1. OVERVIEW OF THE 2007 *MARYLAND SCHOOL ASSESSMENT-READING*

In 2002, the Maryland State Department of Education (MSDE), in order to conform to the requirements of the new Federal program "No Child Left Behind," retired its award-winning *Maryland School Performance Assessment Program* and adopted a testing program known as the *Maryland School Assessment* (*MSA*). The new program, like its predecessor, 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.

In 2003, the MSA-Reading was introduced in grades 3, 5, and 8, and grades 4, 6, and 7 were added to the program in 2004. The MSA-Reading included SAT10 as well as Maryland-specific items. SAT10 abbreviated Form A was administered at grades 3 through 8. SAT10 common items aligned to Maryland curriculum played as possible form-to-form and year-to-year linking items. It should be noted that the Rasch difficulty estimates generated in the first year continued to be used in subsequent years' calibration and equating procedures so that all scale scores were on the same scale.

A Bookmark standard setting was conducted in 2003 to set proficiency level cut scores for grades 3, 5, and 8. Because 2004 was the first testing year for grades 4, 6, and 7, a second Bookmark standard setting was held in summer 2004 to set cut scores for these additional grades. The performance level cut scores were used to assign students to three proficiency levels (Basic, Proficient, and Advanced) for AYP reporting under the "No Child Left Behind" act. Information about the Bookmark procedures and results can be found from MSDE. It should be noted that these cut scores have been applied since 2003 (grades 3, 5, and 8) and 2004 (grades 4, 6, and 7).

From March 12 to March 21, 2007, students in grades 3 through 8 took the 2007 *MSA* in reading (MSA-Reading).

1.1 General Overview of the 2007 MSA-Reading

The 2007 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 2007 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.

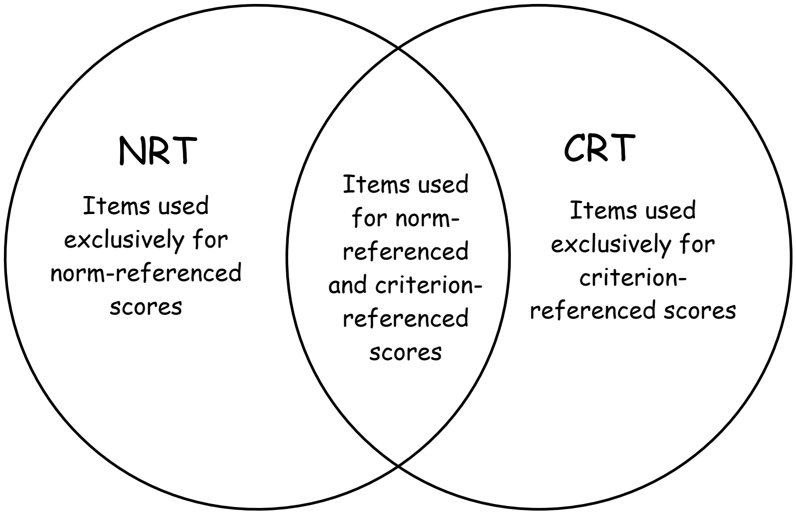
NRT
Items used exclusively for norm-referenced scores
Items used for norm-referenced and criterion-referenced scores
CRT
Items used exclusively for criterion-referenced scores

Figure 1.1 Schematic of the 2007 MSA-Reading

1.2 Purposes/Uses of the 2007 MSA-Reading

By measuring students' achievement against the new academic standards, the 2007 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. 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 2007 MSA-Reading

Developing the 2007 MSA-Reading was a complex process. It required a great deal of involvement from 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 2007 MSA-Reading.

Table 1.1 The 2007 MSA-Reading Responsibility for Test Development

Development of the 2007 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 MSDE on the development and implementation of the 2007 MSA-Reading program. For example, they made recommendations to 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. MSDE adopted their guidelines and recommendations.

Content Review Committee

Content Review Committee members 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 2007 MSA-Reading

2007 MSA-Reading Test Structure

The 2007 MSA-Reading was composed of the *SAT10* items and augmented (Maryland-specific) operational items. In addition, the uniqueness of the MSA-Reading was to spiral a relatively large number of Maryland field test items into multiple test forms (10 forms) for each grade in test administration.

As can be seen from Table 1.2, the 2007 MSA-Reading produced ten test forms for each grade, and there were 2 operational forms within each grade. This means that Forms 1, 3, 5, 7, and 9 (Form A) are identical, and Forms 2, 4, 6, 8, and 10 (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 1.10 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 section 1.8, Operational Test Analyses.

Table1.2 The 2007 MSA-Reading Test Structure: Grades 3 through 8

	Operational Item Sets		Field Test Item Sets									
	A	B	1	2	3	4	5	6	7	8	9	10
Form 1	X		X									
Form 2		X		X								
Form 3	X				X							
Form 4		X				X						
Form 5	X						X					
Form 6		X						X				
Form 7	X								X			
Form 8		X								X		
Form 9	X										X	
Form 10		X										X

Note. Total number of operational test items = 37 (33 *SR* + 4 *BCR*) items. Forms 1, 3, 5, 7, and 9 (Form A) are identical, and Forms 2, 4, 6, 8, and 10 (Form B) are identical in terms of operational test items.

Types of Items

The 2007 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 2007 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 2007 MSA-Reading, these items were scored on a general rubric with maximum values between 0 and 3. For example, score was the higher of the first and the second readers' scores provided they were adjacent. A resolution reader's score was used of two non-adjacent initial scores were received. That is, the resolution reader's score was used in place of both the first and second readers' scores. Detailed information on BCR scoring procedures and rules can be found in section 1.7, MSA-Reading Scoring Procedures.

Table 1.3 The 2007 MSA-Reading Test Design: Grades 3, 5, and 8

Grade	Strand Title	*SAT10* / Augmented	Item Type	No. of Items of Each Form	
				FA	FB
3	Total NRT	*SAT10*	*SR*	70	70
	Word Study	*SAT10*	*SR*	20	20
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	37 (10)	37 (10)
	General Reading	*SAT10*	*SR*	16	16
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	10 (10)	10
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11 (10)
5	Total NRT	*SAT10*	*SR*	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	37 (10)	37 (10)
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	11 (10)	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11(10)
8	Total NRT	*SAT10*	*SR*	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	37 (10)	37 (10)
	General Reading	*SAT10*	*SR*	16	16
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	10 (10)	10
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11 (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 the total number of field test items tested during operational testing. Form A designates the forms 1, 3, 5, 7, and 9. Form B designates the forms 2, 4, 6, 8, and 10.

Table 1.4 The 2007 MSA-Reading Test Design: Grades 4, 6, and 7

Grade	Strand Title	*SAT10* / Augmented	Item Type	No. of Items of Each Form	
				FA	FB
4	Total NRT	*SAT10*	*SR*	70	70
	Word Study	*SAT10*	*SR*	20	20
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	37 (10)	37 (10)
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	11 (10)	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11 (10)
6	Total NRT	*SAT10*	*SR*	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	37 (10)	37 (10)
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	11(10)	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11(10)
7	Total NRT	*SAT10*	*SR*	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT	*SAT10*, Augmented	*SR*, *BCR*	37 (10)	37 (10)
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	11 (10)	11
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	11	11 (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 the total number of field test items tested during operational testing. Form A designates the forms 1, 3, 5, 7, and 9. Form B designates the forms 2, 4, 6, 8, and 10.

Table 1.5 The 2007 MSA-Reading Item Distribution of Each Strand: Grade 3 and 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 BCR	No. of Items	No. of SR	No. of *BCR*	No. of Items	No. of SR	No. of *BCR*	No. of Items
A	16	4	5	0	0	0	4	2	6	4	2	6
B	16	4	5	0	0	0	4	2	6	4	2	6

Table 1.6 The 2007 MSA-Reading Item Distribution of Each Strand: Grades 5

	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
A	15	5	5	0	0	0	4	2	6	4	2	6
B	15	5	5	0	0	0	4	2	6	4	2	6

Table 1.7 The 2007 MSA-Reading Item Distribution of Each Strand: Grade 4, 6, and 7

	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
A	15	5	5	0	0	0	4	2	6	4	2	6
B	15	5	5	0	0	0	4	2	6	4	2	6

Table 1.8 The 2007 MSA-Reading Total and Strand Scores: Grade 3 and 8

	Total and Each Cluster Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form A	16 (16 MC)	14 (8 MC + 6 BCR)	15 (9 MC + 6 BCR)	45
Form B	16 (16 MC)	14 (8 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

Table 1.9 The 2007 MSA-Reading Total and Strand Scores: Grades 5

	Total and Each Cluster Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form A	15 (15 MC)	15 (9 MC + 6 BCR)	15 (9 MC + 6 BCR)	45
Form B	15 (15 MC)	15 (9 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

Table 1.10 The 2007 MSA-Reading Total and Strand Scores: Grade 4, 6, and 7

	Total and Each Cluster Scores			
	General Reading	Literary Reading	Informational Reading	Total Score
Form A	15 (15 MC)	15 (9 MC + 6 BCR)	15 (9 MC + 6 BCR)	45
Form B	15 (15 MC)	15 (9 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

1.6 Test Administration

Test Materials

All test materials had to be stored in a secure location prior to test administration. The School Test Coordinator (STC) provided test administration training and test materials to the test examiners. Pre-test workshops were held in Baltimore for all Local Accountability Coordinators in Maryland. These workshops provided the representatives of all 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 Manuals
- Preprinted and generic labels, which were applied to the Test/Answer Books by or under the direct supervision of the STC.
- Scoring Service Identification sheets
- Student Roster

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

- Test/Answer Book
- Special accommodations testing materials, if necessary

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 also needed the following additional materials:

- A sign for the door, "Testing: Do not Disturb"
- A digital clock or a watch, or clock with a second hand
- Copies of the STOP and GO ON sample pages

Two test related examiner's manuals (EM) were developed for the 2007 MSA; one version for reading and the other for mathematics for use in all grades 3-8. Developed in partnership with MSDE, the EMs contained instructions for preparation and administration of the test. In addition to the EMs, one Test Administration and Coordination Manual (TACM) was developed for use by the Local Accountability Coordinators (LAC) and building-level School Test Coordinators (STC). Included in this manual were instructions for preparation of materials for testing, monitoring of testing, and packaging of materials for return to Harcourt for scoring. The TACM was distributed and reviewed during a workshop in January for STCs and LACs with duplicates sent to each school with its testing materials.

Test Administration Schedule

The overall test window for MSA was established by MSDE (March 12-21, 2007, with make-up testing held March 22-27, 2007). However, each Local Education Agency (LEA) set a specific schedule for administration of the MSA within that window for their district. Each LEA schedule was submitted to MSDE in advance and proved for each district by the State. For a given grade and content area, all testing had to take place on the same schedule. In addition, each content area at each grade was tested on two days during the window. For the 2007 MSA-Reading, the primary testing days were as follows:

• Test materials delivered to schools (Examiner's Manuals, Test/Answer Books, and Test Coordinator's Kit)	On or Before February 26, 2007
• Reading Primary Testing Window	March 12 - March 21, 2007
• Make-up Testing Window	March 22 - March 27, 2007

Students and parents should be reminded of the importance of students attending school during the administration of the MSA and the importance of student participation in MSA testing. Maryland was held to the 95% participation requirement under NCLB by the US Department of Education, and schools should do all they could to test all students on MSA or Alt-MSA (as applicable).

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 2007 MSA-Reading, 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 was in the building.

Student Participation

All students in grades 3 through 8 had to participate in the 2007 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. The criteria that students should need to be tested in the Alt-MSA program instead of the MSA-Reading could be viewed in the section 2, Appendix C of the TACM.

The U.S. Department of Education was developing specific guidance related to Modified Assessment, but that guidance, as yet, had not been issued. Students might have been identified through the Individualized Education Program (IEP) process in the current school year as takers of the Mod-MSA. However, since the Mod-MSA was not available, those students had to be assessed using the regular MSA-Reading.

Testing Accommodations

Testing accommodations for students with disabilities (i.e., students having an Individualized Education Program or a Section 504 Plan) and students for English Language Learners (ELL) had to be approved and documented according to the procedures and requirements outlined in the document entitled "*Maryland Accommodations Manual: A Guide to Selecting, Administrating, and Evaluating the Use of Accommodations for Instruction and Assessment*," (MAM). A copy of the most recent edition of this document is available electronically on the LAC and STC web pages at https://docushare.msde.state.md.us/docushare.

No accommodations might 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 MSDE upon request. Please refer to Section 1 of the 2007 TACM for further information regarding testing accommodations.

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

MSA-Reading was administered to those requiring (1) large-print Student Test/Answer Books or (2) Braille Test Books, or (3) Kurzweil^{TM} Test Forms on CD. For large-print Test/Answer Books, Braille Test Books, and Kurzweil^{TM} Test Forms on CD, student responses were transcribed into the standard-size Test/Answer Book following testing.

The pre-printed student ID label was affixed to the standard-size Test/Answer Book containing the transcribed responses, not to the large-print Test/Answer Book or Braille books.

An eligible Test Examiner (TE) transcribed the student responses into a standard-size Test/Answer Book exactly as given by the student. Any original student Test/Answer Books which were used as source documents for transcription was invalidated by drawing a large slash across the student demographic page with a black permanent marker.

Once the student responses had been transcribed, the transcribed Test/Answer Book was returned for scoring with the standardized materials. Specific packing instructions are provided in the TACM in section 3 and 4.

Verbatim Reading Accommodation and Kurzweil^{TM} Test Forms on CD

Students who had a verbatim reading accommodation documented in their Individual Education Plan (IEP), ELL Plan, or Section 504 Plan and who received that accommodation in regular instruction might receive the accommodation on the 2007 MSA-Reading. The accommodation might be provided by a live reader or through technology. Section 2, Appendix F of TACM provided information on verbatim reading instruction for reading items. Technology was used to provide the verbatim reading accommodation, and the software was Kurzweil reading software. Official, secure electronic copies of the test were ordered through the LAC directly from MSDE. MSDE encouraged the use of Kurzweil^{TM} software to ensure uniformity in the delivery of the verbatim reading accommodation throughout the state.

Students using Kurzweil^{TM} software had to familiarity with its operation prior to the test administration.

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 (MSDE, 2007):

> 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 (LEA) or MSDE disciplinary action. (p. 9)

The Test/Answer Books for the 2007 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 (MSDE, 2007):

> 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. 9)

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

- Secure materials: Test/Answer Books (including large-print and Braille), Kurzweil^{TM} test forms on CD, and used scratch paper
- Non-secure materials: TACM, Examiner's Manuals, unused pre-printed student and generic ID labels, unused FedEx return shipping labels, and unused green/orange shipping labels.

Test Format

In 2007, there were 10 forms of MSA-Reading. Different test forms were administered to students in each classroom participating in reading tests, and each test form was identified by color and form number/letter. All forms of the MSA Test/Answer Books for each grade had the same grade designation and picture on the front cover.

The Test/Answer Books were spiraled within a classroom, and each student used a combined Test/Answer Book. Since the Test/Answer Books were scanned for scoring, students were encouraged not to use highlights in any part of the book. Although students might be accustomed to using highlighters in daily instruction, highlighting in the Test/Answer Book could obliterate information in a student's book when it was scanned for scoring. As an alternative to highlighting, students were allowed to lightly circle or underline information in test items or perform calculations to help them in responding, as long as markings did not interfere with the bubbled answer choice area and/or the track marks along the outside margins of each page.

1.7 MSA-Reading Scoring Procedures

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

Once received by Harcourt, Test/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, blank responses (omits), and responses with multiple marks were also assigned a score of zero.

Hand-Scored Items

Test/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. Detailed information on MSA scoring specification can be found in a document, Performance Assessment Scoring Center: Spring 2007 Scoring Specification for MSA-Reading and Math which is available from MSDE.

Scoring Staff

The MSDE had one Room Director (RD) dedicated for each grade level, domain (Reading), and site. The RD worked closely with the PASC Training Supervisor and the PASC Language Arts. The PASC Training Supervisor, Language Arts Specialist, and RDs participated in the anchor-pulling sessions in Maryland. The Room Director/Training Team Leader was responsible for maintaining annotations and meeting minutes from all sessions. These notes were a record of the comments and decisions made by the MSDE personnel and members of the Maryland teacher committee. These notes were utilized by the RD responsible for training the Team Leaders (TLs) and Readers for the respective Maryland prompts. For MSDE scoring projects, PASC had qualified alternate RDs available at the beginning of the project to ensure a timely start of training in the event that the primary RD was unavailable to start as scheduled. The alternate RD acted as a TL unless the RD couldn't fulfill his/her duties.

1) Reader/Scorer

A graduate of a four-year accredited college or university who had successfully passed the PASC new reader exam and new reader training. The readers were eligible to score custom programs for which they have been trained and successfully qualify.

2) Team Leader (TL)

An experienced reader who directly monitored the scoring of a team of readers and retrains as needed. The reader had successfully completed the PASC TL training program.

3) Room Director (RD)

A knowledgeable team leader who had been selected to work with team leaders and the training supervisor to oversee the scoring of several teams. An RD's main duty was to rule on validity of questionable papers and to maintain consistency in scoring decisions. RDs also served as trainers.

4) Reader's Aide (RA)

PASC storeroom personnel whose main responsibilities during scoring were to do copying and printing for the PASC materials center. During anchor pulling, RA responsibility might include duplicating student papers. They might also be assigned a variety of clerical duties.

5) Developers

An experienced PASC reader that was responsible for selecting a wide variety of student responses for such activities as benchmarking, anchor pulling range finding, and training materials. Selected papers were then submitted to MSDE for comment and approval. Developers remained on the project as anchor pulling participants and trainers whenever possible.

6) Trainers

Experienced personnel who were TLs or RDs and selected by the Training Supervisor to train and qualify readers for Maryland. Additionally these experienced personnel might also train new readers and do domain specific training.

Reader Recruitment and Qualifications

All Readers for MSDE had to provide Harcourt's staffing vendor their resume and documentation of a four-year, college degree. As part of the initial screening process for recruiting Readers into Harcourt's general pool, applicants had to respond to an open-ended prompt. This writing sample ensured that all applicants were able to perform the kinds of tasks they would assess. The writing sample was intended to screen out those who couldn't write standard, idiomatically correct English or who couldn't organize their thoughts clearly. The writing prompt was scored by a qualified PASC staff member. If successful on the preliminary screening, applicants then participated in a one-day general introductory training workshop presented by a PASC staff member. These workshops allowed Harcourt to eliminate potential Readers who might seem qualified according to their educational and professional experience but who couldn't learn to score to a scale consistently or who were otherwise unsuitable for assignment to large-scale scoring projects. The PASC staff member who presented the workshop evaluated each potential Reader and submitted these evaluations to the Training Supervisor/Site Supervisor with his/her recommendations. Those who successfully completed the workshop were to Harcourt's general pool of Readers who were potential scorers of Reading assessments. This addition to the general pool did not necessarily qualify these Readers for scoring the MSDE program.

Team Leader Selection and Qualification

The training for new TLs consisted of a two day course focusing on the duties and responsibilities necessary to successfully manage a team of Readers. The workshop was led by two PASC Training Supervisors. The instruction included a review of PASC policies and procedures, sessions on use of the Reader monitoring reports to track a Reader's speed and accuracy, practice annotating anchors and simulated training of the annotated papers, role playing activities which explored various situations that could occur with Readers during the scoring of a project, and Reader counseling and retraining guidelines. Hands-on training on the various TL computer applications were also covered in the work shop. Upon completion of the workshop, the two PASC Training Supervisors reviewed each participant's performance making sure that each had a complete understanding of the TL role and its responsibilities. Any participant they found who had not performed to their satisfaction was not added to the qualified TL list.

Team Leader Project Training

Project-specific TL training for MSDE was conducted in the days immediately preceding scoring and Reader training. This training begun with the RD reading the rubrics aloud and answering any questions the TL or assistant RD might have regarding the rubric. The RD then read each anchor paper aloud to the TLs. Each response in the anchor set was thoroughly explained including the notes and comments of the anchor-pulling committee. Training set A was reviewed next. The TLs scored the training set individually, recorded the scores on the answer sheet and then waited for all TLs to complete the scoring. When everyone had completed scoring the training set, the RD discussed the answers one-by-one, focusing on why it was that score and not another. The RD reviewed with the group the reason for assigning each score point and discussed each paper in its entirety. The TLs were then ready to score Training set B. Training set B was scored and reviewed exactly as Training set A.

Having thoroughly discussed both training sets with the group, the RD explained that in order for a participant to qualify as a TL, it was required that the TL should score at least an 80% perfect match on both of the qualifying sets (Qualification Rules, Attachment M). The TLs scored the first qualifying set individually and recorded their scores on the appropriate answer sheet. As each TL finished scoring, he/she brought the answer sheet to the RD for grading. Each answer was reviewed and any questions the TL had were addressed before the TL attempted the next qualifying set; the TL followed the same procedure with Qualifying set 2. Upon completing the second qualifying set, the TL submitted the answer sheet to the RD for grading. The TL had to achieve at least an 80% perfect match on two of the three Reading sets as specified in the qualification rules or they would be released from the MSDE project.

After the qualification process, the RD continued the training process with the decision set. This set was read aloud and each paper thoroughly explained and discussed. By following these procedures, the RD ensured that the anchor-pulling committees' notes and comments were completely understood.

Team Leader Duties

TLs were responsible for monitoring the training and qualifying of the Readers assigned to their team. The TLs assisted the RD, if requested, during the training of the Readers. The TL was responsible for grading the Readers' qualifying sets and discussing the results with the Readers so everyone received the same direction. The TL certified to the RD and Training Supervisor that the Reader was qualified and recorded the scores under Qualification scores on the Reader evaluation form. The TL was also responsible for monitoring each Reader's assignment of scores to the responses. Additionally, the TL reviewed the daily Reader statistical reports with each individual on the team. The TL consulted the RD regarding variations by the team members from the acceptable standards (80% perfect match for Reading). The TL had the initial responsibility to see that the Reader maintained the set standards through individual retraining. The RD monitored the TL by reviewing team statistics and working one on one with the TL.

Room Director Selection and Qualification

The candidates for RD had been recommended by the PASC Managers or Training Supervisors. The recommendations were based upon the evaluations the candidates received as Readers and TLs and were part of their personnel file. The Training Supervisors met as a group to discuss who might be considered for the position of RD. The Training Supervisor group reviewed the evaluations and the duties that the potential RDs had performed. The candidates generally had been TLs on large-scale projects for multiple teams, and/or they had served as TLs on small-scale projects where TLs trained their individual teams. They had been evaluated on their ability to train Readers as well as their ability to monitor the scoring accuracy and consistency of Readers. These evaluations were submitted in writing at the end of each scoring project by the Readers and RDs that had observed the work of the RD candidates.

Room Director Project Training

The RDs familiarized themselves with the rubric. Any questions regarding the rubric were addressed by the PASC Language Arts, or MSDE. The next step was for the RD/TTL to prepare the anchors by annotating each response to all score points in the Anchor Set utilizing the notes from the anchor-pulling session. The MSDE approved the anchor-pulling notes and the Training Supervisor confirmed that the RD had accurately added the anchor-pulling notes to the training materials. The RD continued the process by annotating the training sets and decision sets with all notes and comments from the anchor-pulling session. Additionally, the RDs became familiar with the wording of all of the other prompts for the administration on which they are assigned.

