Functions
- Standard Error Curves

Test Construction Specifications – 3

Anchor Item Guidelines

- Number and position of questions
- Content Representation – Mini Test
- Performance Characteristics (range of difficulty)
- Previous use as a Core or Anchor
- No change in wording
- Passage position

Test Construction
Review and Approval Process

- 1st Draft of Content – Harcourt Content Staff
- Review of Content – DOE Content Staff
- Review of Statistics – Harcourt Psychometric Staff
- Review of Statistics – DOE Psychometric Staff
- Approval by DOE FCAT team leadership

Psychometric Primer -1

Classical Item Statistics:

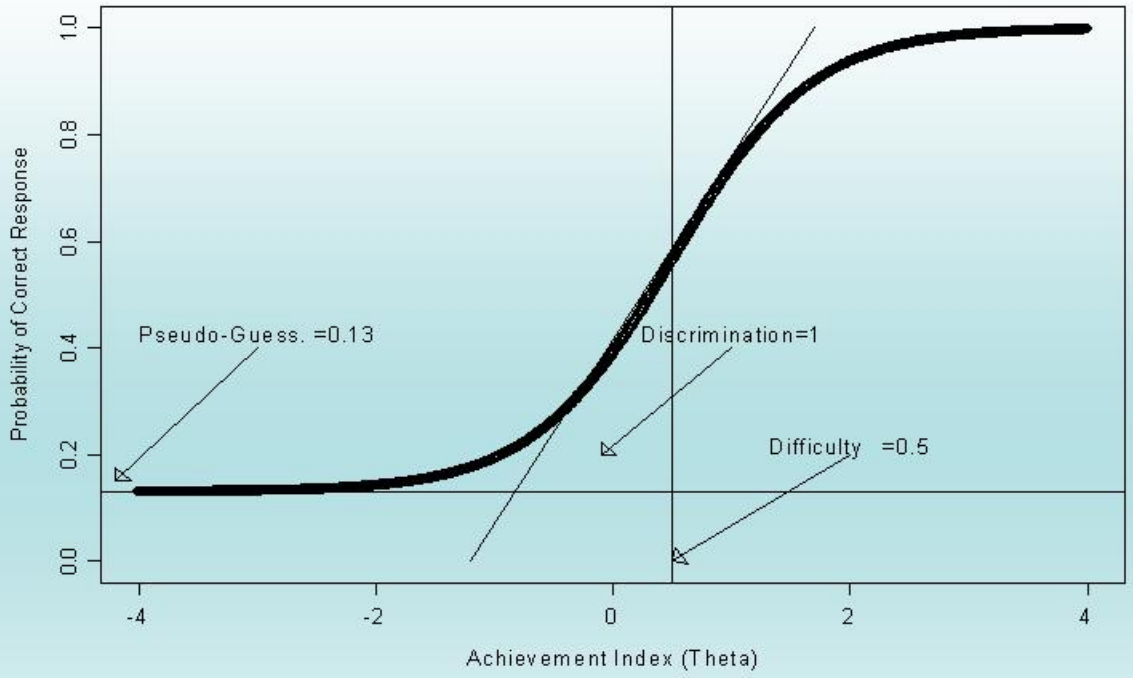
- P-value or difficulty - the percent (P) who answer the question correctly.
- Discrimination (point-biserial) - the degree to which students who get high scores answer the question correctly and vice versa (similar to correlation).

