1. OVERVIEW OF THE 2005 *MARYLAND SCHOOL ASSESSMENT-READING*

In 2002, the Maryland State Department of Education (MSDE) took an important step toward raising learning expectations for all students in public schools. The State Board of Education retired the *Maryland School Performance Assessment Program (MSPAP)* and adopted a new testing program known as the *Maryland School Assessment* (*MSA*). The *MSA* was based on the *Voluntary State Curriculum*, which set reasonable academic standards for what teachers were expected to teach and for what students were expected to learn in schools.

From March 1 to March 16, 2005, students in grades 3 through 8 took the 2005 *MSA* in reading (MSA-Reading).

1.1 General Overview of the 2005 MSA-Reading

The 2005 MSA-Reading was designed to provide two types of information. First, *norm-referenced* information was provided by the items from the abbreviated form of the *Stanford Achievement Test Series, Tenth Edition (SAT10)*. For third and fourth grades, for example, the *SAT10* consisted of *Word Study*, *Reading Vocabulary*, and *Reading Comprehension* items. For fifth through eighth grades, on the other hand, the *SAT10* consisted of *Reading Vocabulary* and *Reading Comprehension* items. Second, to produce *criterion-referenced* information, additional items, called augmented items, were written for the *Maryland Reading Standards* (*MRS*) in grades 3 through 8 and were organized under the three reading processes: *General Reading, Literary Reading*, and *Informational Reading*.

The 2005 MSA-Reading produced both norm-referenced and criterion-referenced scores for each student. While norm-referenced scores included only the *SAT10* items, both items selected from the *SAT10* and augmented items created for Maryland comprised criterion-referenced scores. Figure 1.1 shows a schematic of the *SAT10* and augmented items that produced these test scores.

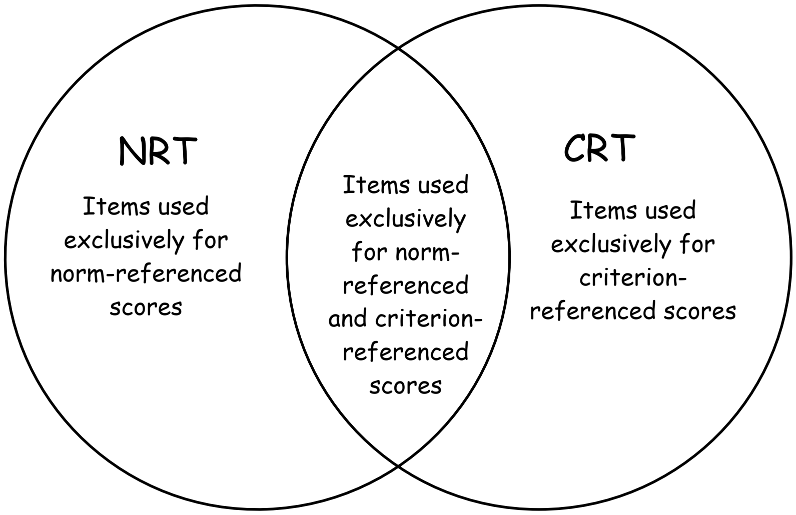

NRT
Items used exclusively for norm-referenced scores
Items used exclusively for norm-referenced and criterion-referenced scores
CRT
Items used exclusively for criterion-referenced scores

Figure 1.1 Schematic of the 2005 MSA-Reading

1.2 Purposes/Uses of the 2005 MSA-Reading

By measuring students' achievement against the new academic standards, the 2005 MSA-Reading provides two main purposes. First, the MSA-Reading was designed to inform parents, teachers, and educators of what students actually learned in schools by providing specific feedback that can be used to improve the quality of schools, classrooms, and individualized instructional programs and to model effective assessment approaches that can be used in classrooms. Second, the MSA-Reading serves as an accountability tool to measure performance levels of individual students, schools, and districts against the new academic standards.

1.3 *The Voluntary State Curriculum*

Federal law requires that states align their tests with their state content standards. The MSDE worked carefully and rigorously to construct new tests to provide a strong alignment as defined by the U.S. Department of Education.

The *Voluntary State Curriculum* (*VSC*), which defined what students should know and be able to do at each grade level, helped schools understand the standards more clearly, and included more specificity with indicators and objectives. The format of the *VSC* specified standards statements, indicators, and objectives. Standards are broad, measurable statements of what students should know and be able to do. Indicators and objectives provide more specific content knowledge and skills that are unique at each grade level.

While 100% of the standards should be tested, it was not the case that every indicator would necessarily be tested each year. Consequently, the *VSC* specified curricular indicators and objectives that contributed directly to measuring content standards, which were aligned to the *Maryland School Assessment (MSA)*.

1.4 Development and Review of the 2005 MSA-Reading

Developing the 2005 MSA-Reading was a complex process. It required a great deal of involvement from the MSDE, Harcourt, and local school systems. In addition, teachers, administrators, and content specialists from all over Maryland were recruited for different test development committees. These individuals reviewed test forms and items to ensure that they measured students' knowledge and skills fairly and without bias. Table 1.1 identifies which groups were responsible for developing the 2005 MSA-Reading.

Table 1.1 The 2005 MSA-Reading Responsibility for Test Development

Development of the 2005 MSA-Reading	Primary Responsibility
Development of Preliminary Blueprints and Item Specifications	Harcourt; MSDE; NPC
Development of Preliminary Brief Constructed Response Rubrics	MSDE
Item Writing	Harcourt
Item Review	Harcourt; MSDE; NPC; Content Review Committee
Bias Review	Harcourt; MSDE; Bias Review Committee
Construction of Field Test Forms	Harcourt; MSDE
Modification of Special Forms	Harcourt; MSDE
Review of Special Forms	MSDE
Pre-Field Test Training Workshops	Harcourt; MSDE; LEAs
Field Test Administrations	MSDE; LEAs
Construction of Operational Test Forms	Harcourt; MSDE; NPC
Review of Operational Test Forms	MSDE
Final Construction of Operational Test Forms	Harcourt; MSDE

National Psychometric Council

The National Psychometric Council (NPC) took a major role in reviewing and recommending to the MSDE on the development and implementation of the 2005 MSA-Reading program. For example, they made recommendations to the MSDE on issues, such as test blueprints, field test design, item analysis, item selection for scoring purposes, linking, equating and scaling issues, standard setting, and other relevant statistical and psychometric issues. They recommended guidelines and accommodations for students with physical disabilities or limited English proficiency. The MSDE adopted their guidelines and recommendations.

Content Review Committee

During the item review process, the Content Review Committee members were briefed on the item review process. They ensured that the MSA-Reading was appropriately difficult and fair. Committee members were either specialists in reading for test items, or experts in test construction and measurement. They represented all levels of education as well as the ethnic and social diversity of Maryland students. Committee members were from different areas of the state.

The educators' understanding of Maryland curriculum and extensive classroom experience made them a valuable source of information. They reviewed test items and forms and took a holistic view to ensure that tests were fair and balanced across reporting categories.

Bias Review Committee

In addition to the Content Review Committee, a separate Bias Review Committee examined each item on reading tests. They looked for indications of bias that would impact the performance of an identifiable group of students. Committee members discussed and, if necessary, rejected items based on gender, ethnic, religious, or geographical bias.

1.5 Test Structure of the 2005 MSA-Reading

2005 MSA-Reading Test Structure

The 2005 MSA-Reading was composed of the *SAT10* items, augmented (Maryland-specific) operational items, and field test items for future augmentation. The uniqueness of the MSA-Reading was to spiral a relatively large number of Maryland field test items into multiple test forms for each grade in test administration.

As can be seen from Table 1.2, the 2005 MSA-Reading produced four test forms for each grade, and there exist 2 operational forms within each grade. This means that Forms 1 and 3 (Form A) are identical, and Forms 2 and 4 (Form B) are identical.

Tables 1.3 and 1.4 provide information concerning the test design of NRT and CRT and the number of operational and field test items included for each test form. Tables 1.5 through 1.12 provide information concerning the number of items that contribute to each strand (e.g., General, Literary, and Informational Reading).

The descriptive statistics of each operational test form can be found in the section 1.8, Operational Test Analyses.

Table1.2 The 2005 MSA-Reading Test Structure: Grades 3 through 8

	Operational Test Item Sets		Field Test Item Sets			
	A	B	1	2	3	4
Form 1	X		X			
Form 2		X		X		
Form 3	X				X	
Form 4		X				X

Note. Total number of operational test items = 37 (33 *SR* + 4 *BCR*) items. Forms 1 and 3 (Form A) are identical, and Forms 2 and 4 (Form B) are identical in terms of operational test items.

Types of Items

The 2005 MSA-Reading contains two types of items: *selected response* (*SR*) and *brief constructed response* (*BCR*) items. *SR* items required students to select a correct answer from several alternatives. For the 2005 MSA-Reading, students selected an answer from four alternatives. Each *SR* item was scored as right or wrong.

BCR items required students to answer a question with a couple of words, a sentence, or a more elaborated way. For the 2005 MSA-Reading, these items were scored on a general rubric with maximum values between 0 and 3.

Table 1.3 The 2005 MSA-Reading Test Design: Grades 3, 5, and 8

Grade	Strand Title	*SAT10* / Augmented	Item Type	No. of Items of Each Form			
				F1	F2	F3	F4
3	Total NRT	*SAT10*	*SR*	70	70	70	70
	Word Study	*SAT10*	*SR*	20	20	20	20
	Reading Vocabulary	*SAT10*	*SR*	20	20	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	47 (10)	47 (10)	47 (10)	47 (10)
	General Reading	*SAT10*	*SR*	16	16	16	16
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	20 (10)	20 (10)	10	10
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11	21 (10)	21 (10)
5	Total NRT	*SAT10*	*SR*	50	50	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	47 (10)	47 (10)	47 (10)	47 (10)
	General Reading	*SAT10*	*SR*	15	15	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	21 (10)	21 (10)	11	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11	21 (10)	21 (10)
8	Total NRT	*SAT10*	*SR*	50	50	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	47 (10)	47 (10)	47 (10)	47 (10)
	General Reading	*SAT10*	*SR*	16	16	16	16
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	20 (10)	20 (10)	10	10
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11	21 (10)	21 (10)

Note. CRT contains *SAT10* items. *SR* items are selected response items, and *BCR* items are brief constructed response items. The number in parentheses indicates that Literary and Informational Reading include 10 field test items within each reading strand.

Table 1.4 The 2005 MSA-Reading Test Design: Grades 4, 6, and 7

Grade	Strand Title	*SAT10* / Augmented	Item Type	No. of Items of Each Form			
				F1	F2	F3	F4
4	Total NRT	*SAT10*	*SR*	70	70	70	70
	Word Study	*SAT10*	*SR*	20	20	20	20
	Reading Vocabulary	*SAT10*	*SR*	20	20	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	47 (10)	47 (10)	47 (10)	47 (10)
	General Reading	*SAT10*	*SR*	15	15	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	21 (10)	21 (10)	11	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11	21 (10)	21 (10)
6	Total NRT	*SAT10*	*SR*	50	50	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	47 (10)	47 (10)	47 (10)	47 (10)
	General Reading	*SAT10*	*SR*	15	15	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	11	21 (10)	11	21 (10)
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	21 (10)	11	21 (10)	11
7	Total NRT	*SAT10*	*SR*	50	50	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	47 (10)	47 (10)	47 (10)	47 (10)
	General Reading	*SAT10*	*SR*	15	15	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	21 (10)	21 (10)	11	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11	21 (10)	21 (10)

Note. CRT contains *SAT10* items. *SR* items are selected response items, and *BCR* items are brief constructed response items. The number in parentheses indicates that Literary and Informational Reading include 10 field test items within each reading strand.