Room Director Duties

The RD's job was to conduct the training of the TLs and Readers, oversee the actual scoring of the papers, monitor the work of the TL, and act as the decision maker for situations or questions that may arise during the scoring process. For example, all invalid (foreign language, off-topic, off-mode, etc.) responses were reviewed by the RD, who had to confirm any such decision and ensure consistency of decisions (Blanks were confirmed at the TL level and did not require RD confirmation). Additionally the RD and TL (after approval of Training Supervisor) conducted all resolution readings. Responses for which scores were non-matching or non-adjacent were automatically routed to the RD for an independent resolution scoring. The resolution score became the reported score.

The RD was familiar with all prompts and trained the TLs and Readers to recognize these alternate prompts. Thus, should the student place his/her answer in the wrong place, the answer was recognized by the RD, who could electronically move the response to the appropriate space for scoring by a Reader qualified on the appropriate prompt. The RD also reviewed any potential questionable content responses and forwarded those to the Training Supervisor to consult with the MSDE before processing.

The RD was also responsible for daily statistical review and analysis of all monitoring reports to ensure the quality of the scoring within the room. Review of the data allowed the RD not only to monitor the Reader but also to provide the TL with additional input. Available data included 1) individual Reader agreement rates between two independent scorings; 2) score point distributions by Reader and trend review; 3) prompt statistics for agreement rates and score point distributions; 4) Resolution data.

Project Scoring Parameters

MSDE had a long-standing history of implementing assessments that were composed of multiple item types: selected response (SR), brief constructed response (BCR), extended constructed response (ECR), and gridded or student-produced response (SPR). The MSA contained all such item types for operational scoring and each of the 10 forms per grade/subject also contained field-test items of each of these types. Open-ended items were scored using a generic rubric as follows:

- Reading items were scored on a 0-3 scale (BCRs only in Reading)

All MSA response documents were image scanned at Harcourt's scoring center in San Antonio, Texas. The image scanner captured document identification (ID), demographic information, SR responses, and creates a bi-tonal image of the entire document, allowing images of the BCR responses to be distributed to readers for human scoring while images of the SR, SPR and all other data were made available to Scoring Editing for human review.

All constructed responses were scored by Harcourt's Performance Assessment Scoring Center (PASC). The PASC mission was to provide accurate, reliable, on-time scores for all student responses entrusted to our care. PASC maintained large pools of qualified, trained, professional readers who were well-experienced in scoring a wide range of writing assessments and open-ended assessments in reading, mathematics, science, social science, and other subjects, at each of our scoring sites.

Reader Project Training

Reader training was lead by the RD/TTL and was conducted utilizing our central scoring model. There was one RD responsible for each site, grade and Domain (Reading). After all student responses were scored for the first item, the RD reconvened the group and trained the second item. Training began with the definition and an overview of holistic scoring. Training continued with a reading and discussion of the generic rubric and then the student responses in the anchor set were read and discussed. In the anchor set the scores had been recorded on the student responses and were arranged in ascending point-scale order. Each annotated anchor response was read aloud and discussed thoroughly. Emphasis was placed on the Readers' understanding of how the responses differed from one another in incremental quality and how each response reflected the description of its score point as generalized in the scoring rubric and how each reflected the MSDE's standard for application of each score point.

Once Readers had all their questions answered and the discussion of the anchor set was finished, the Readers began to score the first training set. Each Reader independently read and scored the responses in the training set. The trainer scored and recorded each reader's responses on a training record form. The correct scores were then read to the group when everyone had completed the scoring. In addition, each training paper was discussed as to reasons for applying each given score. At this point, Readers interacted with the RD in discussing the characteristics of each response that earned the assigned score point. The same format was followed for each training set. During this process, the job of the Reader was to internalize the scoring scale and adjust his or her individual scoring to conform to that scale. Once all training papers had been scored and fully discussed, Readers began the qualifying process.

For MSDE, there were three qualifying sets. MSDE informed PASC in writing for each specific administration how many qualifying sets were approved and were available to the Readers. Readers must score at least 80% match on two of three qualifying sets for Reading.

Inter-Rater Agreement

Harcourt's scoring system generated many kinds of internal monitoring reports that enabled the project leadership to monitor the accuracy and consistency of MSDE scoring. These reports were compiled by prompt, listed the entire prompt's Readers and provided the results of their scoring for each day. Information on these reports included the number of responses read by the Readers during the period, the number and percent of invalid responses and the number of responses for which there had been a second reading. The number of responses with second readings provided data that allowed for reporting of the number and percent of responses with perfect agreement; the number and percent of responses on which the first Reader was a point lower than the second Reader; the number and percent of responses on which the first Reader was a point higher than the second Reader (Adjacent) and the number and percent of responses differing by more than one score point (Non-Adjacent/Non-Perfect). The Training Supervisor also reviewed the daily statistical reports to identify individuals or teams who might need retraining in order to provide continuous scoring consistency on the project. MSDE received data summary reports. Statistical summaries of inter-rater reliability can be found in section 3.4, *inter-rater reliability*.

Reader Retraining

When a Reader's performance fell below acceptable parameters for a project, the Reader was retrained. Retraining was the process by which the RD or TL utilized a number of methods such as individual tutoring on problem score points, individual review of selected responses and anchor and rubric review to get a Reader back on track with the guidelines provided by a specific program. Group retraining was conducted by the RD every Monday (or following any extended break) during the scoring project. In addition, daily retraining occurred as deemed necessary by the MSDE representative and Training Supervisor.

Read Behinds

Harcourt's system allowed TLs and/or RDs to conduct read behinds as an additional monitoring method. When conducting read behinds, the TL or RD received images of student responses and the scores assigned by the Reader. Responses selected for read behinds might be randomly selected or might be targeted read behinds (i.e., responses receiving specific scores, etc.). These read behinds were very useful in tracking specific areas of confusion for a given Reader or group of Readers and assisted the TL and RD in knowing just how to direct retraining activities for individual Readers or teams. The initial read behind percentage was set at 50%. This percentage might be adjusted either higher or lower by the TL based upon the performance of the Reader.

Retrain Readers with <80% Agreement rates

It was the responsibility of the Team Leader ("TL") to not only address questions and provide guidance to the Readers, but to also monitor and manage performance; this included Calibrations, Read Behinds, Agreement rates and Resolution rates. At times, TLs could become easily side-tracked and spend more time acting as a resource for Readers more so than managing performance. PASC had identified this issue and planed to allocate additional TLs whose primary job responsibility was to manage/monitor performance. This level of staffing allowed us to monitor each Reader daily and provided retraining when the level of acceptable performance had not been met.

Pre-"Live" training on Field Test prompts

For 2007, PASC used scored student responses from the appropriate field test administration. This allowed the Readers to build familiarity with the program prior to live scoring.

Trainers Earlier and Longer

In addition to increasing the number of TLs dedicated to the program, PASC also felt it more effective to expedite and extend the time the Trainers were onsite. PASC trained a qualified individual at each site to act as the remote Trainer once the primary left. This individual was responsible for re-training Readers as needed.

Technology

PASC utilized the Student Response Window ("SRW") application supplemented with the PASC Performance Monitoring ("PPM") system that provided the Reader and/or client a "real-time" look into the scoring of each item. This system allowed the viewer to filter the information to provide detail down to the prompt, item, domain, site, Reader, etc. level. This helped in reporting results and creating a custom view of the program. The most important attribute of the application was its security features. Even though Readers in the same room could access the SRW application, each Reader could be setup to view different information within a program. This allowed segregation per domain or even grade within a partitioned room. This system greatly enhanced the quality and timeliness of reporting.

Scoring rules for MSA-Reading

The following scoring rules were applied to MSA-Reading BCR items:

- Reading BCR items were scored:

  0, 1, 2, 3 with two readings
- Score were the higher of the 1st and 2nd Readers' scores provided they were adjacent. If they are equal that was the score.
- The score result from adjacent reads was a decimal numeric; round this up to the nearest whole number.
- For example:

1st Reader	2nd Reader	Final Score
1	2	2
2	3	3

- A resolution reader was used if two non-adjacent initial scores were received.
- The resolution reader's score was used in place of both the 1st and 2nd Readers' scores.
- For example:

1st Reader	2nd Reader	Resolution Reader	Final Score
0	2	1	1
0	3	2	2
1	3	3	3
2	0	1	1
3	0	2	2

Development Procedures for Anchor Pulling

For a given reading prompt, the PASC Developers had the following responsibilities (A developer was a PASC Reader who was selected by the PASC Training Supervisor to prepare sets of papers for client approval. These experienced Readers were judged by the Training Supervisor for their ability to recognize and assemble a wide variety of responses. A Material Development Evaluation was completed by the Language Arts Specialist for review by the Training Supervisor. This evaluation was part of the developer's personnel file. The developer also participated with the clients as a facilitator during the anchor-pulling session in order to make notes and be prepared to assemble the finished sets to the client's specifications. In the case of the MSDE, the developer was also the RD):

1) To know the prompt and the rubric thoroughly

2) To read responses

- Looked for responses that seemed to represent the full range of quality as described in the rubric.
- Searched all orders for responses, with particular emphasis on the state's high performing districts.
- Included not only papers that were homogeneous in their level of quality but also papers that differed in quality from variable to variable but which could be given an overall classification of High, Medium, and Low.
- Marked High, Medium, and Low papers—marked especially good ones that might be the potentially top scores.
- Identified and flagged problem papers—off-topic, off-task, verbatim copying, strange, potential teacher interference, etc.
- Marked the flag with score range or the nature of the problem and paper ID.

3) To sort copies

- Copies were sorted into piles, reflecting the nature of the flag—all potential high papers were together, all potential medium papers were together, etc., with all problem papers grouped together.
- For problem or decision papers, duplicates of types of problems were culled. The best example of each problem type was retained; the rest were set aside for possible future use.

4) To develop sets for anchor pulling

- Decided which particular papers from the sorted piles should go into which set for anchor pulling. Each paper selected went into only one set.
- Used the following guidelines in deciding for which set a paper was most appropriate.

  A. ***Anchor set***: At least three examples of each score point, depending upon the score scale (no invalids). These had to be clean papers but should illustrate different types of the same score point, if there were such clear differences. Once completed, this set was submitted to the Training Supervisor and to MSDE for review and approval.

B. ***Decision set***: This had to be a set of whatever size necessary to illustrate the various kinds of problems that might arise with this prompt or item. If the number of such responses was small, these might be incorporated into the first training set instead of being grouped into a separate additional set.

C. ***Training sets***: These were at least two sets of up to 20 papers each (again, this varied according to the score point scale). They had to contain a range of responses including clean papers, line papers, and problem papers. The responses had to be in random order of quality and unmarked.

D. ***Qualifying sets***: There were three sets of these. Generally there were 10 responses per set, but could be fewer, depending upon the score scale. These had to consist heavily of clean papers but not exclusively so. One of the sets might include an example of an invalid response, but it should be clearly so.

E. ***Calibration sets (validity sets)***: These were composed of five responses of mixed quality, arranged in random order. Harcourt created as many different sets as there were expected to be scoring days on a single prompt or group of items—minus one or two for the training day and the initial scoring day.

Comprehensive notes concerning the specific problems presented in these papers (and the solutions as decided by the committee during the anchor-pulling session) were to be recorded by the Harcourt representatives (developers and Training Specialists) and were to be discussed with the Readers during training. Any subsequent notes or communication from MSDE were incorporated into the training material as well.

Anchor Pulling Procedures

The objective of anchor pulling was for the team members to arrive at a consensus as to the score of each paper in the proposed training materials. These sessions were attended by Maryland educators, MSDE and from PASC the Language Arts Specialists, Manager, Training Supervisor and the developers who selected and prepared all of the papers that would be reviewed. These papers and their corresponding scores formed the basis of selecting final Anchor Sets, Decision Sets Training Sets and Qualifying Sets. Discussions among the team members were important, as they revealed what kinds of qualities characterized certain score points. The most difficult aspects involved balancing widely discrepant qualities found in the same paper and defining the line between adjacent scores.

During formal anchor pulling, the procedure for assigning scores to the papers in each set was as follows:

- Papers were read aloud and discussed by the anchor-pulling panel. Reading aloud focused attention on the ideas presented—or what the student had to say—allowing the panel members to divorce themselves from how the paper looked or how well it had been edited.
- After each response was read, each panel member independently assigned a score. An overall tentative score was assigned to each response on which there seemed to be consensus. However, all assigned scores at this point, even those on responses for which there were complete agreement, were provisional and subject to change based on later considerations.

- Each subsequent set was read and scored by each panel member, using the tentative scores on the previous sets as guidelines. After each set had been read, the results were recorded on a consensus sheet and discussed.

The responses in which score points were not in perfect agreement were discussed starting with the lowest, but least controversial, score point. The papers that had the widest discrepancies of assigned scores around this lowest score point were discussed next before moving to the papers whose assigned scores were in the next higher range. There might be frequent reference to previous sets to make sure that decisions on score points were consistent.

This iterative process of reading, charting, and discussing successive sets had three goals:

- It established scores on papers for which there was virtual agreement.
- It identified papers that were on the line between two adjacent scores, forcing the clarification of that line.
- It contributed to understanding the rationale behind scoring decisions.

During this process, the tentative scores assigned to papers in earlier sets became firm.

1.8 Classical Analyses with SAT10 Form-to Form Linking Common Items

The main purpose of this analysis was to check that the groups taking the two operational forms were essentially equivalent. 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 1.11.

Table 1.11 Descriptive Statistics for the 2007 MSA-Reading Form-to-Form Common Items

Grade	Form	No. of Items	*N*	*M*	*SD*
3	A	25	29,732	18.05	4.79
	B	25	29,675	18.09	4.78
4	A	25	30,174	19.66	3.97
	B	25	29,955	19.71	3.92
5	A	25	30,883	17.76	4.60
	B	25	30,693	17.74	4.58
6	A	25	31,339	18.27	4.85
	B	25	31,128	18.35	4.82
7	A	25	32,114	17.43	4.79
	B	25	31,846	17.45	4.80
8	A	25	32,609	17.08	4.47
	B	25	32,452	17.14	4.43

Note. Form A designates the operational portion of Forms 1, 3, 5, 7, and 9, which is identical. Form B designates the operational portion of Forms 2, 4, 6, 8, and 10, which is identical.
Note. Analyses were conducted with a whole population.

1.9 P-Value Check with SAT10 Year-to-Year Linking Common Items

Tables 1.12 through 1.17 provide information about how much the p-value of each *SAT10* common item changed in consecutive years. It should be noted that these analyses conducted with a whole population. The general conclusion could be drawn from the results that most of the p-values in Year 2007 were pretty much the same as those in Year 2006 across all grades.

Table 1.12 Year 2006 vs. Year 2007 Linking Common Item P-Value Comparison: Grade 3

Item Number	Item Type	Y06 FA	Y06 FB	Y07 FA	Y07 FB
2	SR	0.94	0.94	0.93	0.93
5	SR	0.90	0.91	0.88	0.89
6	SR	0.69	0.70	0.66	0.67
9	SR	0.87	0.87	0.85	0.86
11	SR	0.70	0.70	0.68	0.68
15	SR	0.85	0.85	0.83	0.83
18	SR	0.74	0.73	0.71	0.71
20	SR	0.43	0.43	0.40	0.41
23	SR	0.68	0.68	0.67	0.67
30	SR	0.50	0.51	0.52	0.52
31	SR	0.74	0.74	0.72	0.72
32	SR	0.72	0.71	0.69	0.69
34	SR	0.79	0.78	0.76	0.76
41	SR	0.78	0.78	0.75	0.75
44	SR	0.92	0.92	0.91	0.91
49	SR	0.67	0.66	0.66	0.66
55	SR	0.68	0.68	0.65	0.65
56	SR	0.48	0.48	0.46	0.47
57	SR	0.83	0.82	0.82	0.82
58	SR	0.91	0.91	0.90	0.90
59	SR	0.85	0.85	0.83	0.83
61	SR	0.56	0.56	0.55	0.56
68	SR	0.78	0.78	0.77	0.76
69	SR	0.81	0.81	0.81	0.81
70	SR	0.65	0.65	0.62	0.62

Note. Analyses were conducted with a whole population.

Descriptive Statistics for Year-to-Year Linking Common Items: Grade 3

Grade	Form	No. of Items	*M*	*SD*
3	Y06 FA	25	0.74	0.14
	Y06 FB	25	0.74	0.14
	Y07 FA	25	0.72	0.14
	Y07 FB	25	0.72	0.14

Table 1.13 Year 2006 vs. Year 2007 Linking Common Item P-Value Comparison: Grade 4

Item Number	Item Type	Y06 FA	Y06 FB	Y07 FA	Y07 FB
1	SR	0.99	0.99	0.99	0.99
4	SR	0.94	0.94	0.95	0.95
9	SR	0.84	0.83	0.84	0.84
10	SR	0.89	0.89	0.90	0.90
18	SR	0.79	0.80	0.78	0.78
23	SR	0.86	0.86	0.86	0.86
24	SR	0.81	0.81	0.80	0.80
29	SR	0.91	0.92	0.94	0.94
35	SR	0.83	0.83	0.86	0.86
38	SR	0.71	0.72	0.73	0.73
41	SR	0.83	0.83	0.82	0.82
42	SR	0.76	0.76	0.73	0.73
43	SR	0.86	0.86	0.85	0.85
44	SR	0.82	0.82	0.82	0.81
45	SR	0.45	0.45	0.43	0.44
46	SR	0.95	0.95	0.95	0.95
47	SR	0.82	0.82	0.82	0.82
50	SR	0.84	0.83	0.82	0.82
51	SR	0.94	0.94	0.95	0.96
52	SR	0.62	0.62	0.62	0.62
53	SR	0.51	0.51	0.51	0.51
54	SR	0.38	0.38	0.37	0.36
55	SR	0.92	0.92	0.92	0.92
62	SR	0.79	0.78	0.79	0.79
64	SR	0.66	0.66	0.64	0.65

Note. Analyses were conducted with a whole population.

Descriptive Statistics for Year-to-Year Linking Common Items: Grade 4

Grade	Form	No. of Items	*M*	*SD*
4	Y06 FA	25	0.79	0.16
	Y06 FB	25	0.79	0.16
	Y07 FA	25	0.79	0.16
	Y07 FB	25	0.79	0.16

Table 1.14 Year 2006 vs. Year 2007 Linking Common Item P-Value Comparison: Grade 5

Item Number	Item Type	Y06 FA	Y06 FB	Y07 FA	Y07 FB
4	SR	0.61	0.62	0.60	0.59
5	SR	0.57	0.57	0.57	0.56
6	SR	0.64	0.63	0.63	0.64
9	SR	0.91	0.91	0.93	0.93
10	SR	0.91	0.91	0.90	0.90
11	SR	0.84	0.85	0.84	0.84
13	SR	0.85	0.84	0.86	0.86
16	SR	0.83	0.83	0.84	0.85
17	SR	0.80	0.80	0.80	0.80
19	SR	0.75	0.75	0.75	0.75
21	SR	0.82	0.82	0.82	0.82
23	SR	0.59	0.59	0.57	0.57
25	SR	0.73	0.73	0.72	0.72
26	SR	0.70	0.70	0.72	0.72
28	SR	0.54	0.55	0.54	0.54
31	SR	0.60	0.59	0.58	0.58
32	SR	0.70	0.70	0.69	0.70
33	SR	0.81	0.81	0.81	0.81
34	SR	0.44	0.43	0.39	0.39
35	SR	0.67	0.68	0.69	0.68
37	SR	0.68	0.68	0.63	0.63
41	SR	0.77	0.77	0.78	0.78
44	SR	0.66	0.66	0.67	0.66
45	SR	0.57	0.57	0.57	0.56
49	SR	0.85	0.85	0.87	0.87

Descriptive Statistics for Year-to-Year Linking Common Items: Grade 5

Grade	Form	No. of Items	*M*	*SD*
5	Y06 FA	25	0.71	0.12
	Y06 FB	25	0.71	0.13
	Y07 FA	25	0.71	0.14
	Y07 FB	25	0.71	0.14

Table 1.15 Year 2006 vs. Year 2007 Linking Common Item P-Value Comparison: Grade 6

Item Number	Item Type	Y06 FA	Y06 FB	Y07 FA	Y07 FB
1	SR	0.79	0.79	0.77	0.77
5	SR	0.54	0.54	0.51	0.51
8	SR	0.63	0.63	0.59	0.60
9	SR	0.92	0.92	0.94	0.94
10	SR	0.75	0.75	0.74	0.74
14	SR	0.76	0.76	0.77	0.78
16	SR	0.81	0.81	0.79	0.80
18	SR	0.84	0.84	0.83	0.84
21	SR	0.89	0.88	0.88	0.89
22	SR	0.78	0.78	0.78	0.78
23	SR	0.70	0.71	0.69	0.70
24	SR	0.71	0.71	0.69	0.70
25	SR	0.68	0.68	0.68	0.69
28	SR	0.80	0.80	0.79	0.79
29	SR	0.66	0.66	0.63	0.63
30	SR	0.69	0.69	0.69	0.69
32	SR	0.86	0.86	0.86	0.86
33	SR	0.34	0.33	0.32	0.32
34	SR	0.83	0.83	0.81	0.82
35	SR	0.64	0.63	0.64	0.65
36	SR	0.79	0.79	0.79	0.79
37	SR	0.82	0.82	0.82	0.82
38	SR	0.61	0.61	0.58	0.58
39	SR	0.87	0.87	0.88	0.88
40	SR	0.77	0.76	0.78	0.78

Descriptive Statistics for Year-to-Year Linking Common Items: Grade 6

Grade	Form	No. of Items	*M*	*SD*
6	Y06 FA	25	0.74	0.13
	Y06 FB	25	0.74	0.13
	Y07 FA	25	0.74	0.13
	Y07 FB	25	0.74	0.13

Table 1.16 Year 2006 vs. Year 2007 Linking Common Item P-Value Comparison: Grade 7

Item Number	Item Type	Y06 FA	Y06 FB	Y07 FA	Y07 FB
1	SR	0.90	0.90	0.89	0.89
3	SR	0.85	0.85	0.85	0.85
6	SR	0.49	0.49	0.47	0.48
8	SR	0.41	0.41	0.37	0.37
10	SR	0.61	0.62	0.62	0.62
14	SR	0.70	0.70	0.69	0.69
16	SR	0.66	0.65	0.65	0.64
20	SR	0.84	0.84	0.82	0.83
22	SR	0.88	0.88	0.87	0.87
23	SR	0.53	0.54	0.55	0.54
26	SR	0.77	0.77	0.77	0.77
27	SR	0.53	0.53	0.52	0.51
28	SR	0.65	0.65	0.63	0.63
31	SR	0.57	0.57	0.58	0.58
32	SR	0.87	0.87	0.86	0.85
33	SR	0.63	0.62	0.61	0.61
36	SR	0.90	0.90	0.89	0.90
37	SR	0.73	0.73	0.71	0.71
38	SR	0.76	0.76	0.73	0.74
39	SR	0.65	0.65	0.63	0.63
40	SR	0.88	0.88	0.87	0.87
41	SR	0.78	0.77	0.77	0.77
42	SR	0.73	0.73	0.70	0.71
43	SR	0.74	0.74	0.71	0.71
44	SR	0.69	0.69	0.68	0.68

Descriptive Statistics for Year-to-Year Linking Common Items: Grade 7

Grade	Form	No. of Items	*M*	*SD*
7	Y06 FA	25	0.71	0.14
	Y06 FB	25	0.71	0.14
	Y07 FA	25	0.70	0.14
	Y07 FB	25	0.70	0.14

Table 1.17 Year 2006 vs. Year 2007 Linking Common Item P-Value Comparison: Grade 8

Item Number	Item Type	Y06 FA	Y06 FB	Y07 FA	Y07 FB
3	SR	0.66	0.66	0.62	0.63
6	SR	0.54	0.54	0.53	0.53
8	SR	0.56	0.57	0.54	0.54
9	SR	0.93	0.94	0.93	0.93
22	SR	0.97	0.97	0.98	0.98
23	SR	0.56	0.56	0.55	0.55
24	SR	0.59	0.59	0.58	0.58
25	SR	0.82	0.82	0.81	0.81
26	SR	0.63	0.63	0.62	0.63
29	SR	0.74	0.74	0.71	0.72
30	SR	0.62	0.61	0.62	0.61
31	SR	0.65	0.64	0.61	0.62
32	SR	0.49	0.50	0.52	0.52
33	SR	0.65	0.64	0.62	0.62
35	SR	0.78	0.79	0.76	0.76
36	SR	0.53	0.53	0.53	0.53
37	SR	0.73	0.73	0.74	0.74
38	SR	0.75	0.75	0.75	0.75
39	SR	0.53	0.53	0.52	0.52
41	SR	0.84	0.84	0.85	0.85
44	SR	0.74	0.75	0.73	0.73
46	SR	0.79	0.80	0.80	0.80
48	SR	0.73	0.73	0.70	0.70
49	SR	0.74	0.74	0.74	0.75
50	SR	0.74	0.75	0.73	0.74

Descriptive Statistics for Year-to-Year Linking Common Items: Grade 8

Grade	Form	No. of Items	*M*	*SD*
8	Y06 FA	25	0.69	0.13
	Y06 FB	25	0.69	0.13
	Y07 FA	25	0.69	0.13
	Y07 FB	25	0.69	0.13

1.10 Validation Check with Augmented Items

To collect information about how much the same items that appeared on the test forms in consecutive years changed in terms of item difficulty, difficulty indices such as p-value and Rasch difficulty were calculated.

