1. OVERVIEW OF THE 2008 *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 what students were expected to learn in schools.

In 2003, the MSA-Reading was introduced in grades 3, 5, and 8, with grades 4, 6, and 7 being added to the program in 2004. Until 2007 the MSA-Reading was administered along with *Stanford Achievement Test Series, Tenth Edition (SAT10)*, and the SAT10 common items aligned to Maryland curriculum were used exclusively for the purpose of form-to-form and year-to-year linking. In 2007, however, MSDE decided to drop all of the SAT10 items from the 2008 assessment. Due to the decision, MSDE and Pearson team members examined options to replace the SAT10 items removed from the test. The minimum requirement was to develop enough items to cover the same total and subtotal score points that SAT10 common items contributed in previous years (for grade 5, for example, 45 total score points with 15 points each for general reading, literary, and informational reading). In addition, it was decided that only one operational form would be developed for the 2008 administration, and that options for year-to-year equating would focus on items that were originally field-tested in 2006. It should be noted that Maryland-specific selected-response items (i.e., multiple choice items) which appeared both in 2008 and in 2006 were used exclusively for the purpose of year-to-year linking. All scale scores of the 2008 assessment were linked back to the 2003 (for grades 3, 5, and 8) or 2004 (for grades 4, 6, and 7) assessment so that all of the scale scores were on a common scale. It should be noted that more detailed information on the changes to the 2008 reading assessment can be found in section 1.11, *Constructing the 2008 MSA-Reading Operational Forms*.

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 obtained from MSDE. It should be noted that these cut scores have been applied since 2003 (for grades 3, 5, and 8) or 2004 (for grades 4, 6, and 7).

1.1 Purposes/Uses of the 2008 MSA-Reading

By measuring students' achievement against the new academic standards, the 2008 MSA-Reading fulfills 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.

1.2 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.

The objectives assessed by the MSA at each grade level are embedded in the *VSC*. In addition, they are identified with the notation, ***assessment limit***. Assessment limits provide clarification about the specific skills and content that students are expected to have learned for each assessed objective. Even though some objectives in the VSC may not have an Assessment limit at a given grade-level, these non-assessed objectives still must be included in instruction. They introduce important concepts in preparation for assessed skills and content at subsequent grade levels.

The following provides one example of assessment limit of Grade 3 MSA-Reading:

STANDARD 1.0

General Reading Process

TOPIC:

B. VOCABULARY: Students will apply their knowledge of letter/sound relationships and word structure to decode unfamiliar words

INDICATOR:

1. Use a variety of phonetic skills to read unfamiliar words

OBJECTIVES:

a. Apply phonics skills

Assessment limits:

- Hard and soft consonants
- Initial consonant blends (2 letters)
- Open and closed syllables
- Digraphs

It should be noted that it was not the case that every indicator would necessarily be tested each year even if 100% of the standards should be tested. Consequently, the *VSC* specified curricular indicators and objectives that contributed directly to measuring content standards, which were aligned to the *MSA*. More information on assessment limits and standards can be found in appendix D, *The 2008 MSA-Reading Blueprint*.

1.3 Development and Review of the 2008 MSA-Reading Items and Test

The development of the 2008 MSA-Reading test required the involvement of four groups in addition to MSDE and Pearson. These groups are as follows:

National Psychometric Council

The National Psychometric Council (NPC) took a major role in reviewing and making recommendations to MSDE on the development and implementation of the 2008 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 approach 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, passage and art 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.

Vision Review Committee

A Vision Review Committee reviewed the passages, art, and items for bias to the visually impaired. The committee makes their recommendations to NOT put any item they had a concern with on Form 1.

Table 1.1 identifies responsibilities of each group in developing the 2008 MSA-Reading test.

Table 1.1 The 2008 MSA-Reading Responsibility for Test Development

Development of the 2008 MSA-Reading	Primary Responsibility
Development of Preliminary Blueprints and Item Specifications	Pearson; MSDE; NPC
Development of Preliminary Brief Constructed Response Rubrics	MSDE; NPC
Item Writing	Pearson; MSDE
Item Review	Pearson; MSDE; Content Review Committee
Bias Review	Pearson; MSDE; Bias Review Committee
Vision Review	Pearson; MSDE; Vision Review Committee
Construction of Field Test Forms	Pearson; MSDE
Modification of Special Forms	Pearson; MSDE
Review of Special Forms	MSDE
Pre-Field Test Training Workshops	Pearson; MSDE; LEAs
Field Test Administrations	MSDE; LEAs
Construction of Operational Test Forms	Pearson; MSDE; NPC
Review of Operational Test Forms	MSDE
Final Construction of Operational Test Forms	Pearson; MSDE

1.4 Test Form Design, Specifications, and Item Type

Test Form Design

Each test form included both operational and field test items. The 2008 assessment had 10 test forms for each grade. All 10 forms shared a single set of operational items, but contained unique field test items. It should be noted that MSDE administered two operational test forms every year until 2007. More detailed information about the 2008 test form design can be found in chapter 1.11, *Constructing the 2008 MSA-Reading Operational Forms*.

Test Form Specifications and Reporting Category

Tables 1.2 through 1.9 provide information on the total number of operational items included in the 2008 operational test form and how these items were broken down based on each content standard. It should be noted that the test specifications in these tables represent the targeted test design for each grade and show the targeted distribution of each content standard.

Specifically, each standard was used for reporting purposes (i.e., reporting subscale scores). That is, there were three reporting standards for reading across grades: general reading, literary, and informational processes. The number of raw score points for each reporting standard was identical (i.e., 15) for all grades except for grades 3 and 8.

Item Type

The 2008 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 2008 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 or a sentence, or in a more elaborate way. For the 2008 MSA-Reading, these items were scored on a general rubric with maximum values between 0 and 3. For example, the score given was the higher of the first and the second Reader's scores provided the scores were adjacent. A resolution Reader's score was used when two non-adjacent initial scores were received. That is, the resolution Reader's score was used in place of both the first and second Reader's scores. Detailed information on BCR scoring procedures and rules can be found in section 1.6, *MSA-Reading Scoring Procedures.*

Table 1.2 The 2008 MSA-Reading Item Distribution of Each Standard: Grade 3

General Reading			Literary Reading			Informational Reading		
No. of SR	No. of BCR	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items
16	0	16	8	2	10	9	2	11

Table 1.3 The 2008 MSA-Reading Item Distribution of Each Standard: Grade 5

General Reading			Literary Reading			Informational Reading		
No. of SR	No. of BCR	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items
15	0	15	9	2	11	9	2	11

Table 1.4 The 2008 MSA-Reading Item Distribution of Each Standard: Grade 8

General Reading			Literary Reading			Informational Reading		
No. of SR	No. of BCR	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items
16	0	16	8	2	10	9	2	11

Table 1.5 The 2008 MSA-Reading Item Distribution of Each Standard: Grades 4, 6, and 7

General Reading			Literary Reading			Informational Reading		
No. of SR	No. of BCR	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items	No. of SR	No. of *BCR*	No. of Total Items
15	0	15	9	2	11	9	2	11

Table 1.6 The 2008 MSA-Reading Total and Standard Scores: Grade 3

Total and Standard Scores			
General Reading	Literary Reading	Informational Reading	Total Score
16 (16 MC)	14 (8 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

Table 1.7 The 2008 MSA-Reading Total and Standard Scores: Grade 5

Total and Standard Scores			
General Reading	Literary Reading	Informational Reading	Total Score
15 (15 MC)	15 (9 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

Table 1.8 The 2008 MSA-Reading Total and Standard Scores: Grade 8

Total and Standard Scores			
General Reading	Literary Reading	Informational Reading	Total Score
16 (16 MC)	14 (8 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

Table 1.9 The 2008 MSA-Reading Total and Standard Scores: Grades 4, 6, and 7

Total and Standard Scores			
General Reading	Literary Reading	Informational Reading	Total Score
15 (15 MC)	15 (9 MC + 6 BCR)	15 (9 MC + 6 BCR)	45

1.5 Test Administration of the 2008 MSA-Reading

The 2008 MSA-Reading was administered to all students in grades 3 through 8. Pearson coordinated test administration procedures with MSDE prior to implementation. This section was prepared to provide general information about the 2008 test administration. Detailed information about the 2008 test administration can be obtained from the 2008 Test Administration and Coordination Manual (TACM) and Examiners Manual (EM) which are available from either MSDE or Pearson.

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. The Daily Testing Materials Tracking Record (or an equivalent form designed by the LEA) was used to track the distribution and return of Test Books.

Before testing began, the Test Examiners (TEs) carefully inventoried all test materials given to them, as they were accountable for the return of all secure materials at the end of testing. TEs checked to ensure they had all the materials they needed for testing.

For the Test Examiner, Pearson provided the following materials:

- Examiner's Manual- Reading

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

- 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
- Blank scratch paper

Each classroom used for the assessment also needed the following additional materials:

- Sign for the door reading "Testing: Do not Disturb"
- Digital clock or a watch, or clock with a second hand
- Copy of the Scoring Service Identification Document (SSID) Header Sheet

Two test-related Examiners Manuals (EM) were developed for the 2008 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 Pearson for scoring. The

TACM was distributed and reviewed during a workshop in January for STCs and LACs, with duplicates sent to each school along with its testing materials.

Test Administration Schedule

The primary test window for MSA was established by MSDE (April 1-10, 2008, with make-up testing held April 11-16, 2008). However, each Local Education Agency (LEA) set a specific schedule for administration of the MSA within that window for their district. For a given grade and content area, all testing had to take place on the same schedule. Each LEA schedule was submitted to MSDE in advance and approved for each district by the State. For example, all Grade 3 reading had to be administered on the same days throughout the LEA. In addition, each content area at each grade was tested on two days during the window.

The MSA-Reading testing schedule allowed approximately 2 1/2 hours on each of the two days (including preparation time and breaks).

For the 2008 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 March 10, 2008
• Reading Primary Testing Window	April 1 – April 10, 2008
• Make-up Testing Window	April 11 – April 16, 2008

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 were urged to do all they can 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 the 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 by LACs to determine the testing schedule to be followed.

During the administration of the 2008 MSA-Reading, MSDE had testing monitors in selected schools observing administration procedures and testing conditions. All monitors had identification cards for security purposes. There was 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 2008 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 can be viewed in section 2, Appendix C of the TACM.

On May 9, 2007, the U.S. Department of Education issued guidance for the development of Alternative Assessment based on Modified Academic Achievement Standards (also known as AA-MAAS or "Modified Assessments"). Maryland was in the process of developing the Modified Maryland School Assessment (Mod-MSA), but the assessment was not completed in time for the 2008 administration window. Students, however, might have been identified through the Individualized Education Program (IEP) process in the current school year as takers of the Mod-MSA. For 2008, these students were assessed using the regular MSA-Reading.

Accommodations for Assessment

Accommodations for assessment of students with disabilities (i.e., students having an Individualized Education Program or a Section 504 Plan) and students who are 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 could 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 2008 TACM for further information regarding testing accommodations.

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

The 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 a verbatim reading accommodation. 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 student's name, LEA number, and school number were written on the large-print Test/Answer Book for proper transcription into the standard-size Test/Answer Book.

The pre-printed student ID label was affixed to the standard-size Test/Answer Book containing the transcribed responses, and not to the large-print Test/Answer Book or Braille books. The bubbles on the demographic page of the standard-size Test/Answer Book were not filled in if there was a pre-printed student ID label for the student.

A certified Test Examiner (TE) transcribed the student responses into a standard-size Test/Answer Book exactly as given by the student. The standard-size Test/Answer Book with the pre-printed or general label attached was returned to Pearson with all other Test/Answer Books.

Large-Print Test/Answer Books and Braille Test/Answer Books containing the original student responses prior to transcription were to be returned with Non-Scorable materials. Any Test/Answer Books which were used as source documents for transcription were 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 standard-size materials. Specific packing instructions are provided in the 2008 TACM in section 4.

Verbatim Reading Accommodation and Kurzweil™ Test Form 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, received the accommodation on the 2008 MSA-Reading. The accommodation was provided by a live reader or through technology. Section 1 of the 2008 TACM provided information on verbatim reading instruction. Technology used to provide the verbatim reading accommodation was Kurzweil™ reading software. Official, secure electronic copies of the test were ordered through the LAC. MSDE encouraged (but did not require) the use of the Kurzweil™ software to ensure uniformity in the delivery of the verbatim reading accommodation throughout the state.

Students using Kurzweil™ software had to familiarize themselves with its operation prior to the test administration. When there were technical difficulties with Kurzweil™ a certified staff member was used instead. Kurzweil™ Test Form CDs were shipped by Pearson. After testing, schools returned the CDs to Pearson with the non-scorable secure materials.

Administration Procedures for Students with IEP, 504 Plan, or ELL Plan Permitting a Dictated Responses or Use of Word Processor

A student whose IEP, 504 Plan, or ELL Plan permitted a dictated response had his/her responses transcribed at the school level by an eligible TE, or by a staff member working under the direct supervision of a certified TE, into the student’s Test/Answer Book with a pre-printed or generic ID label attached.

