

Determining the Scoring for the Grade 4 and Grade 8 English Language Arts and Mathematics Examinations

Carla Collins, Program Manager for NYSTP

William Lorié, Research Project Manager for NYSTP

Kyoko Ito, Research Manager

CTB / McGraw-Hill

Gerald DeMauro, Coordinator, Office of State Assessment

Virginia Hammer, Testing Specialist

NYSED

Session Outline

- The Structure of the Tests
- A Brief History of the Program
- Types of Scores
- Performance Standards and Standards
Performance Indices

Structure of the Tests - Outline

- Overview
- Item Types and Content Coverage
- Test Development

Structure of the Tests - Overview

- The NYS tests are designed to measure student achievement in English Language Arts (ELA) and Mathematics (MA) in grades 4 and 8.
- The NYS tests reflect New York State content standards for these grades and subject areas.
- Since the program began in 1999, the structure of the tests has remained the same.

Structure of the Tests - General

	ELA	MA
G4 MC items / points	28 / 28	30 / 30
G4 CR items / points	8 / 14	18 / 40
Total	36 / 42	48 / 70
G8 MC items / points	25 / 25	27 / 27
G8 CR items / points	9 / 18	18 / 42
Total	34 / 43	45 / 69

Structure of the Tests - G4 ELA

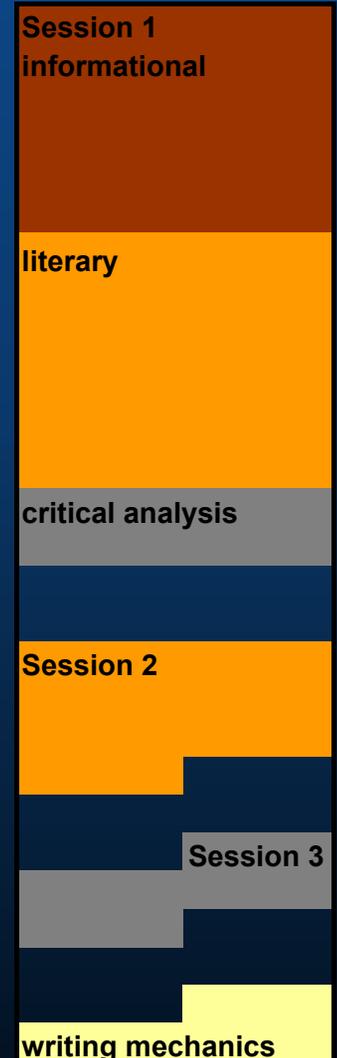
	Session 1	Session 2	Session 3
Standard 1 - informational	15 MC		
Standard 2 - literary	7 MC	2 SR, 1 ER, 1ER	
Standard 3 - critical analysis	6 MC		3 SR, 1 ER

writing mechanics cluster

listening cluster

reading cluster

independent writing



“Reading Score” = All Session 1 + 3 SRs from Session 3

ELA content coverage and subscores

Structure of the Tests - G8 ELA

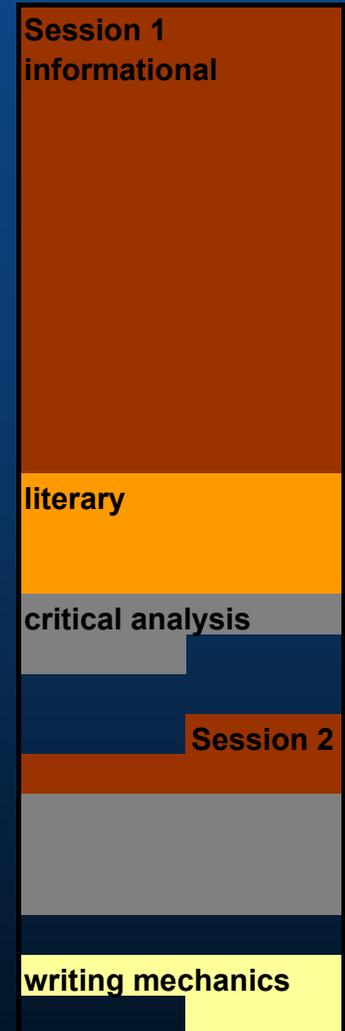
	Session 1	Session 2
Standard 1 - informational	10 MC, 3 SR, 1 ER	1 ER
Standard 2 - literary	10 MC	
Standard 3 - critical analysis	5 MC	3 SR, 1 ER

listening cluster

independent writing

reading cluster

writing mechanics cluster



No separate “Reading Score”