First, it should be noted these items were at first augmented as field test items in Year 2005 and appeared as operational test items in Year 2007 as seen from Table 1.18. Second, Year 2007 Forms 1, 3, 5, 7, and 9 are the same, and Year 2007 Forms 2, 4, 6, 8, and 10 are the same except for the field test portion. Third, in Tables 1.19 through 1.54, item numbers were given by those of Year 2007. Detailed information about the specific test design and construction of Year 2007 can be obtained from section 1.5, Test Structure of the 2007 MSA-Reading.

First of all, it should be noted that Year 2005 p-value was calculated with a field-tested sample and Year 2007 p-value was calculated with a whole population. P-value of BCR item was the item mean score divided by the item score range. In addition, the numbers in “Omits” in each table were very substantial and included students who did not responded at all. Item p-value (easiness) results indicated that in general, most of the p-values in Year 2007 increased somewhat compared to those in Year 2005 for grades 3 through 7. However, most of p-values were much the same as those in Year 2005 for grade 8.

With respect to Rasch difficulty analysis, most of the items in Year 2007 became easier compared to those in Year 2005 for grades 3 though 7. For grade 8, most of the item difficulties in Year 2007 were much the same as those in Year 2005. It should be noted that Rasch difficulties were based on the same scale (e.g., linked to Year 2003 or Year 2004).

In conclusion, both p-value and Rasch difficulty results reflected the same phenomenon, indicating that most of the items became easier.

Table 1.18 Form Identification for Items Appearing Year 2005 and Year 2007: Grades 3 through 8

Grade	Year 2005	Year 2007
3	Form 1, 3	Form A (1, 3, 5, 7, 9)
	Forms 2 and 4 /1 and 3	Form B (2, 4, 6, 8, 10)
4	Form 1, 3	Form A (1, 3, 5, 7, 9)
	Form 2, 4	Form B (2, 4, 6, 8, 10)
5	Form 1, 3	Form A (1, 3, 5, 7, 9)
	Form 2, 4	Form B (2, 4, 6, 8, 10)
6	Form 2, 1	Form A (1, 3, 5, 7, 9)
	Form 4, 3	Form B (2, 4, 6, 8, 10)
7	Form 1, 3	Form A (1, 3, 5, 7, 9)
	Form 2, 4	Form B (2, 4, 6, 8, 10)
8	Form 1, 3	Form A (1, 3, 5, 7, 9)
	Form 2, 3	Form B (2, 4, 6, 8, 10)

Table 1.19 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 3 Form A

Item Number	Item Type	Year 05	Year 07
71	SR	0.75	0.80
72	BCR	0.45	0.55
73	SR	0.85	0.87
74	SR	0.64	0.70
75	BCR	0.51	0.51
76	SR	0.51	0.57
77	SR	0.56	0.57
78	BCR	0.34	0.38
79	SR	0.48	0.45
80	SR	0.33	0.37
81	BCR	0.32	0.36
82	SR	0.56	0.66

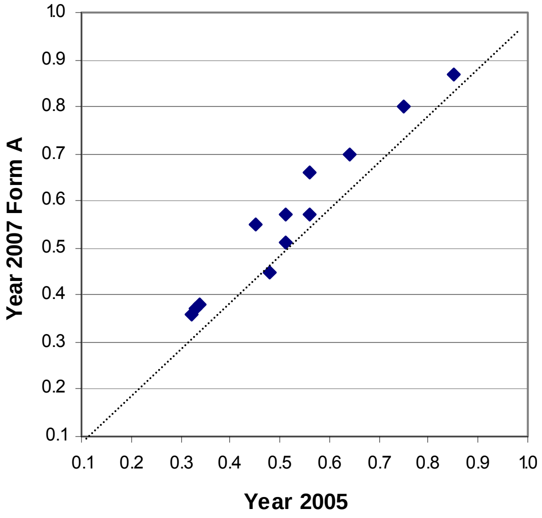

Table 1.20 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 3 Form A

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	72	BCR	2,199	1.36	0.65	5.00	53.00	37.00	3.00	1.00
2005	75	BCR	2,199	1.52	0.67	5.00	40.00	49.00	4.00	1.00
2005	78	BCR	2,308	1.03	0.77	25.00	47.00	25.00	2.00	1.00
2005	81	BCR	2,308	0.97	0.67	22.00	56.00	19.00	1.00	2.00
2007	72	BCR	29,732	1.66	0.66	1.94	36.40	52.76	8.13	0.77
2007	75	BCR	29,732	1.53	0.59	2.20	42.13	53.14	1.57	0.95
2007	78	BCR	29,732	1.15	0.77	20.82	44.24	32.23	2.05	0.66
2007	81	BCR	29,732	1.08	0.68	18.02	55.10	25.00	0.81	1.07

Table 1.21 Augmented IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 3 Form A

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	71	SR	-0.3692			
2005	72	BCR	1.4454	-3.4971	0.1538	3.3433
2005	73	SR	-1.0949			
2005	74	SR	0.3514			
2005	75	BCR	1.1364	-3.1808	-0.2633	3.4441
2005	76	SR	0.9950			
2005	77	SR	0.7788			
2005	78	BCR	2.3113	-2.2846	-0.2140	2.4986
2005	79	SR	1.1558			
2005	80	SR	1.9304			
2005	81	BCR	2.8046	-3.0911	-0.1739	3.2650
2005	82	SR	0.6826			
2007	71	SR	-0.6027			
2007	72	BCR	0.6767	-3.7457	-0.1823	3.9279
2007	73	SR	-1.3357			
2007	74	SR	0.0066			
2007	75	BCR	1.3267	-4.1588	-0.2494	4.4082
2007	76	SR	0.8172			
2007	77	SR	0.7501			
2007	78	BCR	2.1952	-2.4813	-0.5270	3.0083
2007	79	SR	1.4026			
2007	80	SR	1.8600			
2007	81	BCR	2.7146	-3.1880	-0.2086	3.3966
2007	82	SR	0.1785			

Note. These Rasch difficulties were based on a common scale.

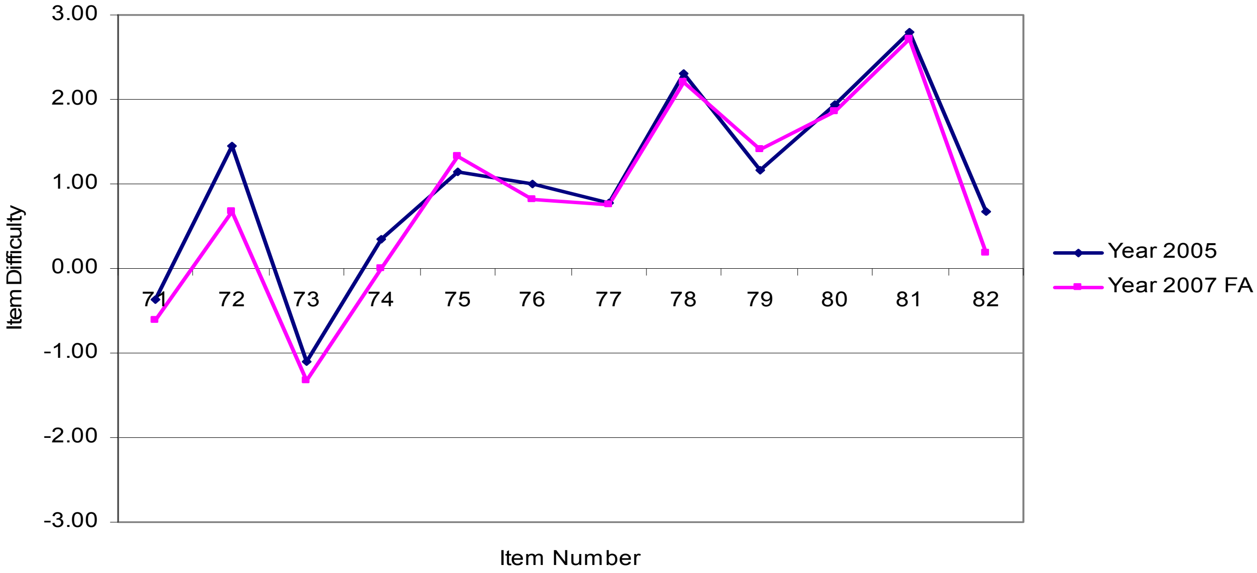

Figure 1.2 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 3 Form A

Table 1.22 Augmented Item P-Value Comparison for Year 2005/2004 vs. Year 2007: Grade 3 Form B

Item Number	Item Type	Year 05/04	Year 07
71	SR	0.79	0.80
72	BCR	0.48	0.56
73	SR	0.77	0.79
74	SR	0.77	0.78
75	BCR	0.56	0.54
76	SR	0.50	0.48
77	SR	0.55	0.54
78	BCR	0.43	0.51
79	SR	0.56	0.58
80	SR	0.44	0.40
81	BCR	0.42	0.50
82	SR	0.58	0.65

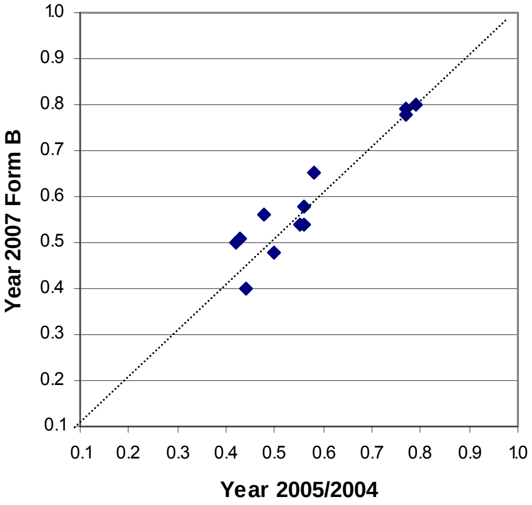

Table 1.23 BCR Item Score-Point Distribution Comparison for Year 2005/04 vs. Year 2007: Grade 3 Form B

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	72	BCR	27,592	1.43	0.89	20.00	23.00	50.00	7.00	0.00
2005	75	BCR	27,592	1.67	0.78	3.00	40.00	41.00	15.00	1.00
2004	78	BCR	28,301	1.29	0.73	10.00	51.00	32.00	4.00	2.00
2004	81	BCR	28,301	1.27	0.70	11.00	52.00	33.00	3.00	1.00
2007	72	BCR	29,675	1.68	0.84	13.60	14.07	61.23	10.45	0.65
2007	75	BCR	29,675	1.61	0.81	4.97	41.12	37.75	14.64	1.52
2007	78	BCR	29,675	1.54	0.76	5.56	43.29	40.36	9.91	0.88
2007	81	BCR	29,675	1.49	0.71	5.18	44.99	42.23	6.52	1.08

Table 1.24 Augment IRT Item Difficulty Comparison for Year 2005/2004 vs. Year 2007: Grade 3 Form B

Year	Item #	Item Type	Rasch Difficulty	Step	Step	Step
				0-1	1-2	2-3
2005	71	SR	-0.5402			
2005	72	BCR	1.4353	-1.2373	-1.1211	2.3583
2005	73	SR	-0.4450			
2005	74	SR	-0.4370			
2005	75	BCR	0.4477	-2.9007	0.4904	2.4104
2005	76	SR	1.1262			
2004	77	SR	0.6985			
2004	78	BCR	1.4359	-2.8478	0.1608	2.6870
2004	79	SR	0.6020			
2004	80	SR	1.2637			
2004	81	BCR	1.3251	-2.8354	-0.0638	2.8993
2004	82	SR	0.4816			
2007	71	SR	-0.5812			
2007	72	BCR	1.0200	-0.9972	-1.3939	2.3911
2007	73	SR	-0.5350			
2007	74	SR	-0.4499			
2007	75	BCR	0.6505	-2.4498	0.3899	2.0599
2007	76	SR	1.1214			
2007	77	SR	0.8924			
2007	78	BCR	1.0854	-2.5958	0.2578	2.3381
2007	79	SR	0.7167			
2007	80	SR	1.6168			
2007	81	BCR	1.0017	-2.8775	0.1426	2.7349
2007	82	SR	0.2386			

Note. These Rasch difficulties were based on a common scale.

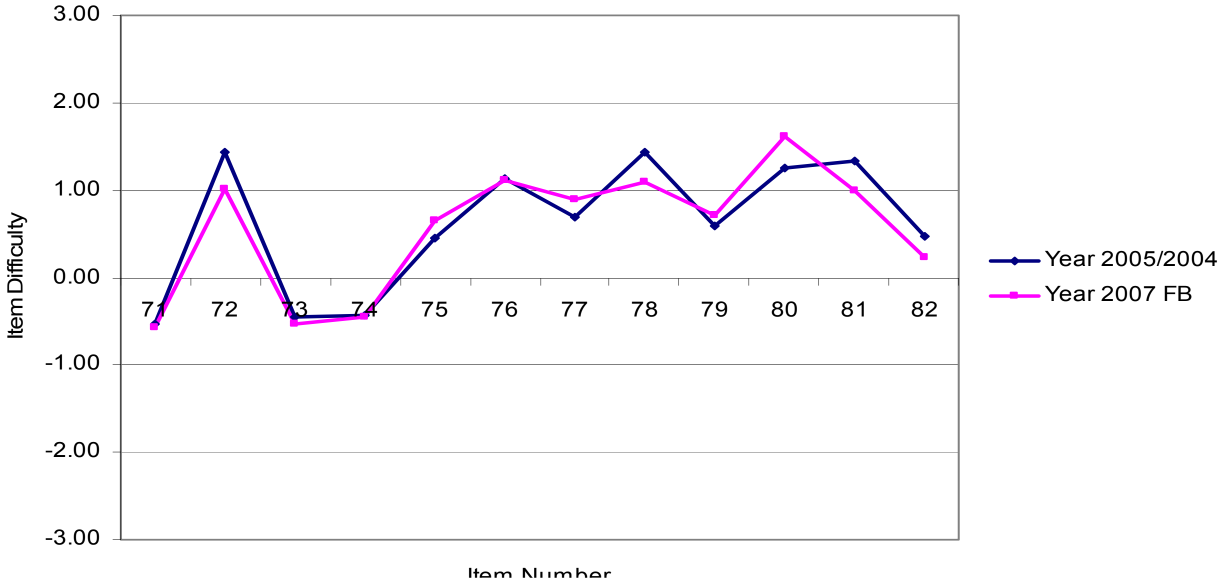

Figure 1.3 Augmented IRT Item Difficulty Comparison Plot for Year 2005/2004 vs. Year 2007: Grade 3 Form B

Table 1.25 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 4 Form A

Item Number	Item Type	Year 05	Year 07
71	SR	0.51	0.56
72	BCR	0.48	0.55
73	SR	0.55	0.61
74	SR	0.79	0.79
75	BCR	0.48	0.52
76	SR	0.76	0.78
77	SR	0.58	0.56
78	BCR	0.47	0.49
79	SR	0.79	0.82
80	SR	0.52	0.56
81	BCR	0.24	0.28
82	SR	0.67	0.73

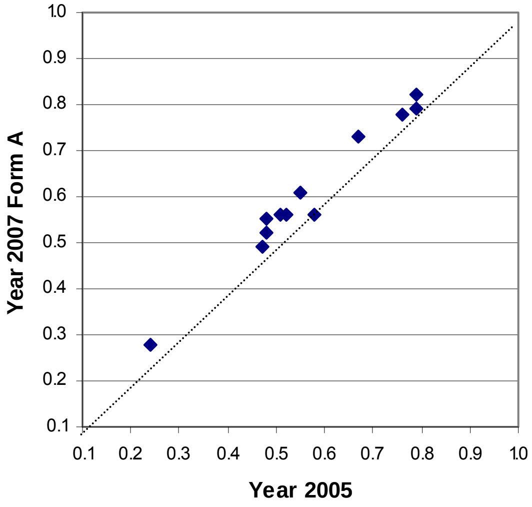

Table 1.26 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 4 Form A

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	72	BCR	2,188	1.45	0.61	5.00	45.00	49.00	1.00	1.00
2005	75	BCR	2,188	1.44	0.60	5.00	45.00	49.00	0.00	1.00
2005	78	BCR	2,266	1.41	0.57	2.00	55.00	41.00	2.00	1.00
2005	81	BCR	2,266	0.71	0.66	39.00	48.00	11.00	0.00	2.00
2007	72	BCR	30,174	1.66	0.53	2.03	30.64	66.52	0.64	0.17
2007	75	BCR	30,174	1.55	0.55	1.59	40.75	56.66	0.47	0.53
2007	78	BCR	30,174	1.46	0.60	3.01	49.35	45.19	1.94	0.51
2007	81	BCR	30,174	0.85	0.66	28.23	55.82	13.94	0.48	1.53

Table 1.27 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 4 Form A

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	71	SR	1.3183			
2005	72	BCR	1.9371	-3.6737	-0.6056	4.2793
2005	73	SR	1.0621			
2005	74	SR	-0.2556			
2005	75	BCR	2.1874	-3.9498	-0.8403	4.7900
2005	76	SR	-0.0642			
2005	77	SR	1.0255			
2005	78	BCR	1.6339	-4.2731	0.1857	4.0874
2005	79	SR	-0.2319			
2005	80	SR	1.3183			
2005	81	BCR	4.0141	-2.9786	-0.4466	3.4251
2005	82	SR	0.3795			
2007	71	SR	1.1275			
2007	72	BCR	1.4246	-3.9432	-1.0716	5.0148
2007	73	SR	0.8360			
2007	74	SR	-0.1767			
2007	75	BCR	1.5790	-4.7955	-0.4851	5.2806
2007	76	SR	-0.1660			
2007	77	SR	1.2465			
2007	78	BCR	1.5483	-3.8291	0.1137	3.7154
2007	79	SR	-0.4818			
2007	80	SR	1.1823			
2007	81	BCR	3.4195	-2.9824	-0.1251	3.1075
2007	82	SR	0.2118			

Note. These Rasch difficulties were based on a common scale.