A student whose IEP, 504 Plan, or ELL plan permitted the use of a word processor had his/her responses transcribed by hand or under the direct supervision of an eligible TE or STC exactly as the student entered his/her responses on the word processor. The student’s responses were always transcribed at the school level into the student’s Test/Answer Book with the pre-printed or generic ID label attached. After the student’s responses were transcribed, the memory of the word processor was cleared. The original word-processed print-out was returned to Pearson with the non-scorable materials.

Test Format

All grade levels of the MSA-Reading used a Test Book format in which students wrote their answers directly in the Test Book. 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 highlighters 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, creating problems 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.

Security of Test Materials

The following code of ethics conforms to the Standards for Educational and Psychological Testing developed by the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education (Harcourt, 2008):

> It is breach of professional ethics for school personnel to provide verbal or nonverbal clues or answers, teach items on the test, share writing prompts, coach, hint, or in any way influence a student's performance during the testing situation. A breach of ethics may result in invalidation of test results and local education agency or MSDE disciplinary action. (p. 13)

The Test/Answer Books for the 2008 MSA-Reading were confidential and kept secure at all times. Unauthorized use, duplication, or reproduction of any or all portions of the assessment was prohibited, which is reflected by the following statement (Harcourt, 2008):

> 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. 13)

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

1.6 Scoring Procedures of the 2008 MSA-Reading

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

Once received by Pearson, 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 the document, *Performance Assessment Scoring Center: Spring 2008 Scoring Specification for MSA-Reading and Math*, which is available from either MSDE or Pearson.

Scoring Staff

The MSDE had one Room Director (RD) dedicated to 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. (Detailed information about anchor-pulling procedures can be found in the following portion of this section: *Development Procedures for Anchor Pulling*.) 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 was unable to 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 had been trained and successfully qualified.

2) Team Leader (TL)

An experienced reader who directly monitored the scoring of a team of Readers and retrained 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 Pearson's staffing vendor their résumé and documentation of a four-year college degree. As part of the initial screening process for recruiting Readers into Pearson'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 were unable to 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 Pearson to eliminate potential Readers who might seem qualified according to their educational and professional experience but who could not 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 added to Pearson's general pool of Readers who were potential scorers of Reading assessments. This addition to the general pool did not necessarily qualify 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 was also provided in the workshop. 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 who did not perform 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 began 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. TLs had to score 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 committee's 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 to which they were 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 have written 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) and brief constructed response (BCR). The MSA-Reading contained all such item types for operational scoring, and each of the 10 forms per grade 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-Reading response documents were image-scanned at Pearson's scoring center in San Antonio, Texas. The image scanner captured document identification (ID), demographic information, SR responses, and created 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 and all other data were made available to Scoring Editing for human review.

All constructed responses were scored by Pearson'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, 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 had to score at least an 80% match on two of three qualifying sets for Reading.

Inter-Rater Agreement

Pearson'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 listing the entire prompt's Readers and providing 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

Pearson'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 (e.g., 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.

Retraining 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 than managing performance. PASC had identified this issue and planned 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 provide retraining when the level of acceptable performance had not been met.

Pre-"Live" training on Field Test prompts

For 2008, 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.

Scoring Rules for MSA-Reading

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

- Reading BCR items were scored:

  0, 1, 2, or 3 with two readings
- Scores given were the higher of the 1st and 2nd Reader's scores provided they were adjacent.
- 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 Reader's 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

A Developer is 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 Specialists 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. For a given reading prompt, the PASC Developers had the following responsibilities:

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, or Low.
- Marked High, Medium, and Low papers—marked especially good ones that might potentially receive 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 needed to 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 there could have been 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 had to be clearly so.

E. ***Calibration sets (validity sets)***: These were composed of five responses of mixed quality, arranged in random order. Pearson 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 Pearson 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 sessions 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, PASC Language Arts Specialists, Managers, Training Supervisors, 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 on 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 results:

- It established scores for papers for which there was virtually unanimous agreement.
- It identified papers that were on the line between two adjacent scores, necessitating 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.

1.7 Classical Analyses for the 2008 MSA-Reading Operational Forms

Table 1.10 shows the descriptive statistics for the 2008 MSA-Reading operational form for each grade. First of all, the following results were obtained with a statewide population, and the total score point of each operational test form was 45 regardless of grade.

Detailed information about the total and subtotal (strand) score points of the 2008 MSA-Reading operational form for each grade can be found in section 1.4, *Test Form Design, Specifications, and Item Type*.

Table 1.10 Classical Descriptive Statistics for the 2008 MSA-Reading: Grades 3 through 8

Grade	N	Total number of Items	Min. Point	Max. Point	Mean	SD	Reliability	SEM
3	58,301	37	0	45	28.69	6.95	0.86	2.60
4	59,697	37	0	45	28.57	6.88	0.87	2.48
5	60,486	37	0	45	29.71	6.81	0.87	2.46
6	61,036	37	0	45	30.14	6.99	0.88	2.42
7	62,513	37	0	45	29.05	7.27	0.88	2.52
8	63,858	37	0	45	29.54	7.26	0.88	2.51

Note. Analysis was conducted with a statewide population.

1.8 P-Value Check with Year-to-Year Linking Common Items

Tables 1.11 through 1.16 provide information about how much the p-values of the 2008 year-to-year linking common items varied from those calculated in previous years. Only SR items were used for the purpose of year-to-year linking. The 2006 p-values were calculated based on a smaller, field-test sample while the 2008 statistics are based on the statewide population. Item sequence numbers appearing the tables were assigned based on the 2008 assessment. Detailed information on the 2008 MSA-Reading test design can be found in chapter 1.11, *Constructing the 2008 MSA-Reading Operational Forms*. In general, we could conclude that most of the 2008 p-values were slightly increased compared to the 2006 p-values across all grades.

Table 1.11 P-Value Comparisons of Linking Common Items for Year 2006 vs. Year 2008: Grade 3

Item Number	Item Type	2006	2008
1	SR	0.89	0.93
14	SR	0.68	0.78
16	SR	0.45	0.46
17	SR	0.49	0.43
19	SR	0.62	0.62
20	SR	0.78	0.85
22	SR	0.76	0.78
23	SR	0.65	0.65
25	SR	0.63	0.64

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Descriptive Statistics of Linking Common Items for Year 2006 vs. Year 2008: Grade 3

Grade	Year	No. of Items	*M*	*SD*
3	2006	9	0.66	0.14
	2008	9	0.68	0.17

Table 1.12 P-Value Comparisons of Linking Common Items for Year 2006 vs. Year 2008: Grade 4

Item Number	Item Type	2006	2008
2	SR	0.80	0.79
6	SR	0.90	0.95
7	SR	0.90	0.96
13	SR	0.64	0.63
15	SR	0.73	0.72
16	SR	0.63	0.68
18	SR	0.61	0.68
19	SR	0.56	0.61
21	SR	0.45	0.49
22	SR	0.85	0.82
24	SR	0.67	0.78

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Descriptive Statistics of Linking Common Items for Year 2006 vs. Year 2008: Grade 4

Grade	Year	No. of Items	*M*	*SD*
4	2006	11	0.70	0.15
	2008	11	0.74	0.14

Table 1.13 P-Value Comparisons of Linking Common Items for Year 2006 vs. Year 2008: Grade 5

Item Number	Item Type	2006	2008
1	SR	0.80	0.86
2	SR	0.73	0.81
3	SR	0.87	0.91
4	SR	0.89	0.94
6	SR	0.88	0.92
7	SR	0.92	0.95
8	SR	0.84	0.89
12	SR	0.70	0.72
14	SR	0.57	0.63
15	SR	0.71	0.80
17	SR	0.70	0.70
18	SR	0.71	0.76
20	SR	0.54	0.62
21	SR	0.69	0.73
23	SR	0.63	0.73

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Descriptive Statistics of Linking Common Items for Year 2006 vs. Year 2008: Grade 5

Grade	Year	No. of Items	*M*	*SD*
5	2006	15	0.75	0.12
	2008	15	0.80	0.11

Table 1.14 P-Value Comparisons of Linking Common Items for Year 2006 vs. Year 2008: Grade 6

Item Number	Item Type	2006	2008
1	SR	0.88	0.92
2	SR	0.93	0.96
4	SR	0.86	0.88
7	SR	0.89	0.92
8	SR	0.51	0.52
9	SR	0.78	0.80
10	SR	0.89	0.93
12	SR	0.57	0.59
14	SR	0.53	0.61
15	SR	0.74	0.73
17	SR	0.74	0.78
18	SR	0.70	0.73
20	SR	0.79	0.84
21	SR	0.29	0.32
23	SR	0.56	0.61

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Descriptive Statistics of Linking Common Items for Year 2006 vs. Year 2008: Grade 6

Grade	Year	No. of Items	*M*	*SD*
6	2006	15	0.71	0.18
	2008	15	0.74	0.18

Table 1.15 P-Value Comparisons of Linking Common Items for Year 2006 vs. Year 2008: Grade 7

Item Number	Item Type	2006	2008
1	SR	0.92	0.95
2	SR	0.89	0.92
3	SR	0.77	0.79
5	SR	0.85	0.90
6	SR	0.91	0.94
8	SR	0.69	0.74
10	SR	0.83	0.87
11	SR	0.55	0.60
13	SR	0.70	0.77
14	SR	0.60	0.59
16	SR	0.63	0.66
17	SR	0.90	0.93
19	SR	0.79	0.82

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Descriptive Statistics of Linking Common Items for Year 2006 vs. Year 2008: Grade 7

Grade	Year	No. of Items	*M*	*SD*
7	2006	13	0.77	0.13
	2008	13	0.81	0.13

Table 1.16 P-Value Comparisons of Linking Common Items for Year 2006 vs. Year 2008: Grade 8

Item Number	Item Type	2006	2008
1	SR	0.91	0.95
2	SR	0.90	0.93
4	SR	0.79	0.84
5	SR	0.80	0.83
6	SR	0.88	0.89
8	SR	0.69	0.76
10	SR	0.71	0.78
11	SR	0.74	0.80
13	SR	0.73	0.73
14	SR	0.51	0.54
16	SR	0.53	0.55
17	SR	0.67	0.65
19	SR	0.55	0.59

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Descriptive Statistics of Linking Common Items for Year 2006 vs. Year 2008: Grade 8

Grade	Year	No. of Items	*M*	*SD*
8	2006	13	0.72	0.13
	2008	13	0.76	0.14

1.9 Validation Check with the 2008 Operational BCR Items

To collect information about how much the same BCR items that appeared in both 2006 and 2008 changed in terms of item difficulty, indices such as the classical p-value and Rasch item difficulty were calculated.

These items were first field-tested on the 2006 assessment and appeared as operational test items on the 2008 assessment, as shown in Table 1.17. There was only one operational test form at each grade in 2008. The item numbers in Tables 1.18 through 1.35 were assigned based on the 2008 assessment. Detailed information about the specific test design and construction of Year 2008 can be obtained from section 1.4, *Test Structure of the 2008 MSA-Reading* and section 1.11, *Constructing the 2008 MSA-Reading Operational Forms*.

While the 2006 p-value was calculated with a field test sample, the 2008 p-value was calculated with a statewide population. The p-value of a BCR item was the mean item score divided by the item score range. The percentage of "Omits" response to each CR item was low and indicated that a small number of students did not respond at all.

Classical item p-value results indicated that, in general, most of the 2008 p-values increased somewhat compared to the 2006 p-values. For grade 8, however, most of the 2008 p-values slightly decreased compared to the 2006 p-values.

With respect to Rasch item difficulty analysis, most of the 2008 items became easier compared to the 2006 items except in grade 8. For grade 8, most of the 2008 Rasch item difficulties slightly increased compared to the 2006 Rasch item difficulties. It should be noted that all Rasch difficulties were put on the base scale.

In conclusion, both p-value and Rasch difficulty results reflected the same phenomenon, indicating that most of the 2008 items became easier than the 2006 items except for in grade 8.

