1. OVERVIEW OF THE 2003 *MARYLAND SCHOOL ASSESSMENT-READING*

In 2002, the Maryland State Department of Education (MSDE) took an important step toward raising learning expectations for all students in public schools. The State Board of Education retired the *Maryland School Performance Assessment Program (MSPAP)* and adopted a new testing program known as the *Maryland School Assessment* (*MSA*). The 2003 *MSA* was based on the *Voluntary State Curriculum*, which set reasonable academic standards for what teachers were expected to teach and for what students were expected to learn in schools.

Beginning in March 2003, students in grades 3, 5, and 8 took the *MSA* in reading (*MSA-Reading*) and mathematics. Students in grade 10 took only the *MSA-Reading*, because high school mathematics achievement was measured by the *Maryland High School Assessment* in geometry. In addition, tests in reading and mathematics will be phased in for students in grades 4, 6, and 7 in February 2004.

1.1 Overview of the 2003 *MSA-Reading*

As can be seen from Table 1.1, the 2003 *MSA-Reading* field tests were designed to provide two kinds of information. First, *norm-referenced* information was provided by the items from the abbreviated form of the *Stanford Achievement Test Series, Tenth Edition (SAT10)*. The *SAT10* consisted of *Word Study*, *Reading Vocabulary*, and *Reading Comprehension* items. To produce *criterion-referenced* information, additional items, called augmented items, were written for the *Maryland Reading Standards* (*MRS*) in grades 3, 5, and 8 and organized under three reading processes: *General Reading*, *Literary Reading*, and *Informational Reading*.

The 2003 *MSA-Reading* produced both norm-referenced and criterion-referenced scores for each student. While norm-referenced scores included only the *SAT10* items, both items selected from the *SAT10* and augmented items created for Maryland comprised criterion-referenced scores. Figure 1.1 shows a schematic of the *SAT10* and augmented items that produced these test scores.

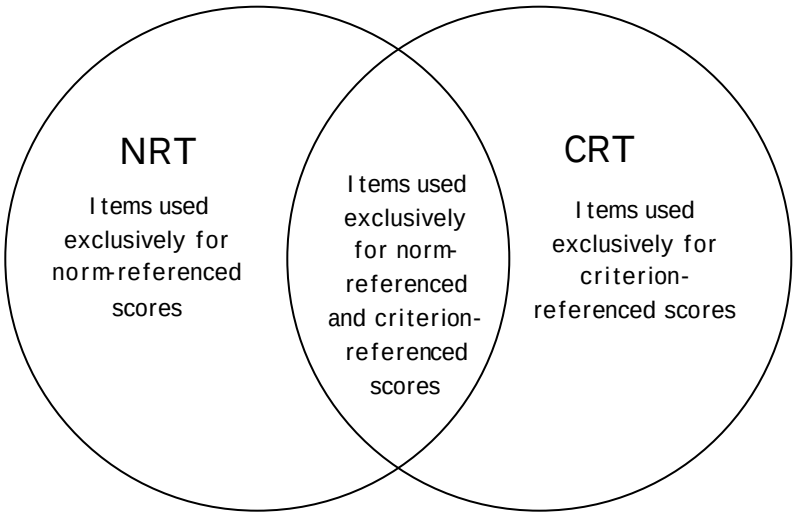

Figure 1.1 Schematic of the 2003 *MSA-Reading*

Table 1.1 The 2003 *MSA-Reading* Field Test Design: Grades 3, 5, and 8

Grade	Strand Title	*SAT10* / Augmented	Item Type	Number of Items	Total Points
3	Total NRT	*SAT10*	*SR*	70	70
	Word Study	*SAT10*	*SR*	20	20
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT*	*SAT10*, Augmented	*SR*, *BCR*	45	57
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	15	21
	Information Reading	*SAT10,* Augmented	*SR*, *BCR*	15	21
5	Total NRT	*SAT10*	*SR*	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT*	*SAT10*, Augmented	*SR*, *BCR*	45	57
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	15	21
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	15	21
8	Total NRT	*SAT10*	*SR*	50	50
	Reading Vocabulary	*SAT10*	*SR*	20	20
	Reading Comprehension	*SAT10*	*SR*	30	30
	Total CRT*	*SAT10*, Augmented	*SR*, *BCR*	45	57
	General Reading	*SAT10*	*SR*	15	15
	Literary Reading	*SAT10,* Augmented	*SR*, *BCR*	15	21
	Informational Reading	*SAT10,* Augmented	*SR*, *BCR*	15	21