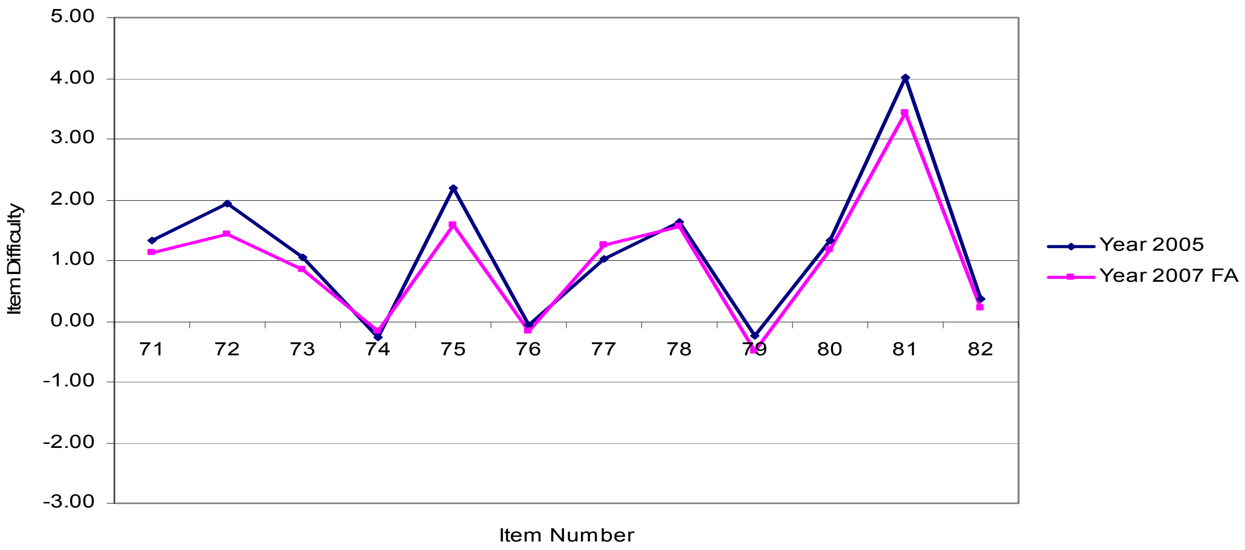

Figure 1.4 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 4 Form A

Table 1.28 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 4 Form B

Item Number	Item Type	Year 05	Year 07
71	SR	0.79	0.84
72	BCR	0.46	0.44
73	SR	0.59	0.57
74	SR	0.63	0.68
75	BCR	0.52	0.56
76	SR	0.73	0.78
77	SR	0.55	0.62
78	BCR	0.43	0.48
79	SR	0.49	0.57
80	SR	0.48	0.49
81	BCR	0.33	0.50
82	SR	0.53	0.58

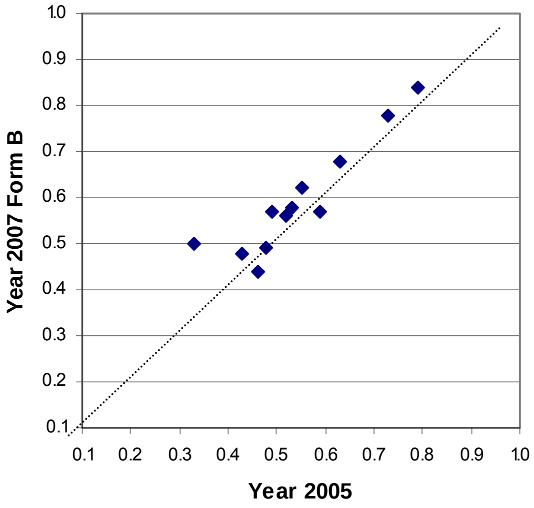

Table 1.29 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 4 Form B

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	72	BCR	2,235	1.38	0.65	6.00	51.00	39.00	3.00	1.00
2005	75	BCR	2,235	1.56	0.65	5.00	31.00	60.00	1.00	2.00
2005	78	BCR	2,215	1.29	0.56	4.00	63.00	32.00	1.00	0.00
2005	81	BCR	2,215	0.99	0.75	25.00	48.00	23.00	2.00	2.00
2007	72	BCR	29,955	1.31	0.61	6.48	55.99	35.88	1.24	0.40
2007	75	BCR	29,955	1.69	0.59	4.26	23.29	69.91	1.81	0.72
2007	78	BCR	29,955	1.45	0.55	2.09	49.99	47.37	0.17	0.39
2007	81	BCR	29,955	1.49	0.63	6.11	37.55	54.85	0.64	0.85

Table 1.30 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 4 Form B

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	71	SR	-0.2477			
2005	72	BCR	1.7832	-3.3964	-0.021	3.4174
2005	73	SR	0.9410			
2005	74	SR	0.6725			
2005	75	BCR	1.7672	-3.209	-1.119	4.3280
2005	76	SR	-0.0593			
2005	77	SR	1.0932			
2005	78	BCR	2.0176	-4.1043	0.1823	3.922
2005	79	SR	1.2683			
2005	80	SR	1.3979			
2005	81	BCR	2.6567	-2.3317	-0.2162	2.5479
2005	82	SR	1.1418			
2007	71	SR	-0.5343			
2007	72	BCR	2.0834	-3.7861	-0.0235	3.8096
2007	73	SR	1.1757			
2007	74	SR	0.5294			
2007	75	BCR	1.6275	-3.1782	-1.3278	4.5059
2007	76	SR	-0.0504			
2007	77	SR	0.9068			
2007	78	BCR	2.4055	-5.0568	-0.7227	5.7795
2007	79	SR	1.1552			
2007	80	SR	1.6008			
2007	81	BCR	2.3188	-3.6205	-1.0745	4.6950
2007	82	SR	1.0512			

Note. These Rasch difficulties were based on a common scale.

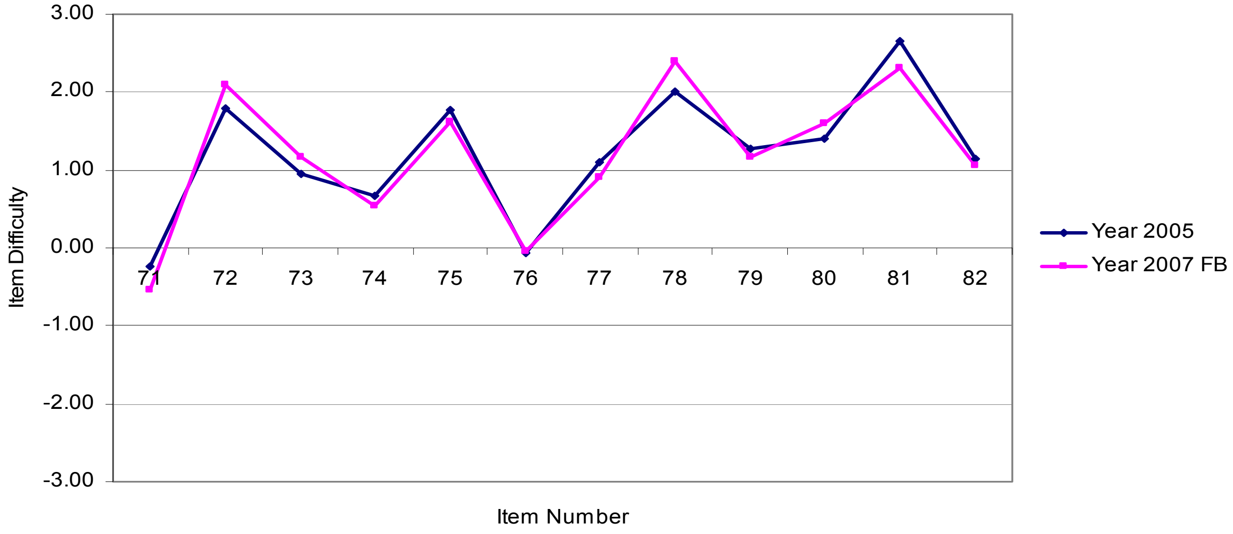

Figure 1.5 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 4 Form B

Table 1.31 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 5 Form A

Item Number	Item Type	Year 05	Year 07
61	SR	0.75	0.72
62	BCR	0.36	0.42
63	SR	0.53	0.59
64	SR	0.81	0.83
65	BCR	0.37	0.38
66	SR	0.56	0.67
67	SR	0.61	0.64
68	BCR	0.35	0.41
69	SR	0.34	0.42
70	SR	0.37	0.45
71	BCR	0.35	0.43
72	SR	0.63	0.76

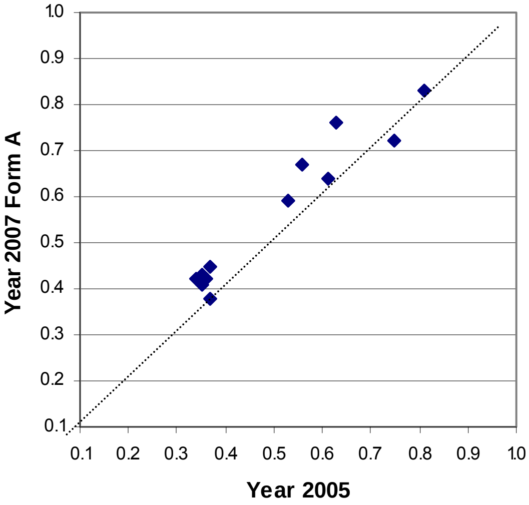

Table 1.32 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 5 Form A

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,163	1.07	0.51	8.00	75.00	15.00	0.00	1.00
2005	65	BCR	2,163	1.10	0.61	11.00	64.00	23.00	1.00	2.00
2005	68	BCR	2,257	1.05	0.60	13.00	65.00	20.00	0.00	2.00
2005	71	BCR	2,257	1.06	0.55	9.00	71.00	17.00	0.00	3.00
2007	62	BCR	30,883	1.25	0.55	4.27	65.63	28.76	0.50	0.84
2007	65	BCR	30,883	1.13	0.54	7.42	71.43	19.70	0.76	0.70
2007	68	BCR	30,883	1.24	0.67	11.27	53.33	33.24	1.43	0.73
2007	71	BCR	30,883	1.28	0.62	8.08	55.76	34.84	0.82	0.51

Table 1.33 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 5 Form A

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	-0.3477			
2005	62	BCR	2.5179	-4.2835	0.4316	3.8519
2005	63	SR	0.8127			
2005	64	SR	-0.8112			
2005	65	BCR	2.3627	-3.6517	-0.0064	3.6581
2005	66	SR	0.6417			
2005	67	SR	0.3405			
2005	68	BCR	2.6820	-3.8218	-0.2301	4.0519
2005	69	SR	1.7021			
2005	70	SR	1.5429			
2005	71	BCR	2.3054	-3.8853	0.4131	3.4722
2005	72	SR	0.1902			
2007	61	SR	-0.2445			
2007	62	BCR	1.8768	-4.3444	0.6386	3.7058
2007	63	SR	0.4821			
2007	64	SR	-0.9831			
2007	65	BCR	1.9237	-3.9951	0.5322	3.4629
2007	66	SR	0.018			
2007	67	SR	0.1235			
2007	68	BCR	1.8519	-3.0717	-0.412	3.4836
2007	69	SR	1.2965			
2007	70	SR	1.2079			
2007	71	BCR	1.8363	-3.3608	-0.0997	3.4605
2007	72	SR	-0.6218			

Note. These Rasch difficulties were based on a common scale.

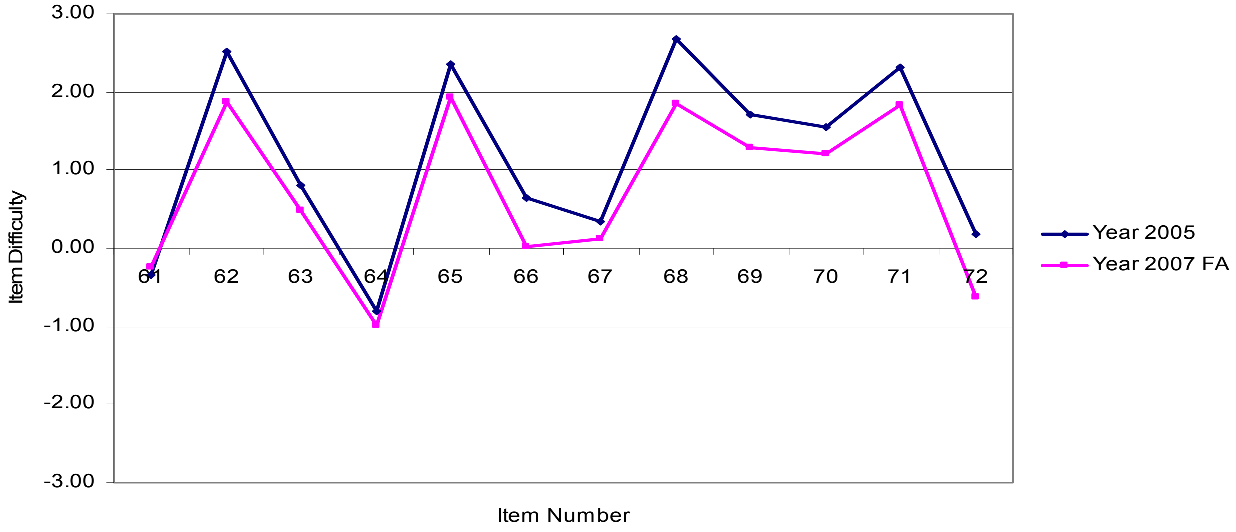

Figure 1.6 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 5 Form A

Table 1.34 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 5 Form B

Item Number	Item Type	Year 05	Year 07
61	SR	0.51	0.55
62	BCR	0.43	0.48
63	SR	0.72	0.82
64	SR	0.48	0.51
65	BCR	0.37	0.41
66	SR	0.75	0.81
67	SR	0.63	0.68
68	BCR	0.38	0.42
69	SR	0.71	0.77
70	SR	0.68	0.68
71	BCR	0.41	0.47
72	SR	0.55	0.63

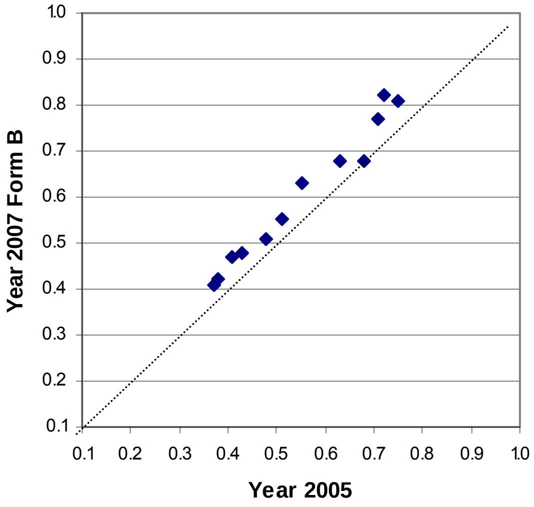

Table 1.35 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 5 Form B

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,231	1.30	0.70	10.00	47.00	39.00	2.00	2.00
2005	65	BCR	2,231	1.11	0.55	10.00	68.00	21.00	0.00	1.00
2005	68	BCR	2,221	1.15	0.60	10.00	63.00	26.00	0.00	1.00
2005	71	BCR	2,221	1.22	0.74	15.00	43.00	39.00	1.00	3.00
2007	62	BCR	30,693	1.44	0.59	3.32	49.64	45.39	1.33	0.32
2007	65	BCR	30,693	1.22	0.57	6.19	65.86	26.47	0.94	0.54
2007	68	BCR	30,693	1.27	0.53	2.70	66.86	28.96	0.72	0.77
2007	71	BCR	30,693	1.40	0.70	9.92	41.92	45.21	2.41	0.54

Table 1.36 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 5 Form B

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	0.9315			
2005	62	BCR	1.7668	-3.0373	-0.4671	3.5044
2005	63	SR	-0.3614			
2005	64	SR	1.0777			
2005	65	BCR	3.1486	-4.6852	-0.6436	5.3288
2005	66	SR	-0.4182			
2005	67	SR	0.2996			
2005	68	BCR	2.2817	-3.6787	-0.1882	3.8668
2005	69	SR	-0.2080			
2005	70	SR	-0.0461			
2005	71	BCR	2.2445	-2.8682	-1.0489	3.9172
2005	72	SR	0.5447			
2007	61	SR	0.7126			
2007	62	BCR	1.2424	-3.8438	-0.1524	3.9962
2007	63	SR	-0.9256			
2007	64	SR	0.9681			
2007	65	BCR	1.9756	-3.9258	0.2455	3.6803
2007	66	SR	-0.7653			
2007	67	SR	0.033			
2007	68	BCR	1.6086	-3.8547	0.2498	3.6048
2007	69	SR	-0.4724			
2007	70	SR	-0.0251			
2007	71	BCR	1.4035	-2.6719	-0.4491	3.121
2007	72	SR	0.1946			

Note. These Rasch difficulties were based on a common scale.

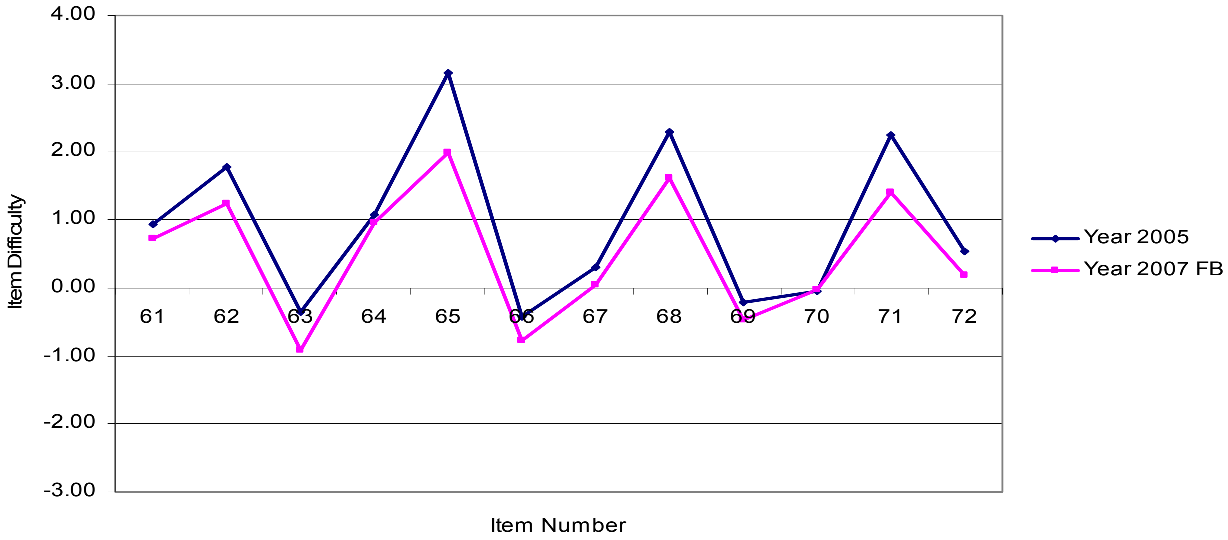

Figure 1.7 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 5 Form B

Table 1.37 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 6 Form A

Item Number	Item Type	Year 05	Year 07
61	SR	0.41	0.46
62	BCR	0.43	0.47
63	SR	0.62	0.70
64	SR	0.68	0.78
65	BCR	0.42	0.46
66	SR	0.68	0.70
67	SR	0.46	0.49
68	BCR	0.42	0.40
69	SR	0.86	0.89
70	SR	0.58	0.56
71	BCR	0.40	0.40
72	SR	0.54	0.61

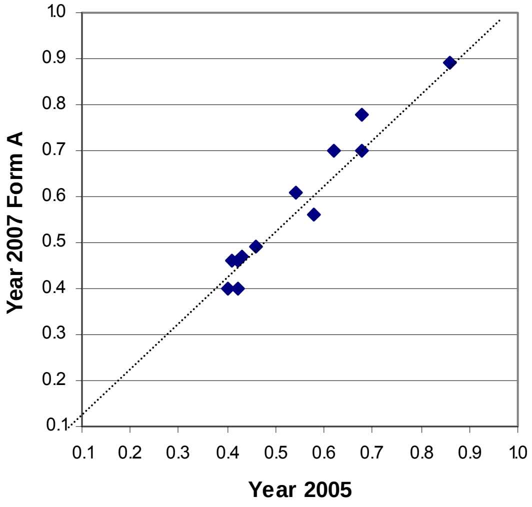

Table 1.38 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 6 Form A

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,262	1.28	0.65	5.00	58.00	31.00	3.00	3.00
2005	65	BCR	2,262	1.25	0.69	7.00	56.00	29.00	3.00	4.00
2005	68	BCR	2,213	1.25	0.71	12.00	49.00	35.00	2.00	2.00
2005	71	BCR	2,213	1.21	0.82	15.00	46.00	29.00	6.00	4.00
2007	62	BCR	31,339	1.42	0.62	2.75	53.26	39.93	2.84	1.22
2007	65	BCR	31,339	1.39	0.65	4.98	52.11	38.73	3.11	1.07
2007	68	BCR	31,339	1.19	0.70	12.62	54.55	29.06	2.27	1.51
2007	71	BCR	31,339	1.20	0.89	22.99	36.93	31.27	6.90	1.90

Table 1.39 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 6 Form A

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	1.6154			
2005	62	BCR	1.5753	-3.5694	0.4231	3.1463
2005	63	SR	0.5317			
2005	64	SR	0.1910			
2005	65	BCR	1.6099	-3.2673	0.4131	2.8542
2005	66	SR	0.1524			
2005	67	SR	1.2876			
2005	68	BCR	1.9383	-2.9117	-0.3301	3.2419
2005	69	SR	-1.1948			
2005	70	SR	0.6606			
2005	71	BCR	1.6137	-2.2421	0.1446	2.0975
2005	72	SR	0.8837			
2007	61	SR	1.2967			
2007	62	BCR	0.9838	-3.9978	0.1002	3.8976
2007	63	SR	0.0096			
2007	64	SR	-0.5493			
2007	65	BCR	1.3890	-3.3172	0.0212	3.2961
2007	66	SR	0.0128			
2007	67	SR	1.1631			
2007	68	BCR	2.0624	-3.1815	0.1260	3.0555
2007	69	SR	-1.4390			
2007	70	SR	0.7441			
2007	71	BCR	1.5694	-1.9073	-0.0550	1.9623
2007	72	SR	0.4720			

Note. These Rasch difficulties were based on a common scale.

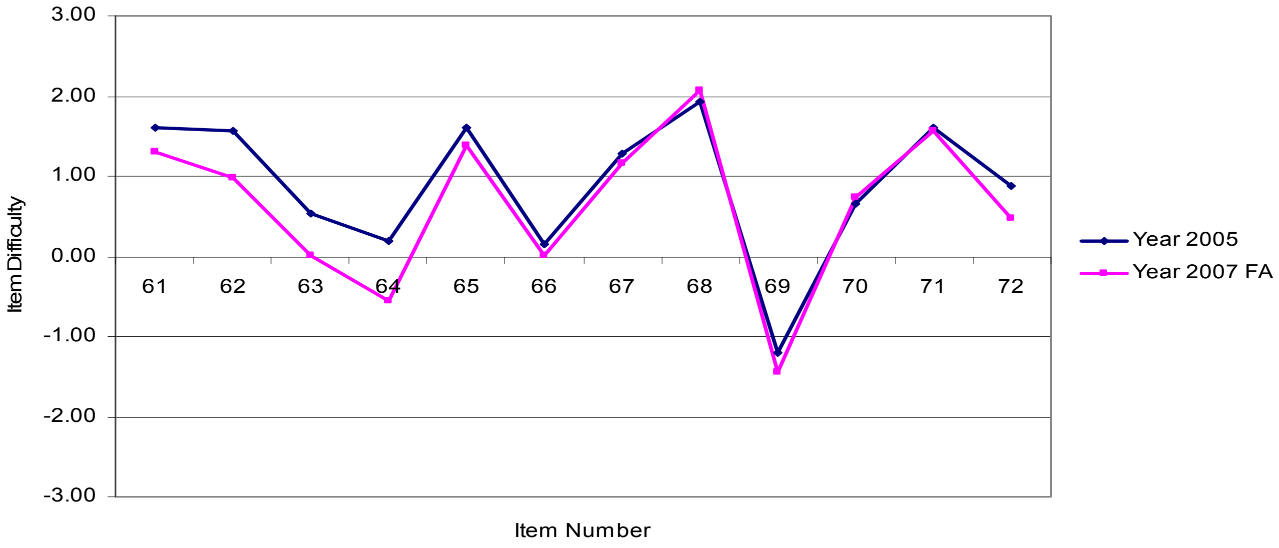

Figure 1.8 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 6 Form A

Table 1.40 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 6 Form B

Item Number	Item Type	Year 05	Year 07
61	SR	0.48	0.50
62	BCR	0.39	0.39
63	SR	0.44	0.47
64	SR	0.37	0.38
65	BCR	0.31	0.38
66	SR	0.66	0.71
67	SR	0.42	0.47
68	BCR	0.45	0.40
69	SR	0.58	0.56
70	SR	0.45	0.45
71	BCR	0.48	0.54
72	SR	0.43	0.54

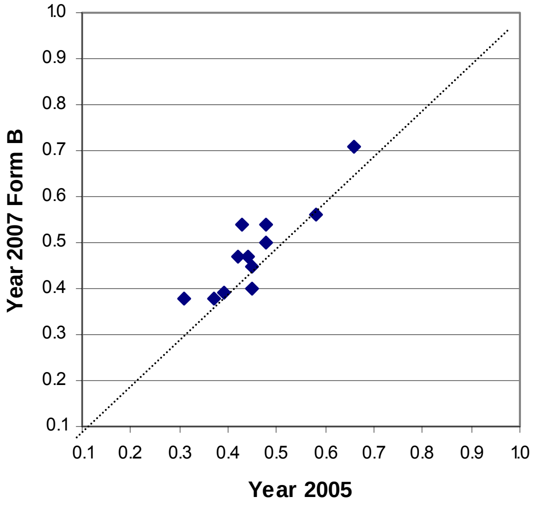

Table 1.41 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 6 Form B

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,286	1.18	0.71	14.00	53.00	29.00	2.00	2.00
2005	65	BCR	2,286	0.94	0.57	17.00	67.00	13.00	0.00	3.00
2005	68	BCR	2,278	1.35	0.73	10.00	46.00	38.00	4.00	1.00
2005	71	BCR	2,278	1.44	0.75	9.00	38.00	46.00	4.00	3.00
2007	62	BCR	31,128	1.18	0.78	19.28	44.04	32.56	2.89	1.23
2007	65	BCR	31,128	1.13	0.62	10.98	63.18	23.16	1.30	1.38
2007	68	BCR	31,128	1.20	0.82	19.56	42.22	32.21	4.53	1.49
2007	71	BCR	31,128	1.61	0.87	12.40	23.76	50.19	12.28	1.37

Table 1.42 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 6 Form B

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	1.1655			
2005	62	BCR	1.9738	-2.7592	-0.0462	2.8054
2005	63	SR	1.3777			
2005	64	SR	1.7278			
2005	65	BCR	3.0781	-3.7831	0.1486	3.6345
2005	66	SR	0.0572			
2005	67	SR	1.4703			
2005	68	BCR	1.4382	-2.6269	-0.0810	2.7079
2005	69	SR	0.6581			
2005	70	SR	1.2828			
2005	71	BCR	1.3191	-2.4929	-0.4440	2.9369
2005	72	SR	1.2718			
2007	61	SR	1.1203			
2007	62	BCR	1.6868	-2.1096	-0.3222	2.4318
2007	63	SR	1.2314			
2007	64	SR	1.6474			
2007	65	BCR	1.9829	-3.1114	0.1952	2.9162
2007	66	SR	-0.0760			
2007	67	SR	1.1471			
2007	68	BCR	1.6954	-1.8039	-0.2437	2.0476
2007	69	SR	0.8038			
2007	70	SR	1.2334			
2007	71	BCR	1.0008	-1.5360	-0.7039	2.2399
2007	72	SR	0.8124			

Note. These Rasch difficulties were based on a common scale.