ELA content coverage and subscores

Structure of the Tests - G4 & 8 MA

Key Idea	Percent Emphasis	
	Grade 4	Grade 8
Mathematical Reasoning	13%	13%
Number and Numeration	20%	12%
Operations	21%	19%
Modeling / Multiple Representation	9%	16%
Measurement	16%	12%
Uncertainty	9%	9%
Patterns / Functions	13%	20%

Structure of the Tests - Development

- All field test items are reviewed extensively by New York State teachers and community representatives.
- Test forms are assembled from field test items so that they resemble earlier versions of the test and meet content and statistical constraints.

A Brief History of the Program

- 1999 - 2001
- 2002
- 2003

History of the Program - 1999-2001

➤ Operational testing

- All or most of the multiple-choice items were taken from CTB's *TerraNova* tests.
- However, only the items that reflected the State standards were selected from *TN*.
 - Custom items + cost savings.
- Comparability of different-year forms was achieved through the *TN* items.
- NYS's highest and lowest scale scores (H/LOSS) established in 1999.

History of the Program - 1999-2001, cont.

➤ Operational testing (cont.)

- Non-*TN* items were released after the administration.
- Scales scores based on response patterns (pattern scoring).

➤ Field testing

- The 1998 field-testing provided enough items through 2001.

History of the Program - 1999-2001, conc.

➤ Field testing (cont.)

- 2001 field-testing
 - 2 forms of new items (both MC and CR),
 - Separate field-testing using samples.
 - New items placed on the NYS scales.

➤ Operational form selection

- Reviews (Bias & content)
- Content coverage, item difficulties, discrimination, statistical bias, etc.

History of the Program - 2002

➤ Operational testing

- All items, MC or CR, were custom-written to the NYS content standards (no change).
- All items were released after the admin.
- Change from pattern scoring to number-correct scoring (based on a raw-score-to-scale-score conversion table).
- The conversion table based on the 2001 field-test data (Regents model).

History of the Program - 2002

➤ Field-testing

- 3 forms of new items (again both MC and CR),
- Separate field-testing using samples.
- New items placed on the NYS scales.

History of the Program - 2003

➤ Operational testing

- Same as 2002.

➤ Field testing

- Embedded administration of field-test items.
 - Each student will take a small number of FT items → Much less demanding on the students, schools, and teachers.
 - No need to acquire separate FT samples.
 - Can still release all op. items after the admin.

History of the Program - 2003

➤ Field testing (continued)

- Separate field-testing of ELA constructed-response items
 - Not breakable into small segments (clusters).
 - Usual FT using separate samples necessary.

Types of Scores - Outline

- Raw Scores
- Scale Scores
- Raw Score to Scale Score (RS-SS) Tables
 - Item Response Theory (IRT)
 - “Number Correct” Scoring
 - Pattern Scoring

Types of Scores - Raw Scores

- As noted earlier, the total number of possible raw score points for each of the four tests does not change from administration to administration.
- However, raw scores are comparable only within administration, not between administrations.

Types of Scores - Scale Scores

- Scale scores are reported on a metric that ranges from a lowest obtainable scale score (LOSS) to a highest obtainable scale score (HOSS).
- Scale scores are comparable both within and between administrations.