*Note: 1. CRT contains *SAT10* items
2. *SR* items are selected response items, and *BCR* items are brief constructed response items

1.2 Purposes/Uses of the 2003 *MSA-Reading*

By measuring students' achievement against the new academic standards, the 2003 *MSA-Reading* provides two main purposes. First, the *MSA-Reading* was designed to inform parents, teachers, and educators of what students actually learned in schools by providing specific feedback that can be used to improve the quality of schools, classrooms, and individualized instructional programs and to model effective assessment approaches that can be used in classrooms. Second, the *MSA-Reading* serves as an accountability tool to measure performance levels of individual students, schools, and districts against the new academic standards.

1.3 The *Voluntary State Curriculum*

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

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

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

1.4 Development and Review of the 2003 *MSA-Reading*

Developing the 2003 *MSA-Reading* was a complex process. It required a great deal of involvement from the MSDE, Harcourt, and local school systems. In addition, teachers, administrators, and content specialists from all over Maryland were recruited for different test development committees. These individuals reviewed test forms and items to ensure that they measured students' knowledge and skills fairly and without bias. Table 1.2 identifies which groups were responsible for developing the 2003 *MSA-Reading*.

National Psychometric Council

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

Content Review Committee

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

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

Bias Review Committee

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

Table 1.2 The 2003 *MSA-Reading* Responsibility for the Test Development

Development of the 2003 *MSA-Reading*	Primary Responsibility
Development of Preliminary Blueprints and Item Specifications	Harcourt; MSDE; NPC
Development of Preliminary Brief Constructed Response Rubrics	MSDE
Item Writing	Harcourt
Item Review	Harcourt; MSDE; NPC; Content Review Committee
Bias Review	Harcourt; MSDE; Bias Review Committee
Construction of Field Test Forms	Harcourt; MSDE
Modification of Special Forms	Harcourt; MSDE
Review of Special Forms	MSDE
Pre-Field Test Training Workshops	Harcourt; MSDE; LEAs
Field Test Administrations	MSDE; LEAs
Construction of Operational Test Forms	Harcourt; MSDE; NPC
Review of Operational Test Forms	MSDE
Final Construction of Operational Test Forms	Harcourt; MSDE
Setting Standards for the 2003 *MSA-Reading*	Standard Setting Committee for the 2003 *MSA-Reading;* CTB/McGraw-Hill

1.5 Structure of the 2003 *MSA-Reading*

Structure of Field Test Forms

The 2003 *MSA-Reading* was composed of the *SAT10*, augmented, and field test items for future augmentation. The design of the *MSA-Reading* was to spiral a relatively large number of Maryland augmented field test items into multiple test forms for each grade in the 2003 test administration. Thus, the 2003 *MSA-Reading* produced 7 test forms. For forms 1 through 6, the order of the *SAT10* items was the same across the test forms, and Maryland-specific (augmented) items were located after the *SAT10* items. Form 7 was a re-ordered version of form 1 and was produced in order to examine the effects of item re-ordering. The descriptive statistics of each field test form can be found in section 1.8.

Types of Items

The 2003 *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 2003 *MSA-Reading*, students selected an answer from four alternatives. Each *SR* item was scored as right or wrong.

BCR items required students to answer a question with a couple of words, a sentence, or a more elaborated way. For the 2003 *MSA-Reading*, these items were scored on a general rubric with maximum values between 0 and 3.

1.6 Test Administration

Test Administration Preparation and Materials

Pre-test workshops were held across the state prior to field test. These workshops provided the representatives of all the local school divisions with an overview of the tests' content, security expectations, and procedures for completing the answer documents. They also considered the receipt, distribution, and return of test materials.