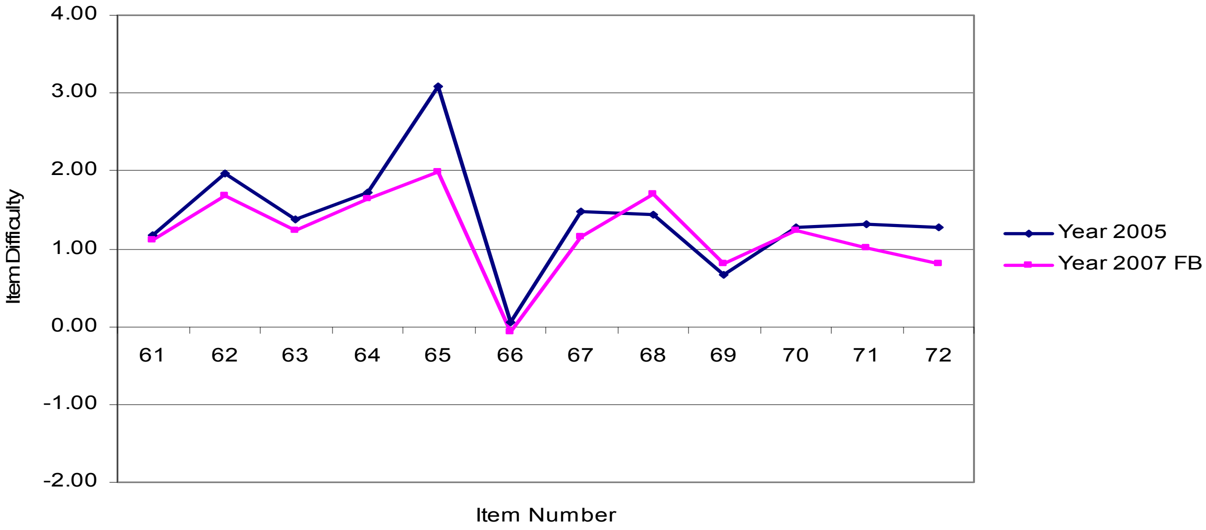

Figure 1.9 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 6 Form B

Table 1.43 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 7 Form A

Item Number	Item Type	Year 05	Year 07
61	SR	0.63	0.66
62	BCR	0.52	0.55
63	SR	0.63	0.65
64	SR	0.66	0.65
65	BCR	0.45	0.42
66	SR	0.68	0.72
67	SR	0.60	0.60
68	BCR	0.40	0.47
69	SR	0.88	0.88
70	SR	0.75	0.77
71	BCR	0.43	0.44
72	SR	0.53	0.54

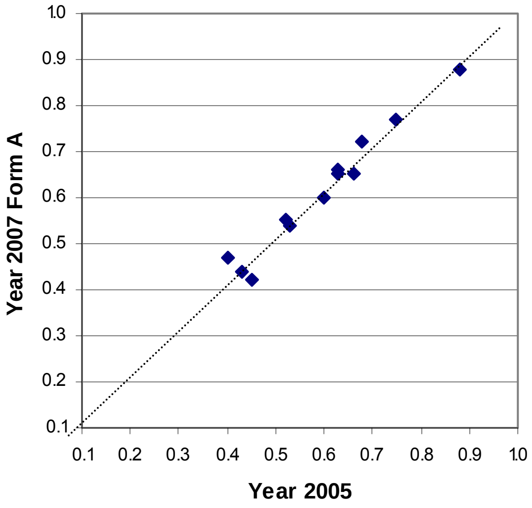

Table 1.44 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 7 Form A

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,194	1.55	0.70	4.00	38.00	50.00	6.00	2.00
2005	65	BCR	2,194	1.34	0.77	9.00	43.00	39.00	4.00	5.00
2005	68	BCR	30,234	1.21	0.62	8.00	62.00	27.00	2.00	1.00
2005	71	BCR	30,234	1.29	0.82	15.00	44.00	32.00	7.00	2.00
2007	62	BCR	32,114	1.65	0.62	2.61	31.34	60.81	3.97	1.26
2007	65	BCR	32,114	1.26	0.75	14.03	42.60	39.14	1.80	2.43
2007	68	BCR	32,114	1.41	0.67	5.23	49.68	39.75	4.11	1.24
2007	71	BCR	32,114	1.33	0.66	4.96	58.21	30.68	4.35	1.80

Table 1.45 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 7 Form A

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	0.2584			
2005	62	BCR	0.8135	-2.9830	-0.2325	3.2155
2005	63	SR	0.2397			
2005	64	SR	0.0934			
2005	65	BCR	1.3223	-2.7247	-0.2025	2.9272
2005	66	SR	-0.0798			
2005	67	SR	0.3810			
2005	68	BCR	1.8123	-3.5048	0.3619	3.1429
2005	69	SR	-1.5336			
2005	70	SR	-0.5920			
2005	71	BCR	1.4011	-2.2542	0.0299	2.2243
2005	72	SR	0.6634			
2007	61	SR	0.0086			
2007	62	BCR	0.5418	-3.2089	-0.8134	4.0224
2007	63	SR	0.0527			
2007	64	SR	0.0459			
2007	65	BCR	1.7340	-2.6962	-0.5614	3.2576
2007	66	SR	-0.2786			
2007	67	SR	0.3226			
2007	68	BCR	0.9831	-3.2780	0.1664	3.1116
2007	69	SR	-1.5373			
2007	70	SR	-0.6676			
2007	71	BCR	1.1365	-3.2492	0.4938	2.7554
2007	72	SR	0.6768			

Note. These Rasch difficulties were based on a common scale.

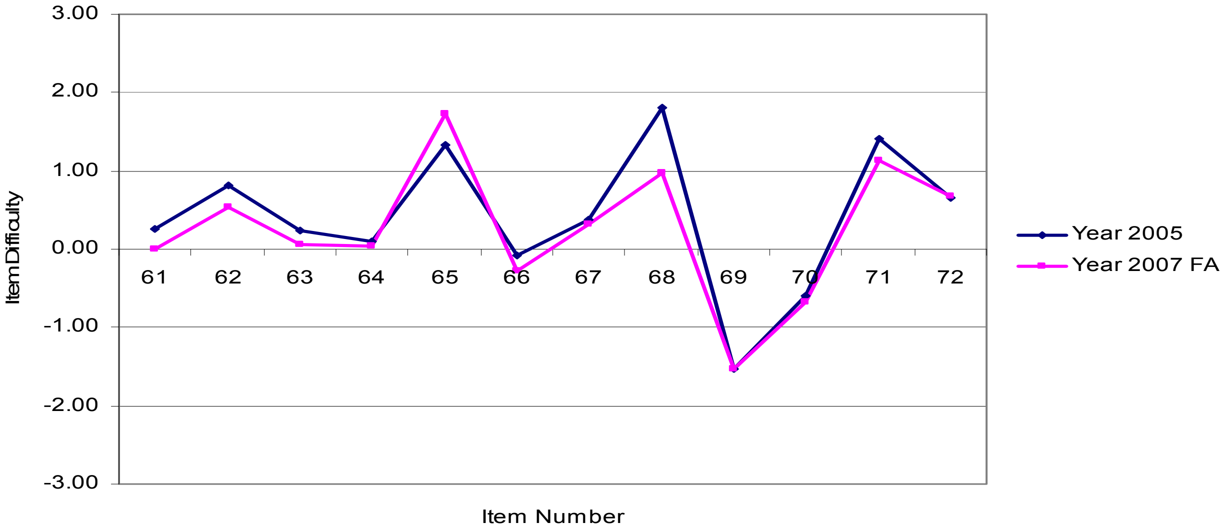

Figure 1.10 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 7 Form A

Table 1.46 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 7 Form B

Item Number	Item Type	Year 05	Year 07
61	SR	0.60	0.59
62	BCR	0.55	0.53
63	SR	0.72	0.71
64	SR	0.42	0.53
65	BCR	0.51	0.52
66	SR	0.55	0.59
67	SR	0.72	0.83
68	BCR	0.41	0.47
69	SR	0.29	0.27
70	SR	0.60	0.73
71	BCR	0.49	0.50
72	SR	0.55	0.54

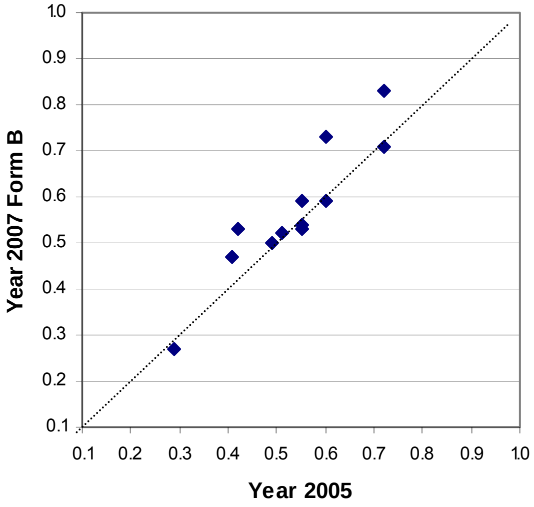

Table 1.47 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 7 Form B

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,269	1.64	0.57	1.00	33.00	62.00	2.00	1.00
2005	65	BCR	2,269	1.54	0.63	3.00	42.00	51.00	4.00	1.00
2005	68	BCR	2,282	1.24	0.66	10.00	52.00	35.00	1.00	2.00
2005	71	BCR	2,282	1.46	0.67	7.00	36.00	54.00	1.00	2.00
2007	62	BCR	31,846	1.59	0.63	1.47	40.37	52.05	4.73	1.38
2007	65	BCR	31,846	1.57	0.63	2.30	40.08	52.08	4.11	1.42
2007	68	BCR	31,846	1.40	0.65	4.57	48.91	42.09	2.44	1.99
2007	71	BCR	31,846	1.49	0.69	5.74	38.67	49.91	3.67	2.02

Table 1.48 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 7 Form B

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	0.4186			
2005	62	BCR	0.6991	-4.0223	-0.5227	4.5450
2005	63	SR	-0.2446			
2005	64	SR	1.3314			
2005	65	BCR	0.8247	-3.6511	-0.1183	3.7693
2005	66	SR	0.5738			
2005	67	SR	-0.4375			
2005	68	BCR	2.1609	-3.4798	-0.5650	4.0448
2005	69	SR	2.1106			
2005	70	SR	0.4515			
2005	71	BCR	1.8958	-3.3291	-1.2878	4.6168
2005	72	SR	0.7476			
2007	61	SR	0.3754			
2007	62	BCR	0.5535	-3.6912	0.0403	3.6509
2007	63	SR	-0.2669			
2007	64	SR	0.7089			
2007	65	BCR	0.7465	-3.4363	-0.1985	3.6348
2007	66	SR	0.3701			
2007	67	SR	-0.9994			
2007	68	BCR	1.1703	-3.7139	0.0922	3.6216
2007	69	SR	2.1429			
2007	70	SR	-0.4029			
2007	71	BCR	1.1420	-2.8994	-0.4808	3.3802
2007	72	SR	0.6241			

Note. These Rasch difficulties were based on a common scale.

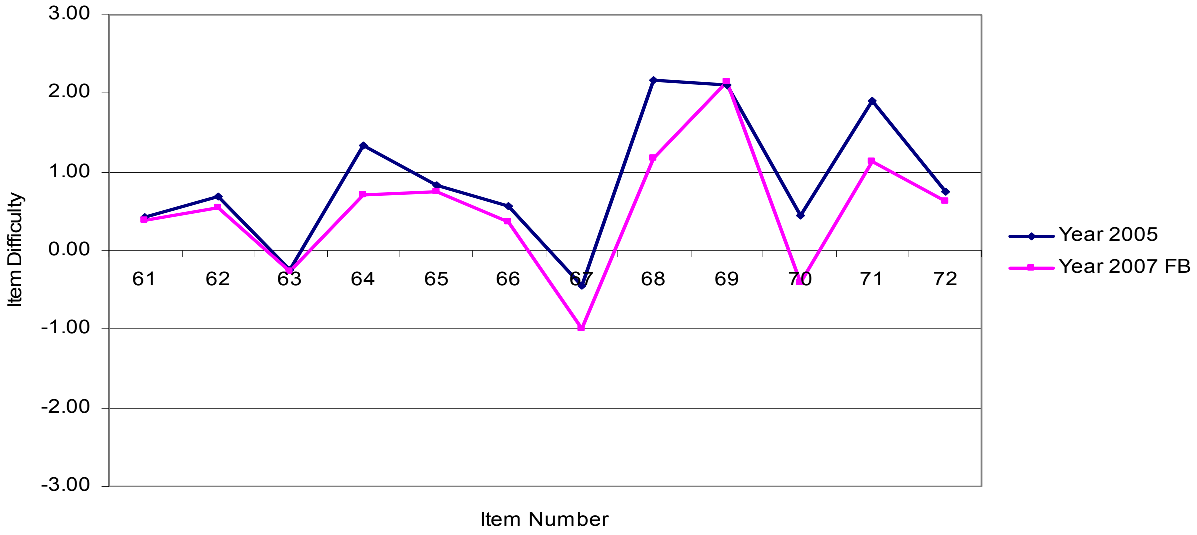

Figure 1.11 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 7 Form B

Table 1.49 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 8 Form A

Item Number	Item Type	Year 05	Year 07
61	SR	0.28	0.34
62	BCR	0.56	0.55
63	SR	0.79	0.82
64	SR	0.85	0.86
65	BCR	0.52	0.50
66	SR	0.79	0.80
67	SR	0.57	0.56
68	BCR	0.47	0.45
69	SR	0.42	0.41
70	SR	0.76	0.75
71	BCR	0.42	0.43
72	SR	0.62	0.69

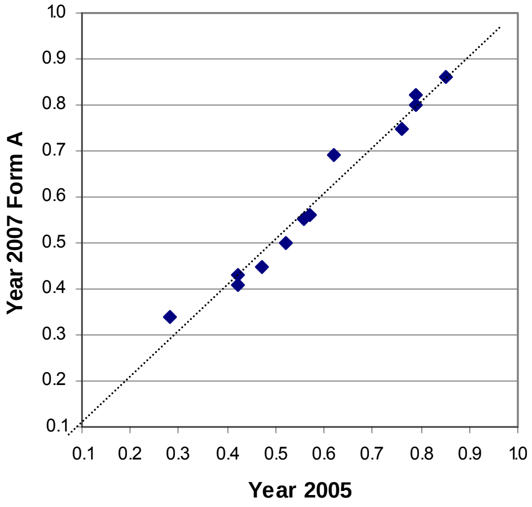

Table 1.50 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 8 Form A

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,195	1.67	0.74	5.00	32.00	52.00	10.00	1.00
2005	65	BCR	2,195	1.57	0.71	4.00	36.00	51.00	6.00	3.00
2005	68	BCR	2,270	1.42	0.77	9.00	40.00	43.00	5.00	3.00
2005	71	BCR	2,270	1.26	0.71	8.00	49.00	36.00	2.00	5.00
2007	62	BCR	32,609	1.65	0.69	3.10	34.21	53.08	8.27	1.34
2007	65	BCR	32,609	1.50	0.73	6.50	40.06	45.75	5.99	1.70
2007	68	BCR	32,609	1.36	0.71	6.88	49.54	37.06	4.03	2.49
2007	71	BCR	32,609	1.30	0.65	3.79	57.47	31.61	3.01	4.12

Table 1.51 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 8 Form A

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	2.0677			
2005	62	BCR	0.5928	-2.2842	-0.2361	2.5202
2005	63	SR	-0.5742			
2005	64	SR	-1.1145			
2005	65	BCR	0.7904	-2.6349	-0.2166	2.8515
2005	66	SR	-0.6172			
2005	67	SR	0.6604			
2005	68	BCR	1.2768	-2.2997	-0.2838	2.5835
2005	69	SR	1.3912			
2005	70	SR	-0.3607			
2005	71	BCR	1.6618	-2.9262	-0.1769	3.1032
2005	72	SR	0.4188			
2007	61	SR	1.7184			
2007	62	BCR	0.4281	-2.9192	-0.0803	2.9995
2007	63	SR	-0.8649			
2007	64	SR	-1.2399			
2007	65	BCR	0.8175	-2.7566	-0.1481	2.9046
2007	66	SR	-0.8170			
2007	67	SR	0.6456			
2007	68	BCR	1.2712	-2.7023	-0.0236	2.7259
2007	69	SR	1.4585			
2007	70	SR	-0.3658			
2007	71	BCR	1.3742	-3.8384	0.4270	3.4114
2007	72	SR	-0.1350			

Note. These Rasch difficulties were based on a common scale.

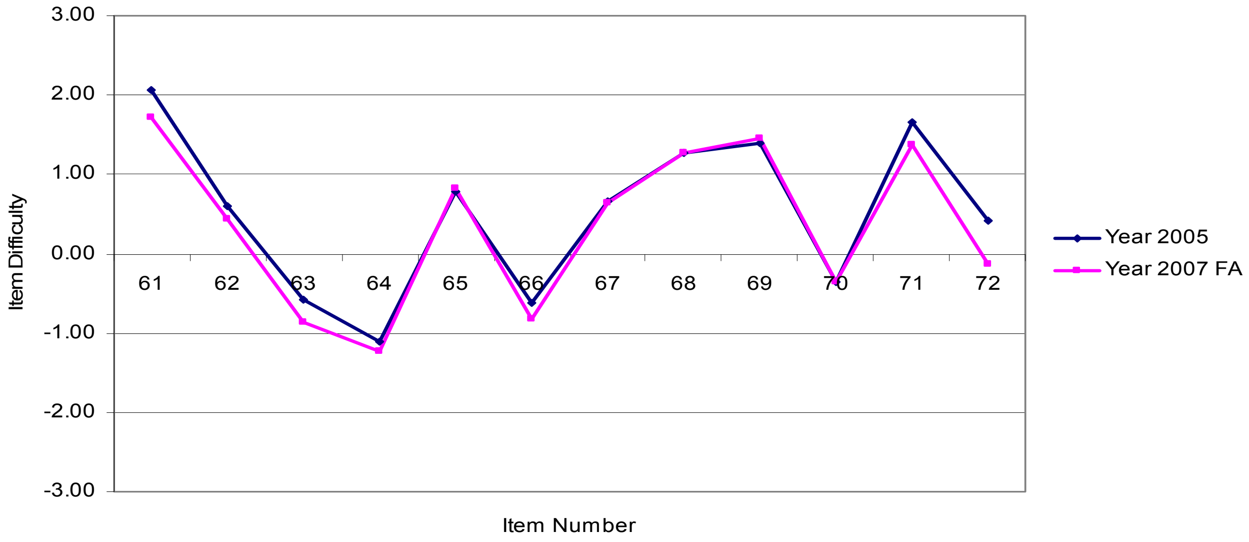

Figure 1.12 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 8 Form A

Table 1.52 Augmented Item P-Value Comparison for Year 2005 vs. Year 2007: Grade 8 Form B

Item Number	Item Type	Year 05	Year 07
61	SR	0.63	0.64
62	BCR	0.55	0.56
63	SR	0.91	0.92
64	SR	0.69	0.69
65	BCR	0.54	0.50
66	SR	0.56	0.53
67	SR	0.62	0.60
68	BCR	0.48	0.52
69	SR	0.81	0.81
70	SR	0.65	0.65
71	BCR	0.52	0.51
72	SR	0.58	0.57

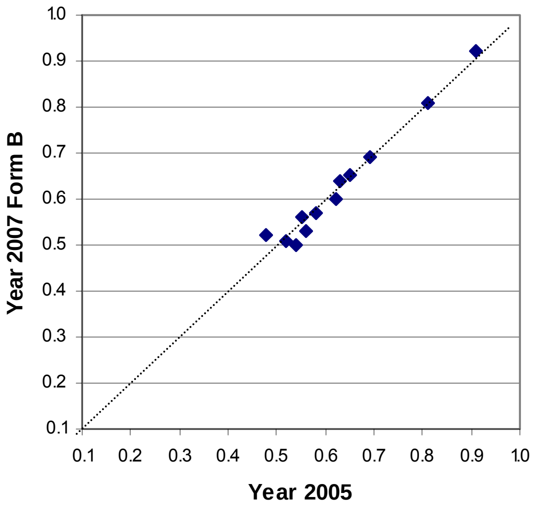

Table 1.53 BCR Item Score-Point Distribution Comparison for Year 2005 vs. Year 2007: Grade 8 Form B

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2005	62	BCR	2,276	1.65	0.66	3.00	32.00	58.00	6.00	1.00
2005	65	BCR	2,276	1.61	0.67	2.00	39.00	51.00	7.00	2.00
2005	68	BCR	30,460	1.44	0.71	7.00	45.00	42.00	5.00	1.00
2005	71	BCR	30,460	1.55	0.79	10.00	28.00	53.00	7.00	2.00
2007	62	BCR	32,452	1.68	0.63	2.03	31.12	59.66	5.88	1.30
2007	65	BCR	32,452	1.51	0.69	4.17	42.42	46.28	5.38	1.75
2007	68	BCR	32,452	1.56	0.74	4.75	39.52	45.61	8.54	1.59
2007	71	BCR	32,452	1.54	0.79	9.56	30.11	51.05	7.34	1.95

Table 1.54 Augment IRT Item Difficulty Comparison for Year 2005 vs. Year 2007: Grade 8 Form B

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2005	61	SR	0.3788			
2005	62	BCR	0.7000	-2.8688	-0.3921	3.2609
2005	63	SR	-1.7352			
2005	64	SR	0.0105			
2005	65	BCR	0.5321	-3.2655	0.1298	3.1357
2005	66	SR	0.6932			
2005	67	SR	0.3867			
2005	68	BCR	1.0702	-2.5919	-0.0272	2.6191
2005	69	SR	-0.7417			
2005	70	SR	0.1154			
2005	71	BCR	1.1735	-1.8262	-0.8182	2.6444
2005	72	SR	0.5270			
2007	61	SR	0.2266			
2007	62	BCR	0.4596	-3.0584	-0.2356	3.2940
2007	63	SR	-1.9561			
2007	64	SR	-0.0410			
2007	65	BCR	0.8603	-2.7890	-0.0510	2.8400
2007	66	SR	0.7404			
2007	67	SR	0.5508			
2007	68	BCR	0.7356	-2.4394	0.0959	2.3435
2007	69	SR	-0.7751			
2007	70	SR	0.2330			
2007	71	BCR	0.9936	-2.0250	-0.3500	2.3750
2007	72	SR	0.5625			

Note. These Rasch difficulties were based on a common scale.