Table 1.17 Form Identification for Items Appearing in both 2006 and 2008: Grades 3 through 8

Grade	Year 2006	Year 2008
3	Form 1,3	Form 1-10
4	Form 1,2	Form 1-10
5	Form 1,2	Form 1-10
6	Form 1,4	Form 1-10
7	Form 1,2	Form 1-10
8	Form 1,2	Form 1-10

Table 1.18 P-Value Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 3

Item Number	Item Type	Year 06	Year 08
15	BCR	0.39	0.39
18	BCR	0.46	0.44
21	BCR	0.41	0.46
24	BCR	0.31	0.36

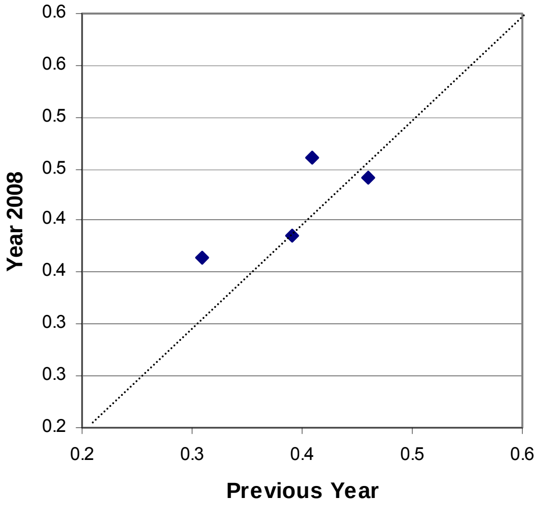

Table 1.19 Score-Point Distribution Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 3

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2006	15	BCR	2,197	1.18	0.68	12.92	54.71	30.00	1.23	1.14
2006	18	BCR	2,450	1.37	0.62	4.93	49.47	43.06	0.41	2.12
2006	21	BCR	2,514	1.24	0.76	16.19	43.79	36.75	2.15	1.11
2006	24	BCR	2,525	0.93	0.76	28.87	47.21	20.63	1.66	1.62
2008	15	BCR	58,301	1.16	0.82	23.92	36.43	36.29	2.19	1.16
2008	18	BCR	58,301	1.33	0.58	3.30	58.51	35.67	0.95	1.57
2008	21	BCR	58,301	1.38	0.75	12.24	39.43	43.96	3.71	0.67
2008	24	BCR	58,301	1.09	0.75	21.44	47.66	28.62	1.37	0.91

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Table 1.20 Rasch Item and Step Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 3

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2006	15	BCR	2.1690	-3.0299	-0.2015	3.2314
2006	18	BCR	2.0716	-3.9839	-0.6980	4.6819
2006	21	BCR	1.9855	-2.3612	-0.5222	2.8835
2006	24	BCR	2.4801	-2.1261	-0.1422	2.2684
2008	15	BCR	2.2206	-1.9595	-0.8455	2.8050
2008	18	BCR	1.5812	-4.1468	0.1837	3.9631
2008	21	BCR	1.6154	-2.4425	-0.5135	2.9561
2008	24	BCR	2.3142	-2.4282	-0.3975	2.8256

Note. Rasch item and step difficulties were placed on a common scale.

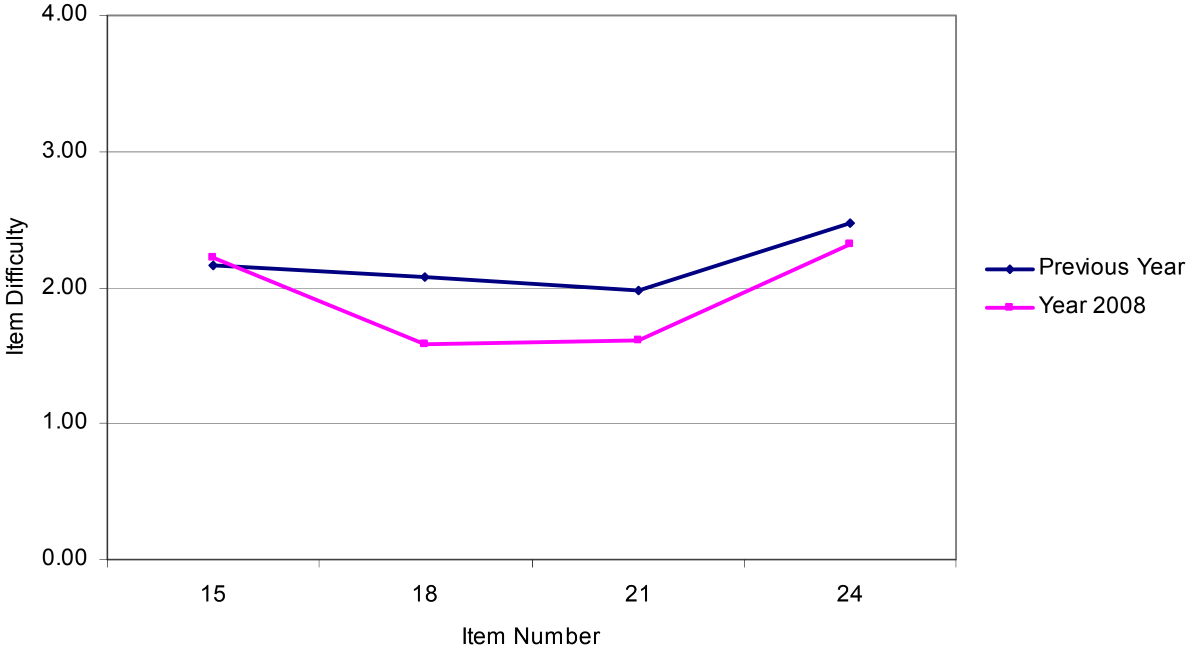

Figure 1.1 Rasch Item Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 3

Table 1.21 P-Value Comparisons of BCR items for Year 2006 vs. Year 2008: Grade 4

Item Number	Item Type	Year 06	Year 08
14	BCR	0.45	0.54
17	BCR	0.42	0.49
20	BCR	0.48	0.52
23	BCR	0.47	0.49

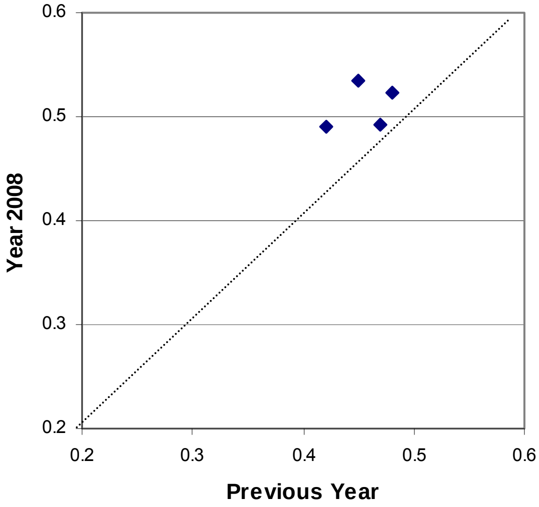

Table 1.22 Score-Point Distribution Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 4

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2006	14	BCR	2,297	1.35	0.55	3.00	58.86	37.35	0.48	0.30
2006	17	BCR	2,458	1.25	0.54	2.97	65.50	29.62	0.24	1.67
2006	20	BCR	2,164	1.45	0.61	5.03	43.39	50.28	0.42	0.88
2006	23	BCR	2,381	1.41	0.55	2.10	54.01	42.80	0.38	0.71
2008	14	BCR	59,697	1.61	0.58	2.10	37.14	57.99	2.46	0.31
2008	17	BCR	59,697	1.47	0.56	1.61	48.10	48.63	0.53	1.13
2008	20	BCR	59,697	1.57	0.58	2.82	37.87	57.33	1.39	0.59
2008	23	BCR	59,697	1.48	0.56	1.91	48.26	48.45	0.88	0.51

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Table 1.23 Rasch Item and Step Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 4

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2006	14	BCR	2.2261	-4.5927	-0.1531	4.7459
2006	17	BCR	2.5343	-4.9127	-0.0318	4.9444
2006	20	BCR	2.3989	-3.8405	-0.9598	4.8003
2006	23	BCR	2.1120	-4.7233	-0.3093	5.0326
2008	14	BCR	1.4129	-3.7214	-0.5028	4.2243
2008	17	BCR	2.4090	-5.5103	-0.8980	6.4084
2008	20	BCR	1.8767	-3.7237	-0.7714	4.4951
2008	23	BCR	2.1074	-4.7107	-0.5710	5.2817

Note. Rasch item and step difficulties were placed on a common scale.

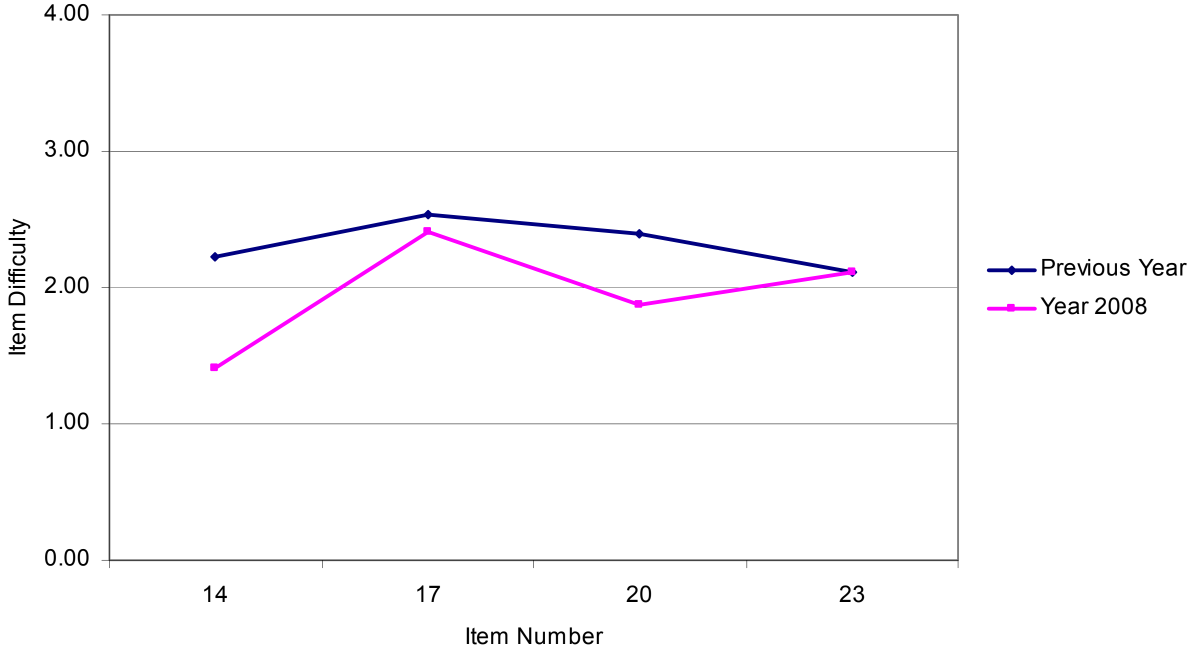

Figure 1.2 Rasch Item Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 4

Table 1.24 P-Value Comparisons of BCR items for Year 2006 vs. Year 2008: Grade 5

Item Number	Item Type	Year 06	Year 08
13	BCR	0.52	0.61
16	BCR	0.48	0.51
19	BCR	0.40	0.45
22	BCR	0.24	0.26

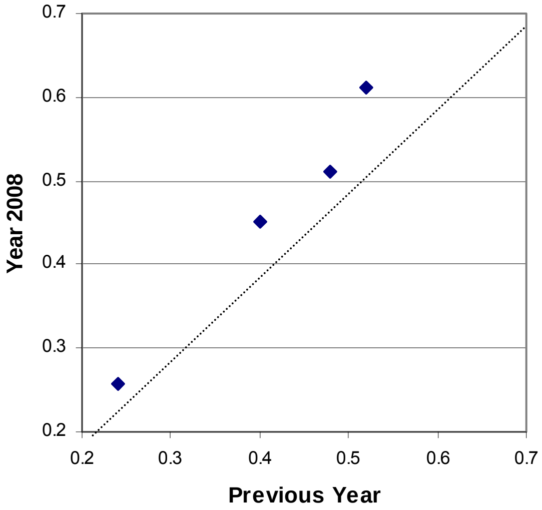

Table 1.25 Score-Point Distribution Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 5

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2006	13	BCR	2,479	1.56	0.62	4.24	33.64	59.46	1.05	1.61
2006	16	BCR	2,423	1.43	0.63	4.29	44.41	48.20	0.70	2.39
2006	19	BCR	2,209	1.20	0.83	21.46	39.11	34.45	4.12	0.86
2006	22	BCR	2,496	0.73	0.68	2.28	37.46	47.40	12.54	0.32
2008	13	BCR	60,486	1.84	0.46	0.66	17.23	78.89	2.89	0.33
2008	16	BCR	60,486	1.54	0.57	2.11	42.52	53.69	1.25	0.43
2008	19	BCR	60,486	1.36	0.78	14.10	39.38	41.57	4.38	0.58
2008	22	BCR	60,486	0.77	0.76	40.15	41.72	15.66	1.44	1.02

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Table 1.26 Rasch Item and Step Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 5

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2006	13	BCR	1.3615	-3.2989	-1.036	4.3349
2006	16	BCR	1.6034	-3.7081	-0.6771	4.3852
2006	19	BCR	1.7248	-1.8343	-0.4197	2.2540
2006	22	BCR	3.1844	-2.5412	-0.2708	2.8119
2008	13	BCR	0.8128	-3.5853	-1.1317	4.7170
2008	16	BCR	1.6931	-4.1387	-0.5787	4.7175
2008	19	BCR	1.9023	-2.1832	-0.5677	2.7509
2008	22	BCR	3.2443	-1.9905	-0.2633	2.2539

Note. Rasch item and step difficulties were placed on a common scale.