For the test examiner, Harcourt provided the following materials:

- Test examiner's manual
- One set of pre-printed student ID labels and one set of generic ID labels
- Paper bands for used Answer Books
- Student roster

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

- Test Book
- Answer Book
- Two No.2 pencils with erasers (by school or student)

Two test-related manuals were developed for the administration of the 2003 *MSA-Reading*: Test Administration and Coordination Manual (TACM) and Examiner's Manual for Test Administration (EMTA). The TACM was developed and distributed by Harcourt. This manual provided Local Accountability Coordinators (LACs) and building level School Test

Coordinators (STCs) with information about the administration, packaging, and return of test materials. The TACM also described any issues specific to grades 3, 5, and 8. One TACM was produced for all administrations in grades 3, 5, and 8. The TACM was distributed one per school at the pre-test workshops and again included in the shipping materials.

The EMTA was developed for each grade by Harcourt and provided directions for administering the 2003 *MSA-Reading* at each grade level.

Test Administration Schedule

Specific dates were designated for each content area test. For the 2003 *MSA-Reading*, students were required to take Part I on Day 1 and Part II on Day 2, and the primary testing days were as follows:

- Test materials delivered to schools (including pre-print Student ID labels) — February 10-14, 2003
- Reading testing days — March 3 and 4, 2003
- Make-up testing days — March 7, and 10-14, 2003

Sessions were scheduled at any convenient time during the school day, but testing had to be scheduled to allow sufficient time to complete the test. Table 1.3 shows timing sessions allowed for the 2003 *MSA-Reading*.

Table 1. 3 The 2003 *MSA-Reading* Timing Sessions: Grades 3, 5, and 8

Grade	Form	Session				
		1	2	3	4	5
3	1-6	Q1-Q20	Q21-Q40	Q41-Q70	Q71-Q80	Q81-Q90
		12 min.	14 min.	30 min.	35 min.	35 min.
	7	Q1-Q20	Q21-Q40	Q41-Q55	Q56-Q65	Q66-Q90
		12 min.	14 min.	45 min.	20 min.	60 min.
5	1-6	Q1-Q20	Q21-Q50	Q51-Q60	Q61-Q70	
		14 min.	30 min.	35 min.	35 min.	
		Q1-Q20	Q21-Q35	Q36-Q50	Q51-Q60	Q61-Q70
		26 min.	40 min.	25 min.	35 min.	20 min.
8	1-6	Q1-Q20	Q21-Q50	Q51-Q60	Q61-Q70	
		14 min.	30 min.	35 min.	35 min.	
	7	Q1-Q20	Q21-Q40	Q41-Q50	Q51-Q60	Q61-Q70
		26 min.	50 min.	20 min.	35 min.	20 min.

If a student was absent on the testing days, a make-up test was administered on one of the subsequent days within the testing window (March 7 or March 10-14, 2003). If a school had an unscheduled closing or delayed opening that prohibited the administration from occurring on the scheduled testing dates, the STCs were consulted with LACs to determine the testing schedule to be followed. The LACs addressed all questions to the Assessment Branch in the Division of Planning, Results, and Information Management at the MSDE.

Any student who was tested during the make-up period was to continue to test in the original book. There was not a separate make-up book. Therefore, if a student unexpectedly left during a session, the remaining time was noted so that the student might finish that section with the proper amount of remaining time. If a known absence occurred during testing, splitting a session between original testing and make-up testing was avoided.

During the administration of the 2003 *MSA-Reading*, the MSDE had testing monitors in selected schools observing administration procedures and testing conditions. All monitors had identification cards for security purposes.

Testing Accommodations

Testing accommodations for Special Education students, English Language Learners (ELL), and students with disabilities covered under Section 504 had to be approved and documented according to the procedures and requirements outlined in the document entitled "Requirements for Accommodating, Excusing, and Exempting Students in Maryland Assessment Programs," as revised December 20, 2002. (A copy of the most recent edition of this document is available electronically on the LAC and STC web pages at http://docushare.msde.state.md.us)

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

Braille and Large-Print Test Books

The 2003 *MSA-Reading* was administered to those requiring large-print Test Books and Answer Books or Braille Test Books. For both large-print and Braille Test Books, students' responses were transcribed into the regular Answer Books following testing. The pre-printed student ID labels were affixed to the regular Answer Books containing the transcribed responses.