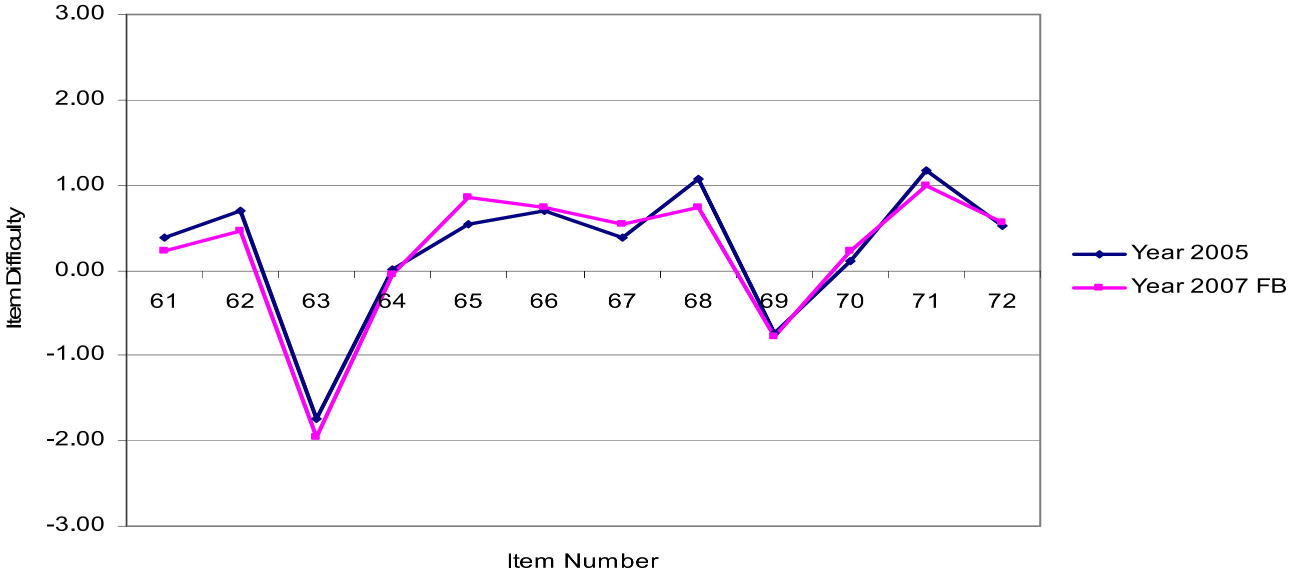

Figure 1.13 Augmented IRT Item Difficulty Comparison Plot for Year 2005 vs. Year 2007: Grade 8 Form B

1.11 Field Test Analyses

All field test items embedded in operational forms were subjected to rigorous analyses for their properties because these analyses will provide information about which items would be included as operational items in the future. All statistical results concerning field test items were reserved in the 2007 item bank. Information on item bank can be found in the section 1.17, Item Bank Construction. 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 were flagged for further scrutiny if:

- An item distractor was not selected 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 were flagged for further scrutiny 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.

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 as a possible operational item. If the item represented important content that had not been extensively taught, however, it would be justified to be included in operational test form.

Differential Item Functioning Analyses

Analyses of *Differential item functioning* (*DIF*) are intended to compare the performance of different subgroups of the population on specific items, when the group have been statistically matched on their tested proficiency.

In present analyses, the gender reference group was males, and the ethnic group was Caucasians. The gender focal group was females and the ethic focal group was African-Americans. Because the 2007 MSA-Reading included both the *SAT10* items and the "Maryland-specific" items on each field test form, the total score as the matching variable consisted of selected SAT items and Maryland-specific items.

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 differential difficulty of the item is unfairly related to group membership, then the item should not be used at all.
- If the differential 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 2007 MSA-Reading, please see section 3.7.

Item Response Theory (IRT) Analyses

To put field test items on the same scale of the operational test items, field test items were calibrated by fixing the parameters of the operational test items within each test form. Then, item difficulties, step difficulties, and fit statistics were stored in the 2007 item bank.

1.12 Operational Test Construction Using IRT Methods

The selection of items to be included in the final operational test forms of the 2007 MSA-Reading required a careful consideration based on test blueprints, classical item analyses, *DIF* analyses, and IRT analyses. Specifically, IRT method played a major role in constructing 2007 operational forms. First, Harcourt suggested the following guidelines:

- Do not include the items with too easy or too hard items
- Do not include the *BCR* items with score distributions that do not elicit the full range of rubric scores
- Do not include the 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
- Finally, do not include the items which have Rasch *Infit* and *Outfit* mean-squares lower than .5 or higher than 1.5. More specific information on Rasch *Infit* and *Outfit* mean-squares can be found in Chapter 3.

A procedure for using IRT methods to build tests that meet any desired set of test specifications was outlined by Lord (1977). The procedure utilizes an item bank with item parameter estimates available for the IRT model of choice, with accompanying information functions. The steps in the procedure suggested by Lord (1977) are as follows:

- First, the shape of desired test information needs to be decided. This was termed as the target information function by Lord (1977).
- Second, specific items need to be selected from the item bank with item information functions that will fill up hard-to-fill areas under the target information function.
- Third, the test information function after test items are added needs to be recalculated.
- Fourth, until the test information function approximates the target information function to a satisfactory degree, test items need to keep on being selected.

It should be noted that these steps were implemented within a framework defined by the content specification of the test. In addition, reading specialists from MSDE reviewed the final test forms of the 2007 MSA-Reading. The following table and figure show the results of constructing grade 3 operational form A using IRT method. Further information on other grades can be obtained from MSDE.

Table 1.55 Grade 3 Form A Construction Using IRT Method

CID	Item Type	P-value	b1	b2	b3
3283234	BCR	0.51	-2.0444	0.8731	4.5805
3283230	BCR	0.45	-2.0517	1.5992	4.7887
3293093	BCR	0.34	0.0267	2.0973	4.8099
3293091	BCR	0.32	-0.2865	2.6307	6.0696
SAT10	SR	0.93	-2.3300		
SAT10	SR	0.85	-1.1000		
SAT10	SR	0.65	0.1500		
SAT10	SR	0.82	-0.9300		
SAT10	SR	0.54	0.9300		
SAT10	SR	0.85	-1.0800		
SAT10	SR	0.63	0.4500		
SAT10	SR	0.30	2.2900		
SAT10	SR	0.65	-0.0800		
SAT10	SR	0.46	1.0000		
SAT10	SR	0.68	-0.1100		
SAT10	SR	0.64	0.2400		
SAT10	SR	0.72	-0.1500		
SAT10	SR	0.69	-0.0300		
SAT10	SR	0.90	-1.8000		
SAT10	SR	0.65	0.0000		
SAT10	SR	0.65	0.0300		
SAT10	SR	0.45	0.9300		
SAT10	SR	0.82	-1.0600		
SAT10	SR	0.88	-1.4300		
SAT10	SR	0.81	-0.9100		
SAT10	SR	0.54	0.6500		
SAT10	SR	0.73	-0.4017		
SAT10	SR	0.78	-0.4000		
SAT10	SR	0.60	0.5900		
3283204	SR	0.85	-1.0949		
3283207	SR	0.75	-0.3692		
3283216	SR	0.64	0.3514		
3283222	SR	0.51	0.9950		
3293081	SR	0.56	0.7788		
3293085	SR	0.33	1.9304		
3293089	SR	0.56	0.6826		

Note. a: item discrimination; b1: step 1 difficulty; b2: step 2 difficulty; b3: step 3 difficulty

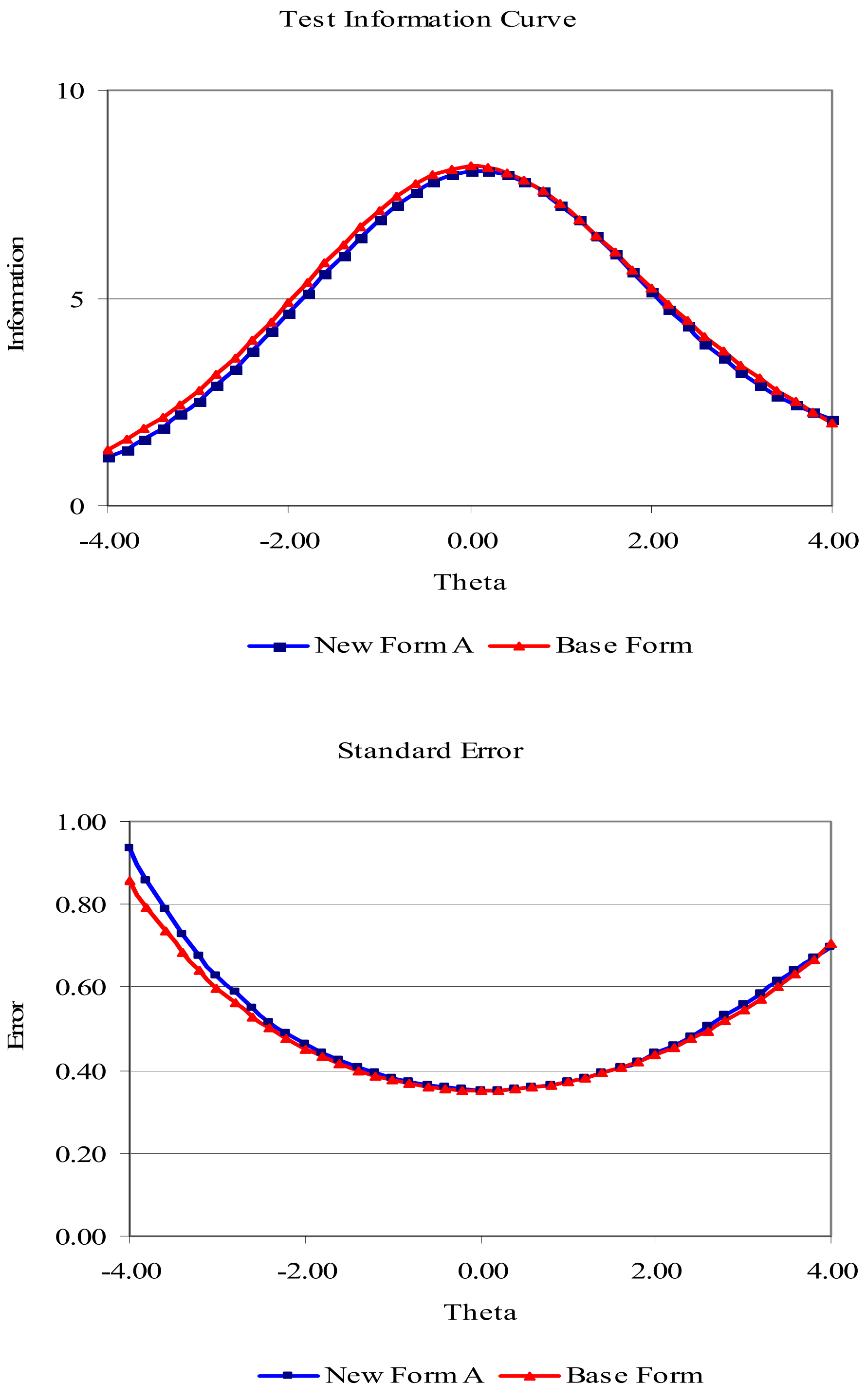

Figure 1.14 Grade 3 Form A IRT Form Construction

1.13 Linking, Equating, and Scaling Procedures

The 2007 MSA-Reading was calibrated, equated, and scaled using the same statistical methods and procedures that were employed in 2006. It should be noted that only SR items were considered as potential year-to-year linking items.

Stratified Random Sampling Procedures

To select equating samples to conduct linking and equating with, stratified random sampling procedures were used in 2007. To verify that the sample was representative of the statewide examinee population in terms of gender and ethnicity, the distributions of gender and ethnicity in the 2007 sample were compared with the total 2007 MSA population distributions. Appendix A, The 2007 MSA-Reading Stratified Random Sampling provides the results of sampling. The results indicated that the calibration sample were representative of the statewide examinee population in terms of gender and ethnicity.

Robust Z Procedures

Robust z values were calculated by the following calculations (South Carolina Department of Education, 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).

Guidelines for Possible Linking Items

Once the above calculations were made, the following guidelines were taken in determining possible sets of common items to be used for the Rasch equating (SCDE, 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.

The reason to apply these guidelines was to exclude items that changed in difficulty more than the other items.

Form-to-Form Linking Procedures

The stability of SAT10 common items appearing on both form A and form B was verified at each grade level:

- Calibrate the two operational test forms separately
- Calculate robust z with Rasch difficulties for form A and form B
- Correlate Rasch difficulties for form A and form B

After examining the robust z and correlations from the two separate calibration, it was determined that the common item difficulties were consistent across the two forms for all items and could be included as form-to-form linking items in the fixed calibration of the two forms.

Year-to-Year Linking Procedures

Each test form contained a set of SAT 10 common items, and these items were used to equate the item parameters and place the 2007 tests on the previous years' scale using the fixed method. The stability of the equating common items was evaluated using robust z, correlation coefficients, and standard deviations.

Tables 1.56 through 1.61 included Rasch item difficulties used for calculating robust z values, correlation coefficients, and standard deviations. Figures 1.21 through 1.38 depicts common item difficulty between the base form (2003 or 2004) and either 2007 form A or B. It should be noted that the item difficulties in 2007 form A or B were obtained from independent calibration, and those in base form were on a common scale (e.g., linked to 2003 or 2004 item parameters).

Table 1.56 Common Linking Item Difficulties of Year 2003 vs. Year 2007 MSA-Reading: Grade 3

Item No.	Y2003 Base (F1)	Y2007 Form A	Y2007 Form B
2	-2.33	-2.0116	-1.9111
5	-1.10	-1.5020	-1.3905
6	.15	.1247	.2107
9	-.93	-1.2383	-1.1024
11*	.93	.0472	.1322
15	-1.08	-.9700	-.9601
18	.45	-.1186	-.1211
20	2.29	1.5432	1.4237
23	-.08	.0244	.1241
30	1.00	.8939	1.0052
31	-.11	-.1938	-.2065
32	.24	.0244	.0629
34	-.15	-.4170	-.3393
41	-.03	-.4343	-.2580
44	-1.80	-1.8931	-1.7596
49	.00	.1565	.2461
55	.03	.2055	.3363
56	.93	1.2442	1.2505
57	-1.06	-.8837	-.7833
58	-1.43	-1.7904	-1.6420
59	-.91	-.9732	-.8421
61	.65	.7536	.8248
69	-.40	-.9128	-.6992
70	.59	.3635	.4434

Form Statistics

Mean	-.173	-.332	-.248
SD	1.030	.965	.928

Comparison of each Form with Base Form (Form 1)

Corr w Base	1.000	.950	.947
SD ratio	100%	94%	90%

Mean Diff	.000	-.159	-.075
Median Diff	.000	-.100	-.046
IQR Diff	.000	.477	.398

Based on robust z and item difficulty plot, item 11 on both Form A and Form B was dropped from the possible linking item pool.

The following correlation and SD ratio are based on dropping the item:

Comparison of each Form with Base Form (Form 1)

Corr w Base	1.000	.959	.956
SD ratio	100%	96%	92%

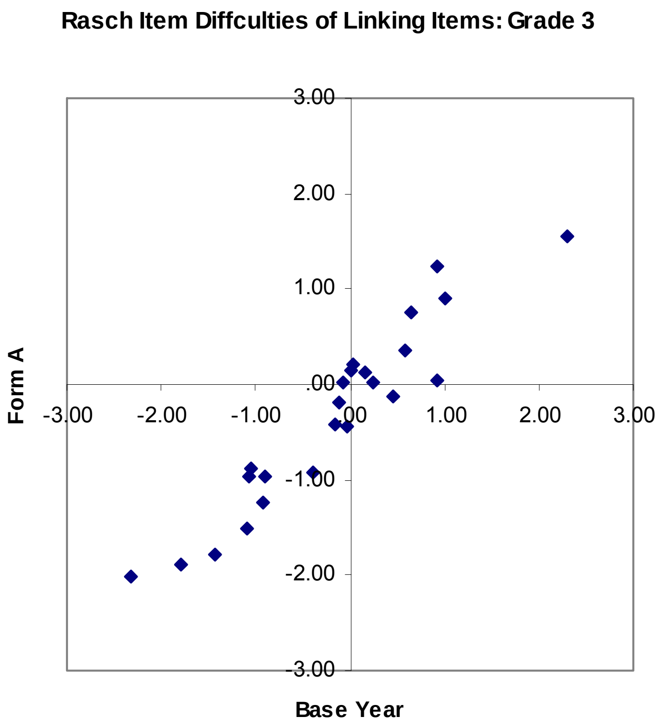

Figure 1.15 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 3 Form A

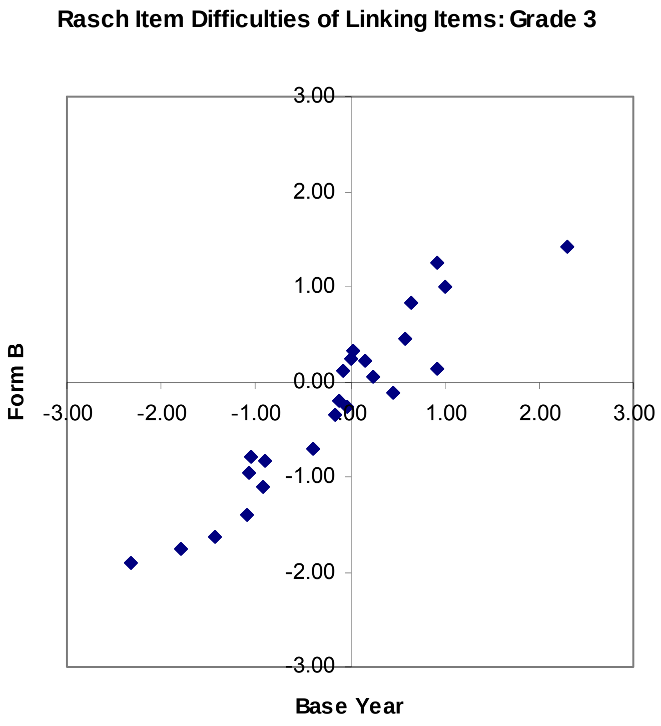

Figure 1.16 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 3 Form B

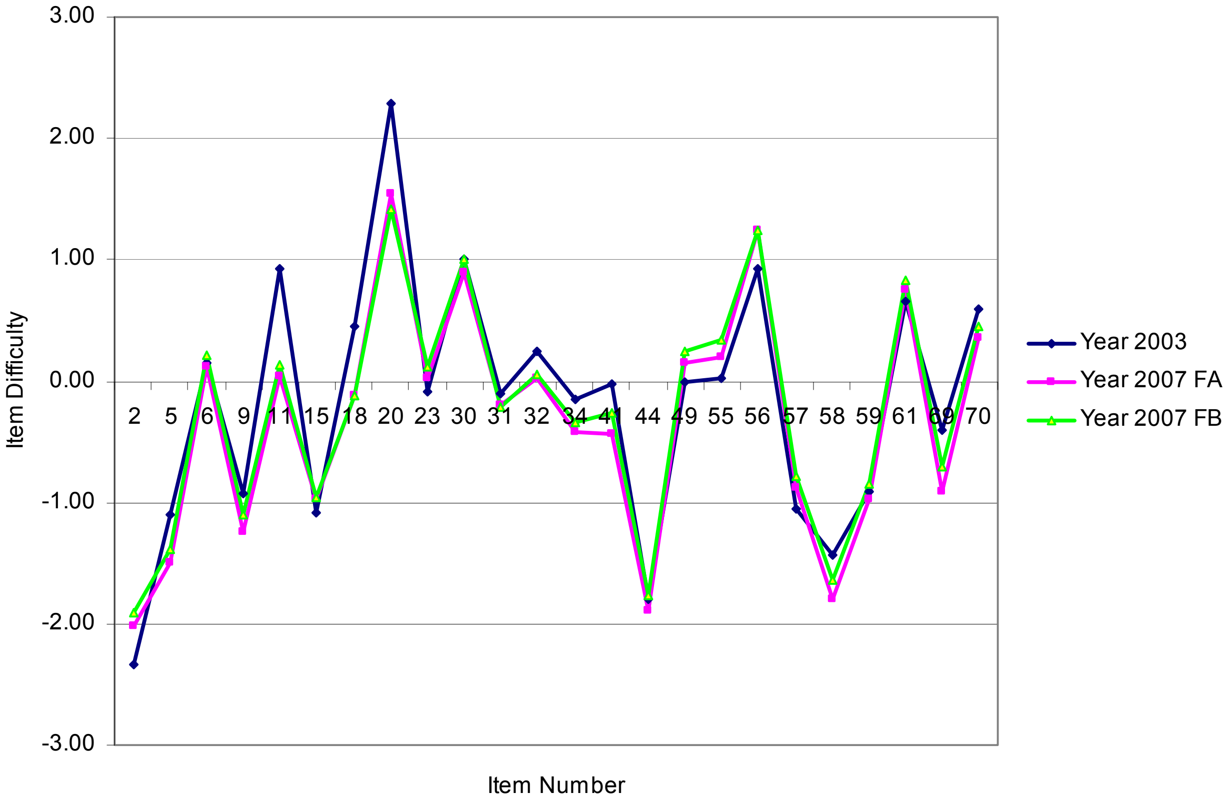

Figure 1.17 Free Calibration Item Difficulty Comparison of Year 2003 vs. Year 2007: Grade 3

Table 1.57 Common Linking Item Difficulties of Year 2004 vs. Year 2007 MSA-Reading: Grade 4

Item No.	Y2004 Base (F6)	Y2007 Form A	Y2007 Form B
1	-3.9886	-3.4568	-3.7872
4	-1.7739	-2.0327	-2.1327
9	-.1033	-.5780	-.7021
10	-1.2892	-1.1162	-1.3136
18	.0403	-.1440	-.2316
23	-.6252	-.7542	-.8282
24	-.2092	-.2200	-.2422
29	-1.4440	-1.9159	-1.9301
35	-.6775	-.7943	-.9849
38	.2123	.2054	.1749
41	-.3429	-.3952	-.5062
42	.2842	.2192	.0280
43	-.7393	-.6942	-.7154
44	-.3247	-.4678	-.3573
45	1.5832	1.8424	1.7750
46	-1.9501	-1.9318	-1.9374
47	-.4109	-.3338	-.4134
50	-.5286	-.3941	-.4587
51	-1.8443	-2.0487	-2.3046
52	.8212	.8822	.8904
53	1.3188	1.4533	1.3810
54	2.0024	2.1056	2.1922
55	-1.4991	-1.6235	-1.5584
62	-.1689	-.1775	-.3074
64	.7087	.6377	.6875

Form Statistics			
Mean	-.438	-.469	-.543
SD	1.278	1.291	1.342
Comparison of each Form with Base Form (Form 6)			
Corr w Base	1.000	.986	.988
SD ratio	100%	101%	105%

Mean Diff	.000	-.031	-.105
Median Diff	.000	-.011	-.033
IQR Diff	.000	.206	.280

None of items was dropped for this grade based on robust z and item difficulty plot.