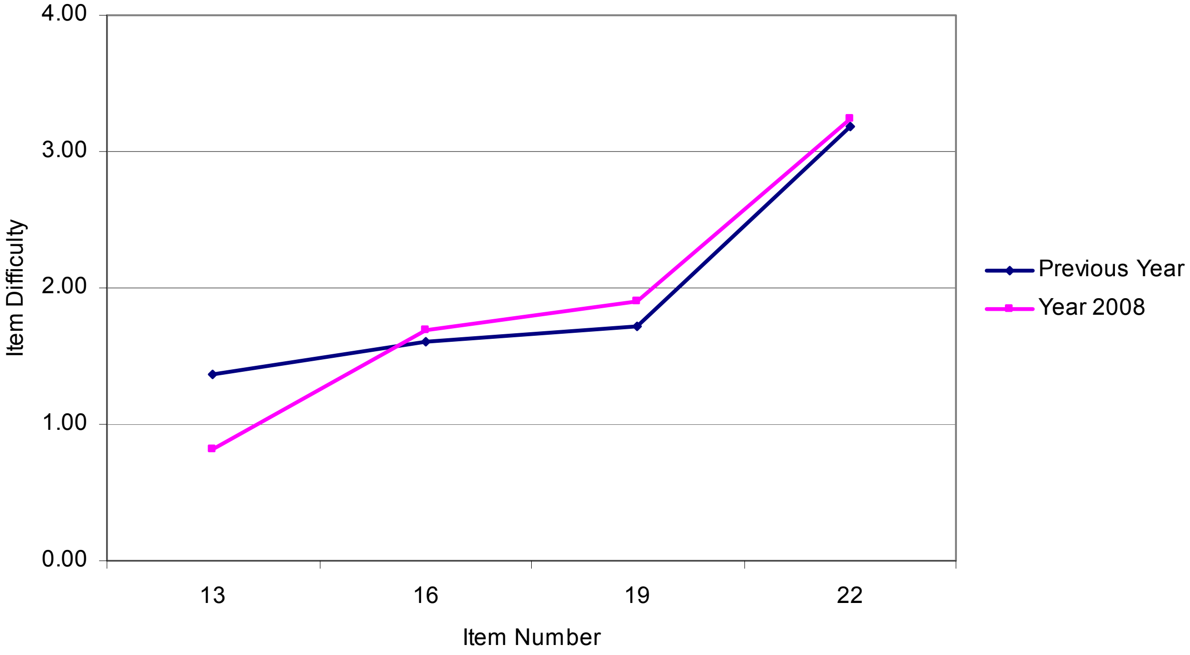

Figure 1.3 Rasch Item Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 5

Table 1.27 P-Value Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 6

Item Number	Item Type	Year 06	Year 08
13	BCR	0.47	0.51
16	BCR	0.44	0.50
19	BCR	0.47	0.48
22	BCR	0.48	0.55

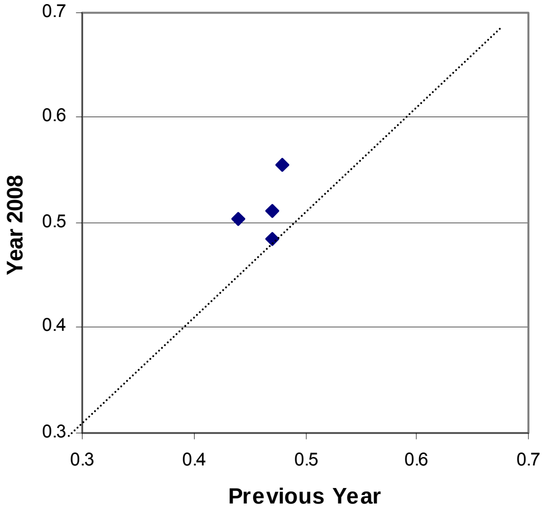

Table 1.28 Score-Point Distribution Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 6

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2006	13	BCR	2,269	1.41	0.64	4.71	52.01	39.40	3.26	0.62
2006	16	BCR	2,493	1.31	0.65	6.82	53.19	36.54	1.60	1.85
2006	19	BCR	2,366	1.42	0.63	2.92	52.83	39.22	3.63	1.39
2006	22	BCR	2,452	1.44	0.71	3.51	45.07	42.01	5.02	4.40
2008	13	BCR	61,036	1.54	0.56	1.17	44.92	51.59	1.83	0.50
2008	16	BCR	61,036	1.51	0.60	2.45	44.39	50.26	2.15	0.75
2008	19	BCR	61,036	1.45	0.59	2.95	49.47	45.35	1.68	0.55
2008	22	BCR	61,036	1.66	0.66	2.37	33.84	55.46	7.20	1.14

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Table 1.29 Rasch Item and Step Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 6

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2006	13	BCR	1.3979	-3.4419	0.1636	3.2783
2006	16	BCR	1.7989	-3.4763	-0.1102	3.5866
2006	19	BCR	1.1271	-3.8493	0.3761	3.4732
2006	22	BCR	1.0466	-3.4049	0.2009	3.2040
2008	13	BCR	1.2440	-4.3701	0.0005	4.3697
2008	16	BCR	1.3674	-3.7982	-0.3367	4.1348
2008	19	BCR	1.6025	-3.9479	-0.1895	4.1374
2008	22	BCR	0.7994	-3.3192	-0.0990	3.4182

Note. Rasch item and step difficulties were placed on a common scale.

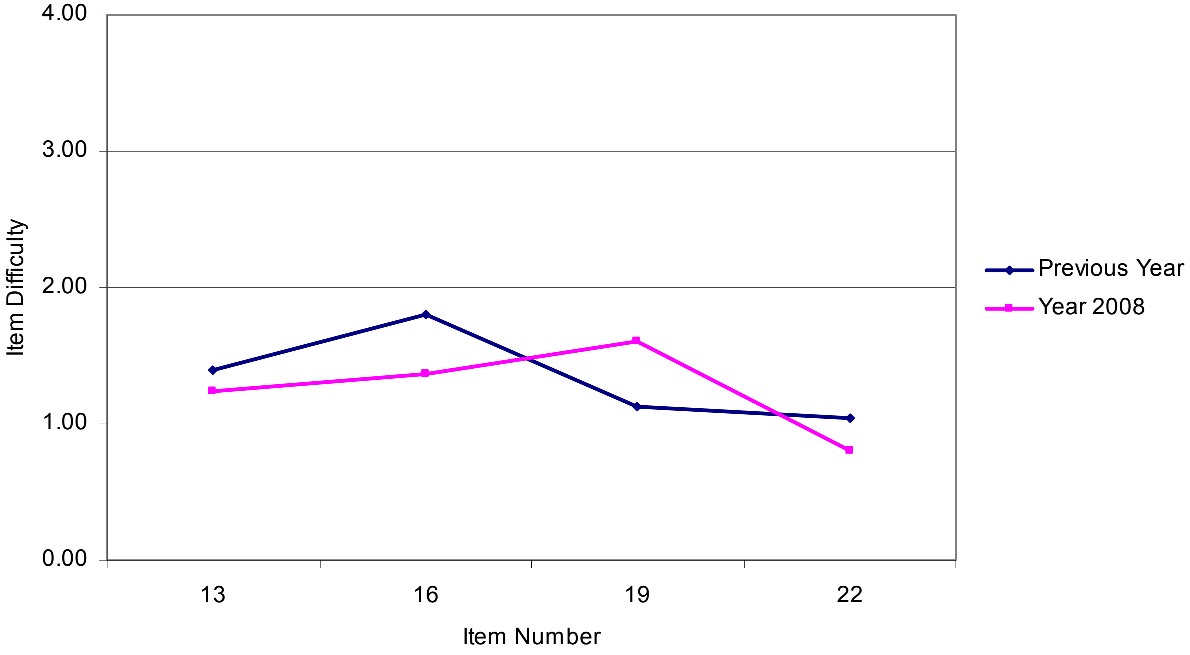

Figure 1.4 Rasch Item Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 6

Table 1.30 P-Value Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 7

Item Number	Item Type	Year 06	Year 08
9	BCR	0.42	0.48
12	BCR	0.47	0.54
15	BCR	0.46	0.49
18	BCR	0.34	0.41

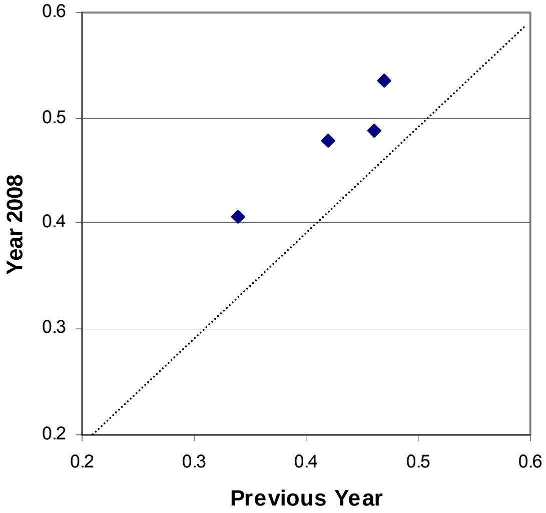

Table 1.31 Score-Point Distribution Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 7

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2006	9	BCR	2,339	1.26	0.72	11.63	47.07	36.47	2.05	2.78
2006	12	BCR	2,417	1.41	0.61	2.36	54.65	39.10	2.57	1.32
2006	15	BCR	2,266	1.38	0.71	9.70	43.95	42.28	3.09	0.97
2006	18	BCR	2,387	1.03	0.65	14.29	61.25	19.15	1.26	4.06
2008	9	BCR	62,513	1.44	0.73	6.79	45.47	40.41	5.86	1.47
2008	12	BCR	62,513	1.61	0.65	1.93	39.76	51.01	6.36	0.95
2008	15	BCR	62,513	1.46	0.76	12.49	31.00	52.25	3.62	0.65
2008	18	BCR	62,513	1.22	0.68	10.80	56.13	29.49	2.32	1.25

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Table 1.32 Rasch Item and Step Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 7

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2006	9	BCR	1.7444	-2.7903	-0.3255	3.1158
2006	12	BCR	1.0754	-4.0074	0.3505	3.6568
2006	15	BCR	1.4679	-2.6513	-0.3766	3.0280
2006	18	BCR	2.2241	-3.1453	0.2696	2.8757
2008	9	BCR	1.3861	-2.8846	0.0487	2.8359
2008	12	BCR	0.5998	-3.4107	0.1328	3.2779
2008	15	BCR	1.8303	-2.1114	-1.141	3.2524
2008	18	BCR	2.0556	-3.0554	0.0723	2.9830

Note. Rasch item and step difficulties were placed on a common scale.

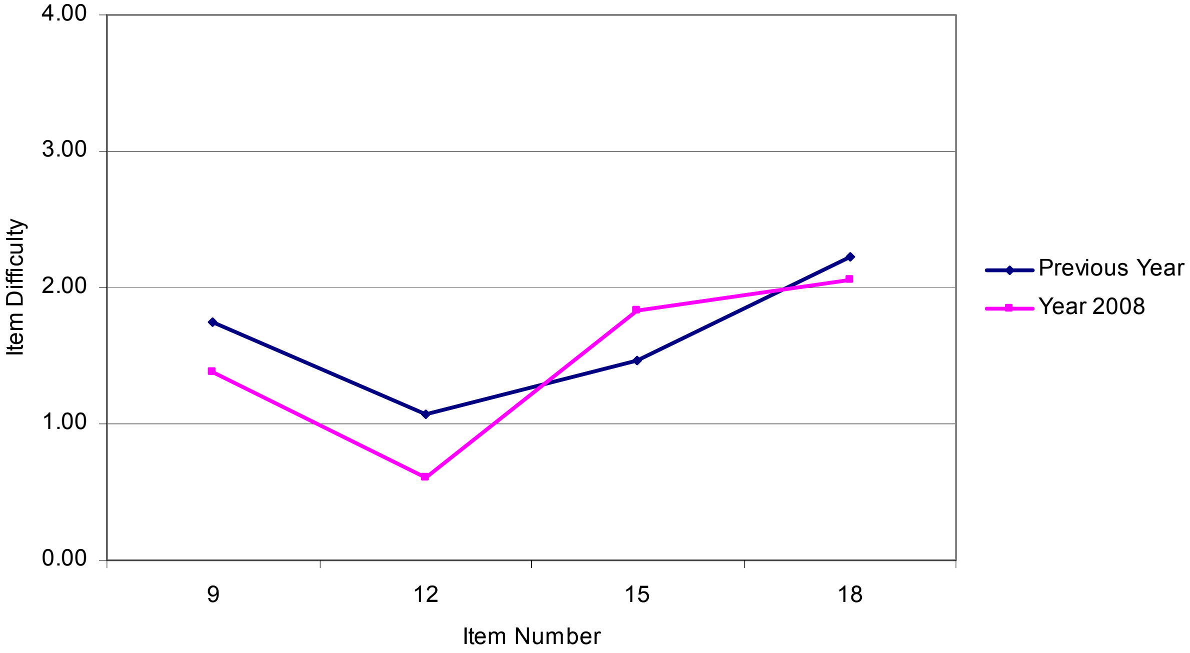

Figure 1.5 Rasch Item Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 7

Table 1.33 P-Value Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 8

Item Number	Item Type	Year 06	Year 08
9	BCR	0.60	0.56
12	BCR	0.47	0.53
15	BCR	0.54	0.52
18	BCR	0.51	0.48