Table 1.5 The 2005 MSA-Reading Item Distribution of Each Strand: Grade 3

	25 Common Items (*SAT10* / Maryland)			Augmented Maryland Items (12 items)								
	GR.	Lit.	Inf.	General Reading			Literary Reading			Informational Reading		
	No. of SR	No. of SR	No. of SR	No. of SR	No. of BCR	No. of Items	No. of SR	No. of *BCR*	No. of Items	No. of SR	No. of *BCR*	No. of Items
F1	16	4	5	0	0	0	4	2	6	4	2	6
F2	16	4	5	0	0	0	4	2	6	4	2	6
F3	16	4	5	0	0	0	4	2	6	4	2	6
F4	16	4	5	0	0	0	4	2	6	4	2	6

Table 1.6 The 2005 MSA-Reading Item Distribution of Each Strand: Grades 4, 6, and 7

	25 Common items (SAT10 / Maryland)			Augmented Maryland Item (12 items)								
	GR.	Lit.	Inf.	General Reading			Literary Reading			Informational Reading		
	No. of SR	No. of SR	No. of SR	No. of SR	No. of BCR	No. of Items	No. of SR	No. of *BCR*	No. of Items	No. of SR	No. of *BCR*	No. of Items
F1	15	5	5	0	0	0	4	2	6	4	2	6
F2	15	5	5	0	0	0	4	2	6	4	2	6
F3	15	5	5	0	0	0	4	2	6	4	2	6
F4	15	5	5	0	0	0	4	2	6	4	2	6

Table 1.7 The 2005 MSA-Reading Item Distribution of Each Strand: Grade 5

	25 Common items (*SAT10* / Maryland)			Augmented Maryland Item (12 items)								
	GR.	Lit.	Inf.	General Reading			Literary Reading			Informational Reading		
	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR
F1	15	5	5	0	0	0	4	2	6	4	2	6
F2	15	5	5	0	0	0	4	2	6	4	2	6
F3	15	5	5	0	0	0	4	2	6	4	2	6
F4	15	5	5	0	0	0	4	2	6	4	2	6

Table 1.8 The 2005 MSA-Reading Item Distribution of Each Strand: Grade 8

	25 Common Items (*SAT10* / Maryland)			Augmented Maryland Items (12 items)								
	GR.	Lit.	Inf.	General Reading			Literary Reading			Informational Reading		
	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR	No. of SR
F1	16	4	5	0	0	0	4	2	6	4	2	6
F2	16	4	5	0	0	0	4	2	6	4	2	6
F3	16	4	5	0	0	0	4	2	6	4	2	6
F4	16	4	5	0	0	0	4	2	6	4	2	6

Table 1.9 The 2005 MSA-Reading Total and Strand Scores: Grade 3

	Total and Each Strand Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form 1	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 2	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 3	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 4	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45

Table 1.10 The 2005 MSA-Reading Total and Strand Scores: Grades 4, 6, and 7

	Total and Each Strand Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form 1	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 2	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 3	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 4	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 SR + 6 BCR)	45

Table 1.11 The 2005 MSA-Reading Total and Strand Scores: Grade 5

	Total and Each Strand Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form 1	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 MC + 6 BCR)	45
Form 2	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 MC + 6 BCR)	45
Form 3	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 MC + 6 BCR)	45
Form 4	15 (15 SR)	15 (9 SR + 6 BCR)	15 (9 MC + 6 BCR)	45

Table 1.12 The 2005 MSA-Reading Total and Strand Scores: Grade 8

	Total and Each Strand Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form 1	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 2	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 3	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45
Form 4	16 (16 SR)	14 (8 SR + 6 BCR)	15 (9 SR + 6 BCR)	45

1.6 Test Administration

Test Administration Preparation and Materials

Pre-test workshops were held in Baltimore for all Local Accountability Coordinators in Maryland prior to the test administration. These workshops provided the representatives of all the local school divisions with an overview of the test's content, security expectations, and procedures for completing the answer documents. They also considered the receipt, distribution, and return of test materials.

For the test examiner, Harcourt provided the following materials:

- Examiner's Manual
- One set of pre-printed student ID labels and one set of generic ID labels for those students who do not have a pre-printed label or who have one with incorrect information. The generic student ID label is to be used in the event that pre-printed labels are damaged. The pre-printed or generic labels are placed on the Answer Book in the area that says "Place Pre-ID Label Here." The label must be applied prior to testing by or under the supervision of the STC.
- Paper bands for used Answer Books
- Student Roster

For each student, the following materials were provided by Harcourt:

- Test Book
- Answer Book

  Note: For Grade 3, the Test Book and Answer Book are combined into a single book.

For each student, the following additional materials were provided by school or student:

- Two No.2 pencils with erasers
- Scratch paper for pre-writing

Each classroom used for the assessment will also need the following additional materials:

- A sign for the door that says "Testing: Do not Disturb"
- A digital clock or a watch, or clock with a second hand

Two test-related manuals were developed for the administration of the 2005 MSA-Reading: Test Administration and Coordination Manual (TACM) and Examiner's Manual for Test Administration (EMTA). For the 2005 testing season, the TACM contents pertaining to Harcourt were developed by Harcourt and produced by MSDE. This manual provided Local Accountability Coordinators (LACs) and building level School Test Coordinators (STCs) with information about the administration, packaging, and return of test materials. The TACM also described any issues specific to grades 3 through 8. One TACM was produced for all administrations in grades 3 through 8. The TACM was distributed one per school at the pre-test workshops and was again included in the shipping materials.

The EMTA was developed for each grade by Harcourt and provided directions for administering the 2005 MSA-Reading at each grade level. It contains information with regards to general information of the test, before testing, during testing, and after testing.

Test Administration Schedule

Specific dates were designated for each content area test. For the 2005 MSA-Reading, the primary testing days were as follows:

• Test materials delivered to schools (Examiner's Manual and Test Books)	February 9 - February 15, 2005
• Reading Primary Testing Window	March 1 - March 10, 2005
• Make-up Testing Window	March 11 - March 16, 2005

If a student was absent on the testing days, a make-up test was administered on any two consecutive days within testing window. If a school had an unscheduled closing or delayed opening that prohibited the administration from occurring on the scheduled testing dates, the STCs were consulted with LACs to determine the testing schedule to be followed.

During the administration of the 2005 MSA-Reading, the MSDE had testing monitors in selected schools observing administration procedures and testing conditions. All monitors had identification cards for security purposes. There were no prior notification of which schools would be monitored, but monitors followed local procedures for reporting to the school's main office and giving proper notification that an MSDE monitor is in the building.

The following sessions were scheduled at any convenient time during the school day, but testing had to be scheduled to allow sufficient time to complete the test. Table 1.13 shows timing sessions allowed for the 2005 MSA-Reading.

Table 1.13 The 2005 MSA-Reading Timing Sessions: Grades 3 through 8

Grade	Form	Session					
		1	2	3	4	5	6
3 & 4	1-4	Q1-Q20 20 min.	Q21-Q40 18 min.	Q41-Q70 45 min.	Q71-Q76 30 min.	Q77-Q82 30 min.	Q83-Q92 35 min.
5 through 8	1-4	Q1-Q20 20 min.	Q21-Q50 45 min.	Q51-Q60 35 min.	Q61-Q66 35 min.	Q67-Q72 35 min.	X

Student Participation

All students in grades 3 through 8 must participate in the 2005 MSA-Reading. The only exception was that students with severe cognitive disabilities were assessed by the *Alternate Maryland School Assessment* (ALT-MSA) instead of the regular MSA-Reading.

Testing Accommodations

Testing accommodations for Special Education students, English Language Learners (ELL), and students with disabilities covered under Section 504 had to be approved and documented according to the procedures and requirements outlined in the document entitled "Requirements for Accommodating, Excusing, and Exempting Students in Maryland Assessment Programs" (the "Accommodations Document"), as revised August 18, 2003. (A copy of the most recent edition of this document is available electronically on the LAC and STC web pages at http://docushare.msde.state.md.us.

No accommodations may be made for students merely because they were members of an instructional group. Any accommodation had to be based on individual needs and not on a category of disability area, level of instruction, environment, or other group characteristics. Responsibility for confirming the need and appropriateness of an accommodation rested with the LAC and school-based staff involved with each student's instructional program. A master list of all students and their accommodations had to be maintained by the principal and submitted to the LAC, who provided a copy to the MSDE upon request. Please refer to Section 1 of the 2005 TACM for further information regarding testing accommodations.

Large-Print and Braille Test Books and Kurzweil™ Test Forms on CD

The 2005 MSA-Reading was administered to those requiring (1) large-print Student Test Books and Answer Books or (2) Braille Test Books, or (3) Kurzweil™ Test Forms on CD. For large-print and Braille Test Books, and Kurzweil™ Test Forms on CD, student responses were transcribed into the regular Answer Book following testing. The pre-printed student ID label was affixed to the regular Answer Book containing the transcribed responses, not the large-print Answer Book or Braille books. If there is no pre-printed student ID label, a generic ID label was applied to the regular Answer Book containing the transcribed responses.

Once the student responses had been transcribed, the transcribe Answer Book was returned for scoring with the regular material. Specific packing instructions are provided in the TACM in section 4 and 7.

Verbatim Reading Accommodation and Kurzweil™ Test Forms on CD

Students who have a verbatim reading accommodation documented in their Individual Education Plan (IEP), ELL Plan, or Section 504 Plan and who receive that accommodation in regular instruction may receive the accommodation on the 2005 MSA-Reading. The accommodation may be provided by a live reader or through technology. If technology is used to provide the verbatim reading accommodation, the software used must be Kurzweil reading software, and official, secure electronic copies of the test must be ordered through the LAC directly from MSDE. MSDE encourages the use of Kurzweil™ software to ensure uniformity in the delivery of the verbatim reading accommodation throughout the state.

Students using Kurzweil™ software must have familiarity with its operation prior to the test administration. Please consult with LAC for the further information on Kurzweil™ and the verbatim reading accommodation.

Security of Test Materials

The following code of ethnics conforms to the Standards for Educational and Psychological Testing developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (Harcourt, 2005):

> It is breach of professional ethics for school personnel to provide verbal or nonverbal clues or answers, teach items on the test, share writing prompts, coach, hint, or in any way influence a student's performance during the testing situation. A breach of ethics may result in invalidation of test results and local education agency or MSDE disciplinary action. (p. 7)

The Test Books and all used Answer Books for the 2005 MSA-Reading were confidential and kept secure at all times. Unauthorized use, duplication, or reproduction of any or all portions of the assessment was prohibited, which is reflected by the following statement (Harcourt, 2005):

> Violation of security can result in prosecution and/or penalties as imposed by the Maryland State Board of Education and/or State Superintendent of Schools in accordance with the COMAR 13A.03.04 and 13A.12.05. (p. 7)

All materials were treated as confidential and placed in locked areas. Secure and non-secure test materials were as follows:

- Secure materials: Test Books and Answer Books
- Non-secure materials: Test Administration and Coordination Manual, Examiner's Manual for Test Administration, unused Answer Books, return address labels, pre-printed student ID labels, and instructions for applying ID labels

Distribution of Materials

Different test forms were administered to students in each classroom participating in reading tests, and each test form was identified by a cover of a different color and number. In addition, the Test Books and Answer Books were spiraled within a classroom. Each student must receive a Test Book and Answer Book that are the same color and have the same form number on the cover (except for Grade 3 where the Test Book and Answer Book are combined in the same document).

1.7 Scoring Procedures

Students' responses to *SR* items were machine-scored, and their responses to *BCR* items were individually read and scored by Harcourt in San Antonio.

Once received by Harcourt, Answer Books were scanned into an electronic imaging system so that the information necessary to score responses was captured and converted into an electronic format. Students' identification and demographic information, school information, and answers to *SR* items were converted to alphanumeric format; hand-written responses were captured in digital image format.

Machine-Scored Items

After students' responses to *SR* items were converted to text format, the scoring key was applied to the captured item responses. Correct answers were assigned a score of one point; incorrect answers were assigned zero points. Students' responses with multiple marks and blank responses (omits) were also assigned zero points.