Once the grades 3, 5, and 8 reading answers had been transcribed, large-print and Braille Test Books were returned for scoring with the regular materials. The Answer Books were bounded with a paper band, and the bundle was labeled "transcribed books."

Security of Test Materials

The Test Books and all used Answer Books for the 2003 *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.

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

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

Distribution of Materials

Different test forms were administered to students in each classroom participating in reading tests, and each test form was identified by a cover of a different color and number. In addition, the Test Books and Answer Books were spiraled within a classroom.

1.7 Scoring Procedures

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

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

Machine-Scored Items

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

Hand-Scored Items

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

Scorer Qualifications

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

performance assessments and to provide a final screening activity before they were added to the overall pool of potential scorers for the *MSA* project.

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

Methodology for Scoring the 2003 *MSA-Reading BCR* Items

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

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

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

Training for Scoring Accuracy

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

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

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

Inter-Rater Agreement

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

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

1.8 Item Selection for Scoring Purposes

All items of the 2003 *MSA-Reading* were subjected to rigorous analyses for their properties. These analyses provided statistical information about test items that would be included as a part of scoring (operational) test forms. The following analyses were conducted:

- Overall statistical analyses for each field test form
- Classical item analyses for *SR* and *BCR* items
- *Differential item functioning* (*DIF*)
- "Not-reached" item analyses

Descriptive Statistical Analyses for Each Field Test Form

To ascertain whether or not each field test showed statistical abnormalities, descriptive statistics, such as means, standard deviations, standard errors, and reliability were used with the common items of each test form. These analyses also provided statistical information about determining whether later calibration and equating were successful for field test forms 1 through 7. As can be seen from Table 1.4, there are no significant differences across six test forms except form 7.

A separate check on test forms 1 and 7 to determine if the test form with the mixed item order (i.e., form 7) exhibited the same properties as the form with the original item order (i.e., form 1) indicated that there were some differences between form 1 and form 7. The "omits" in form 7 dominated the discussions with respect to the possible reason for higher mean of form 7. In addition, different session times allowed for form 7 gave some answers about the different means of these two forms.

Table 1.4 The 2003 *MSA-Reading* Descriptive Statistics for Each Field Test Form (Common Items)

Grade	Form	Number of Items	N	Mean	SD	Reliability	SEM
3	1	25	8,298	15.43	5.07	0.83	2.09
	2	25	8,615	15.58	5.06	0.83	2.09
	3	25	8,562	15.59	5.08	0.83	2.09
	4	25	8,466	15.59	5.11	0.83	2.11
	5	25	8,597	15.55	5.11	0.84	2.04
	6	25	8,291	15.54	5.17	0.84	2.07
	7	25	7,619	16.11	4.83	0.82	2.05
5	1	25	8,414	16.49	5.20	0.84	2.08
	2	25	8,888	16.51	5.23	0.84	2.09
	3	25	8,874	16.56	5.20	0.84	2.08
	4	25	8,833	16.54	5.24	0.84	2.10
	5	25	8,757	16.56	5.24	0.84	2.10
	6	25	8,488	16.55	5.19	0.84	2.08
	7	25	7,947	17.71	4.60	0.81	2.01
8	1	22	8,326	15.17	4.37	0.80	1.95
	2	22	8,799	15.02	4.54	0.82	1.93
	3	22	8,824	15.17	4.44	0.81	1.94
	4	22	8,874	14.99	4.59	0.82	1.95
	5	22	8,787	15.02	4.49	0.81	1.96
	6	22	8,604	15.20	4.35	0.80	1.95
	7	22	4,013	16.13	3.86	0.76	1.89

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

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

SR items for further scrutiny were flagged if:

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

BCR items for further scrutiny were flagged if:

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

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

Differential Item Functioning Analyses

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

For the 2003 *MSA-Reading DIF* analyses, the reference group was either male or Caucasian students, and the focal group was either female or African-American students. Because the 2003 *MSA-Reading* included both the *SAT10* items and the "Maryland-specific" items on each field test form, the total item score on a collection of items was used as the matching variable.