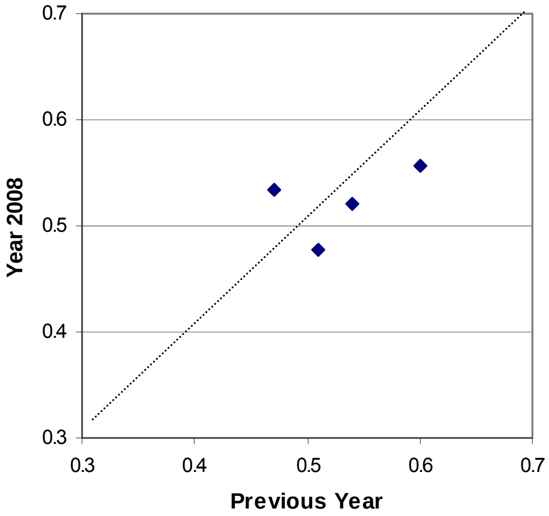

Table 1.34 Score-Point Distribution Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 8

Year	Item #	Item Type	N	Mean	SD	Score-Point Distribution (%)				
						0	1	2	3	Omit
2006	9	BCR	2,311	1.79	0.69	2.38	25.75	58.76	11.81	1.30
2006	12	BCR	2,465	1.41	0.60	1.74	51.60	42.23	1.58	2.84
2006	15	BCR	2,288	1.61	0.79	6.82	32.34	48.25	10.62	1.97
2006	18	BCR	2,456	1.52	0.75	8.23	34.32	50.16	5.86	1.43
2008	9	BCR	63,858	1.67	0.62	0.79	35.56	56.53	6.21	0.91
2008	12	BCR	63,858	1.60	0.59	0.59	39.35	55.46	3.40	1.20
2008	15	BCR	63,858	1.56	0.74	8.54	27.83	56.97	4.78	1.87
2008	18	BCR	63,858	1.43	0.82	13.39	34.08	44.69	6.61	1.23

Note. The 2006 analysis was conducted with a field test sample.
Note. The 2008 analysis was conducted with a statewide population.
Note. Item sequence numbers were assigned based on the 2008 assessment.

Table 1.35 Rasch Item and Step Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 8

Year	Item #	Item Type	Rasch Difficulty	Step 0-1	Step 1-2	Step 2-3
2006	9	BCR	0.2609	-2.5232	-0.3082	2.8314
2006	12	BCR	1.1084	-4.1786	0.1288	4.0498
2006	15	BCR	0.7647	-2.0470	-0.2576	2.3046
2006	18	BCR	1.0691	-2.1787	-0.5151	2.6938
2008	9	BCR	0.4280	-4.0698	0.1405	3.9293
2008	12	BCR	0.5900	-4.3977	0.2564	4.1412
2008	15	BCR	1.2904	-2.0894	-1.0192	3.1087
2008	18	BCR	1.5867	-1.9730	-0.6017	2.5747

Note. Rasch item and step difficulties were placed on a common scale.

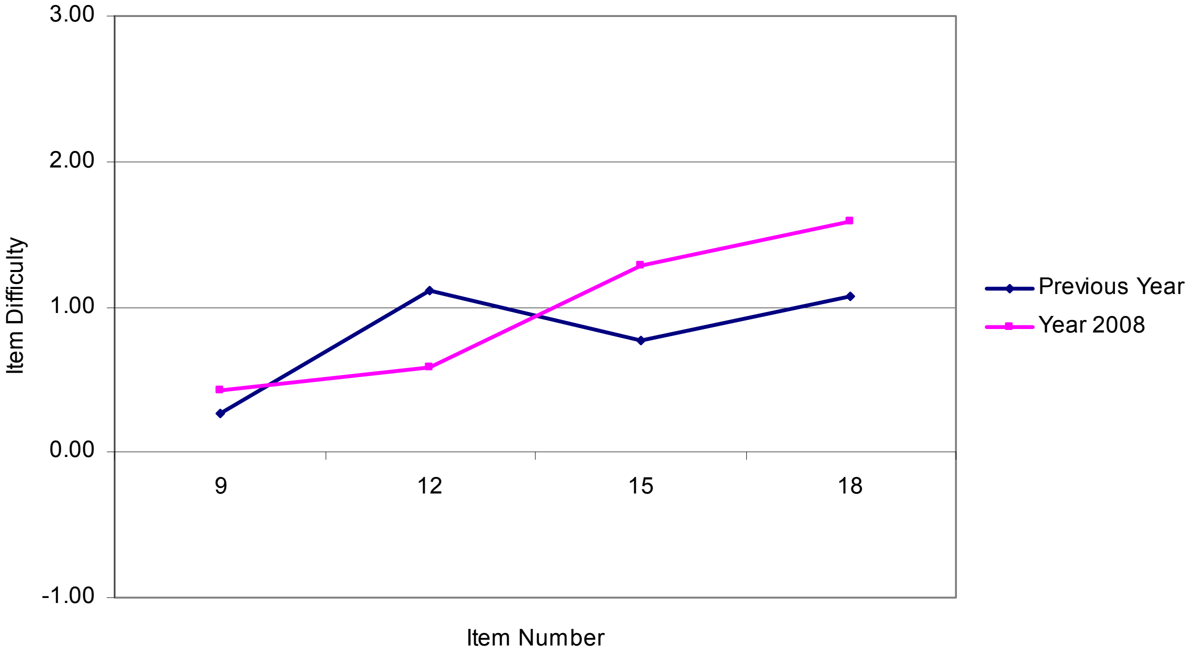

Figure 1.6 Rasch Item Difficulty Comparisons of BCR Items for Year 2006 vs. Year 2008: Grade 8

1.10 Field Test Analyses

All field test items embedded in operational forms were subjected to rigorous analyses for their properties in order to provide information about which items may be included as operational items in the future. All statistical results concerning field test items were preserved in the 2008 item bank. Information on the item bank can be found in section 1.16, *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 any students (i.e., nonfunctional distractor)
- An item was selected by a high proportion of high-ability students while being selected by a low proportion of low-ability students (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.

All items required 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 being dropped as a possible operational item. However, if the item represented important content that had not been extensively taught, a justification could have been made for including it in an 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 groups have been statistically matched on their tested proficiency.

In present analyses, the gender reference group was males, and the ethnic reference group was Caucasians. The gender focal group was females and the ethnic focal group was African-

Americans. For each operational form, the student's total score was used as the matching variable.

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

- If the differential difficulty of the item is related to group membership, and the difference is deemed unfair, then the item should not be used at all.
- If the differential difficulty of the item is related to group membership, but the difference is not deemed unfair, then the item should only be used if there is no other item matching the test blueprint.

It should be noted that DIF analysis results on all the field test items were archived in the 2008 Maryland item bank. In addition, detailed information about the *DIF* procedures can be found in section 3.7, *Differential Item Functioning*..

Item Response Theory (IRT) Analyses

To put the 2008 field test items on the base scale (i.e., the 2003 scale for grades 3, 5, and 8 and the 2004 scale for grades 4, 6, and 7), each field test item was freely calibrated while the Rasch item and step parameters of the 2008 operational items, which has been already placed on the base scale during the 2008 operational calibration and equating, were fixed to their post-equated values.

It should be noted that all the Rasch item difficulties, step difficulties, and fit statistics (i.e., Rasch Infit and Outfit indices) of the field test items were archived in the 2008 Maryland item bank. These field test items are eligible to be used as operational items in subsequent years.

1.11 Constructing the 2008 MSA-Reading Operational Forms

Due to the decision to remove all of the SAT10 items starting with the 2008 administration, MSDE and Pearson team members examined options to replace the 25 SAT10 items removed from the test.

The minimum requirement was to develop enough items to cover the same total and subtotal score points that SAT10 common items contributed in previous years (for grade 5, for example, 45 total score points with 15 points each for general reading, literary, and informational reading). In addition, it was decided that only one operational form would be developed for the 2008 administration and that options for year-to-year equating would focus on items that were originally field-tested in 2006.

General Overview of the 2003 through 2007 MSA-Reading

- Both NRT and CRT: SAT10 was utilized both as the *norm-referenced test* (NRT) and *curriculum-referenced test* (CRT) for reading assessment. For example, 25 out of 50 SAT10 selected-response (SR) items contributed to the Maryland CRT total score. This is 56% of the 45 total score points for the CRT.
- Reading Test Form: Each reading test form included SAT10 (NRT and CRT) items, Maryland-specific (CRT) operational items, and Maryland-specific (CRT) field test items. For Grade 5, for example, fifty SAT10 items, six (4 SR and 2 *brief constructed-response, or* BCR) operational Literary passage-based items, six (4 SR and 2 BCR) operational Informational passage-based items, and ten (7 SR and 3 BCR items) field-testing Literary or Informational passage-based items appeared on each test form.
- Each Strand Score Point (Subtotal Score Point) of SAT10 Common Items: Content strands covered by the 25 SAT10 common items included General, Literary, and Informational Reading. These common items met the requirement that a possible linking pool should be a mini-version of the whole test. For Grade 5, for example, 15 out of 25 SAT10 items contributed to General Reading (GR), 5 to Literary, and 5 to Informational Reading.
- Common Linking items: Between 2003 and 2007, SAT10 common items were exclusively used for both form-to-form linking and year-to-year linking.
- Continuity and Stability: SAT10 was administered to every student with Sessions 1 and 2 on Day 1 without any changes every year between 2003 and 2007. The test had a total of five sessions and was administered over two days.

Two Operational Forms for the 2003 through 2007 MSA-Reading

- Test Security Issues: Two operational forms (Forms A and B) were developed and administered due to test security concerns. These forms had some operational items in common and some items unique to each form.
- Different Set of Literary and Informational Passages: Different operational forms were implemented by having a different Literary passage (4 SR and 2 BCR items) and Informational passage (4 SR and 2 BCR items) appear on each operational form. In other words, one Literary passage and one Informational passage appears on Form A while another set of passages appear on Form B. It should be noted that these passages were originally developed and field-tested with 7 SR and 3 BCR items. In addition, the

location of these passages when field-tested was in the very last session of either Day 1 or Day 2. For Grade 5, for example, either the Literary or Informational passage was field-tested in Session 3 on Day 1.

Session Design for the 2003 through 2007 MSA-Reading

- Days 1 and 2: The first testing day consisted of 3 sessions and the second day of 2 sessions. For Grade 5, for example, Day 1 consisted of Sessions 1, 2, and 3 and Day 2 of Sessions 4 and 5.
- Sessions 1 and 2: Administered all SAT10 items (e.g., 50 items for Grade 5).
- Session 3: Field-tested either a Literary or Informational passage. Each passage was originally developed with 7 SR and 3 BCR items.
- Session 4: Administered 1 operational Literary passage. The best 4 SR and 2 BCR items were selected from 7 SR and 3 BCR items which were field-tested in Session 3 in previous years.
- Session 5: Administered 1 operational Informational passage. The best 4 SR and 2 BCR items were selected from 7 SR and 3 BCR items which were field-tested in Session 3 in previous years.

General Overview of 2003 through 2007 Linking and Equating Design

- 25 SAT10 SR Items: 25 SAT10 SR items were exclusively used for the purpose of both form-to-form and year-to-year linking and equating.
- Mini-Version of the Whole Test: The 25 SAT10 SR common items met the requirement that a possible linking pool should be a mini version of the whole test: 15 contributed to GR, 5 to Literary, and 5 to Informational.
- Few Context Effects on SAT10 items: Every year the SAT10 common linking items appeared in Sessions 1 and 2 without any changes. Consequently, there was little opportunity for context effects (such as item position, intact reading passages) to be introduced into common item performance from year-to-year.
- Field-Testing Session: Session 3 was assigned to field-test Literary or Informational passages. Each passage included 7 SR and 3 BCR items. All field test items were calibrated together with operational items during field test analysis to put them on the same scale as the operational items, although only a subset of the items field-tested with a passage would subsequently make it into an operational form.
- Uniqueness of Sessions 4 and 5: To enhance test security, operational form A had a different set of Literary and Informational passages than operational form B in Sessions 4 and 5. Each passage was originally field-tested with 7 SR and 3 BCR, but only 4 SR and 2 BCR items were included with operational passages. In addition, the same amount of time was given to students regardless of whether the items were in field-testing or operational sessions. It should be noted that the second day started with Sessions 4 and 5.
- Maryland-Specific Item Parameters: Item parameters of the 25 SAT10 common items were obtained from either the 2003 (Grades 3, 5, and 8) or 2004 (Grades 4, 6, and 7) calibration based on Maryland population. In addition, these item parameters were used to link any reading assessment back to the base year (i.e., 2003 or 2004). For Grade 5, for example, Rasch item difficulties of the 25 SAT10 common items were

generated based on Maryland population during the 2003 calibration and have been used exclusively through the 2007 calibrations. Please refer to Figure 1.7 for the general overview of 2003 through 2007 linking and equating.

- Mean = 400 and SD = 40 of Population: In 2003 (Grades 3, 5, and 8) or 2004 (Grades 4, 6, and 7), item parameters of SAT10 common linking items and equating constants were generated to center 2003 or 2004 populations with *Mean* = 400 and *SD* = 40.