Hand-Scored Items

Answer Books were scanned into the electronic imaging system, allowing scorers to score these responses online at all scoring sites while maintaining the live documents at the contractor's facility. The imaging system randomly distributed responses, ensuring no one scorer scored a disproportionate number of responses from any one school. This online scoring system maintained a database of actual student responses and the scores associated with those responses. An off-site backup of all images and scores was maintained as well to guard against potential loss of data and images due to system failure. The system also provided continuous, up-to-date monitoring of all scoring activities.

Scorer Qualifications

BCR items were scored by scorers who were trained to stringent requirements and procedures. All applicants for *MSA* scorer positions were required to provide resumes and documentation of completed higher education. They were required to have earned a four-year college degree or higher. As part of the initial recruiting and screening process, applicants responded to a writing prompt and several content specific, open-response questions. The writing sample ensured that all applicants were fluent in writing and reading standard English. If successful on the preliminary screening, applicants participated in introductory workshops. The purpose of these workshops was to familiarize the applicants with general processes and procedures for scoring performance assessments and to provide a final screening activity before they were added to the overall pool of potential scorers for the *MSA* project.

From that pool, potential scorers were assigned to the *MSA* project. *MSA*-specific training and qualifying consisted of having each scorer respond to actual *MSA* items or prompts prior to actual training. Using anchor papers and training sets, scorers then internalized the standards and the scoring scale for the item they were to score and were given qualifying sets. Those who met the qualifying standard were then allowed to score.

Methodology for Scoring the 2005 MSA-Reading *BCR* Items

For the *MSA*, each domain/level had a room director to direct scoring activities. The room director worked closely with the training supervisor and the content training specialist. The room director conducted training to ensure that scorers became experts in their scoring assignment. The main job of the room director was to oversee the actual scoring of the papers, acting as the decision maker for situations in which questions arise during the scoring process. The room director was also responsible for the quality of the scoring within the room. For the MSA-Reading program, those who served as room directors were usually active members of the training material development team, worked with MSDE staff and selected Maryland teachers to finalize scoring guides and training materials, and benchmarked student work.

For each item, scorers were trained to use the same scale to ensure accurate, consistent, and reliable scoring. All *BCR* items received a 0-3 score point range from two independent scorers. Equal or adjacent scores were acceptable. Readers were trained on and scored one item at a time. If the two readers did not assign equal or adjacent scores, the response was routed to a team leader for a third, independent reading to resolve the anomalous scores.

The read-behind application was also used to monitor reader performance. The team leader was provided a random selection of responses from each reader, distributed randomly across all readers. Although it could be tailored for each reader, by default, three percent of all responses scored appeared in the read-behind application. The team leader could agree with the scores and confirm them, disagree and send them back to the reader, or change them.

Training for Scoring Accuracy

The key to accurate scoring of *BCR* items is to train scorers appropriately. The following procedures were employed for training *MSA* project scorers.

Project-specific team leader training was conducted in the days immediately preceding scoring. Team leaders experienced in the scoring process helped train and retrain their team members. In addition, the logistics of the scoring sessions and the routines for resolution reading were discussed. All team leaders were also required to meet the qualifying standards set for the project. These standards were determined in conjunction with the MSDE.

Scorer training for *MSA* scoring began with an overview of the project and continued with the reading and discussion of selected student responses. The training utilized anchor sets, training sets, and qualifying sets, all of which contained MSDE reviewed and approved responses in addition to the *MSA* scoring rubric. Emphasis was placed on the scorer’s understanding of how the responses differed from one another in quality and how each response represented the description of its score point as generalized in the scoring guidelines.

Inter-Rater Agreement

The scoring system generated many different kinds of internal monitoring reports that enabled accuracy of *MSA* scoring to be monitored. Teams produced the reports listing team scorers and providing the results of their scoring on an ongoing basis. Information on these reports included the number of responses read by the scorers during the period, the number and percentage of invalid responses (i.e., off-topic or blank responses, refusals to respond, responses in foreign languages), and the number of responses for which there was a subsequent reading. To illustrate, the number of responses with a second reading provided data that allowed for reporting the number and percentage of responses with perfect agreement, the number and percentage of responses for which the first scorer was a point lower than the second scorer, the number and percentage of responses for which the first scorer was a point higher than the second scorer, and the number and percentage of responses differing by more than one score point.

In addition to the scorer reports described above, a daily order status report was generated each day to monitor the progress, logistically, of the overall scoring process through the system. This report was given at the individual, team, and room levels, and showed, by order of completion and prompt, the number and percentage of responses for which first and second (check score) readings were required and completed for each item. These reports were available to team leaders, room directors, and training supervisors. They were also calculated and reported cumulatively for the day, the week, and the entire project. All reports were made available to the *MSA* supervisor every morning, and several of these monitoring reports could be called up online anytime throughout the scoring day. Statistical summaries of inter-rater reliability can be found in section 3.4.

1.8 Operational Test Analyses

To ascertain whether or not two operational test forms generated statistically significant discrepancy, descriptive statistics, such as mean (*M*), standard deviation (*SD*) were calculated for the *SAT10* common items (e.g., 25 items included in the operational test forms). The statistical results of the two test forms were almost identical across all grades, as can be seen from Table 1.14.

Table 1.14 The 2005 MSA-Reading Common Item Descriptive Statistics: Grades 3, 5, and 8

Grade	Form	No. of Items	*N*	*M*	*SD*
3	A	25	27,284	17.98	4.78
	B	25	27,620	18.08	4.77
4	A	25	27,790	19.52	4.05
	B	25	28,017	19.53	4.01
5	A	25	28.432	17.57	4.66
	B	25	28,806	17.62	4.61
6	A	25	29,282	18.22	4.91
	B	25	29,417	18.33	4.88
7	A	25	30,423	17.57	4.84
	B	25	30,435	17.56	4.85
8	A	25	30,685	17.20	4.59
	B	25	30,903	17.20	4.59

Note. Form A designates the operational portion of Forms 1 and 3, which is identical. Form B designates the operational portion of Forms 2 and 4, which is identical.

Common Item P-Value Check

Tables 1.15 through 1.20 and Figures 1.2 through 1.7 provide information about how much the item difficulty (p-value) of the *SAT10* common items changed in consecutive years. The general conclusion can be drawn from the results that most of the p-values in Year 2005 increased compared to those in Year 2004 across all grades except for grades 6 through 8.

Table 1.15 Common Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 3

Item Number	Sequence Number	Y04 FA	Y04 FB	Y05 FA	Y05 FB
2	1	0.93	0.93	0.93	0.94
5	2	0.85	0.85	0.89	0.89
6	3	0.65	0.65	0.68	0.68
9	4	0.82	0.83	0.85	0.86
11	5	0.54	0.55	0.64	0.65
15	6	0.85	0.85	0.86	0.86
18	7	0.63	0.63	0.70	0.71
20	8	0.30	0.30	0.39	0.40
23	9	0.65	0.64	0.67	0.66
30	10	0.46	0.46	0.49	0.48
31	11	0.68	0.68	0.71	0.72
32	12	0.64	0.65	0.70	0.70
34	13	0.72	0.73	0.77	0.77
41	14	0.69	0.69	0.76	0.76
44	15	0.90	0.90	0.91	0.92
49	16	0.65	0.66	0.65	0.66
55	17	0.65	0.65	0.68	0.68
56	18	0.45	0.45	0.48	0.49
57	19	0.82	0.82	0.81	0.81
58	20	0.88	0.88	0.90	0.90
59	21	0.81	0.81	0.84	0.84
61	22	0.54	0.54	0.56	0.57
68	23	0.73	0.73	0.76	0.77
69	24	0.78	0.77	0.80	0.80
70	25	0.60	0.60	0.63	0.64

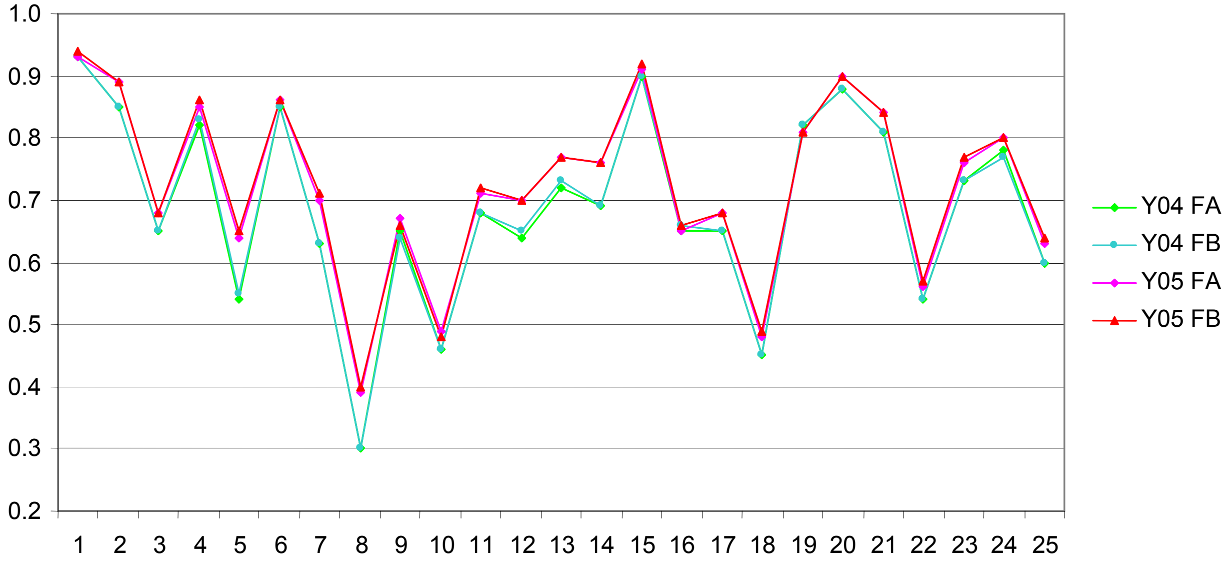


Figure 1.2 Common Item P-value Plots for Year 2004 vs. Year 2005: Grade 3

Table 1.16 Common Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 4

Item Number	Sequence Number	Y04 F1	Y04 F2	Y04 F3	Y04 F4	Y04 F5	Y04 F6	Y05 FA	Y05 FB
1	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
4	2	0.92	0.92	0.92	0.93	0.92	0.92	0.93	0.94
9	3	0.74	0.74	0.75	0.75	0.75	0.75	0.80	0.81
10	3	0.89	0.89	0.89	0.89	0.88	0.89	0.89	0.90
18	5	0.73	0.73	0.72	0.73	0.73	0.72	0.76	0.78
23	6	0.82	0.81	0.82	0.82	0.82	0.82	0.83	0.85
24	7	0.76	0.76	0.76	0.76	0.76	0.76	0.77	0.79
29	8	0.89	0.89	0.91	0.90	0.91	0.90	0.91	0.92
35	9	0.81	0.82	0.82	0.83	0.82	0.83	0.83	0.85
38	10	0.68	0.68	0.70	0.69	0.70	0.70	0.70	0.71
41	11	0.78	0.78	0.78	0.78	0.78	0.78	0.79	0.80
42	12	0.68	0.68	0.68	0.68	0.69	0.68	0.76	0.78
43	13	0.82	0.82	0.83	0.82	0.83	0.83	0.84	0.86
44	14	0.78	0.77	0.78	0.77	0.78	0.78	0.81	0.82
45	15	0.44	0.44	0.43	0.42	0.45	0.44	0.43	0.45
46	16	0.93	0.94	0.93	0.94	0.93	0.93	0.94	0.95
47	17	0.79	0.80	0.80	0.79	0.80	0.79	0.81	0.83
50	18	0.80	0.80	0.80	0.81	0.80	0.81	0.81	0.83
51	19	0.93	0.92	0.93	0.93	0.93	0.93	0.94	0.95
52	20	0.59	0.59	0.59	0.59	0.58	0.58	0.61	0.61
53	21	0.48	0.49	0.50	0.49	0.50	0.49	0.51	0.52
54	22	0.37	0.37	0.37	0.37	0.36	0.36	0.38	0.37
55	23	0.90	0.90	0.90	0.91	0.91	0.91	0.91	0.91
62	24	0.75	0.75	0.76	0.76	0.76	0.76	0.76	0.78
64	25	0.60	0.60	0.61	0.61	0.61	0.61	0.62	0.65