Psychometric Primer -2

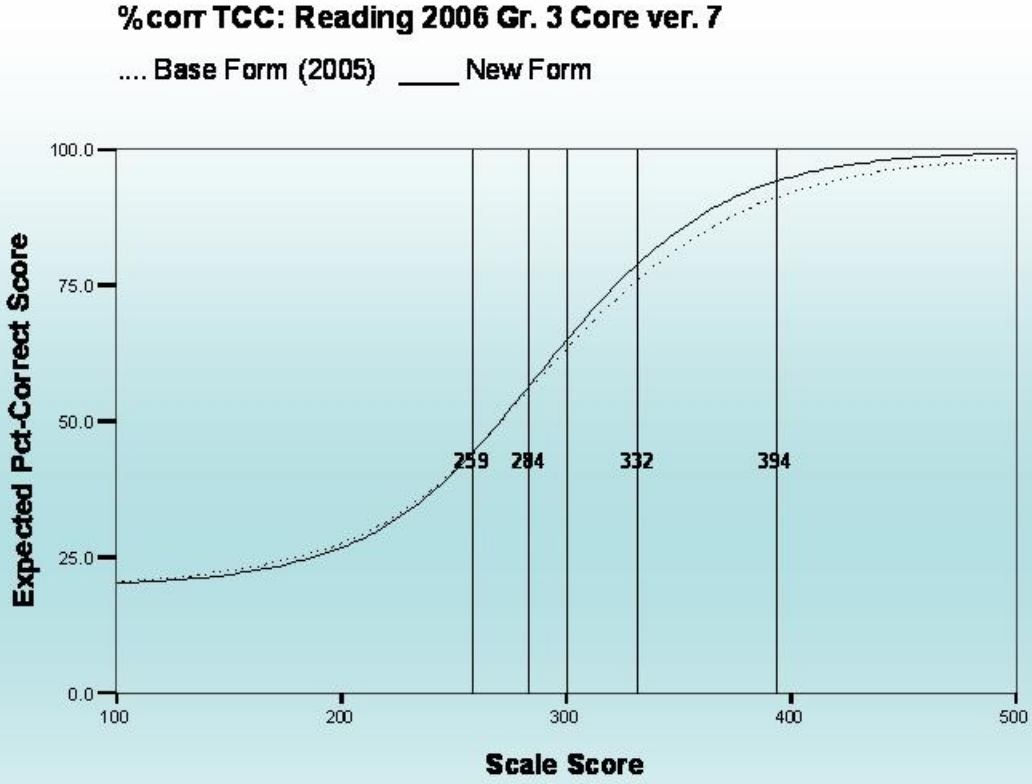
Item Response Theory (IRT) Statistics –

- A-parameter - discrimination or how well the question differentiates between lower and higher performing students.
- B-parameter - difficulty or the level of ability on the 100-500 scale required to answer the question correctly.
- Guessing - the probability of examinees with extremely low ability levels getting a correct answer.
- FIT - how well the scores for a given item fit, or match, the expected distribution for the model.
- DIF (Differential Item Functioning) - the degree to which the question performs similarly for all demographic groups based on ability.