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

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

For further information about the *DIF* procedures used for the 2003 *MSA-Reading*, please see section 3.7.

"Not-Reached" Item Analyses

An important consideration in the *item response theory* (*IRT*) analyses employed for the 2003 *MSA-Reading* was the treatment of missing responses to test items. Specifically, these procedures drew a distinction between items that were considered to be intentionally missing (*omits*) and missing items that occurred at the end of a block of items (*not-reached*). Researchers have suggested that "ignoring not-reached items introduces slight biases into item parameter estimates when not-reached items are present and speed is correlated with ability (Allen, Donoghue, & Schoeps, 2001, p. 232)."

For the 2003 *MSA-Reading*, this analysis was performed for each session within each day of the assessment. "Not-reached" items were treated as missing values for the purposes of calibration. In addition, "omit" items were scored as wrong for *SR* items and scored in the lowest category for *BCR* items. In addition, if the "not-reached" rate for an item exceeds 10% (i.e., an item completion rate of less than 90%), additional discussion was required with the MSDE to decide whether or not the item should be dropped to alleviate test "speededness".

Item Selection for Scoring Purposes and 2004 Operational Forms

The selection of items to be included in the final test forms of the 2003 *MSA-Reading* required a careful consideration based on test blueprints, classical item analyses, and *DIF* analyses, and "not-reached" item analyses. Harcourt suggested the following guidelines to choose items included in the final test forms:

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

In applying these guidelines, a balance should be made between being too harsh and thus dropping items that may affect the content representativeness of the entire set of field test items and being too lenient and allowing items with poor model fit that might affect resulting measures. In addition, reading specialists from the MSDE reviewed the final test forms of the 2003 *MSA-Reading*.

For the 2004 *MSA-Reading*, four operational test forms were constructed and reviewed by reading specialists from the MSDE. They determined the content validity and equivalency of the test forms for each grade level.

1.9 Linking, Equating, and Scaling Procedures

Linking Procedures

To link different test forms at each grade level, linking steps recommended by the National Psychometric Council were taken into consideration. For the 2003 *MSA-Reading*, items that appeared on each test form were included as potential linking items, but only *SR* items were considered as potential linking items.

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

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

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

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

Toward this end, Harcourt provided Rasch item difficulty lots and identified items that were to be deleted based on the robust Z statistics. Figure 1.2 provides the Rasch item difficulty of each item for each of the two forms and the robust Z calculated by the definition. The item difficulty plot between form 1 and form 2 indicates that there exist no extreme outliers. The correlation coefficient of the two test forms, 1.00, also indicates a very strong relation between the item parameter estimates of the two test forms.

Item Number	Form 1	Form 2	1 vs 2	Robust Z
1	-2.33	-2.31	0.02	1.577
2	-1.10	-1.19	-0.09	-.901
3	.15	.11	-0.04	.225
4	-.93	-.94	-0.01	.901
5	.93	.84	-0.09	-.901
6	-1.08	-1.07	0.01	1.351
7	.45	.41	-0.04	.225
8	2.29	2.20	-0.09	-.901
9	-.08	-.17	-0.09	-.901
10	1.00	.90	-0.10	-1.126
11	-.11	-.16	-0.05	.000
12	.24	.17	-0.07	-.450
13	-.15	-.23	-0.08	-.676
14	-.56	-.60	-0.04	.225
15	-.03	-.11	-0.08	-.676
16	-1.80	-1.80	0.00	1.126
17	.00	-.03	-0.03	.450
18	.03	-.06	-0.09	-.901
19	.93	.82	-0.11	-1.351
20	-1.06	-1.04	0.02	1.577
21	-1.43	-1.48	-0.05	.000
22	-.91	-.93	-0.02	.676
23	.65	.61	-0.04	.225
24	-.40	-.48	-0.08	-.676
25	.59	.50	-0.09	-.901