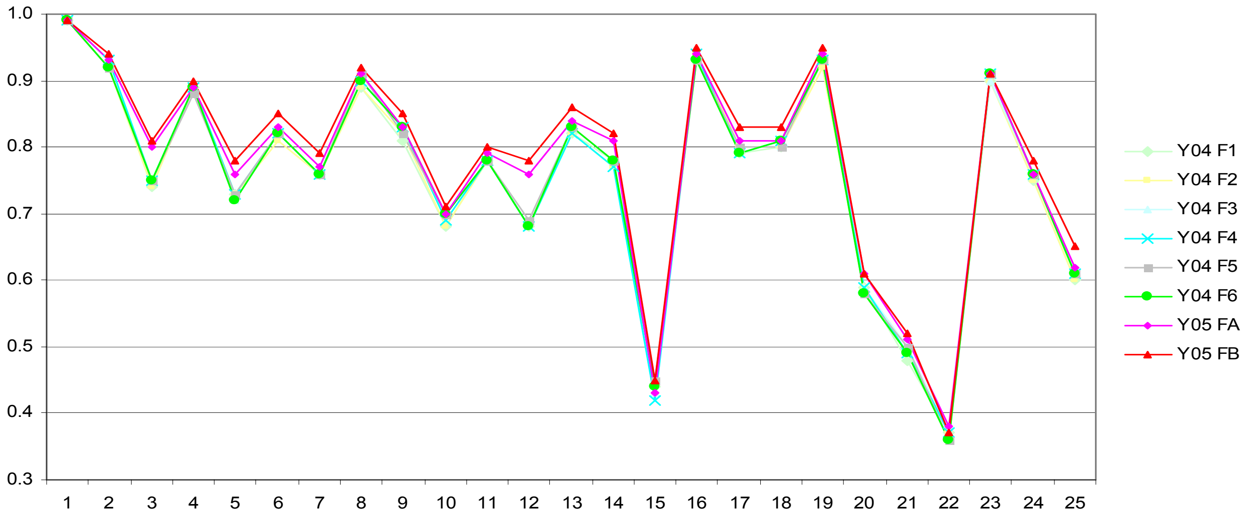


Figure 1.3 Common Item P-value Plots for Year 2004 vs. Year 2005: Grade 4

Table 1.17 Common Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 5

Item Number	Sequence Number	Y04 FA	Y04 FB	Y05 FA	Y05 FB
4	1	0.57	0.57	0.58	0.58
5	2	0.54	0.54	0.55	0.55
6	3	0.59	0.59	0.61	0.61
9	4	0.88	0.88	0.91	0.91
10	5	0.85	0.85	0.88	0.88
11	6	0.83	0.84	0.84	0.84
13	7	0.85	0.85	0.85	0.85
16	8	0.79	0.79	0.82	0.82
17	9	0.77	0.77	0.79	0.79
19	10	0.73	0.73	0.75	0.76
21	11	0.81	0.81	0.81	0.81
23	12	0.56	0.57	0.59	0.59
25	13	0.72	0.71	0.73	0.72
26	14	0.71	0.72	0.71	0.72
28	15	0.60	0.60	0.54	0.54
31	16	0.57	0.57	0.58	0.59
32	17	0.72	0.71	0.70	0.69
33	18	0.79	0.79	0.80	0.80
34	19	0.39	0.39	0.38	0.38
35	20	0.63	0.62	0.65	0.66
37	21	0.69	0.70	0.68	0.67
41	22	0.76	0.76	0.76	0.76
44	23	0.63	0.62	0.64	0.63
45	24	0.55	0.55	0.55	0.56
49	25	0.83	0.84	0.85	0.84

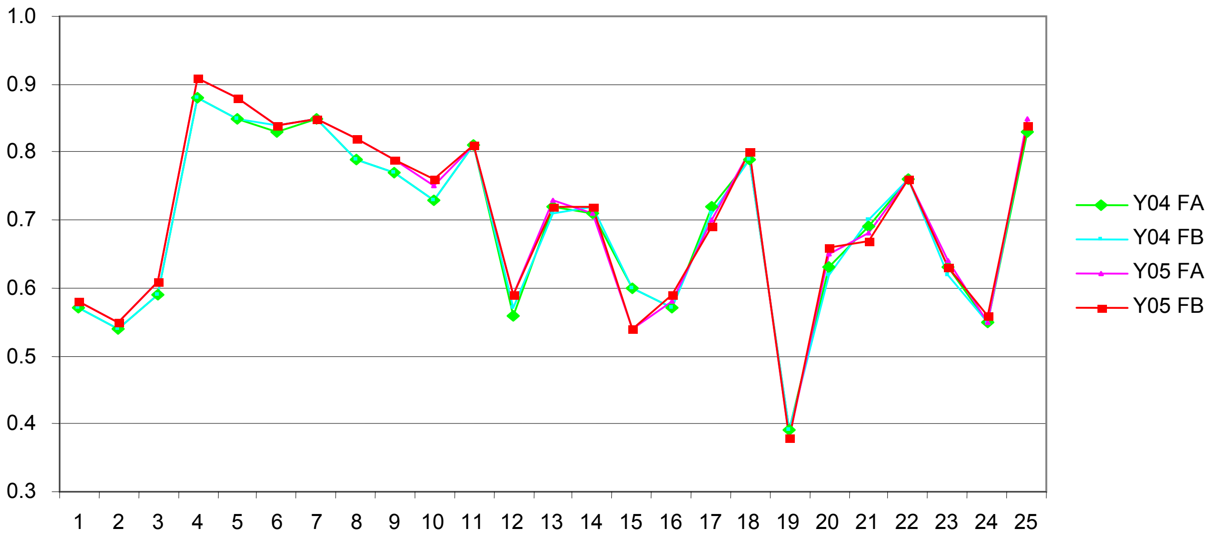


Figure 1.4 Common Item P-value Plots for Year 2004 vs. Year 2005: Grade 5

Table 1.18 Common Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 6

Item Number	Sequence Number	Y04 F1	Y04 F2	Y04 F3	Y04 F4	Y04 F5	Y04 F6	Y05 FA	Y05 FB
1	1	0.80	0.80	0.80	0.80	0.80	0.80	0.78	0.79
5	2	0.54	0.53	0.53	0.54	0.54	0.52	0.52	0.53
8	3	0.56	0.58	0.58	0.57	0.58	0.56	0.59	0.60
9	3	0.90	0.91	0.92	0.92	0.91	0.92	0.91	0.92
10	5	0.75	0.75	0.75	0.75	0.75	0.75	0.74	0.74
14	6	0.74	0.74	0.73	0.74	0.73	0.73	0.73	0.74
16	7	0.79	0.79	0.79	0.78	0.79	0.78	0.78	0.80
18	8	0.84	0.84	0.84	0.84	0.84	0.84	0.83	0.84
21	9	0.89	0.91	0.90	0.91	0.91	0.90	0.88	0.89
22	10	0.78	0.78	0.78	0.78	0.78	0.78	0.77	0.78
23	11	0.71	0.71	0.72	0.71	0.71	0.71	0.69	0.71
24	12	0.70	0.70	0.71	0.70	0.70	0.69	0.69	0.70
25	13	0.65	0.64	0.66	0.65	0.65	0.64	0.64	0.65
28	14	0.75	0.76	0.75	0.76	0.76	0.75	0.77	0.78
29	15	0.66	0.65	0.65	0.65	0.66	0.65	0.64	0.65
30	16	0.63	0.64	0.63	0.63	0.64	0.64	0.65	0.66
32	17	0.87	0.87	0.88	0.87	0.87	0.86	0.86	0.86
33	18	0.34	0.35	0.34	0.35	0.35	0.36	0.34	0.34
34	19	0.84	0.84	0.84	0.85	0.84	0.84	0.83	0.84
35	20	0.60	0.60	0.61	0.60	0.60	0.60	0.62	0.63
36	21	0.78	0.77	0.78	0.78	0.78	0.77	0.78	0.79
37	22	0.82	0.83	0.83	0.83	0.84	0.83	0.81	0.82
38	23	0.58	0.59	0.58	0.59	0.58	0.58	0.59	0.60
39	24	0.89	0.89	0.89	0.89	0.90	0.89	0.87	0.88
40	25	0.78	0.79	0.79	0.78	0.80	0.78	0.76	0.77

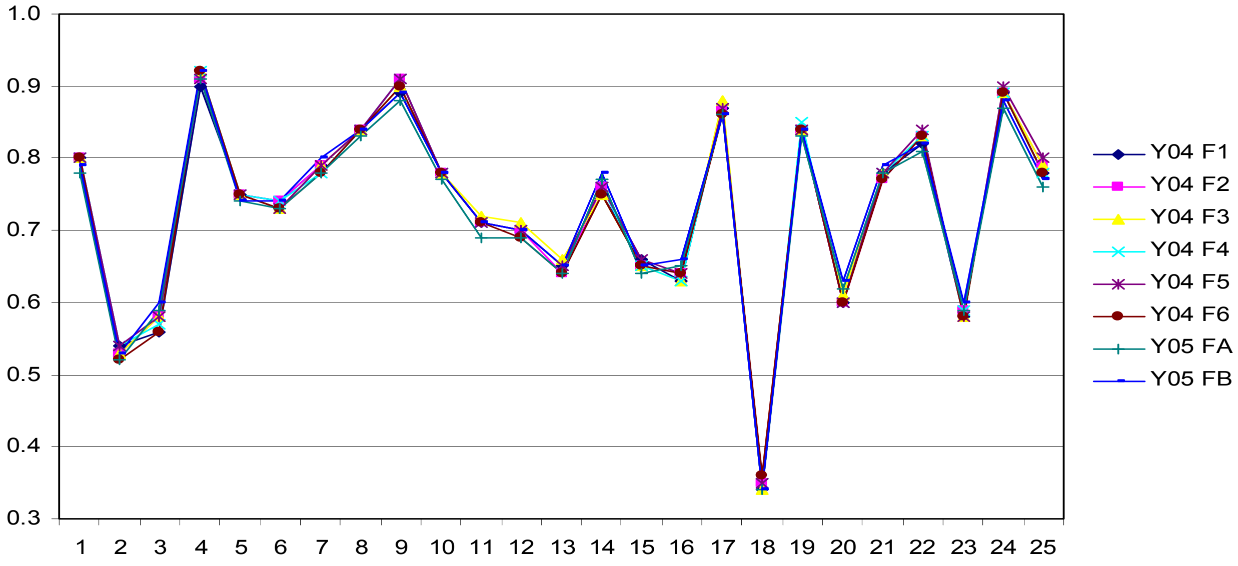


Figure 1.5 Common Item P-value Plots for Year 2004 vs. Year 2005: Grade 6

Table 1.19 Common Items P-Value Comparison for Year 2004 vs. Year 2005: Grade 7

Item Number	Sequence Number	Y04 F1	Y04 F2	Y04 F3	Y04 F4	Y04 F5	Y04 F6	Y05 FA	Y05 FB
1	1	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.89
3	2	0.84	0.84	0.84	0.85	0.85	0.84	0.84	0.84
6	3	0.44	0.44	0.45	0.46	0.44	0.44	0.44	0.45
8	3	0.46	0.46	0.47	0.46	0.46	0.46	0.42	0.42
10	5	0.61	0.60	0.61	0.62	0.61	0.61	0.60	0.60
14	6	0.67	0.67	0.67	0.68	0.66	0.67	0.68	0.68
16	7	0.65	0.65	0.66	0.64	0.65	0.64	0.64	0.63
20	8	0.85	0.84	0.86	0.86	0.85	0.85	0.84	0.84
22	9	0.88	0.88	0.89	0.89	0.89	0.88	0.88	0.88
23	10	0.56	0.56	0.57	0.57	0.56	0.56	0.54	0.54
26	11	0.75	0.75	0.75	0.76	0.76	0.75	0.77	0.76
27	12	0.50	0.51	0.51	0.51	0.50	0.49	0.54	0.53
28	13	0.64	0.64	0.65	0.65	0.64	0.63	0.64	0.63
31	14	0.57	0.58	0.57	0.57	0.57	0.57	0.58	0.57
32	15	0.86	0.87	0.87	0.87	0.86	0.85	0.86	0.85
33	16	0.58	0.58	0.58	0.59	0.57	0.58	0.61	0.61
36	17	0.90	0.91	0.91	0.91	0.90	0.90	0.90	0.89
37	18	0.73	0.73	0.72	0.74	0.71	0.72	0.72	0.71
38	19	0.77	0.77	0.78	0.77	0.77	0.77	0.76	0.75
39	20	0.65	0.63	0.64	0.65	0.65	0.64	0.61	0.61
40	21	0.87	0.87	0.88	0.88	0.87	0.87	0.86	0.86
41	22	0.77	0.77	0.79	0.78	0.78	0.78	0.78	0.77
42	23	0.74	0.74	0.74	0.74	0.74	0.74	0.73	0.72
43	24	0.74	0.74	0.75	0.74	0.74	0.73	0.73	0.73
44	25	0.69	0.69	0.70	0.70	0.69	0.69	0.68	0.68