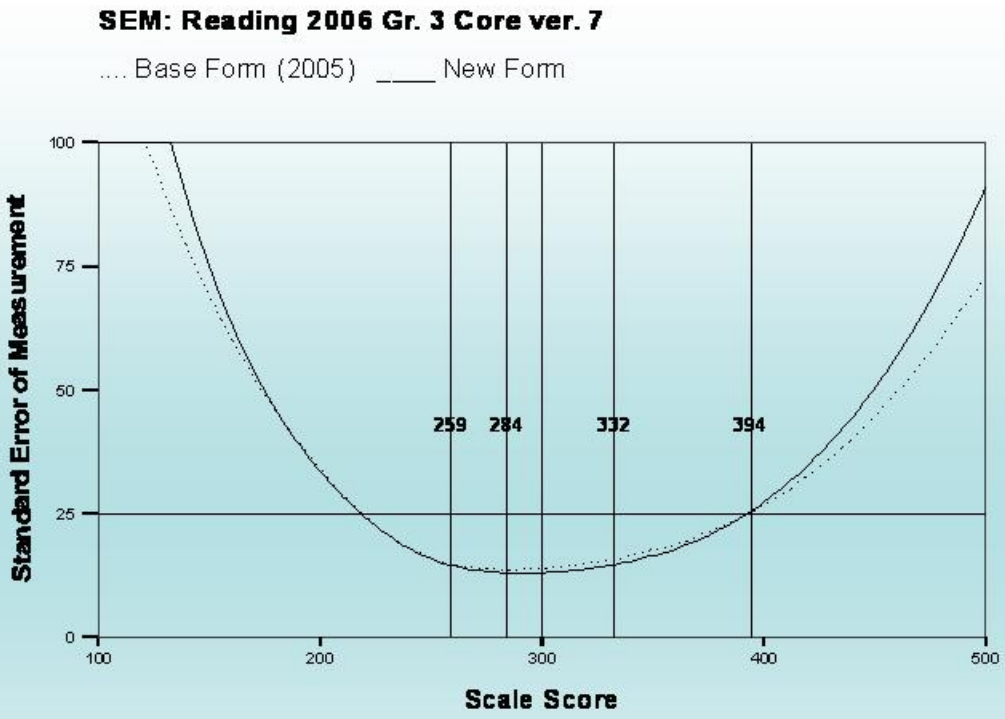
Item Characteristic Curve – Figure 1



Test Characteristic Curve - Figure 2



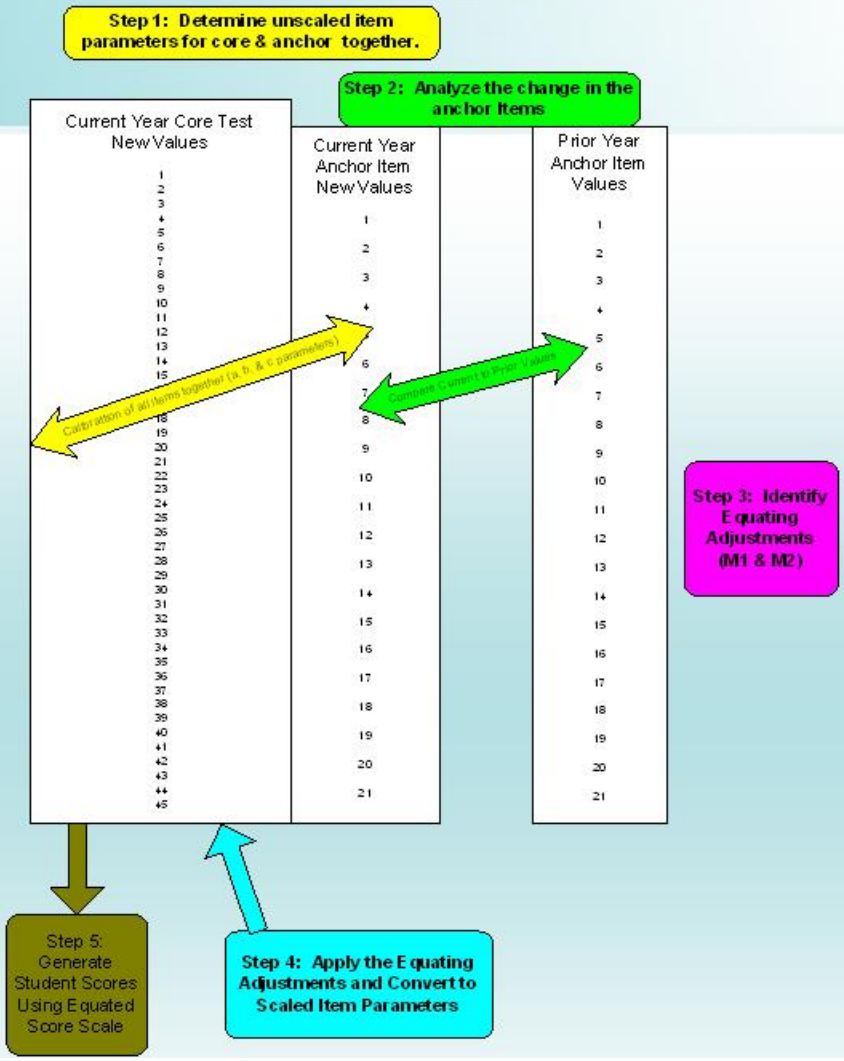
Standard Error Curve - Figure 3



Test Calibration and Equating

- Calibration - Converting from Raw Scores to IRT scores
- Equating - Making Scores Comparable Across Years
- Florida uses Item Response Theory (IRT) to score and equate FCAT results from year to year.

FCAT Calibration and Equating Process



Equating Solutions

- 2006 equating solution – anchor questions ???
- Identify a "better" equating solution
- Define "better"
- Process considerations
  - Select anchor questions
  - Follow the guidelines
  - Evaluate the quality of the anchor