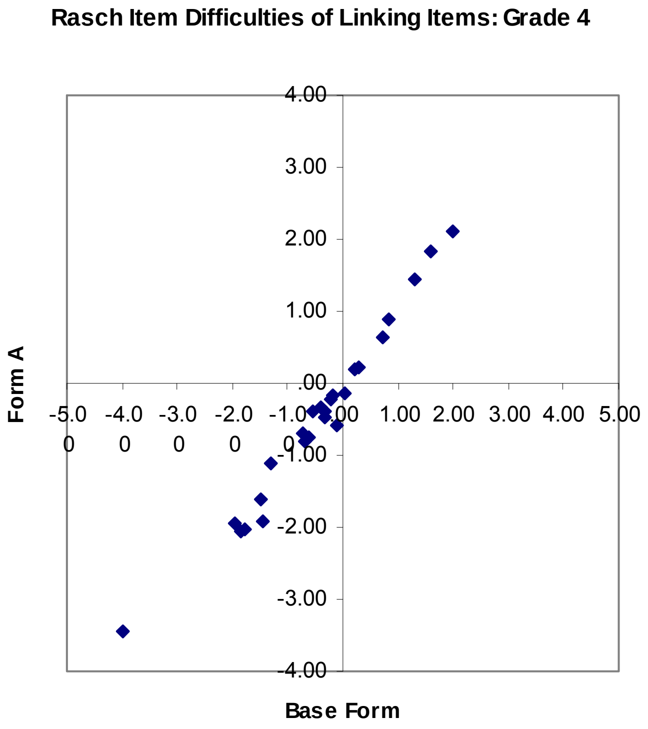

Figure 1.18 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 4 Form A

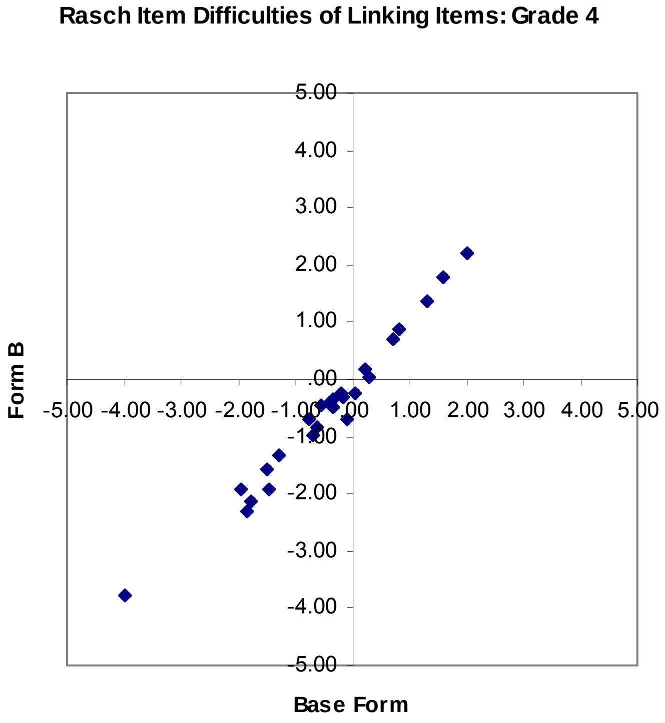

Figure 1.19 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 4 Form B

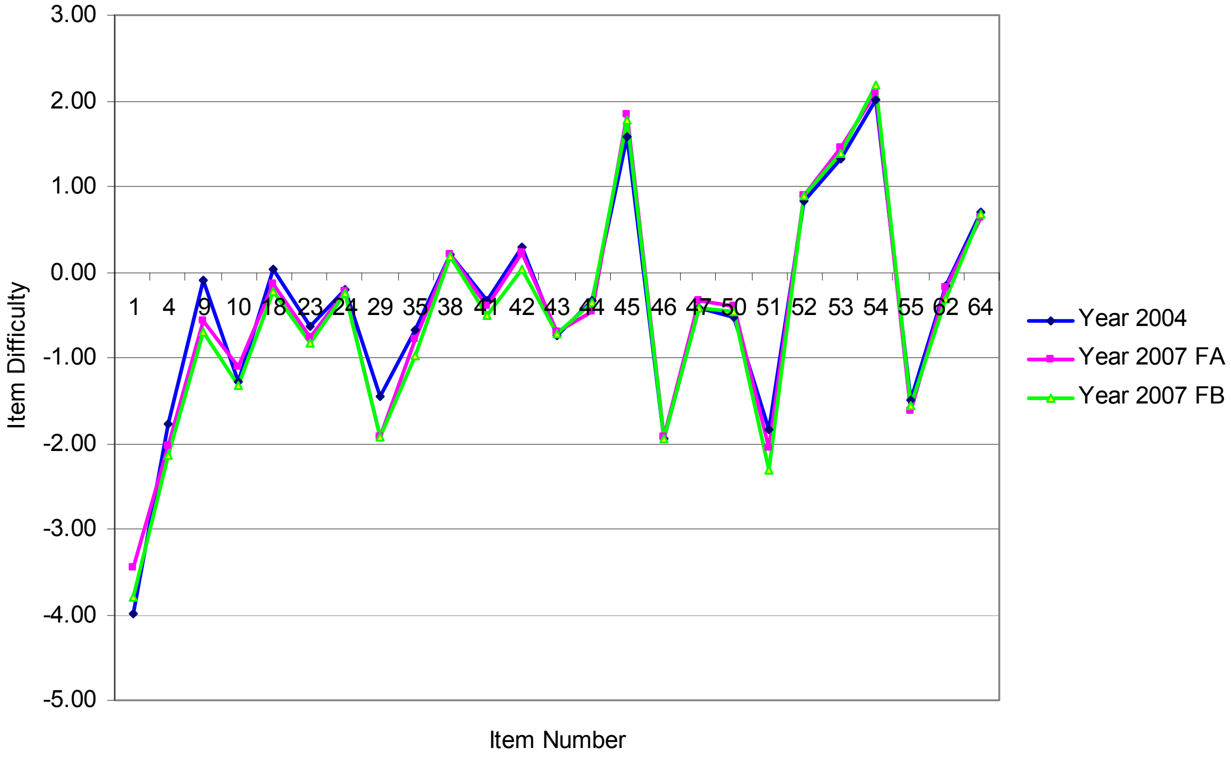

Figure 1.20 Free Calibration Item Difficulty Comparison of Year 2004 vs. Year 2007: Grade 4

Table 1.58 Common Linking Item Difficulties of Year 2003 vs. Year 2007 MSA-Reading: Grade 5

Item No.	Y2003 Base (F1)	Y2007 Form A	Y2007 Form B
4	.51	.4068	.4889
5	.56	.5716	.6842
6	.37	.2230	.2532
9	-1.52	-1.9459	-1.9853
10*	-.96	-1.6921	-1.6186
11	-1.11	-1.0302	-1.0361
13	-1.14	-1.2331	-1.2519
16*	-.71	-1.1025	-1.2959
17	-.47	-.9109	-.6542
19	-.37	-.4540	-.3882
23	.49	.5333	.5929
25	-.19	-.2383	-.1902
28*	.28	.6878	.6488
31	.45	.4814	.5991
33	-.65	-.8598	-.8254
34	1.44	1.4661	1.6013
37	-.02	.1745	.3657
41	-.38	-.6709	-.6000
45	.69	.5597	.6073
49*	-.92	-1.5282	-1.4448

Form Statistics

Mean	-.183	-.328	-.272
SD	.767	.944	.976
Comparison of each Form with Base Form (Form 1)			
Corr w Base	1.000	.970	.972
SD ratio	100%	**123%**	**127%**

Mean Diff	.000	-.146	-.090
Median Diff	.000	-.098	-.052
IQR Diff	.000	.344	.301

Items 10, 16, 28, and 49 on both Form A and Form B were dropped from the possible item linking pool based on robust z and item difficulty plot.

The following correlation and SD ratio are based on dropping those items.

Comparison of each Form with Base Form (Form 1)			
Corr w Base	1.000	.984	.984
SD ratio	100%	111%	116%

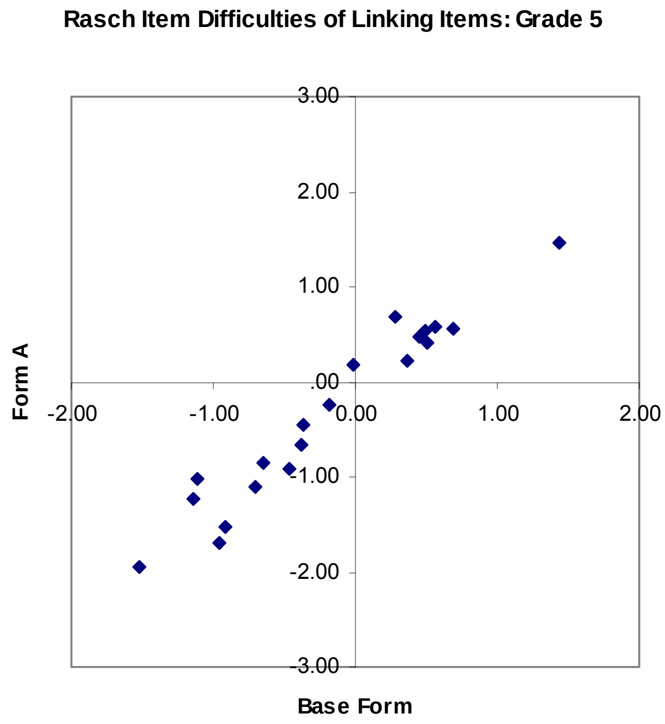

Figure 1.21 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 5 Form A

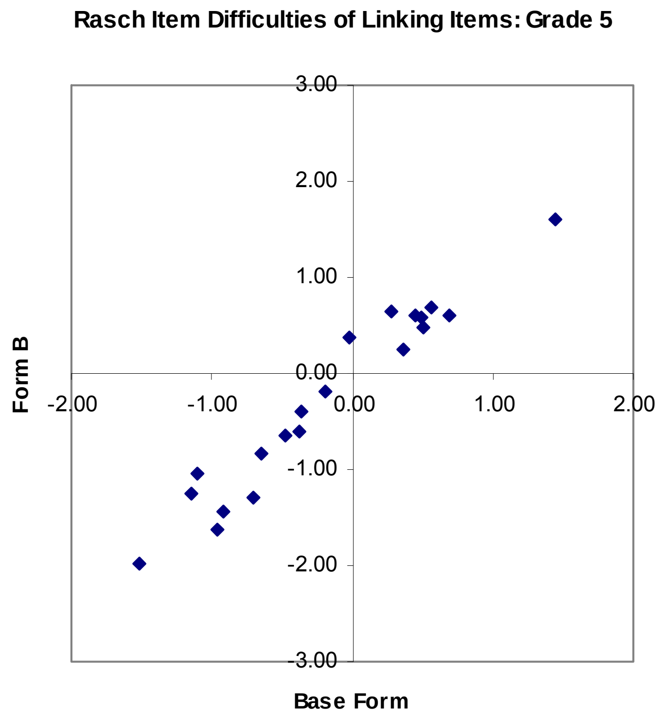

Figure 1.22 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 5 Form B

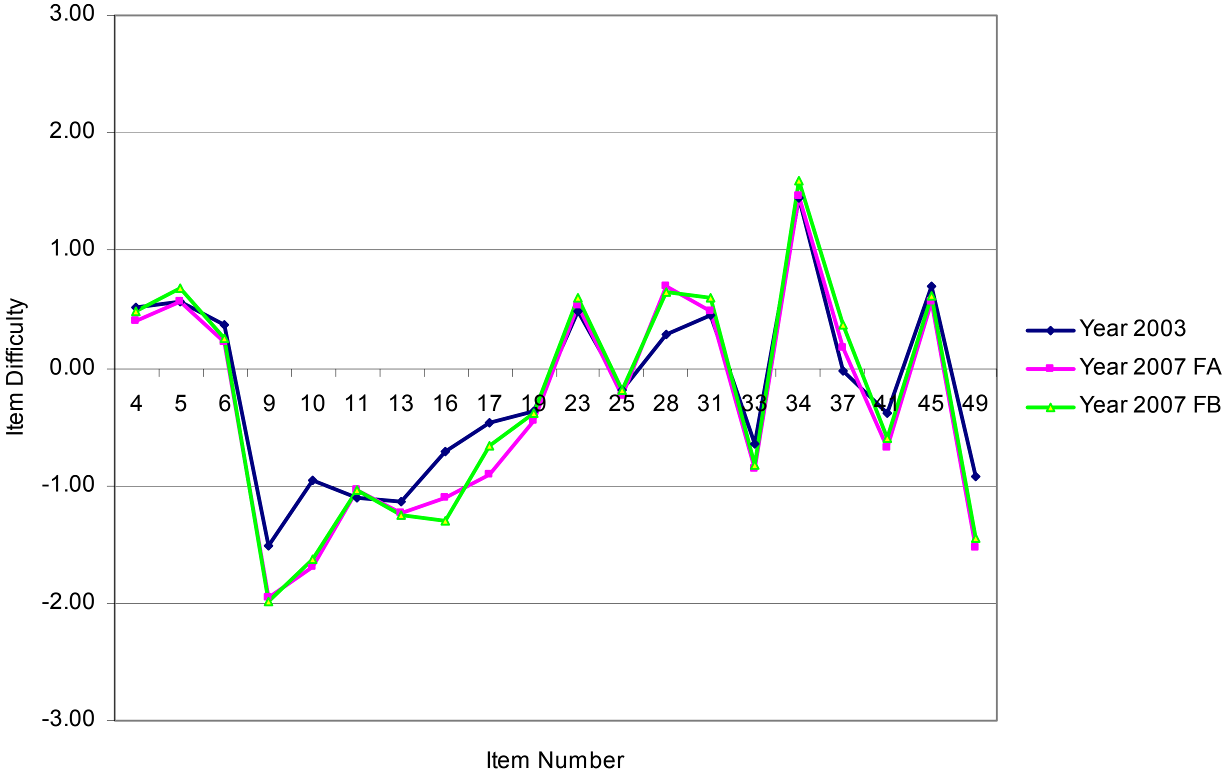

Figure 1.23 Free Calibration Item Difficulty Comparison of Year 2003 vs. Year 2007: Grade 5

Table 1.59 Common Linking Item Difficulties of Year 2004 vs. Year 2007 MSA-Reading: Grade 6

Item No.	Y2004 Base (F4)	Y2007 Form A	Y2007 Form B
1	-.6467	-.3203	-.4886
5	.9241	1.0997	.8664
8	.7190	.7219	.5152
9	-1.8289	-2.2677	-2.4338
10	-.2987	-.2210	-.4330
14	-.2270	-.5556	-.6750
16	-.5273	-.5483	-.7529
18	-.9466	-.9592	-1.2262
21	-1.6635	-1.5110	-1.6005
22	-.4965	-.4737	-.6717
23	-.0437	-.0322	-.1279
24	.0022	.0630	-.1349
25	.2939	.1171	-.0587
28	-.3341	-.6885	-.7565
29	.2820	.4548	.2567
30	.3824	.0415	-.0695
32	-1.2626	-1.1742	-1.4729
33	1.8873	1.9872	1.7651
34	-1.0083	-.7443	-1.0129
35	.5459	.3136	.0784
36	-.4554	-.7186	-.7070
37	-.8703	-.8354	-.9141
38	.6399	.6742	.3943
39	-1.4312	-1.5053	-1.6044
40	-.4922	-.5068	-.7985

Form Statistics			
Mean	-.274	-.304	-.483
SD	.869	.904	.885
Comparison of each Form with Base Form (Form 4)			
Corr w Base	1.000	.975	.979
SD ratio	100%	104%	102%

Mean Diff	.000	-.029	-.208
Median Diff	.000	.012	-.204
IQR Diff	.000	.265	.222

None of items was dropped from the possible item linking pool based on robust z and item difficulty plot.

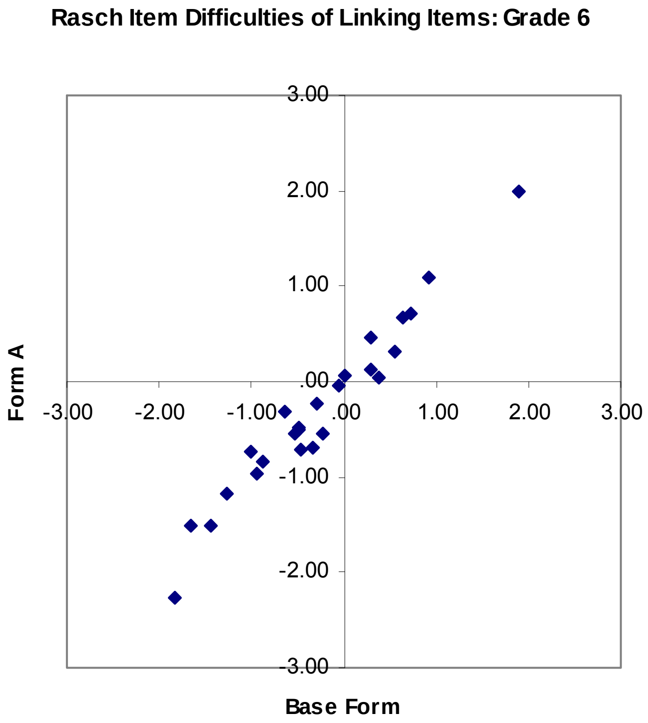

Figure 1.24 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 6 Form A

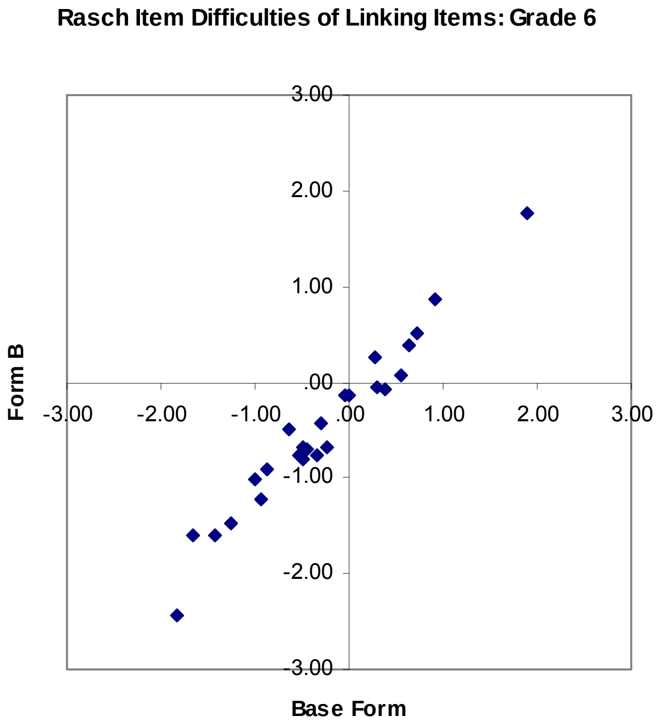

Figure 1.25 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 6 Form B

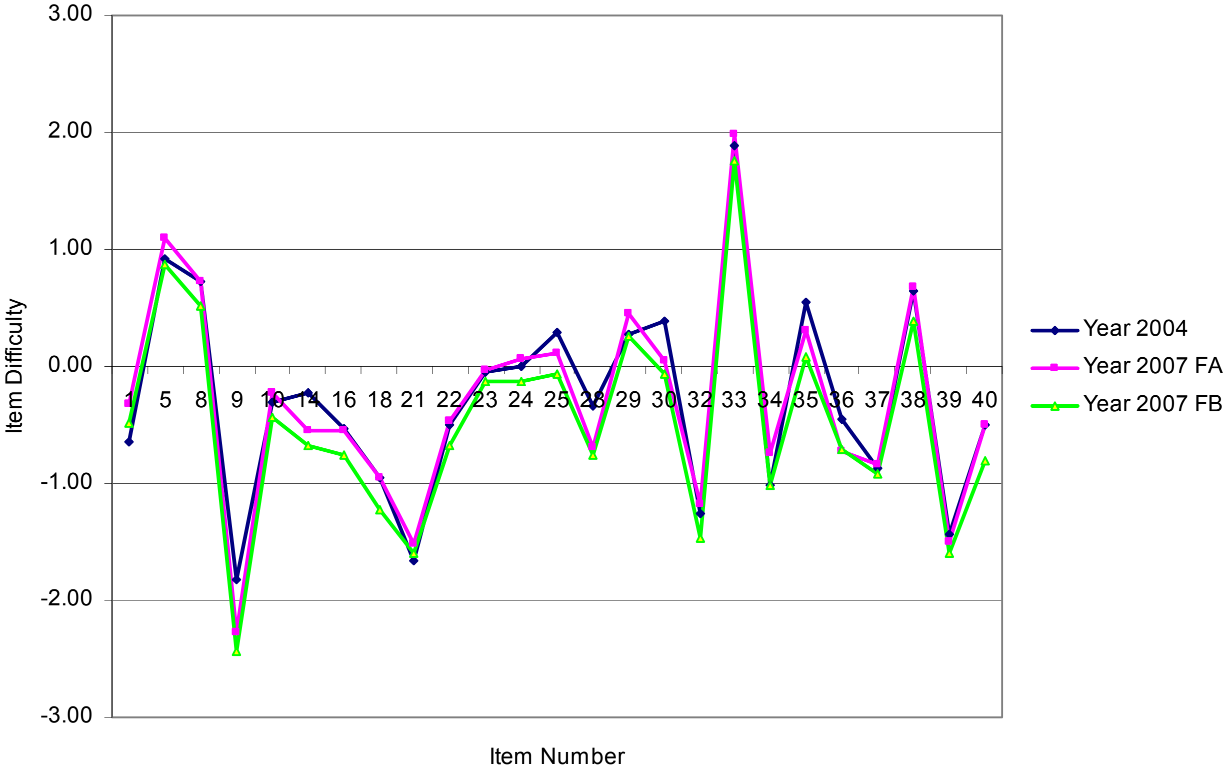

Figure 1.26 Free Calibration Item Difficulty Comparison of Year 2004 vs. Year 2007: Grade 6

Table 1.60 Common Linking Item Difficulties of Year 2004 vs. Year 2007 MSA-Reading: Grade 7

Item No.	Y2004	Y2007	Y2007
	Base (F2)	Form A	Form B
1	-1.6474	-1.5909	-1.6809
3	-1.1065	-1.1286	-1.0808
6	1.2004	1.1664	.9751
8	1.1216	1.7129	1.6403
10	.3792	.3062	.3172
14	.0457	-.0861	-.0639
16	.1649	.2432	.1392
20	-1.1073	-.9055	-1.0695
22	-1.5119	-1.3739	-1.4477
23	.6159	.7216	.6684
26	-.4347	-.4310	-.5532
27	.8787	.9416	.8447
28	.2107	.3233	.2289
31	.5308	.6525	.5086
32	-1.3415	-1.2924	-1.2032
33	.5246	.4380	.3169
36	-1.8027	-1.5741	-1.7298
37	-.2783	-.1860	-.3069
38	-.5500	-.3652	-.4049
39	.2337	.3399	.2653
40	-1.3703	-1.4191	-1.4585
41	-.5760	-.5573	-.6003
42	-.3503	-.1484	-.2518
43	-.3690	-.1239	-.3002
44	-.0528	.0571	-.1153

Form Statistics

Mean	-.264	-.171	-.254
SD	.878	.898	.882

Comparison of each Form with Base Form (Form 2)

Corr w Base	1.000	.987	.987
SD ratio	100%	102%	101%

Mean Diff	.000	.093	.009
Median Diff	.000	.092	-.022
IQR Diff	.000	.134	.126

None of items was dropped from the possible item linking pool based on robust z and item difficulty plot.