One Operational Form for the 2008 MSA-Reading

- 2006 Field-Tested Session 1 Items: MSDE decided to replace SAT10 SR items of Session 1 with items field-tested in 2006. For Grade 5, for example, 9 SAT10 GR items were replaced with 9 items field-tested in 2006. These items were multiple-meaning words or words in context and were called *stand-alone* items because these items were not based on passages. Consequently, these items were able to be embedded and field-tested with SAT10 items in 2006.
- 2008 Session 4 Items: To replace the other SAT10 items (i.e., the 16 SAT10 SR items that appeared in Session 2), Pearson and MSDE content specialists developed 16 items (6 GR, 5 LT and 5 Informational) plus 4 extra items (as an overage) using 2 LT and 2 Informational passages. Each passage was developed with 5 SR items and some of the items for each type of passage were GR items even if the passage was LT or Informational. All of these passages appeared in Session 4 as shown in Table 1,36.
- Procedures for Session 4 Items: The procedures for selecting the best items that replaced the SAT10 items were as follows: 1) In April 2008, Pearson analyzed Session 4 SR items and submitted both classical and IRT-based statistical results to MSDE; 2) MSDE chose the best 16 SR items.
- 2006 Literary Passages: One 2006 field-tested Literary passage (originally developed with 7 SR and 3 BCR items) was chosen as the operational passage (with 4 SR and 2 BCR items). This operational passage was assigned to Session 2 in 2008.
- 2006 Informational Passages: One 2006 field-tested Informational passage (originally developed with 7 SR and 3 BCR items) was chosen as the operational passage (with 4 SR and 2 BCR items). This operational passage was assigned to Session 3 in 2008.

New Session Design for the 2008 MSA-Reading

- Session 1: This session included operational GR items that were originally field-tested in 2006. These items were multiple-meaning words or words in context. For Grade 5, for example, 9 GR items were administered in Session 1. It should be noted that 2 new items were embedded as field test items in this session. These items will be used if some of the 9 items need to be refreshed in the future. Please refer to Table 1.36 for the 2008 MSA-Reading session information.
- Session 2: This session included one operational Literary passage (with 4 SR and 2 BCR items). This passage was originally developed and field-tested (in Session 3) with 7 SR and 3 BCR items in 2006. When administered operationally, 4 SR and 2 BCR items were selected.
- Sessions 3: This session included one operational Informational passage (with 4 SR and 2 BCR items). This passage was originally developed and field-tested (in Session 3)

with 7 SR and 3 BCR items in 2006. When administered operationally, 4 SR and 2 BCR items were selected

- Session 4: This session included 2 Literary passages and 2 Informational passages to replace the SAT10 SR items. Each passage included 5 SR items; some of these items were GR items. When statistics from the operational administration became available, the best 4 items were chosen from these 5 items. For Grade 5, for example, 2 Literary and 2 Informational passages were developed with 20 SR items (5 items for each passage). However, only 16 out of these 20 items (6 GR, 5 Literary, and 5 Informational items) were selected to replace SAT10 items for operational scoring.
- Slot in the middle of Session 4: This slot was assigned to field-test one of 4 field-testing passages (2 Literary and 2 Informational Reading passages). These passages will appear in Session 4 of the 2009 administration with a subset of those items originally field-tested.
- Session 5: This session was assigned to field-test one of 10 passages (5 Literary and 5 Informational Reading). Each passage was developed with 7 SR and 3 BCR items.

General Overview of 2008 Linking and Equating

- Year-to-Year Linking: Only SR items appearing in Sessions 1, 2, and 3 which appeared in both 2006 and 2008 were considered for the purpose of year-to-year linking.
- Item Position of Linking Common Items: Session 1 linking items were embedded and field-tested with SAT10 items in Session 1 in 2006. Session 2 (Literary) and Session 3 (Informational) SR linking items were field-tested in Session 3 in 2006.
- Selection of Linking Common Items: Common items belonging to Literary (Session 2) and Informational (Session 3) passages were originally developed and field-tested with 7 SR and 3 BCR items in 2006 and appeared with 4 SR and 2 BCR items in 2008.

Table 1.36 An Example of the 2008 MSA-Reading Session Table: Grade 5

Day 1		Day 2	
Session 1: General Reading (Stand-Alone) Items (15 min.)	1) 9 operational GR items 2) Questions such as multiple-meaning words or words in context 3) These 9 items were embedded and field-tested with SAT10 items in 2006 Session 1. 4) 2 new field test items were embedded.	**Session 4**: Two Operational Literary and Two Operational Informational Passages Plus One of Four Field-Testing Passages (2 LT and 2 Informational Passages) **(47 min.)**	1) 2 operational Literary passages: Some of the items contributed to LT and others to GR. 2) Each OP LT passage includes 5 SR items. 3) One slot between 2 OP LT and 2 OP Informational passages was assigned to field-test one of 4 passages (2 LT and 2 Informational passages). These passages will be used as 2009 Session 4 operational passage. 4) Each field-testing passage includes 6 SR items. 5) 2 operational Informational passages: Some of the items contributed to Informational and others to GR. 6) Each OP Informational passage includes 5 SR items.
Session 2: One Operational Literary Passage (35 min.)	1) 1 operational Literary passage 2) This operational passage includes 4 SR and 2 BCR items. 3) Original passage was field-tested in Session 3 of the first day. 4) Original passage was developed with 7 SR and 3 BCR items in 2006.	**Session 5**: One of Ten Field-Testing Passage (5 LT and 5 Informational Passages) (35 min.)	1) This session was assigned to field-test one of 10 passages (5 Literary and 5 Informational passages) 2) Each passage was developed with 7 SR and 3 BCR items.
Session 3: One Operational Informational Passage (35 min.)	1) 1 operational Informational passage 2) This operational passage includes 4 SR and 2 BCR items. 3) Original passage was field-tested in Session 3 of the first day. 4) Original passage was developed with 7 SR and 3 BCR items in 2006.		

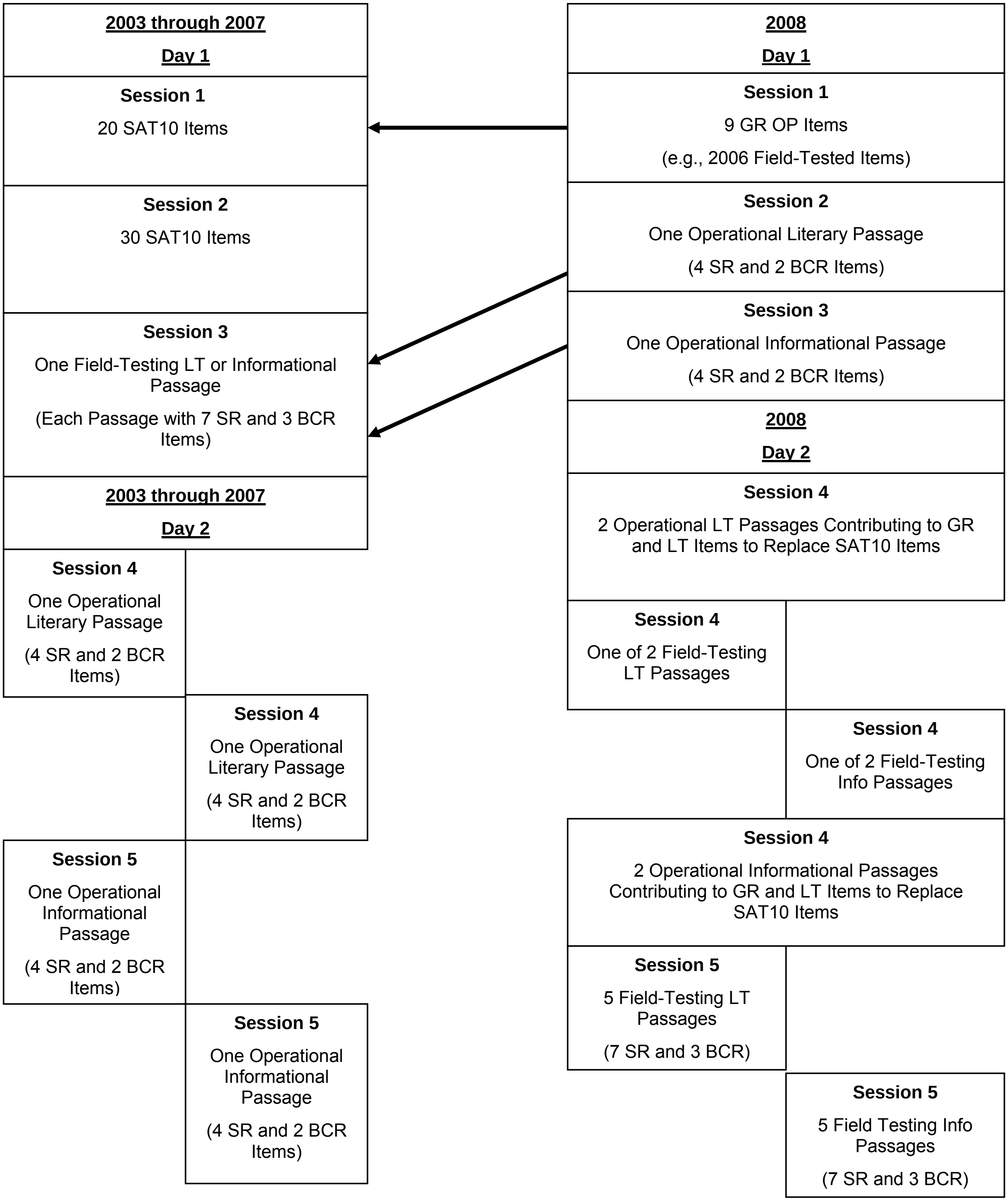
2003 through 2007
Day 1
Session 1
20 SAT10 Items
Session 2
30 SAT10 Items
Session 3
One Field-Testing LT or Informational Passage
(Each Passage with 7 SR and 3 BCR Items)
2003 through 2007
Day 2
Session 4
One Operational Literary Passage
(4 SR and 2 BCR Items)
Session 4
One Operational Literary Passage
(4 SR and 2 BCR Items)
Session 5
One Operational Informational Passage
(4 SR and 2 BCR Items)
Session 5
One Operational Informational Passage
(4 SR and 2 BCR Items)
2008
Day 1
Session 1
9 GR OP Items
(e.g., 2006 Field-Tested Items)
Session 2
One Operational Literary Passage
(4 SR and 2 BCR Items)
Session 3
One Operational Informational Passage
(4 SR and 2 BCR Items)
2008
Day 2
Session 4
2 Operational LT Passages Contributing to GR and LT Items to Replace SAT10 Items
Session 4
One of 2 Field-Testing LT Passages
Session 4
One of 2 Field-Testing Info Passages
Session 4
2 Operational Informational Passages Contributing to GR and LT Items to Replace SAT10 Items
Session 5
5 Field-Testing LT Passages
(7 SR and 3 BCR)
Session 5
5 Field Testing Info Passages
(7 SR and 3 BCR)

Figure 1.7 An Example of the 2008 MSA-Reading Linking and Equating: Grade 5

1.12 Linking, Equating, and Scaling Procedures

The 2008 reading assessment was calibrated, equated, and scaled by fixing the item parameters of the 2008 operational items with those of the 2006 field test items (i.e., the Rasch item fixed method). This means that the 2006 Rasch item difficulty parameters which were put on a common scale to either the 2003 (for grades 3, 5, and 8) or the 2004 (for grades 4, 6, and 7) assessment were carried and fixed during the 2008 linking and equating process.

Stratified Random Sampling Procedures

To select equating samples, a stratified random sampling method was applied to the 2008 state examinee population. To verify that the sample was representative of the statewide examinee population in terms of school district, gender, and ethnicity, the distributions of LEA, gender, and ethnicity of the 2008 sample were compared with those of the 2008 population. Appendix A, *The 2008 MSA-Reading Stratified Random Sampling* provides the results of the 2008 sampling. These results indicated that the equating samples were well representative of the statewide examinee population in terms of LEA, gender, and ethnicity.

Robust Z Procedures

Robust z values were calculated using the following calculations (South Carolina Department of Education, 2001):

- The mean and standard deviation of the linking pool's item difficulties for each operational form
- The ratio of the standard deviations between operational form A and form F
- The correlation between operational form A and F item difficulties
- The difference between operational form A and F for each item in the linking pool
- The mean of the differences calculated above
- The median of the differences calculated above
- The interquartile range of the differences calculated above
- 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 Selecting Year-to-Year Linking Items

Once the above calculations were made, the following guidelines were followed in determining year-to-year common items used for Rasch linking and equating (SCDE, 2001):

- Try not to include items with an absolute value of robust z exceeding 1.645.
- Should not eliminate more than 20 percent of the linking pool items.
- Try to maintain that the ratio of the standard deviations between two operational forms is in the 90 to 110 percent range.
- Try to maintain the correlation between two operational forms is greater than .95.

Year-to-Year Linking Procedures

The 2008 operational form included a set of year-to-year linking common items that appeared on both 2006 and 2008 test forms. First of all, it should be noted that while the 2006 Rasch item difficulties were generated with a field test sample, the 2008 Rasch item difficulties were generated using the 2008 live, operational data. Second, we utilized the Rasch item fixed equating method for all of the operational items to be put on a common scale within each grade.