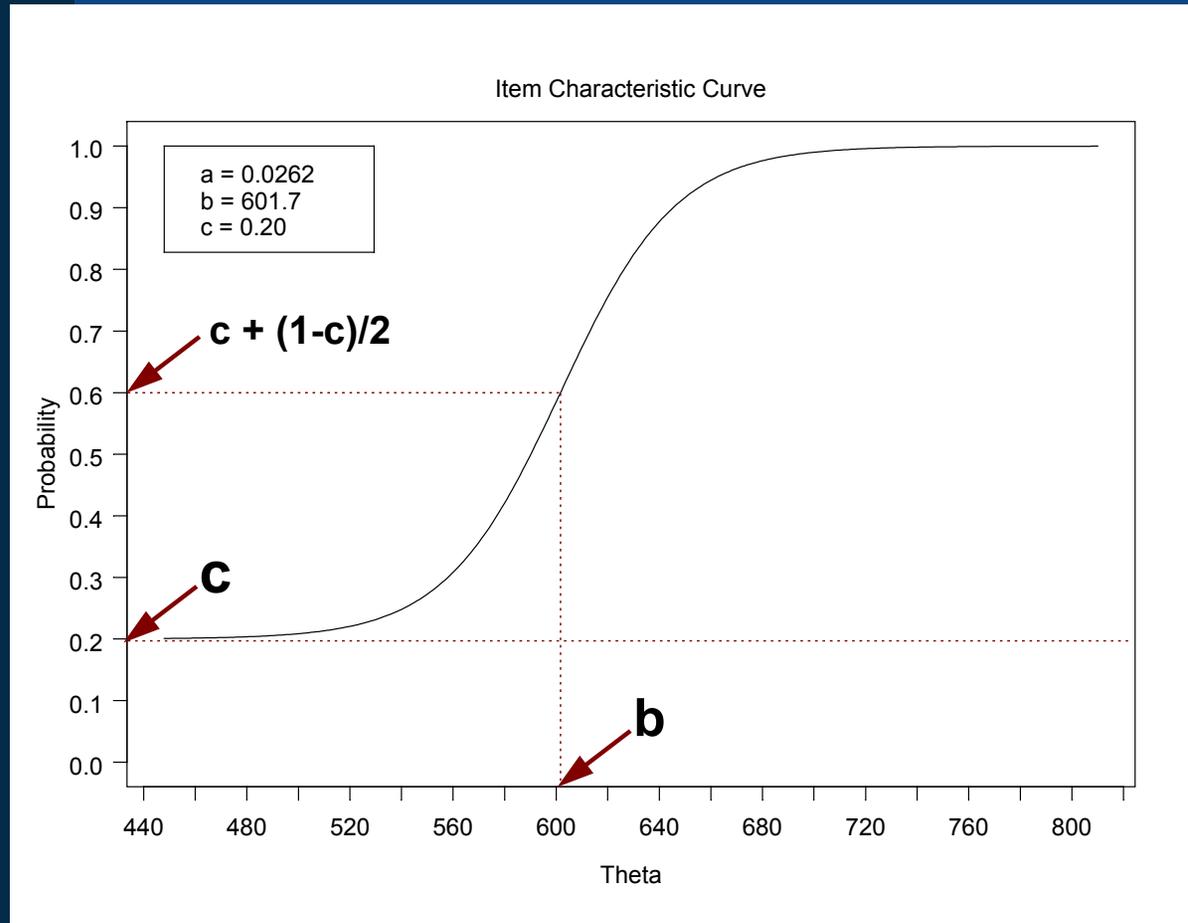
Types of Scores - RS-SS Tables

- A student's scale score is determined by looking up his or her raw score in a RS-SS conversion table.
- RS-SS tables are produced from item parameters and the HOSS and LOSS.
- RS-SS tables change from administration to administration.