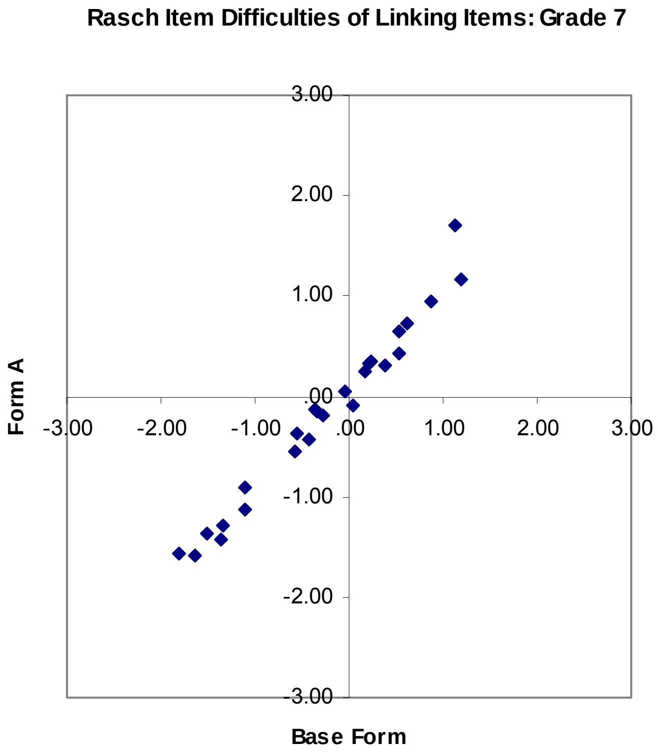

Figure 1.27 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 7 Form A

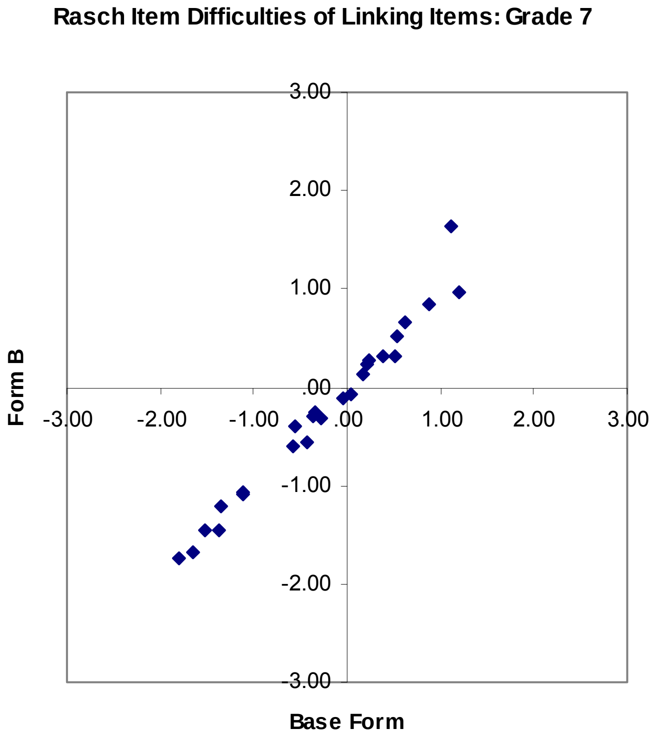

Figure 1.28 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 7 Form B

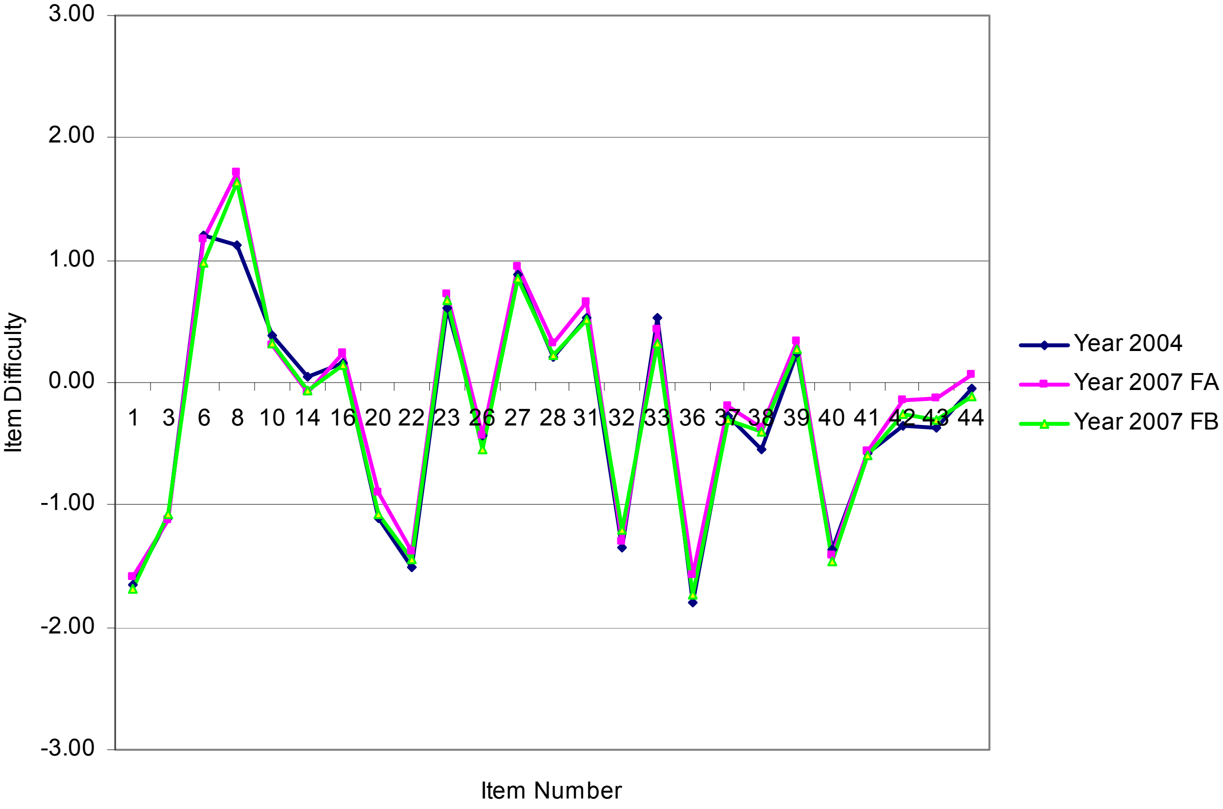

Figure 1.29 Free Calibration Item Difficulty Comparison of Year 2004 vs. Year 2007: Grade 7

Table 1.61 Common Linking Item Difficulties of Year 2003 vs. Year 2007 MSA-Reading: Grade 8

Item No.	Y2003 Base (F1)	Y2007 Form A	Y2007 Form B
3	.07	.3181	.3589
6	.96	.7690	.7176
8	.51	.7486	.7836
9	-1.57	-2.3643	-2.0075
22	-3.60	-3.4811	-3.6271
23	.64	.7257	.7390
25	-.80	-.7794	-.7988
26	.39	.3374	.4941
29	-.19	-.1813	-.1528
31	.20	.4182	.4217
32	.98	.8316	.9377
33	.29	.3873	.3967
35	-.46	-.5108	-.5112
37	-.24	-.3098	-.3326
38	-.49	-.4935	-.4056
41	-1.02	-1.1797	-1.0413
44	-.16	-.3377	-.2593
46	-.22	-.7805	-.7419
48	-.05	-.0791	-.0436
49	-.16	-.2493	-.3353
50	-.18	-.3400	-.2427

Form Statistics

Mean	-.243	-.312	-.269
SD	.984	1.052	1.049

Comparison of each Form with Base Form (Form 1)

Corr w Base	1.000	.973	.982
SD ratio	100%	107%	107%

Mean Diff	.000	-.069	-.026
Median Diff	.000	-.051	-.021
IQR Diff	.000	.245	.192

None of items was dropped from the possible item linking pool based on robust z and item difficulty plot.

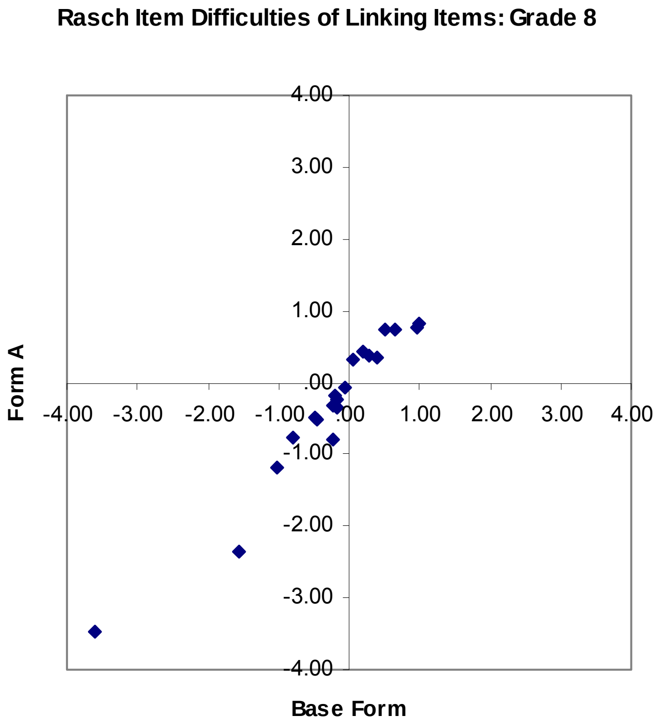

Figure 1.30 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 8 Form A

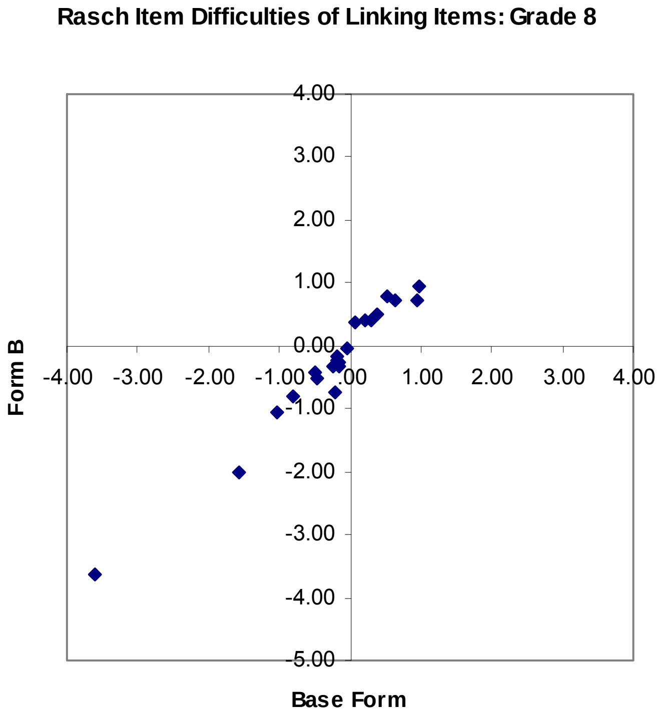

Figure 1.31 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 8 Form B

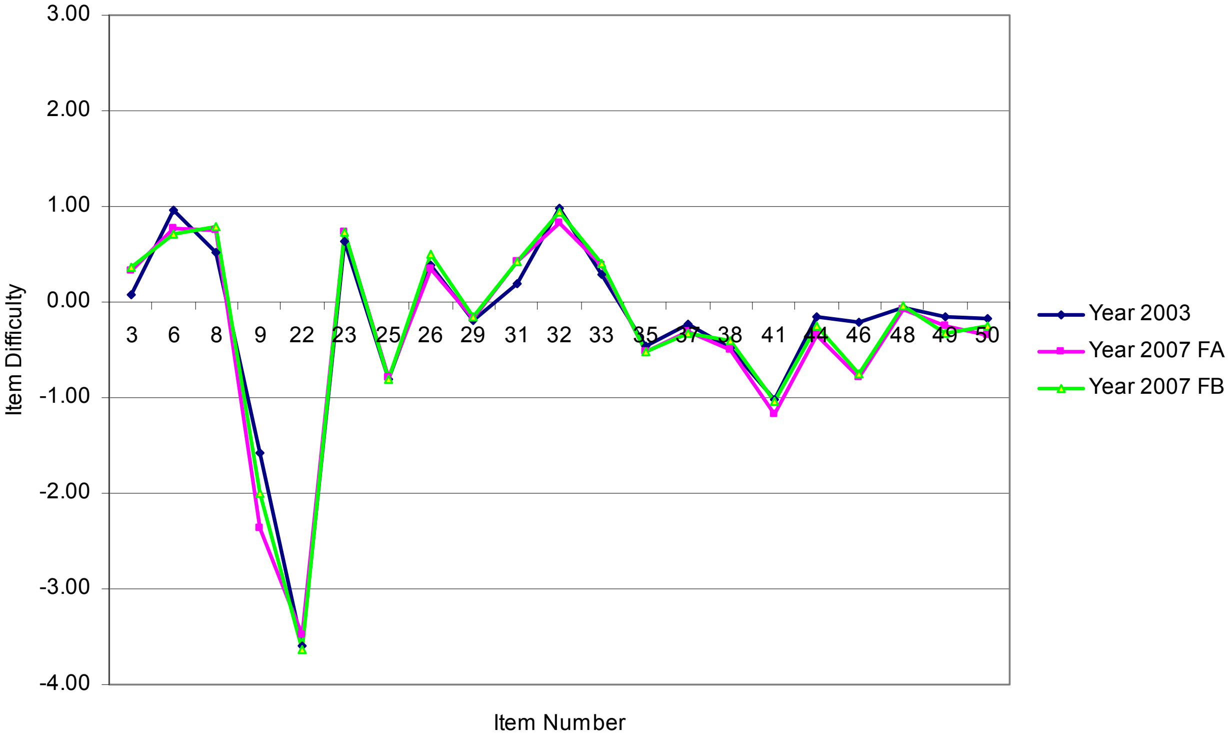

Figure 1.32 Free Calibration Item Difficulty Comparison of Year 2003 vs. Year 2007: Grade 8

Reporting Scale Scores

In order to facilitate the use and interpretation of the results of the 2007 MSA-Reading, MSDE provided Harcourt with specifications about the score scale (Mean = 400, SD = 40, LOSS = 240, HOSS = 650). 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.

The following table depicts the slope and intercept to use for each grade. It should be noted that these same slops and intercepts have been used since Year 2003 (grades 3, 5, and 8) and Year 2004 (grades 4, 6, and 7).

Table 1.62 The 2007 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

1.14 Score Interpretation

To help provide appropriate interpretation of the 2007 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 1.13, Linking, Equating, and Scaling, the 2007 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 Level Descriptors

As previously explained, performance levels and descriptions provide specific information about students' performance levels and help interpret the 2007 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 2007 MSA-Reading score interpretation can be obtained from the MSDE.

1.15 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 essential 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 2007 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 2007 MSA-Reading blueprints provide extensive evidence regarding the alignment between the content in the 2007 MSA-Reading and the *VSC*. The 2007 MSA-Reading operational test forms were created from the pool of item that had been field-tested in 2006 and before. The item composition of these tests is reported in section 1.5, Test Structure of the 2007 MSA-Reading. In addition, 2007 MSA-Reading blueprints are presented in Appendix D.

Item Development

Test development for MSA-Reading is ongoing and continuous. Content specialists, teachers all over Maryland, Harcourt, and MSDE were greatly involved in developing and reviewing test items. Committees such as content review, bias review, and vision review reviewed all of the items which were finally stored in the item bank. Specifically, an internal review by MSDE and Harcourt staff for alignment and quality required a great deal of time and energy. More specific information on item (test) development and review can be obtained in section 1.4, Development and Review of the 2007 MSA-Reading.

Field testing was conducted within a test window scheduled. Once field-test items were scored, MSDE and Harcourt conducted additional item analysis and content review. Any field-test items that exhibited statistics that suggested potential problems were carefully reviewed by content specialists within MSDE and Harcourt. A determination was then made as to whether the item should be eliminated or revised and field-tested again. Information on statistical analyses for field test items can be obtained in section 1.9, Field Test Analyses.

Differential Item Functioning (DIF)

1) Bias Review of Field Test Items

A separate Bias Review Committee examined each item on reading tests looking for indications of bias that would impact the performance of an identifiable group of students. They discussed or rejected items biased on gender, ethnic, religious, or geographical bias.

2) DIF Statistics

For DIF analyses, subgroups were first identified to either reference or focal groups. For 2007 MSA-Reading, males and whites were assigned to the reference group and females and African-Americans were assigned to the focal group.

For SR items, Harcourt applied Mantel-Haenszel procedure, and standardized mean difference (SMD) and standard deviation (SD) were calculated for BCR item analyses. All items were placed in severity classifications base don Educational Testing Service (ETS) guidelines. More information on DIF analyses can be obtained in section 3.7, Differential Item Functioning.

Evidence from Internal Structure

The 2007 MSA-Reading has three reading processes: *General Reading*, *Literary Reading*, and *Informational Reading*. Tables 4.3 through 4.8 show correlations among the reading processes.

1.16 Unidimensionality Analyses

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 2007 MSA-Reading, polychoric correlation coefficients were computed with *LISREL 8.5* (Jöreskog & Sörbom, 1993). Principal component analysis was then applied to produce eigenvalues. The first and the second principal component eigenvalues were compared without rotation. Table 1.63 summarizes the results of the first and second principal component eigenvalues of the 2007 MSA-Reading.

In general, the first factor extracted somewhat large amount of eigenvalues across all grades. With regard to factor analysis and eigenvalues, there is one unit of information per item so that the eigenvalues sum to the number of items. 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 2007 MSA-Reading was met.

Table 1.63 The 2007 MSA-Reading Eigenvalues between the First and Second Components

Grade	Form	Number of Items	First Eigenvalue	Second Eigenvalue
3	A	37	12.47	1.50
	B	37	11.39	1.47
4	A	37	11.56	1.36
	B	37	12.17	1.47
5	A	37	10.15	1.34
	B	37	10.80	1.43
6	A	37	12.33	1.48
	B	37	11.62	1.44
7	A	37	11.82	1.36
	B	37	11.21	1.46
8	A	37	10.46	1.43
	B	37	10.15	1.48

Note. Form A designates the operational portion of Forms 1, 3, 5, 7, and 9, which is identical. Form B designates the operational portion of Forms 2, 4, 6, 8, and 10, which is identical.

1.17 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 2007 MSA-Reading item bank continued to be maintained by Harcourt as computer files and paper copies. This enabled test items to be readily available to both Harcourt and MSDE staff for reference, test construction, test book design, and printing.

Harcourt maintained a computerized statistical item bank to store supporting and identification information on each item. The information stored in this item bank for each item was 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

In terms of Rasch item statistics data, all field test items were calibrated by fixing the parameters of the operational test items within each operational test form. For example, each unique field test items of reading test forms A, B, C, D, and E were independently calibrated after fixing the same operational items appearing across the field test forms with the same Rasch difficulties because these field test forms belonged to the same operational form A (e.g., contained the same operational items on each field test form). Then, item difficulties, step difficulties, and fit statistics were stored in the 2007 item bank.

1.18 Quality Control Procedures

A standard quality procedure at Harcourt Assessment, Inc. was to create a test deck for MSA programs. The test deck began when Quality Assurance entered mock data into the enrollment system, which was transferred to the materials requisition system; the order was packaged by our Distribution Center, and shipped to the Quality Assurance Department. We then reviewed 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 were then used to create a test deck of mock data along with advance copies of documents that were received from the printer. Advance printer copies were inclusive of documents throughout the print run to assure we were randomly testing printed documents. The Maryland test deck was a comprehensive set of all documents that:

- 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 was 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 were included in the test deck to assure the programs were processing all data to meet the requirements of the program with zero defects. Processing of live orders couldn't proceed until each phase of the test deck had been approved by our Quality Assurance Department. An Issues Log with sign-off approvals was utilized to assure we were addressing any issues that arose in the review of the test deck data across all functional groups at Harcourt.

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

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

Upon completion of the processing of all district level data, Harcourt Scoring Operations provided the Quality Assurance Department with a state level data file(s) and state data for review and approval. Harcourt Quality Assurance programmers duplicated all data independently to assure accurate interpretation of the expected results. A series of SAS programs were run on these files to assure 100% accuracy. These included but were 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 posted the statewide student files to a secure FTP site for review by MSDE.