	Form 1	Form 2
Mean	-.188	-.242
SD	1.012	.985

	1 vs 1	1 vs 2
Correlation	1.000	1.000
SD ratio	100%	97%

	1 vs 1	1 vs 2
Mean of Difference	.000	-.053
Median of Difference	.000	-.050
Interquartile Range of Difference	.000	.060

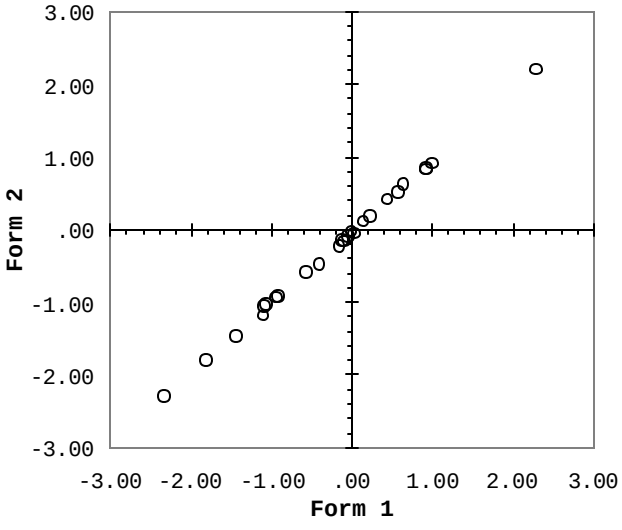

Figure 1.2 Example of Parameters Used to Link Items

Equating Procedures

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

For the 2003 *MSA-Reading*, items selected through the linking procedures were used to equate all different test forms of each grade. Because each test form included a subset of unique items, linking items served as anchor items. Thus, whenever a new test form is constructed in the future, the new form will be equal in difficulty to the previous form via linking items. The design to collect data for the 2003 *MSA-Reading* was common item, non-equivalent groups.

In order to obtain parameter estimates for both the unique items on each form and the linking items, the Rasch model (or Partial Credit Model for *BCR* items) was used. For the 2003 *MSA-Reading*, the common items whose calibrations were known were anchored or fixed to their known estimates during calibration of other forms that were to be put on the scale of the first form. In treating these common item parameters as known they were fixed, and the remaining item parameters (for the unique items of each form) were also forced onto the same scale as the anchored (fixed) items.