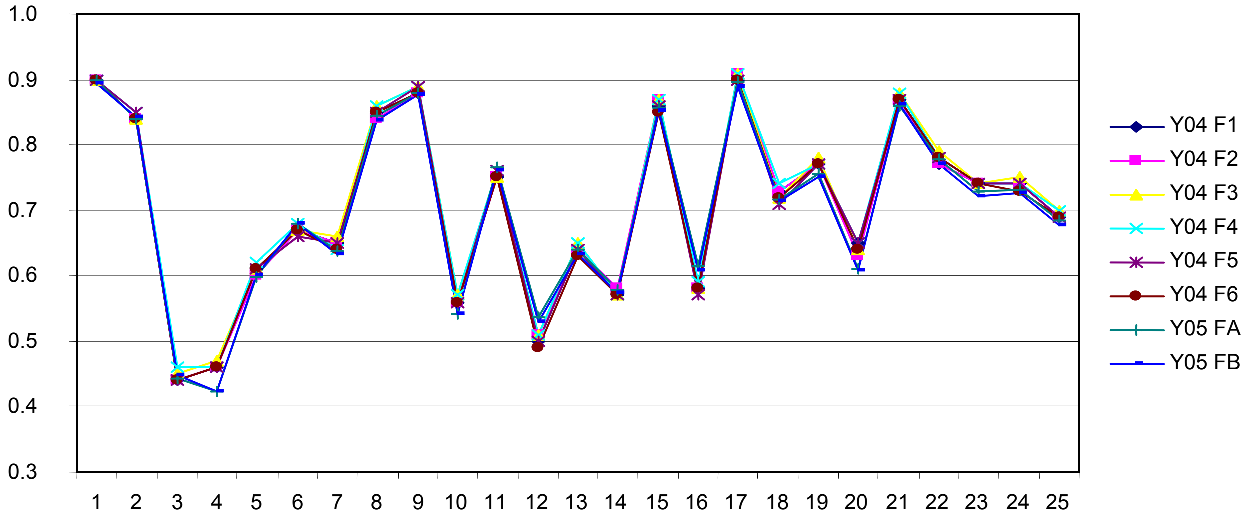


Figure 1.6 Common Item P-value Plots for Year 2004 vs. Year 2005: Grade 7

Table 1.20 Common Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 8

Item Number	Sequence Number	Y04 FA	Y04 FB	Y05 FA	Y05 FB
3	1	0.66	0.66	0.65	0.65
6	2	0.50	0.51	0.50	0.50
8	3	0.61	0.61	0.58	0.57
9	4	0.91	0.91	0.92	0.92
22	5	0.97	0.97	0.97	0.97
23	6	0.57	0.57	0.56	0.57
24	7	0.57	0.57	0.57	0.57
25	8	0.81	0.81	0.80	0.80
26	9	0.63	0.63	0.64	0.63
29	10	0.73	0.74	0.72	0.72
30	11	0.56	0.56	0.60	0.59
31	12	0.66	0.66	0.65	0.64
32	13	0.49	0.49	0.49	0.50
33	14	0.65	0.65	0.65	0.64
35	15	0.77	0.77	0.78	0.78
36	16	0.57	0.56	0.55	0.54
37	17	0.74	0.74	0.73	0.73
38	18	0.78	0.78	0.76	0.76
39	19	0.50	0.50	0.52	0.52
41	20	0.83	0.83	0.83	0.83
44	21	0.73	0.73	0.74	0.74
46	22	0.81	0.81	0.80	0.80
48	23	0.73	0.74	0.73	0.72
49	24	0.75	0.75	0.74	0.74
50	25	0.76	0.75	0.74	0.74

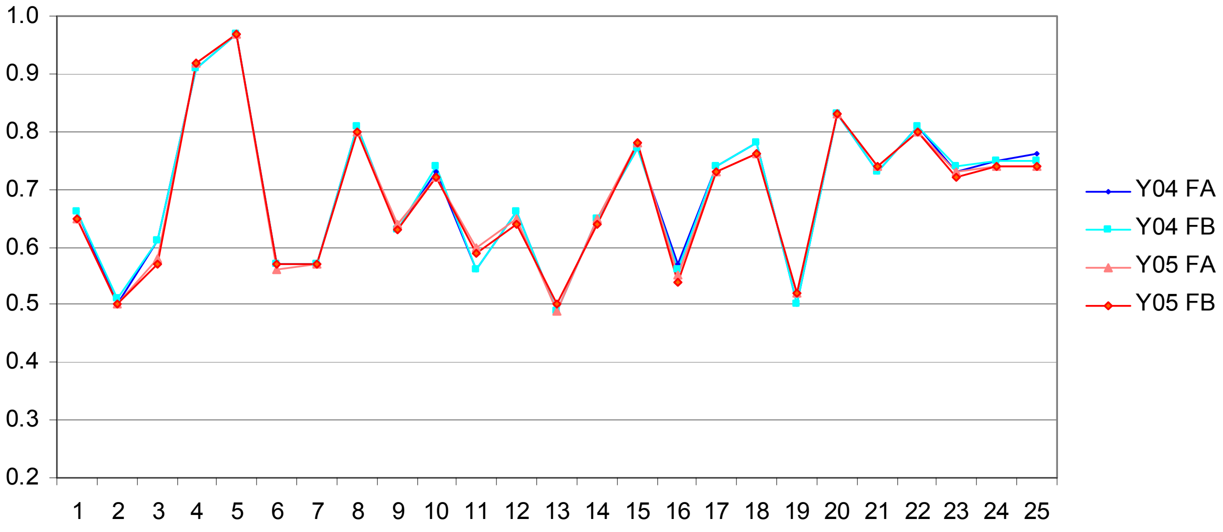


Figure 1.7 Common Item P-value Plots for Year 2004 vs. Year 2005: Grade 8

Validation Check with Augmented Items

To collect information about how much the same items that appear on the test forms in consecutive years (one year as field test items and the next year as operational test items) changed in terms of item difficulty, the p-values of those items were calculated.

Table 1.21 and Table 1.22 depict which field test forms in previous year corresponds to which operational test forms in 2005. It should be noted that Year 2005 Forms 1 and 3 are the same, and Year 2005 Forms 2 and 4 are the same except for the field test portion. In Tables 1.23 through 1.28, item numbers are given by those of Year 2005, and the boldfaced items are brief constructed response (*BCR*) items. More detailed information about the specific test design and construction of Year 2005 can be obtained from section 1.5.

In general, we can conclude that most of the p-values in Year 2005 increased compared to those in previous year across all grades except for grades 6 through 8.

Table 1.21 Form Identification for Items Appearing Year 2003 and Year 2005: Grades 3, 5, and 8

Grade	Year 2003	Year 2005
3	Form 4, 5	Form A (1, 3)
	Form 3, 4	Form B (2, 4)
5	Form 5	Form A (1, 3)
	Form 2, 3	Form B (2, 4)
8	Form 2, 4	Form A (1, 3)
	Form 3, 2	Form B (2, 4)

Note. Form A designates the operational portion of Forms 1 and 3, which is identical. Form B designates the operational portion of Forms 2 and 4, which is identical.

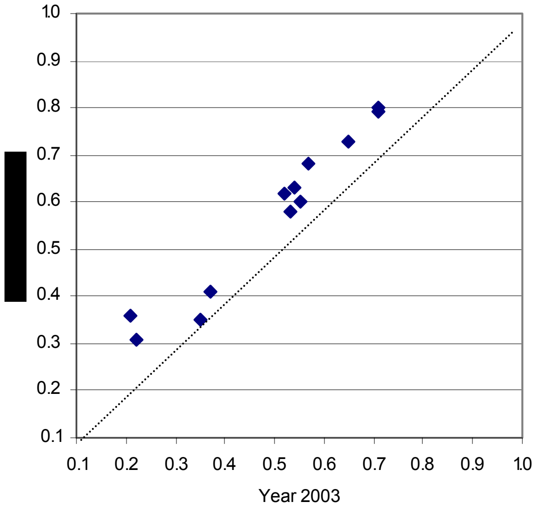
Table 1.22 Form Identification for Items Appearing Year 2004 and Year 2005: Grades 4, 6, and 7

Grade	Year 2004	Year 2005
4	Form 1	Form A (1, 3)
	Form 6	Form B (2, 4)
6	Form 1	Form A (1, 3)
	Form 5	Form B (2, 4)
7	Form 3	Form A (1, 3)
	Form 6	Form B (2, 4)

Note. Form A designates the operational portion of Forms 1 and 3, which is identical. Form B designates the operational portion of Forms 2 and 4, which is identical.

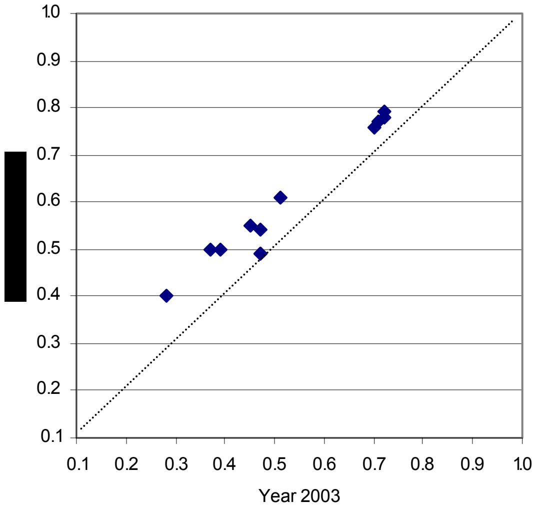
Table 1.23 Augmented Item P-Value Comparison for Year 2003 vs. Year 2005: Grade 3

Item Number	Year 03	Year 05 Form A
71	0.71	0.80
72	0.54	0.63
73	0.22	0.31
74	0.57	0.68
75	0.21	0.36
76	0.65	0.73
77	0.53	0.58
78	0.71	0.79
79	0.37	0.41
80	0.55	0.60
81	0.52	0.62
82	0.35	0.35



Note. Boldfaced number indicates that it is BCR item.