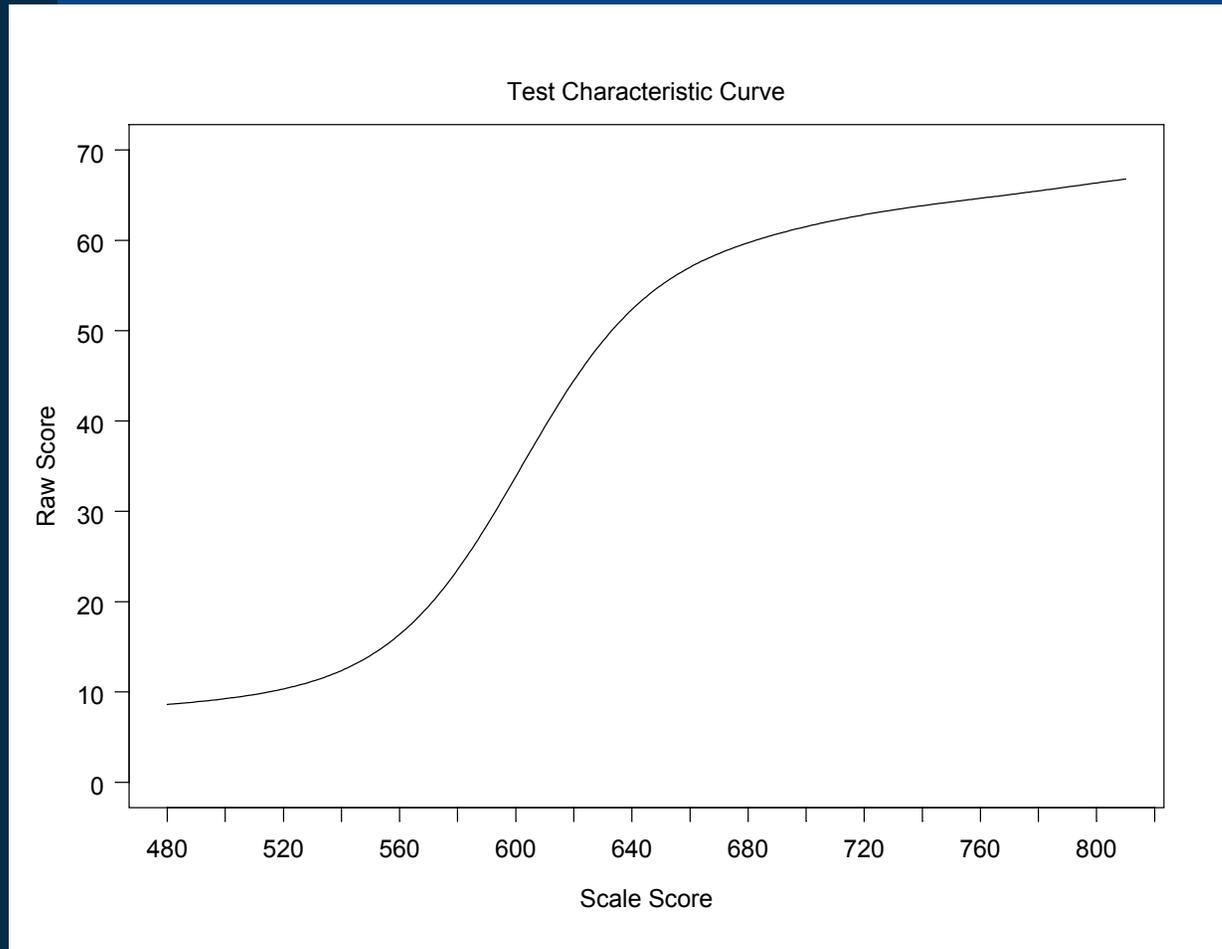
RS-SS Tables - IRT

- Item response theory (IRT) models the probability of getting an item correct as a function of underlying ability.
- In IRT, each item is described through parameters that are estimated from student data.
- With MC items, 3 parameters describe each item:
 - a - discrimination (how well does it measure ability?)
 - b - location (how hard is it?)
 - c - guessing parameter (for MC items only)

RS-SS Tables - IRT, cont.



RS-SS Tables - IRT, cont.



Types of Scores - RS-SS Tables, conc.

- By using the inverse of the TCC, an RS-SS table can be constructed for each operational form.
 - This is known as “number correct” scoring.
- This type of table was used to compute scores in 2002, and NYS will be using RS-SS tables in 2003.
- “Item pattern scoring” is not currently used for the NYS tests.

Performance Standards

- Cut scores on the NYS tests were established using the Bookmark standard setting procedure in 1999.
 - Item-based method.
 - Has been implemented successfully by CTB in a number of states.
 - Encourages content-based judgments.
- Students are classified into performance levels (1-4) based on their scale scores.

Standards Performance Indices (SPIs)

- An index of how well a student is performing on a particular standard or key idea.
- Can be interpreted as the percent of items in that standard that the student would be expected to answer correctly.
- Should be referenced to expected SPIs for further interpretability.
- Designed for within-standard within-administration diagnostic use.