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

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

where

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

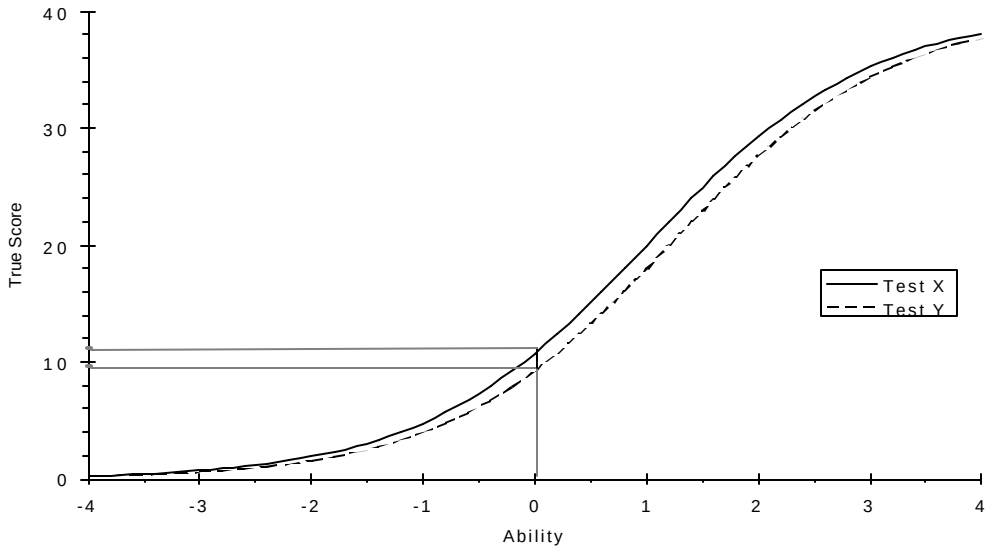

Figure 1.3 True Score Equating

Figure 1.3 illustrates these ideas for two hypothetical test forms, X and Y. In the figure, the true scores on each of the forms are plotted against ability using the true score equation. By drawing a line from the ability (here shown for an ability of 0) to each of the respective curves and moving across to the true score scale, one can find the pairs of true scores that are equated to one another. According to Lord and Wingersky (1984), the procedure applied to true scores can be transferred to observed scores without any major anomalies in the resulting outcomes.

Reporting Scale Scores

In order to facilitate the use and interpretation of the results of the 2003 *MSA-Reading*, scale scores were created based on the information given by both the MSDE and the NPC. The following is the formula to covert each student' ability or theta to scale scores:

$$ReportingAbilityScaleScore = 40 \cdot theta + 400$$

$$ReportingSEM = 40 \cdot SEM$$

where

theta = the IRT ability estimate, and

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

Note that the minimum of the scale was set to 0 and the maximum of the scale was set to 800.

1.10 Standard Setting

Crocker and Algina (1986, p. 410) pointed out, "(m)any situations require the setting of cutoff scores before test performance is interpreted. ... The practice of setting cutoff scores is commonly called *standard setting*."

For the 2003 *MSA-Reading*, the *Bookmark procedure* was used to set cutoff scores. CTB/McGraw-Hill and standard setting committees set two cutoff scores for grades 3, 5 and 8, and the following performance categories were created:

- *Advanced*
- *Proficient*
- *Basic*

Integrating *SR* and *BCR* items

During the *Bookmark procedure*, the key material presented to committee members was the ordered item booklet, in which items were ordered by their scale locations as determined by *item response theory* (*IRT*) calibrations (Mitzel, Lewis, Patz, & Green, 2001).

It has been recognized that standard setting methods traditionally work better on one or the other of the item types, but not on both. Thus, large-scale assessment programs have applied one standard setting procedure to the *SR* items and another to the *BCR* items. However, this creates additional steps of resolving different results from different methods, and potentially raises questions around the validity of final cut scores if those methods produce highly different outcomes (Mitzel, Lewis, Patz, & Green, 2001).

These days, *IRT* methodology is applied to large-scale assessment programs that include both *SR* and *BCR* item types to define a single underlying trait. From the perspective of the *Bookmark procedure* based on *IRT*, the method of setting performance standards should also reflect the unity of the underlying content if both *SR* and *BCR* item contents are calibrated to establish a single trait. Thus, scaling these two different item types together allows both item types to be placed into a single ordered item booklet and to be considered jointly by panelists (Mitzel, Lewis, Patz, & Green, 2001).

In dividing items at a cutoff score between mastery and nonmastery, a response probability (RP) of .67 (i.e., 2/3) is applied during standard setting. This implication means that for a given cutoff score, a student with a test score at that point will have a .67 probability of answering an item correctly at that cutoff score, and this is the technical definition of mastery.

SR item is mapped where the examinee has the probability of 2/3 of a correct answer. For *BCR* items with more than two categories, the *Bookmark procedure* follows the NAEP method of creating pseudo-binary items to map the non-zero categories. Consider a *BCR* item with score categories 0, 1, 2, and 3. The first pseudo-binary item is created by keeping the category 0 as zero and recoding the categories 1, 2, and 3 as one. This binary item then is used to place the score category 1 of the *BCR* item. The second pseudo-binary item is created by recoding the categories 0 and 1 as zero and the category 2 and 3 as one. This binary item is then used to map the category 2 of the *BCR* item. Finally, the third pseudo-binary item is created by recoding the categories 0, 1, and 2 as zero and the category 3 as one. This binary item is then used to map the category 3 of the CR item (Huynh, Meyer III, & Barton, 2000).