The stability of the linking common items was evaluated using robust z values, correlation coefficients, and standard deviation ratios.

Tables 1.37 through 1.42 include Rasch item difficulties used for calculating robust z values, correlation coefficients, and standard deviations. Figures 1.8 through 1.13 depict item difficulty plots between the 2006 and 2008 assessments. It should be noted that the item difficulties of the 2008 operational form were obtained from independent calibration, and those of the 2006 assessment were put on a common scale (i.e., linked back to the 2003 or the 2004 assessment).

Table 1.37 Rasch Item Difficulties and Robust Z Values for 2006 vs. 2008: Grade 3

Item Number	Item Type	Year 2006	Year 2008	Robust Z Value
1	SR	-1.3708	-1.8968	-1.2278
14	**SR**	**.1027**	**-0.4693**	**-1.3558**
16	SR	1.3168	1.3588	.3529
17	SR	1.1695	1.4781	1.0948
19	SR	.4932	0.4367	.0788
20	SR	-.5386	-0.9822	-.9985
22	SR	-.3016	-0.3864	.0000
23	SR	.2707	0.3884	.5635
25	SR	.4743	0.3061	-.2321

One SR item (Item 14) was dropped from the 2008 linking pool based on correlation coefficient, SD ratio, robust z values, and item difficulty plot.

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

With Year 2006	Year 2008
Correlation Coefficient	.992
Standard Deviation Ratio	127%

Form Statistics

Form Statistics	Year 2006	Year 2008
Mean	.180	.026
SD	.838	1.084

Correlation and Standard Deviation Ratio

With Year 2006	Year 2008
Correlation	.982
SD ratio	129%

Values Used for Robust Z Statistics

With Year 2006	Year 2008
Mean Diff	-.154
Median Diff	-.085
IQR Diff	.486

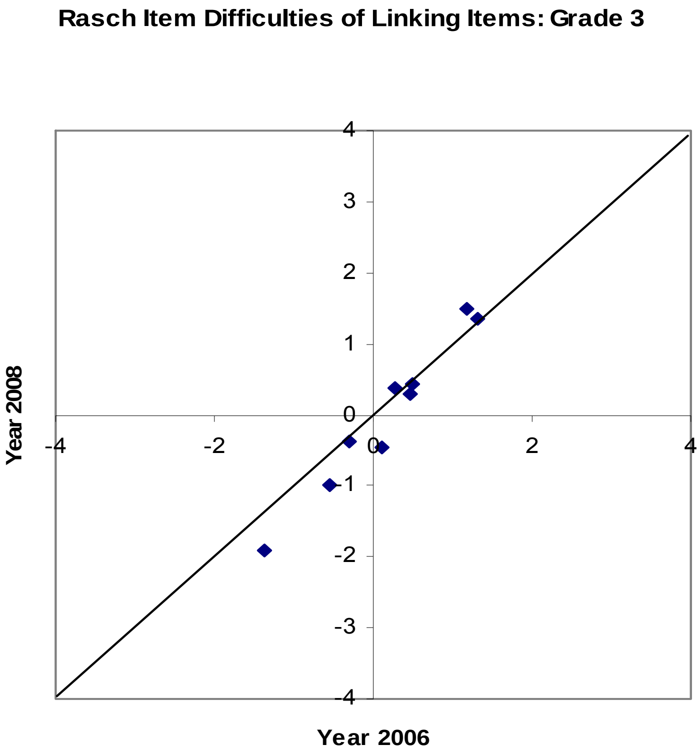

Figure 1.8 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 3

Table 1.38 Rasch Item Difficulties and Robust Z Values for 2006 vs. 2008: Grade 4

Item Number	Item Type	Year 2006	Year 2008	Robust Z Value
2	SR	-0.2773	-0.6769	.8047
6	SR	-1.3091	-2.7012	-1.4350
7	**SR**	**-1.4119**	**-3.0542**	**-1.9996**
13	SR	0.7080	0.3306	.8548
15	SR	0.2293	-0.1458	.8600
16	SR	0.7692	0.013	.0000
18	SR	0.8319	0.0729	-.0063
19	SR	1.1764	0.3449	-.1699
21	SR	1.7387	1.0164	.0765
22	SR	-0.6023	-0.9809	.8521
24	**SR**	**0.5815**	**-0.5629**	**-.8760**

Two SR items (Items 7 and 24) were dropped from the 2008 linking pool based on correlation coefficient, SD ratio, robust z values, and item difficulty plot.

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

With Year 2006	Year 2008
Correlation Coefficient	.952
Standard Deviation Ratio	113%

Form Statistics

Form Statistics	Year 2006	Year 2008
Mean	.221	-.577
SD	1.009	1.265

Correlation and Standard Deviation Ratio

With Year 2006	Year 2008
Correlation	.952
SD Ratio	125%

Values Used for Robust Z Statistics

With Year 2006	Year 2008
Mean Diff	-.798
Median Diff	-.756
IQR Diff	.599

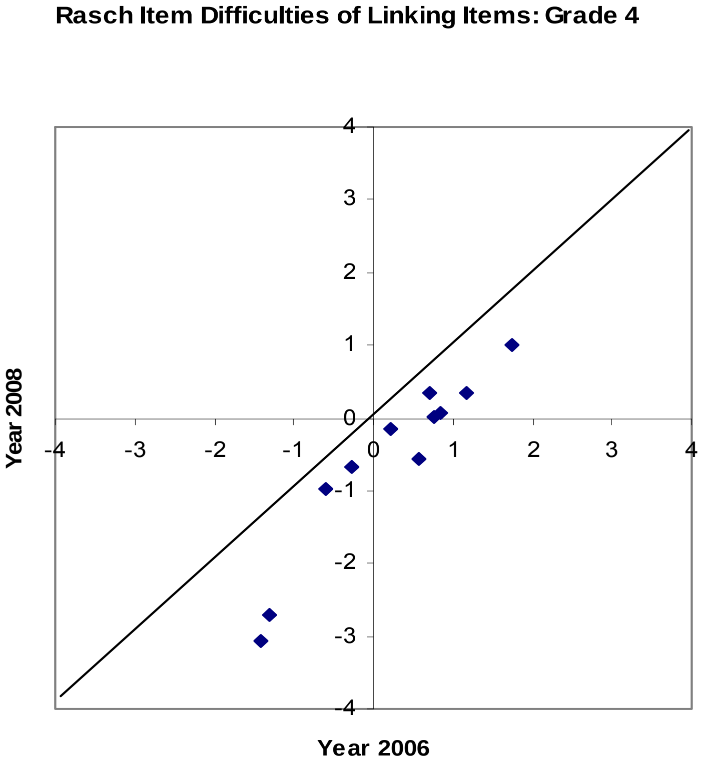

Figure 1.9 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 4

Table 1.39 Rasch Item Difficulties and Robust Z Values for 2006 vs. 2008: Grade 5

Item Number	Item Type	Year 2006	Year 2008	Robust Z Value
1	SR	-0.637	-0.891	.0817
2	SR	-0.209	-0.564	-.4714
3	SR	-1.226	-1.588	-.5123
4	SR	-1.483	-1.923	-.9445
6	SR	-1.321	-1.59	.0000
7	SR	-1.871	-2.206	-.3671
8	SR	-1.012	-1.341	-.3307
12	**SR**	**-0.021**	**0.0957**	**2.1264**
14	SR	0.6569	0.6015	1.1786
15	SR	-0.094	-0.422	-.3229
17	**SR**	**-0.094**	**0.1435**	**2.7943**
18	SR	-0.093	-0.282	.4427
20	SR	0.8069	0.6622	.6856
21	SR	0.0189	0.0141	1.4579
23	SR	0.2838	-0.091	-.5846

Two SR items (Items 12 and 17) were dropped from the 2008 linking pool based on correlation coefficient, SD ratio, robust z values, and item difficulty plot.

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

With Year 2006	Year 2008
Correlation Coefficient	.994
Standard Deviation Ratio	110%

Form Statistics

Form Statistics	Year 2006	Year 2008
Mean	-.420	-.625
SD	.800	.916

Correlation and Standard Deviation Ratio

With Year 2008	Year 2008
Correlation	.982
SD ratio	114%

Values Used for Robust Z Statistics

With Year 2006	Year 2008
Mean Diff	-.206
Median Diff	-.269
IQR Diff	.245

Rasch Item Difficulties of Linking Items: Grade 5

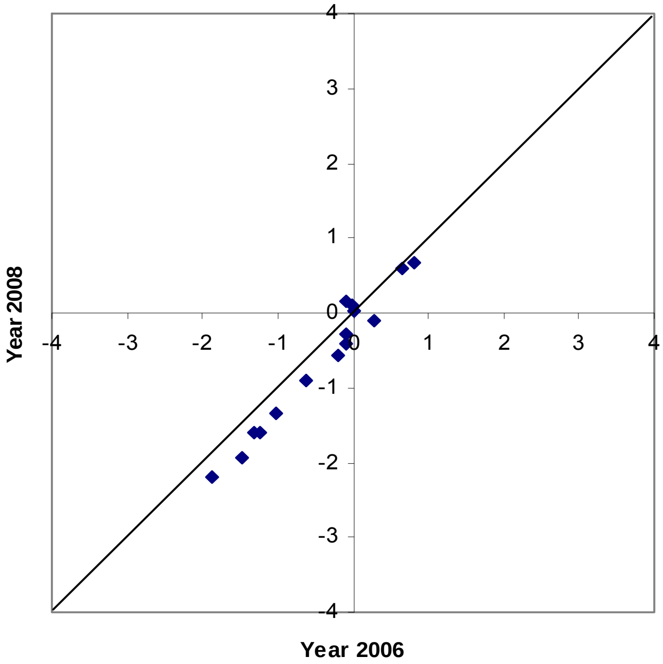

Figure 1.10 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 5

Table 1.40 Rasch Item Difficulties and Robust Z Values for 2006 vs. 2008: Grade 6

Item Number	Item Type	Year 2006	Year 2008	Robust Z Value
1	SR	-1.3336	-1.6505	-.861
2	**SR**	**-2.0006**	**-2.8083**	**-3.133**
4	SR	-1.1479	-1.2788	.000
7	SR	-1.4246	-1.6603	-.485
8	SR	1.0944	1.1986	1.088
9	SR	-.4850	-0.4621	.712
10	SR	-1.5147	-1.8875	-1.120
12	SR	.7875	0.7958	.644
14	SR	.9960	0.6738	-.886
15	SR	-.2046	-0.025	1.437
17	SR	-.2090	-0.3122	.128
18	SR	.0837	0.0224	.322
20	SR	-.5605	-0.7016	-.047
21	SR	2.2706	2.3416	.935
23	SR	.8285	0.6361	-.285

One SR item (Item 2) was dropped from the liking pool based on correlation coefficient, SD ratio, robust z values, and item difficulty plot.

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

With Year 2006	Year 2008
Correlation Coefficient	.993
Standard Deviation Ratio	108%

Form Statistics

Form Statistics	Year 2006	Year 2008
Mean	-.188	-.341
SD	1.201	1.364

Correlation and Standard Deviation Ratio

With Year 2006	Year 2008
Correlation	.990
SD Ratio	114%

Values Used for Robust Z Statistics

With Year 2006	Year 2008
Mean Diff	-.153
Median Diff	-.131
IQR Diff	.292

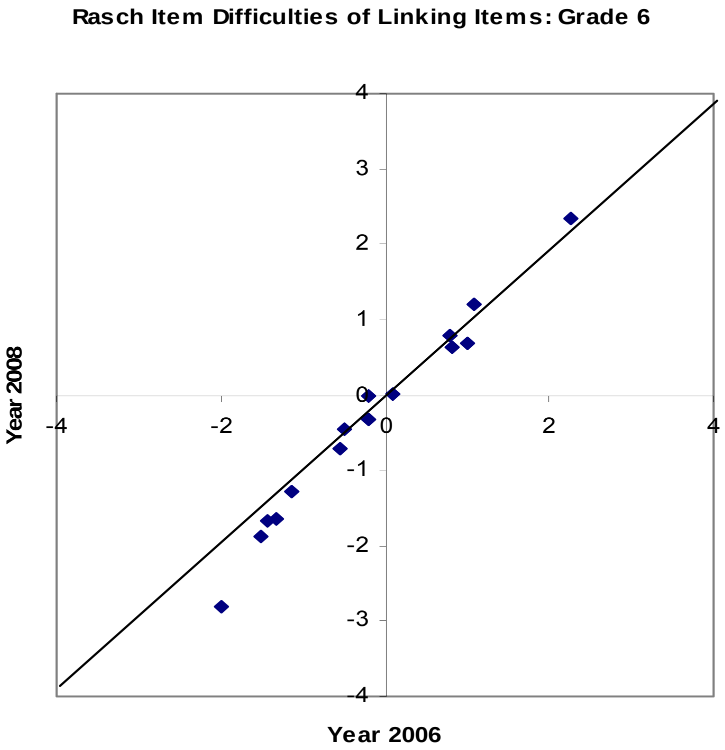

Figure 1.11 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 6

Table 1.41 Rasch Item Difficulties and Robust Z Values for 2006 vs. 2008: Grade 7

Item Number	Item Type	Year 2006	Year 2008	Robust Z Value
1	SR	-1.9151	-2.1521	-.391
2	SR	-1.5468	-1.9957	-1.729
3	SR	-.5743	-0.4446	1.923
5	SR	-1.1399	-1.6623	-2.193
6	SR	-1.8025	-2.1441	-1.052
8	SR	.0008	-0.1328	.261
10	SR	-.9289	-1.1039	.000
11	SR	.7491	0.6216	.300
13	SR	-.0919	-0.3838	-.738
14	**SR**	**.4417**	**0.708**	**2.785**
16	SR	.3333	0.2715	.714
17	**SR**	**-1.7057**	**-2.2645**	**-2.422**
19	SR	-.6797	-0.8117	.271

Two SR items (e.g., Items 14 and 17) were dropped from the liking pool based on correlation coefficient, SD ratio, robust z values, and item difficulty plot.