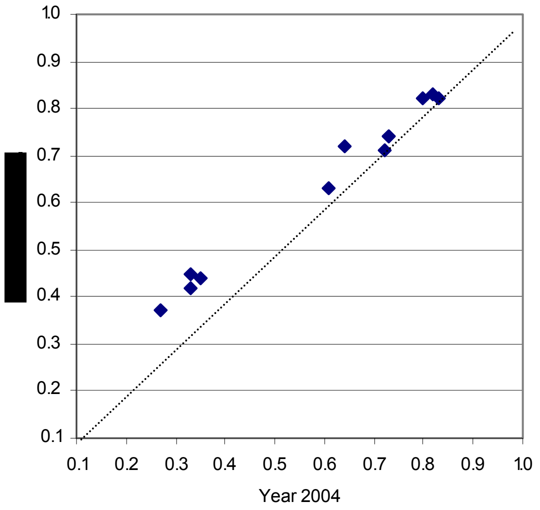
Item Number	Year 03	Year 05 Form B
71	0.72	0.79
72	0.72	0.78
73	0.47	0.49
74	0.71	0.77
75	0.45	0.55
76	0.47	0.49
77	0.28	0.40
78	0.51	0.61
79	0.37	0.50
80	0.70	0.76
81	0.47	0.54
82	0.39	0.50



Note. Boldfaced number indicates that it is BCR item.

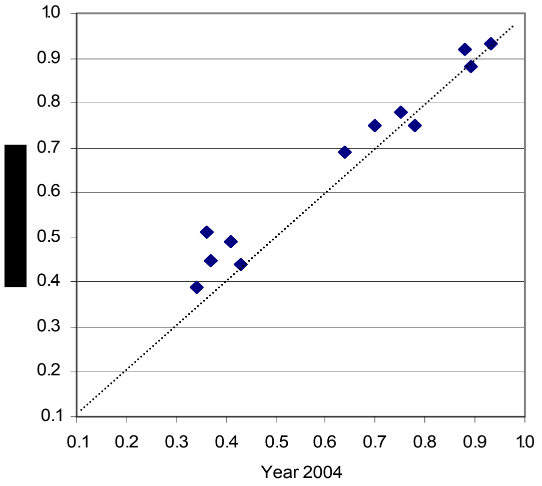
Table 1.24 Augmented Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 4

Item Number	Year 04	Year 05 Form A
71	0.73	0.74
72	0.80	0.82
73	0.33	0.42
74	0.72	0.71
75	0.33	0.45
76	0.73	0.74
77	0.83	0.82
78	0.64	0.72
79	0.35	0.44
80	0.82	0.83
81	0.27	0.37
82	0.61	0.63



Note. Boldfaced number indicates that it is BCR item.

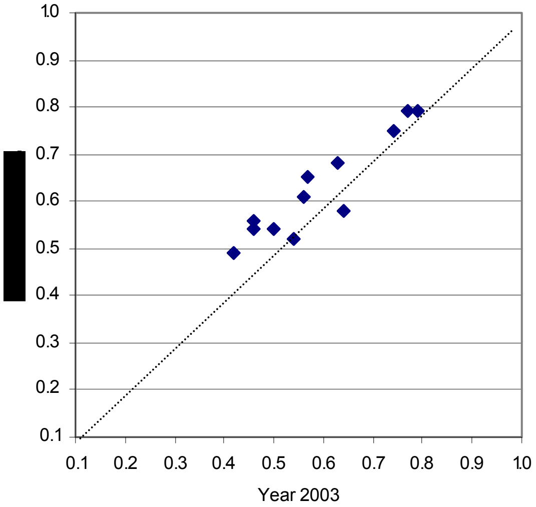
Item Number	Year 04	Year 05 Form B
71	0.93	0.93
72	0.88	0.92
73	0.37	0.45
74	0.89	0.88
75	0.36	0.51
76	0.75	0.78
77	0.78	0.75
78	0.70	0.75
79	0.41	0.49
80	0.64	0.69
81	0.34	0.39
82	0.43	0.44



Note. Boldfaced number indicates that it is BCR item.

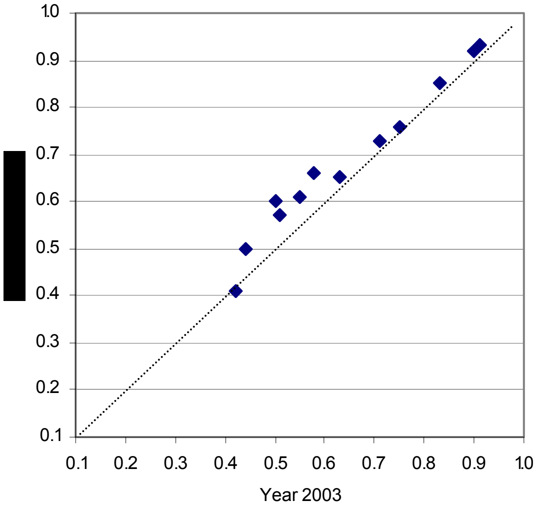
Table 1.25 Augmented Item P-Value Comparison for Year 2003 vs. Year 2005: Grade 5

Item Number	Year 03	Year 05 Form A
61	0.74	0.75
62	0.64	0.58
63	0.46	0.56
64	0.57	0.65
65	0.50	0.54
66	0.56	0.61
67	0.79	0.79
68	0.77	0.79
69	0.46	0.54
70	0.54	0.52
71	0.42	0.49
72	0.63	0.68



Note. Boldfaced number indicates that it is BCR item.

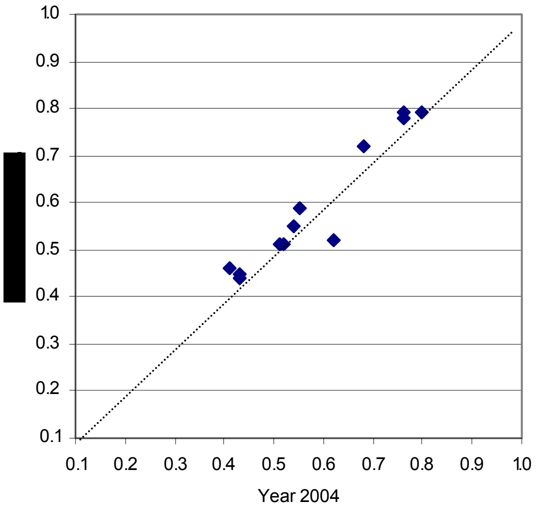
Item Number	Year 03	Year 05 Form B
61	0.75	0.76
62	0.91	0.93
63	0.44	0.50
64	0.71	0.73
65	0.50	0.60
66	0.63	0.65
67	0.90	0.92
68	0.55	0.61
69	0.51	0.57
70	0.83	0.85
71	0.58	0.66
72	0.42	0.41



Note. Boldfaced number indicates that it is BCR item.

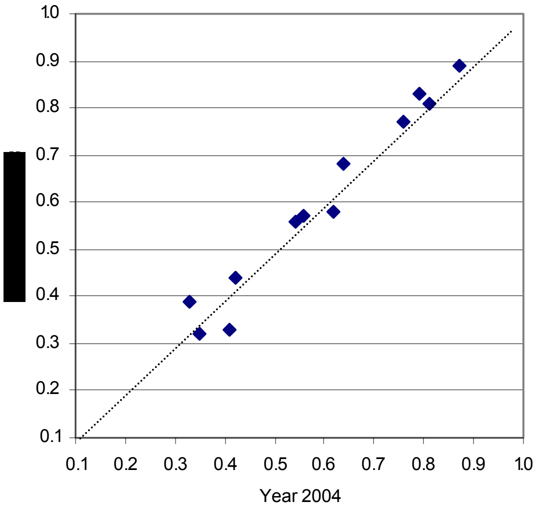
Table 1.26 Augmented Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 6

Item Number	Year 04	Year 05 Form A
61	0.62	0.52
62	0.54	0.55
63	0.51	0.51
64	0.68	0.72
65	0.43	0.44
66	0.76	0.79
67	0.76	0.78
68	0.55	0.59
69	0.43	0.45
70	0.52	0.51
71	0.41	0.46
72	0.80	0.79



Note. Boldfaced number indicates that it is BCR item.

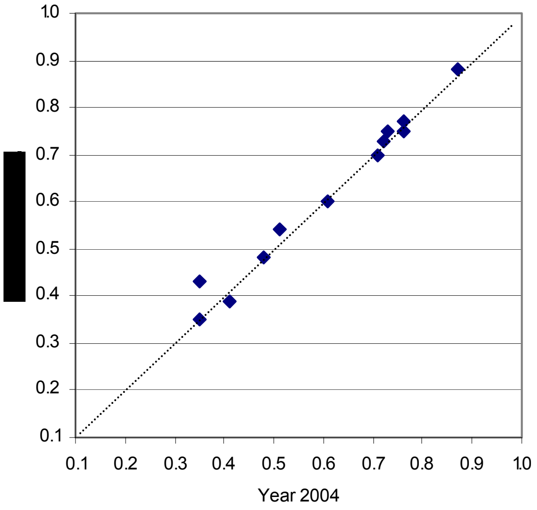
Item Number	Year 04	Year 05 Form B
61	0.87	0.89
62	0.56	0.57
63	0.35	0.32
64	0.64	0.68
65	0.33	0.39
66	0.79	0.83
67	0.76	0.77
68	0.54	0.56
69	0.41	0.33
70	0.62	0.58
71	0.42	0.44
72	0.81	0.81



Note. Boldfaced number indicates that it is BCR item.

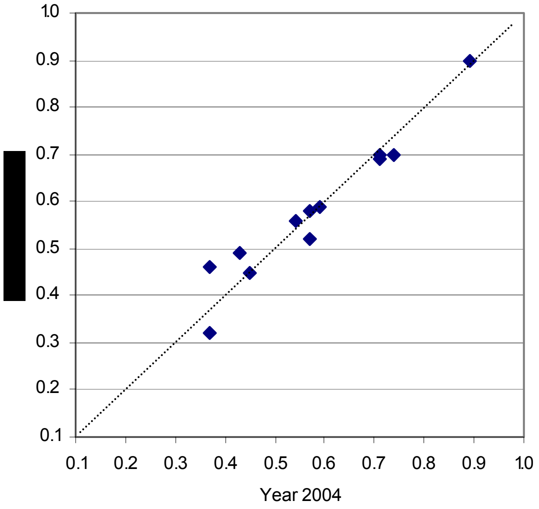
Table 1.27 Augmented Item P-Value Comparison for Year 2004 vs. Year 2005: Grade 7

Item Number	Year 04	Year 05 Form A
61	0.73	0.75
62	0.71	0.70
63	0.48	0.48
64	0.72	0.73
65	0.35	0.35
66	0.76	0.77
67	0.61	0.60
68	0.87	0.88
69	0.41	0.39
70	0.76	0.75
71	0.35	0.43
72	0.51	0.54



Note. Boldfaced number indicates that it is BCR item.

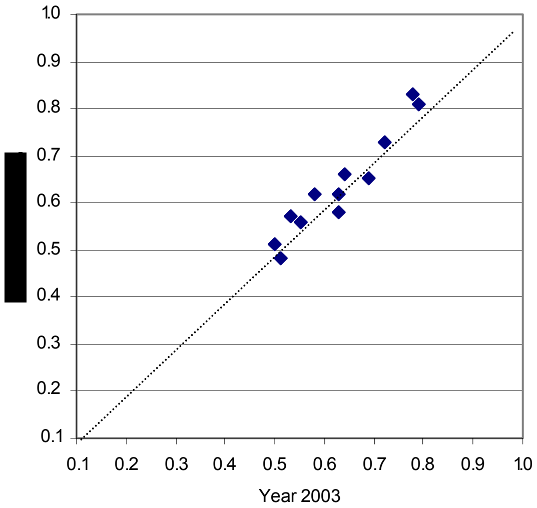
Item Number	Year 04	Year 05 Form A
61	0.89	0.90
62	0.57	0.52
63	0.43	0.49
64	0.57	0.58
65	0.37	0.32
66	0.59	0.59
67	0.54	0.56
68	0.71	0.69
69	0.45	0.45
70	0.71	0.70
71	0.37	0.46
72	0.74	0.70



Note. Boldfaced number indicates that it is BCR item.