For the *2003 MSA-Reading* standard setting, a *BCR* item was placed in the booklet at three locations according to scale scores associated with attaining each additional score point. Scoring rubrics were also placed after each *BCR* item to help participants determine the skills and knowledge required to attain a given score point (Mitzel, Lewis, Patz, & Green, 2001).

Further information about the standard setting can be obtained from the MSDE or the *Maryland Standard Setting Technical Report* of CTB/McGraw-Hill (2003, August).

1.11 Score Interpretation

To help provide appropriate interpretation of the 2003 *MSA-Reading* test scores, two types of scores were created: 0-800 scale scores, and performance levels and descriptions.

0-800 Scale Scores

As explained in section 1.9, Linking, Equating, and Scaling, the 2003 *MSA-Reading* produced scale scores that ranged between 0 and 800. Those scale scores have the same meaning within the same grade, but those scores are not comparable across grade levels.

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

Performance Levels and Descriptions

As previously explained, performance levels and descriptions provide specific information about students' performance levels and help interpret the 2003 *MSA-Reading* scale scores. They describe what students at a particular level generally know and can be applicable to all students within each grade level. As Table 2.1 shows a range of scale scores at each performance level. For example, grade 3 reading scale scores from 404 to 487 indicate the level of *Proficient*, and students at this level can read grade appropriate text and demonstrate the ability to comprehend literature and informational passages. Further information about the 2003 *MSA-Reading* score interpretation can be obtained from the MSDE.

1.12 Test Validity

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

Messick (1989) defined validity as follows:

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

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

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

To investigate validity evidence of the 2003 *MSA-Reading*, content-related evidence, evidence of internal structure, and evidence of unidimensionality were collected.

Content-Related Evidence

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

The 2003 *MSA-Reading* blueprints provide extensive evidence regarding the alignment between the content of the 2003 *MSA-Reading* and the *VSC*. These blueprints are presented in Appendix C.

Evidence of the Internal Structure of the *MSA-Reading*

The 2003 *MSA-Reading* has three reading processes: *General Reading*, *Literary Reading*, and *Informational Reading*. As can be seen from Tables 4.3 through 4.5, there exist moderately strong intercorrelations among these three processes.

Evidence of Unidimensionality

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

To check the unidimensionality of the 2003 *MSA-Reading*, polychoric correlation coefficients were computed with *LISREL 8.5* (Jöreskog & Sörbom, 1993) because they were polytomously scored on reading tests. Principal component analysis was then applied to produce eigenvalues. The first and the second principal component eigenvalues were compared without rotation. Table 1.5 summarizes the results of the first and the second principal component eigenvalues of the 2003 *MSA-Reading*.

The rule of thumb to determine the unidimensionality of a test requires that the eigenvalue of the first component or factor should be at least three times larger than the second one. As can be seen, the size of the eigenvalue of the first component meets the criterion for the unidimensionality. Thus, the assumption of unidimensionality for the 2003 *MSA-Reading* was met.

Table 1.5 The 2003 *MSA-Reading* Eigenvalues between the First and the Second Components

Grade	Form	Number of Items	First Eigenvalue	Second Eigenvalue
3	1	37	11.83	1.59
	2	37	11.53	1.57
	3	37	11.74	1.67
	4	37	12.04	1.53
	5	37	12.50	1.52
	6	37	12.86	1.53
	7	37	11.61	1.65
5	1	37	11.67	1.47
	2	37	11.80	1.48
	3	37	10.98	1.48
	4	37	11.22	1.50
	5	37	11.71	1.50
	6	37	11.65	1.55
	7	37	11.19	1.36
8	1	34	10.94	1.54
	2	34	11.44	1.45
	3	34	11.24	1.59
	4	34	10.86	1.52
	5	34	11.20	1.49
	6	34	10.38	1.49
	7	34	10.22	1.64

1.13 Item Bank Construction

The number of test forms to be constructed each year and the need to replace items that would be released to the public necessitated the availability of a large pool of items. The 2003 *MSA-Reading* item bank continues to be maintained by Harcourt as computer files and paper copies. This enables test items to be readily available to both Harcourt and MSDE staff for reference, test construction, test book design, and printing.

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

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

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