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

With Year 2006	Year 2008
Correlation Coefficient	.987
Standard Deviation Ratio	112%

Form Statistics

Form Statistics	Year 2006	Year 2008
Mean	-.682	-.884
SD	.913	1.086

Correlation and Standard Deviation Ratio

With Year 2006	Year 2008
Correlation	.986
SD Ratio	119%

Values Used for Robust Z Statistics

With Year 2006	Year 2008
Mean Diff	-.203
Median Diff	-.175
IQR Diff	.214

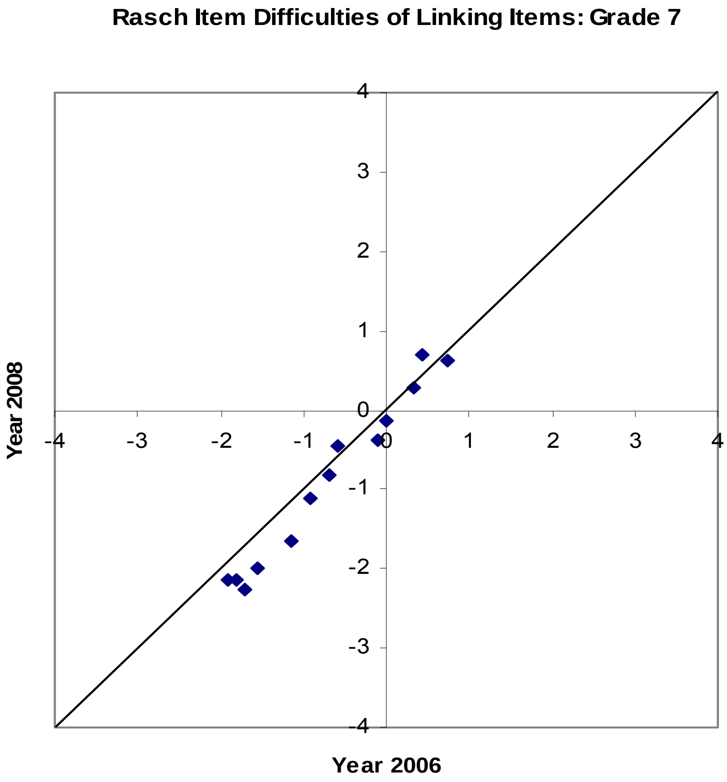

Figure 1.12 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 7

Table 1.42 Rasch Item Difficulties and Robust Z Values for 2006 vs. 2008: Grade 8

Item Number	Item Type	Year 2006	Year 2008	Robust Z Value
1	**SR**	**-1.7533**	**-2.3471**	**-2.0178**
2	SR	-1.6274	-1.9423	-.7925
4	SR	-.6076	-0.7495	-.0325
5	SR	-.6192	-0.7537	.0000
6	SR	-1.3966	-1.4625	.3014
8	SR	.0177	-0.251	-.5896
10	SR	-.0768	-0.3545	-.6291
11	SR	-.3084	-0.5622	-.5241
13	SR	-.2440	-0.1674	.9274
14	SR	.9184	0.9573	.7618
16	SR	.8321	0.8786	.7952
17	**SR**	**.1233**	**0.3392**	**1.5394**
19	SR	.7141	0.6512	.3146

Two SR items (Items 1 and 17) were dropped from the 2008 linking pool based on correlation coefficient, SD ratio, robust z values, and item difficulty plot.

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

With Year 2006	Year 2008
Correlation Coefficient	.991
Standard Deviation ratio	109%

Form Statistics

Form Statistics	Year 2006	Year 2008
Mean	-.310	-.443
SD	.885	1.031

Correlation and Standard Deviation Ratio

With Year 2006	Year 2008
Correlation	.987
SD Ratio	116%

Values Used for Robust Z Statistics

With Year 2006	Year 2008
Mean Diff	-.134
Median Diff	-.135
IQR Diff	.308

Rasch Item Difficulties of Linking Items: Grade 8

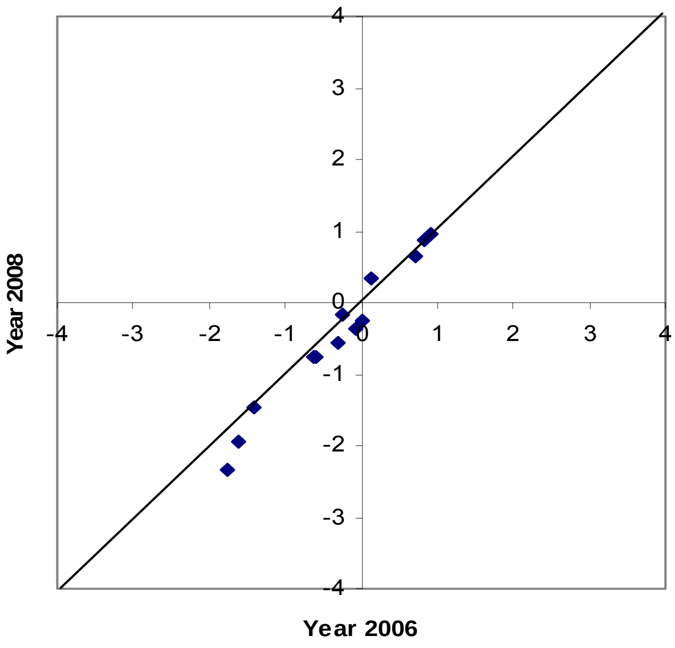

Figure 1.13 Item Difficulty Plot of Base Year Form vs. Current Year Form: Grade 8

Reporting Scale Scores

In order to facilitate the use and interpretation of the results of the 2008 MSA-Reading, the following formula was used to convert each student's ability or theta to the reporting scale score:

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

$$ReportingSE = 32.8271 \cdot SE$$

where

theta = the Rasch (i.e., 1-PL *IRT*) ability estimate, and

SE = the conditional standard error of the ability estimate.

The following table contains information about the slopes and intercepts used to generate the 2008 scale scores. It should be noted that these same slopes and intercepts have been used since the 2003 assessment (for grades 3, 5, and 8) or the 2004 assessment (for grades 4, 6, and 7).

Table 1.43 The 2008 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

1.13 Score Interpretation

To help provide appropriate interpretation of the 2008 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 1.12, *Linking, Equating, and Scaling Procedures*, the 2008 MSA-Reading produced scale scores that ranged between 240 and 650. These scale scores have the same meaning within the same grade, but those scores are not comparable across grade levels.

It should be noted that for scale scores, a higher score simply means a higher performance on 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 level descriptors provide specific information about students' performance levels and help interpret the 2008 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.

Maryland standards are divided into three levels of achievement (*www.marylandpublicshools.org*):

- Advanced is a highly challenging and exemplary level of achievement indicating outstanding accomplishment in meeting the needs of students.
- Proficient is a realistic and rigorous level of achievement indicating proficiency in meeting the needs of students.
- Basic is a level of achievement indicating that more work is needed to attain proficiency in meeting the needs of students.

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*. Students in this level can read grade-appropriate text and demonstrate the ability to comprehend literature and informational passages. Further information about the 2008 MSA-Reading score interpretation can be obtained from the MSDE.

1.14 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 ongoing 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 2008 MSA-Reading, content-related evidence, item development procedures, DIF analysis on gender and ethnicity, and evidence from internal structure 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 2008 MSA-Reading blueprints provide extensive evidence regarding the alignment between the content in the 2008 MSA-Reading and the *VSC*. It should be noted that the 2008 MSA-Reading operational test forms were built exclusively using a Maryland item bank program which contained both content and statistical information about both operational and field-tested items. Detailed information about the item composition of the operational test forms can be obtained from section 1.3, *Test Form Design, Specifications, Item Type* and session 1.11, *Constructing the 2008 Operational Test Form*. In addition, the 2008 MSA-Reading blueprints are presented in Appendix D

Item Development

Test development for MSA-Reading is ongoing and continuous. Content specialists, teachers from across Maryland, Pearson, 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 Pearson 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 1.3, *Development and Review of the 2008 MSA-Reading*.

Field test items were embedded and administered in one of ten test forms. Once these items were scored, MSDE and Pearson conducted additional item analysis and content review. Any field test items that exhibited statistical results that suggested potential problems were carefully reviewed by both MSDE and Pearson content specialists. A determination was then made as to whether an item should be eliminated, revised, or field-tested again. Information on statistical analyses for field test items can be obtained in section 1.10, *Field Test Analyses*.

Differential Item Functioning **(DIF)**

1) Bias Review of Items

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

2) *DIF* Statistics

For DIF analyses, subgroups were first categorized according to either reference or focal groups. For the 2008 MSA-Reading, males and whites were assigned to the reference group and females and African-Americans were assigned to the focal group.

While the Mantel-Haenszel procedure was used for SR items, the standardized mean difference (SMD) and the standard deviation (SD), along with the Mantel statistic, were calculated for BCR items. All of the items were classified based on Educational Testing Service (ETS) guidelines. It should be noted that DIF analyses on the operational items indicated that all the items were satisfactory. All the DIF results were archived in the 2008 Maryland item bank. More information on *DIF* analyses can be obtained in section 3.7, *Differential Item Functioning*.

Evidence from Internal Structure

The 2008 MSA-Reading contains three reading processes: *General Reading*, *Literary Reading*, and *Informational Reading*. Tables 4.3 through 4.8 show correlations among the reading processes.

1.15 Unidimensionality Analyses

Measurement implies order and magnitude along a single dimension (Andrich, 1989). Consequently, in the case of scholastic achievement, a one-dimensional scale is required to reflect this idea of measurement (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 2008 MSA-Reading, we examined the relative sizes of the eigenvalues associated with a principal component analysis of the item set. First, polychoric correlation coefficients were computed with *LISREL 8.5* (Jöreskog & Sörbom, 1993) because of the polytomously scored reading items. Principal component analysis was then applied to produce eigenvalues. The first and the second principal component eigenvalues were compared *without rotation*. Table 1.44 summarizes the results of the first and second principal component eigenvalues of the 2008 MSA-Reading.

A general rule of thumb in exploratory factor analysis suggests that a set of items may represent as many factors as there are eigenvalues greater than 1 in this analysis because there is one unit of information per item and the eigenvalues sum to the total number of items. However, a set of items may have multiple eigenvalues greater than 1 and still be sufficiently unidimensional for analysis with IRT (Loehlin, 1987; Orlando, 2004). As seen from the following table, the first component extracted a substantially larger eigenvalues across all grades: the size of the eigenvalue of the first component was over ten times that of the second eigenvalue for each form at each grade. As a result, we could conclude that the assumption of unidimensionality for the 2008 MSA-Reading was met.

Table 1.44 The 2008 MSA-Reading Eigenvalues between the First and Second Components

Grade	Number of Items	First Eigenvalue	Second Eigenvalue
3	37	11.02	1.44
4	37	11.14	1.57
5	37	11.79	1.55
6	37	11.78	1.48
7	37	11.94	1.45
8	37	11.75	1.41

1.16 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 2008 MSA-Reading item bank continued to be maintained by Pearson in the form of computer files and paper copies. This enabled test items to be readily available to both Pearson and MSDE staff for reference, test construction, test book design, and printing.

Pearson maintained a computerized statistical item bank to store supporting and identification information for 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

It should be noted that each field test item of each form was calibrated by fixing each operational item with its operational Rasch items parameter (i.e., Rasch item fixed equating method). Item difficulties, step difficulties, and infit and outfit fit statistics of all the field test items were stored in the 2008 item bank.

1.17 Quality Control Procedures

A standard quality procedure at Pearson 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:

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

- Verified data descriptions for accuracy against data file
- Created 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 could not 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 Pearson.

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 Pearson staff experts prior to release of the information to MSDE.

Upon completion of the processing of all district-level data, Pearson Scoring Operations provided the Quality Assurance Department with one or more state-level data files, along with state data for review and approval. Pearson Quality Assurance programmers duplicated all data independently to ensure accurate interpretation of the expected results. A series of SAS programs were run on these files to ensure 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 were 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.