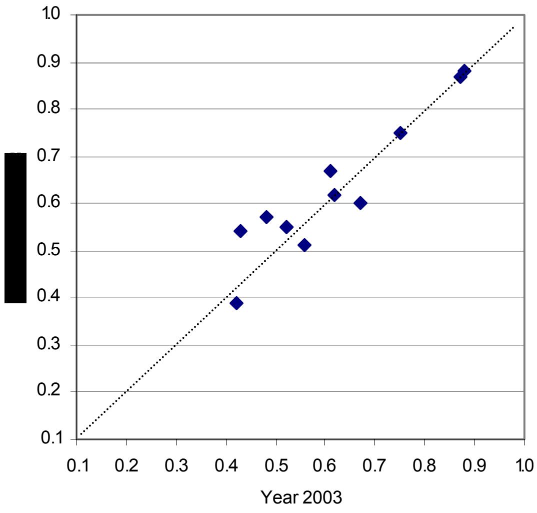
Table 1.28 Augmented Item P-Value Comparison for Year 2003 vs. Year 2005: Grade 8

Item Number	Year 03	Year 05 Form A
61	0.78	0.83
62	0.53	0.57
63	0.58	0.62
64	0.72	0.73
65	0.55	0.56
66	0.64	0.66
67	0.63	0.62
68	0.79	0.81
69	0.51	0.48
70	0.69	0.65
71	0.50	0.51
72	0.63	0.58



Note. Boldfaced number indicates that it is BCR item.

Item Number	Year 03	Year 05 Form B
61	0.75	0.75
62	0.87	0.87
63	0.48	0.57
64	0.88	0.88
65	0.56	0.51
66	0.67	0.60
67	0.62	0.62
68	0.52	0.55
69	0.52	0.55
70	0.42	0.39
71	0.43	0.54
72	0.61	0.67



Note. Boldfaced number indicates that it is BCR item.

1.9 Field Test Analyses

All field test items embedded in operational forms are subjected to rigorous analyses for their properties because these analyses will provide information about which items would be included as a part of operational items in the future. All statistical results concerning field test items were stored in the 2005 item bank. The following field test analyses were conducted:

- Classical item analyses for *SR* and *BCR* items
- *Differential item functioning* (*DIF*) analyses
- *IRT* analyses

Classical Item Analyses for *SR* and *BCR* items

Classical item analyses for *SR* and *BCR* items were conducted within each field test form.

SR items for further scrutiny were flagged if:

- An item distractor was unselected by all students (i.e., nonfunctional distractor), or selected by a large number of high ability students, with low selection from other ability groupings (i.e., ambiguous distractor).
- An item *p*-value was less than .20 or greater than .90.
- An item point-biserial was less than .10 (i.e., poorly discriminating). If an item point-biserial was close to zero or negative, the item was checked for a miskeyed answer.

BCR items for further scrutiny were flagged if:

- An item did not elicit the full range of rubric scores.
- The ratio of mean item score to maximum score was less than .20 or greater than .90.
- An item-total correlation was less than .10.

Dropping any items needed a careful decision. For example, an item that was flagged as being difficult (*p*-value less than .20) and poorly discriminating (point-biserial less than .10) was considered for dropping. If the item represented important content that had not been extensively taught, however, it would be justified to retain the item.

Differential Item Functioning Analyses

Differential item functioning (*DIF*) analyses are primarily designed to detect differential item performance across subgroups of a population while controlled for ability.

For the 2005 MSA-Reading *DIF* analyses, the reference group was either male or Caucasian students, and the focal group was either female or African-American students. Because the 2005 MSA-Reading included both the *SAT10* items and the "Maryland-specific" items on each field test form, the total item score on a collection of items was used as the matching variable.

Any *SR* and *BCR* items that were flagged as showing *DIF* were subjected to further examination. For each of these items, for example, reading experts judged if the differential difficulty of the item was unfairly related to group membership:

- If the difficulty of the item is unfairly related to group membership, then the item should not be used at all.

- If the difficulty of the item is related to group membership, then the item should only be used if there is no other item matching the test blueprint.

For further information about the *DIF* procedures used for the 2005 MSA-Reading, please see section 3.7.

***Item Response Theory* Analyses**

To put field test items on the same scale of the operational test items, they were calibrated with fixing the parameters of the operational test items within each test form. Then, item difficulties, step difficulties, and fit statistics were stored in the 2005 item bank.

Item Selection for Operational Test

The selection of items to be included in the final test forms of the 2005 MSA-Reading required a careful consideration based on test blueprints, classical item analyses and *DIF* analyses. Harcourt suggested the following guidelines to choose items included in the final test forms:

- Avoid the use of the items with *p*-values less than .20 and greater than .90.
- Avoid the use of the *BCR* items with score distributions that do not elicit the full range of rubric scores.
- Avoid the use of items with point-biserial or item-total correlation less than .10.
- Avoid the inclusion of items with *DIF* classifications “C” for the *SR* items and “CC” for the *BCR* items *unless* they have been deemed acceptable by the external review of reading experts.

In applying these guidelines, a balance should be made between being too harsh, and thus dropping items that may affect the content representativeness of the entire set of field test items and being too lenient and allowing items with poor model fit that might affect resulting measures. In addition, reading specialists from the MSDE reviewed the final test forms of the 2005 MSA-Reading.

The 2005 MSA-Reading produced two operational test forms for all grades, and reading specialists from the MSDE reviewed and determined the content validity and equivalency of the test forms for each grade level.

1.10 Linking, Equating, and Scaling Procedures

Linking Procedures

To link different test forms at each grade level, linking steps recommended by the National Psychometric Council were taken into consideration. For the 2005 MSA-Reading, items that appeared on each test form were included as potential linking items, but only *SR* items were considered as potential linking items.

First, the following calculations were made (SDE, 2001):

- The mean and standard deviation of the linking pool's item difficulties for each form
- The ratio of the standard deviations between form 1 and the rest of the forms
- The correlation between test form 1 and other test form item difficulties
- The difference between test form 1 and other test form item difficulties for each item in the linking pool
- The mean of the differences calculated above
- The median of the differences
- The interquartile range of the differences
- The robust Z for each item in the linking pool where the robust Z is defined as (the difference between the test form1 and other test form item difficulty minus the median of the differences) / (interquartile range multiplied by 0.74).

Once the above calculations were made, the following guidelines were taken in determining possible sets of linking items to be used for the Rasch equating (SDE, 2001):

- Do not include those items with an absolute value of robust Z exceeding 1.645. In addition, if one difficulty or step from a *SR* item is eliminated from the pool based on robust Z, all other difficulties are also removed.
- Do not eliminate more than 20 percent of the pool linking items.
- Consider that the ratio of the standard deviations of the test form 1 and other test form item difficulties should be in the 90 to 110 percent range.
- It is assumed that the correlation of the test form 1 and other test form item difficulties is greater than .95.

Toward this end, Harcourt provided Rasch item difficulties, item difficulty plots, and robust Z values and identified items that were to be deleted based on the definition. For example, Figure 1.8 shows those results between the base form and the form A.

	Base Form (Form 6)	Form A	6 vs. A	Robust Z
1	-3.9886	-4.1111	-0.12	-3.551
2	-1.7739	-1.6875	0.09	-.287
3	-.1033	-.2267	-0.12	-3.565
4	-1.2892	-1.0993	0.19	1.329
5	.0403	-.0064	-0.05	-2.367
6	-.6252	-.4871	0.14	.520
7	-.2092	-.0400	0.17	1.006
8	-1.4440	-1.3924	0.05	-.831
9	-.6775	-.5489	0.13	.372
10	.2123	.3484	0.14	.489
11	-.3429	-.2113	0.13	.419
12	.2842	.0396	-0.24	-5.459
13	-.7393	-.6224	0.12	.189
14	-.3247	-.3430	-0.02	-1.923
15	1.5832	1.8030	0.22	1.797
16	-1.9501	-1.8516	0.10	-.098
17	-.4109	-.3101	0.10	-.062
18	-.5286	-.3619	0.17	.967
19	-1.8443	-1.7974	0.05	-.905
20	.8212	.9183	0.10	-.120
21	1.3188	1.4139	0.10	-.152
22	2.0024	2.1072	0.10	.000
23	-1.4991	-1.3484	0.15	.717
24	-.1689	-.0039	0.17	.940
25	.7087	.8240	0.12	.164

	Form 6	Form A
Mean	-.438	-.360
SD	1.278	1.304

	6 vs. 6	6 vs. A
Correlation	1.000	.997
SD ratio	100%	102%

	6 vs. 6	6 vs. A
Mean of Difference	.000	.078
Median of Difference	.000	.105
Interquartile Range of Difference	.000	.087

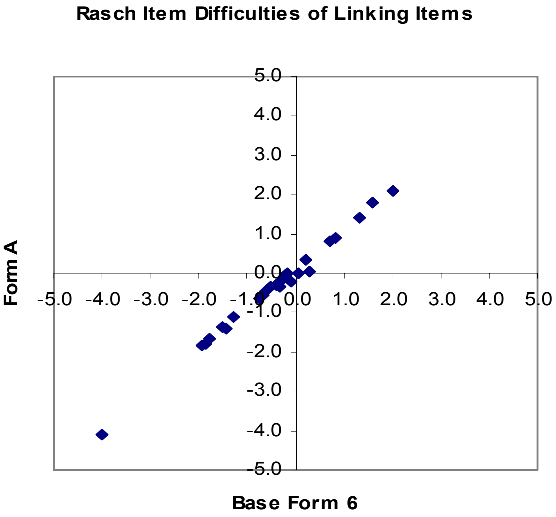


Figure 1.8 Example of Parameters Used to Link Items

Equating Procedures

Equating different test forms ensures that students taking one form of a test are neither advantaged nor disadvantaged when compared to students taking a different form of a test.

For the 2005 MSA-Reading, items selected through the linking procedures were used to equate all different test forms in each grade. Because each test form included a subset of unique items, linking items served as anchor items. Thus, whenever a new test form is constructed in the future, the new form will be equal in difficulty to the previous form via linking items. The design to collect data for the 2005 MSA-Reading was common item, non-equivalent groups.

In order to obtain parameter estimates for both the unique items on each form and the linking items, the Rasch model (or Partial Credit model for *BCR* items) was used. For the 2005 MSA-Reading, the common items whose calibrations were known were anchored or fixed to their known estimates during the calibration of other forms that were to be put on the scale of the first form. In treating these common item parameters as known they were fixed, and the remaining item parameters (for the unique items of each form) were also forced onto the same scale as the anchored (fixed) items.

The final step consisted of obtaining ability score or theta for each raw score point on a form. This was done by iteratively solving the expression:

$$True\,Score = \sum_{i=1}^{I}\sum_{j=0}^{m_i} j \cdot P_{ij}(\theta)$$

where

$P_{ij}(\theta)$ = the probability of a correct response for each of the i = 1, ... , I items given that the item categories are numbered *0, ..., m_i*.

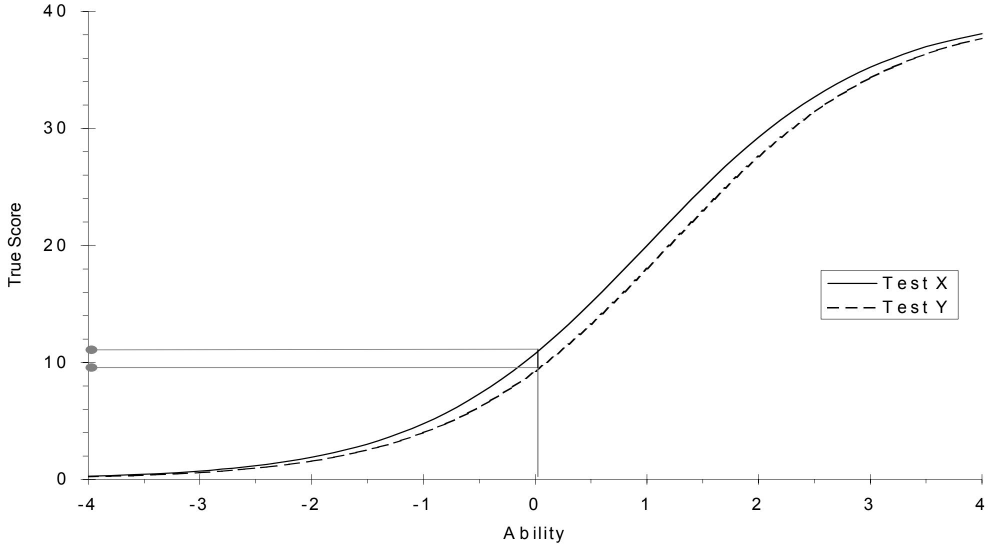


Figure 1.9 True Score Equating

Figure 1.9 illustrates these ideas for two hypothetical test forms, X and Y. In the figure, the true scores on each of the forms are plotted against ability using the true score equation. By drawing a line from the ability (here shown for an ability of 0) to each of the respective curves and moving across to the true score scale, one can find the pairs of true scores that are equated to one another. According to Lord and Wingersky (1984), the procedure applied to true scores can be transferred to observed scores without any major anomalies in the resulting outcomes.

Reporting Scale Scores

In order to facilitate the use and interpretation of the results of the 2005 MSA-Reading, scale scores were generated based on the information given by both the MSDE and the NPC. For grade 4, for example, the following is the formula to convert each student' ability or theta to scale score:

$$ReportingAbilityScaleScore = 32.8271 \cdot theta + 362.7449$$

$$ReportingSEM = 32.8271 \cdot SEM$$

where

theta = the *IRT* ability estimate, and

SEM = the conditional *SEM* of the ability estimate.

Table 1.29 depicts the slope and intercept to use for each grade. It should be noted that the minimum of the scale score was set to 240, and the maximum of the scale score was set to 650.

Table 1.29 The 2005 MSA-Reading Slope and Intercept: Grades 3 through 8

Grade	Slope	Intercept
3	32.4123	384.8579
4	32.8271	362.7449
5	33.0171	380.0082
6	30.4732	373.0575
7	31.9262	377.0054
8	30.3891	376.8316

1.11 Score Interpretation

To help provide appropriate interpretation of the 2005 MSA-Reading test scores, two types of scores were created: 240-650 scale scores, and performance levels and descriptions.

240-650 Scale Scores

As explained in section 1.10, Linking, Equating, and Scaling, the 2005 MSA-Reading produced scale scores that ranged between 240 and 650. Those scale scores have the same meaning within the same grade, but those scores are not comparable across grade levels.

It should be noted that those scale scores have only simple meaning that higher scale scores represent higher performance in reading tests. Thus, performance levels and descriptions can give a specific interpretation other than a simple interpretation because they were developed to bring meaning to those scale scores.

Performance Levels and Descriptions

As previously explained, performance levels and descriptions provide specific information about students' performance levels and help interpret the 2005 MSA-Reading scale scores. They describe what students at a particular level generally know and can be applicable to all students within each grade level. As Table 2.1 shows a range of scale scores at each performance level, for example, grade 4 reading scale scores from 371 to 436 indicate the level of *Proficient*, and students at this level can read grade appropriate text and demonstrate the ability to comprehend literature and informational passages. Further information about the 2005 MSA-Reading score interpretation can be obtained from the MSDE.

1.12 Test Validity

As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), "validity is the most important consideration in test evaluation."

Messick (1989) defined validity as follows:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (p.5)

This definition implies that test validation is the process of accumulating evidence to support intended use of test scores. Consequently, test validation is a series of on-going and independent processes that are essentially independent investigations of the appropriate use or interpretation of test scores from a particular measurement procedure (Suen, 1990).

In addition, test validation embraces all of the experimental, statistical, and philosophical means by which hypotheses and scientific theories can be evaluated. This is the reason that validity is now recognized as a unitary concept (Messick, 1989).

To investigate the validity evidence of the 2005 MSA-Reading, content-related evidence, evidence of internal structure, and evidence of unidimensionality were collected.

Content-Related Evidence

Content validity is frequently defined in terms of the sampling adequacy of test items. That is, content validity is the extent to which the items in a test adequately represent the domain of items or the construct of interest (Suen, 1990). Consequently, content validity provides judgmental evidence in support of the domain relevance and representativeness of the content in the test (Messick, 1989).

The 2005 MSA-Reading blueprints provide extensive evidence regarding the alignment between the content in the 2005 MSA-Reading and the *VSC*. These blueprints are presented in Appendix C.

Evidence of Internal Structure

The 2005 MSA-Reading has three reading processes: *General Reading*, *Literary Reading*, and *Informational Reading*. As can be seen from Tables 4.3 through 4.8, there exist moderately strong intercorrelations among the reading processes.

Evidence of Unidimensionality

Measurement implies order and magnitude on a single dimension (Andrich, 1989). Consequently, in the case of scholastic achievement, this requires a linear scale to reflect this idea of measurement. Such a test is considered to be unidimensional (Andrich, 1988, 1989). However, unidimensionality cannot be strictly met in a real testing situation because students' cognitive, personality, and test-taking factors usually have a unique influence on their test performance to some level (Andrich, 1988; Hambleton, Swaminathan, & Rogers, 1991). Consequently, what is required for unidimensionality to be met is an investigation of the presence of a dominant factor that influences test performance. This dominant factor is considered as the ability measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983).

To check the unidimensionality of the 2005 MSA-Reading, polychoric correlation coefficients were computed with *LISREL 8.5* (Jöreskog & Sörbom, 1993) because they were polytomously scored on reading tests. Principal component analysis was then applied to produce eigenvalues. The first and the second principal component eigenvalues were compared without rotation. Table 1.30 summarizes the results of the first and second principal component eigenvalues of the 2005 MSA-Reading.

The rule of thumb to determine the unidimensionality of a test requires that the eigenvalue of the first component or factor should be at least three times larger than the second one. As can be seen, the size of the eigenvalue of the first component meets the criterion for the unidimensionality. Thus, we can conclude that the assumption of unidimensionality for the 2005 MSA-Reading was met.

Table 1.30 The 2005 MSA-Reading Eigenvalues between the First and Second Components: Grades 3 through 8

Grade	Form	Number of Items	First Eigenvalue	Second Eigenvalue
3	A	37	12.33	1.47
	B	37	12.00	1.44
4	A	37	12.10	1.37
	B	37	11.67	1.26
5	A	37	10.70	1.38
	B	37	10.34	1.45
6	A	37	12.32	1.44
	B	37	12.34	1.31
7	A	37	12.63	1.34
	B	37	12.31	1.41
8	A	37	10.53	1.39
	B	37	11.13	1.39

Note. Form A designates the operational portion of Forms 1 and 3, which is identical. Form B designates the operational portion of Forms 2 and 4, which is identical.

1.13 Item Bank Construction

The number of test forms to be constructed each year and the need to replace items that would be released to the public necessitated the availability of a large pool of items. The 2005 MSA-Reading item bank continues to be maintained by Harcourt as computer files and paper copies. This enables test items to be readily available to both Harcourt and MSDE staff for reference, test construction, test book design, and printing.

Harcourt maintains a computerized statistical item bank to store supporting and identification information on each item. The information stored in this item bank for each item is as follows:

- CID
- Test administration year and season
- Test form
- Grade level
- Item type
- Item stem and options
- Passage code and title
- Subject code and description
- Process code and description
- Standard code and description
- Indicator code and description
- Objective code and description
- Item status
- Item statistics

The item bank Rasch scale statistics were re-calibrated using all of the students' test responses. Thus, the re-calibrated scale would serve as the base scale.

1.14 Quality Control Procedures

A standard quality procedure at Harcourt Assessment is to create a test deck for all programs. The test deck begins when Quality Assurance enters mock data into the enrollment system, which is transferred to the materials requisition system; the order is packaged by our Distribution Center, and shipped to the Quality Assurance Department. We then review the packing list against the data entered, the materials algorithms applied, the materials packaged against the packing list, and the actual packaging of the documents. These documents are then used to create a test deck of mock data along with advance copies of documents that are received from the printer. Advance printer copies are inclusive of documents throughout the print run to assure we are randomly testing printed documents. The Maryland test deck will be a comprehensive set of all documents that will:

- Verify all scan positions for item responses and demographics to verify scanning setup and scan densities
- Verify all constructed response score points, zoning of image, reader scoring, reader resolution, and reader check scores
- Verify the handling of blank documents through the system
- Test all demographic and item edits
- Verify pre-id bar code read, match and no-match
- Verify attemptedness rules applied by subtest
- Verify duplicate student handling (same test duplicate, different test duplicate)
- Verify duplicate student with different demographics rules applied
- Verify the document counts to the enrollment, pre-id and actual document receipt
- Verify pre-id matching and application to student record
- Verify various raw score points and access to dummy and live scoring tables
- Verify cut scores applied
- Verify valid score on one subtest and invalid score on other subtest
- Verify scoring applied to Braille and Large Print
- Verify valid multiple choice and invalid constructed response
- Verify valid constructed response and invalid multiple choice
- Verify all special scoring rules
- Verify all summary programs for rounding
- Verify summary inclusion and exclusion (Braille, standard and non-standard student summarization)
- Verify each scoring level for group reporting
- Verify all reporting programs for accuracy in all text and data presented
- Verify class, school, district, and state summary data on home reports
- Verify all data file programs to assure valid information in every field

- Verify data descriptions for accuracy against data file
- Create compare programs to allow for update of files

The Maryland test deck is the first order processed through the Maryland system to verify all aspects of the materials packaging, scanning, editing, scoring, summary, and reporting. Pre-determined conditions are included in the test deck to assure the programs are processing all data to meet the requirements of the program with zero defects. Processing of live orders cannot proceed until each phase of the test deck has been approved by our Quality Assurance Department. An Issues Log with sign-off approvals is utilized to assure we are addressing any issues that arise in the review of the test deck data across all functional groups at Harcourt.

Prior to release of any order for reporting we will receive a preliminary file from Scoring Operations to run a key check TRIAN to assure that all scoring keys have been determined and applied accurately. Any item that is not performing as expected will be flagged and reviewed by our content specialist and psychometrician. Upon completion of the key check, we will proceed to run the pilot level reports.

We will run the pilot district utilizing live data. The pilot district will include multiple buildings, all grades, and any unique accommodations. A formal pilot review process is conducted with expert Harcourt staff prior to release of the information to the MSDE.

Upon completion of the processing of all district level data, Harcourt Scoring Operations will provide the Quality Assurance Department with a state level data file(s) and state data for review and approval. Harcourt Quality Assurance programmers duplicate all data independently to assure accurate interpretation of the expected results. A series of SAS programs will be run on these files to assure 100% accuracy. These include but are not limited to:

- Statewide Duplicate Student
- Statewide FD of Demographic Variables
- District/Building/N-Count
- Statewide RS/SS/Cut Score tables
- Proc Means to verify summary statistics
- Item Response listing to verify all constructed responses are scored and within the valid range
- Normative data check for all raw scores
- Reader Resolution report to verify all readings and resolution combinations

Upon complete review and approval by Quality Assurance, we will post the statewide student files to a secure FTP site for review by the MSDE. MSDE staff is always welcome to have staff in San Antonio to work directly with our QA staff as they are reviewing the data. We have found this to be very beneficial and also expedite the review of the state level data.

Harcourt understands the importance of providing accurate, reliable, and valid data to the MSDE. We strive to continually improve our processes and verification efficiency to meet our scheduled delivery dates for state reporting.

In addition to the routine procedure from the Quality Assurance Department, Harcourt Psychometric & Research Services purses the complete independent replication policy for equating results in order to maintain zero-defect equating results. The equating results include generation of Raw Score-Scale Score (RS-SS) conversion tables for Maryland students. In generating RS-SS tables, the lead psychometrician first generates them, and then the back-up psychometrician generates the same tables independently. Two results from the lead and the back-up psychometrician are compared. This procedure is repeated until their results match 100%.