2. TEST DESIGN, DEVELOPMENT, AND SCORING OF THE 2011 MOD-MSA: READING

2.1. Test Design and Structure of the 2011 Mod-MSA: Reading

For 2011, a single form in reading was created for each grade level from 3 through 8. However, the operational forms (scoring form) for all grades were selected after administration, during Data Review.

It should be noted that that the total number of items administered to the students in 2011 were 57 each for Grades 3 to 6, and 53 each for Grades 7 and 8. This compares with 69, 68, 67 and 60 for Grades 3 to 6 respectively in 2010. For Grades 7 and 8, a toal of 55 items each was administed in 2010.

The 2011 administration included a mix of operational items (i.e., items on which the students were be scored) and some field-test (FT) items. The selection of the operational items was determined through the data review process only after the test was administered. See Table 2.1.1 for the test design of the Mod-MSA Reading examination

As shown in Table 2.1.1, the Mod-MSA items for reading, Grades 3 to 8 included common items from the 2009/2010 administration (together with newly created items for the Mod-MSA) to help place the 2011 Grades 3 to 8 tests on the established 2009 (Grades 6-8) and 2010 (Grades 3-5) scales. Except for Grades 7 and 8 which had 25 common items, all other grades had 29 common items.

Table 2.1. Test Design for the 2011 Mod-MSA: Reading

Grade	Item Type	Total # of Items	No. of Operational Items After Data Review	No. of Field-Test Items After Data Review
3	Common/Linking Items from previous years	29	29	-
	Modified from Previous MSA and/or Modified MSA Bank Items and/or new 2011 Items	28	16	12
	Total	**57**	**45**	**12**
4	Common/Linking Items from previous years	29	29	-
	Modified from Previous MSA and/or Modified MSA Bank Items and/or new 2011 Items	28	16	12
	Total	**57**	**45**	**12**
5	Common/Linking Items from previous years	29	29	-
	Modified from Previous MSA and/or Modified MSA Bank Items and/or new 2011 Items	28	16	12
	Total	**57**	**45**	**12**
6	Common/Linking Items from previous years	29	29	-
	Modified from Previous MSA and/or Modified MSA Bank Items and/or new 2011 Items	28	16	12
	Total	**57**	**45**	**12**
7	Common/Linking Items from previous years	25	25	-
	Modified from Previous MSA and/or Modified MSA Bank Items and/or new 2011 Items	28	20	8
	Total	**53**	**45**	**8**
8	Common/Linking Items from previous years	25	25	-
	Modified from Previous MSA and/or Modified MSA Bank Items and/or new 2011 Items	28	20	8
	Total	**53**	**45**	**8**

Note: The total number of items is the sum of the operational and field-test items.

2.2. Development and Review of the 2011 Mod-MSA: Reading

Developing the 2011 Mod-MSA: Reading was a complex process. It required a great deal of involvement from MSDE, Pearson, 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 2.2.1 identifies which groups were responsible for developing the 2011 Mod-MSA: Reading.

Table 2.2.1. Responsibility for the 2011 Mod-MSA Test Development

Development of the 2011 Mod-MSA	Primary Responsibility
Development of Preliminary Blueprints and Item Specifications	Pearson, MSDE, NPC
Development of Operational Form Requirements and Blueprint Session	MSDE
Item Writing	MSDE, Pearson
Item Review	Pearson, MSDE, NPC, Content Review Committee
Bias Review	Pearson, MSDE, Bias Review Committee
Vision Review	Pearson, MSDE, Vision Review Committee
Modification of Special Forms	Pearson, MSDE
Review of Special Forms	MSDE
Construction of Operational Test Forms	Pearson, MSDE, NPC
Construction of Field Test Forms	Pearson, MSDE
Review of Operational Test Forms	MSDE
Final Construction of Test Forms	Pearson, MSDE

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 2011 Mod-MSA: Reading program. For example, they made recommendations to MSDE on issues such as test blueprints, operational form construction, field test design, item analysis, item selection for scoring purposes, linking, equating and scaling issues, and other relevant statistical and psychometric issues.

Content Review Committee

Content review committee members ensured that the Mod-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 on the reading tests. They looked for indications of bias that would affect the performance of an identifiable group of students on the test and across the mode of administration (i.e., online, and paper and pencil). Committee members discussed and, if necessary, rejected items based on gender, ethnicity, religious, geographical, or mode of administration bias.

<u>Vision Review Committee</u>

A separate vision review committee examined each item on the reading tests. They looked for indications of bias that would impact the performance of this group of students.

2.3. The 2011 Mod-MSA: Reading: Operational Form Test Structure and Item Types

The 2011 Mod-MSA: Reading tests only had multiple-choice (MC) items with two distractors and a correct answer for each item, except for a few intact items which had three distractors. These items required students to select a correct answer from the available alternatives. Each item was scored dichotomously (i.e., 0 or 1). There was only one form per grade in reading that was administered to the students.

The Mod-MSA Reading has 45 raw score points in each of the six grades. Each strand in Grades 3 to 7 had the same proportionate number of items (total score points) as per those based on the first-year administration that apporoximately corresponded to the MSA examinations. However, Grade 8 had more than the 16 items allocated for General Reading, i.e., 17 items, while the Informational strand had 14 items instead of 15. The test has three reporting strands, i.e., General Reading, Literary and Informational. Table 1.3 provides the score for the reading operational tests based on the number of items used for each strand and grade level.

The Mod-MSA: Reading test was to have the same number of items as the MSA test, i.e., 45 items for each of Grades 3 through 8. These items reflected the same proportion of scores across the different subscales, i.e., the strands.

The Mod-MSA: Reading test was organized under the following content strands for each of the three grades (3–8):

1. Reading Vocabulary (multiple meaning)
2. Reading Vocabulary (words in context)
3. Literary
4. Informational

These strands were combined to match the same three strands as those reported by the reading MSA. For the Mod-MSA: Reading, therefore, the final reporting strands were:

1. General Reading (combination of Reading Vocabulary, i.e., multiple meaning and words in context)
2. Literary
3. Informational

Table 2.3.1 provides the score for the reading operational tests based on the number of items used for each strand and grade level.

Table 2.3.1. The 2011 Mod-MSA: Reading Operational Form with Maximum Points Obtainable Per Strand: Grades 6 to 8

Grade	Strand Title	No. of Items	Item Type	Reporting Strand	Reporting Score
3	Total Test	45	*SR*	Total Test	45
	General Reading	16	*SR*	General	16
	Literary	14	*SR*	Literary	14
	Informational	15	*SR*	Informational	15
4	Total Test	45	*SR*	Total Test	45
	General Reading	15	*SR*	General	15
	Literary	15	*SR*	Literary	15
	Informational	15	*SR*	Informational	15
5	Total Test	45	*SR*	Total Test	45
	General Reading	15	*SR*	General	15
	Literary	15	*SR*	Literary	15
	Informational	15	*SR*	Informational	15
6	Total Test	45	*SR*	Total Test	45
	General Reading	15	*SR*	General	15
	Literary	15	*SR*	Literary	15
	Informational	15	*SR*	Informational	15
7	Total Test	45	*SR*	Total Test	45
	General Reading	15	*SR*	General	15
	Literary	15	*SR*	Literary	15
	Informational	15	*SR*	Informational	15
8	Total Test	45	*SR*	Total Test	45
	General Reading	17	*SR*	General	17
	Literary	14	*SR*	Literary	14
	Informational	14	*SR*	Informational	14

2.4. Item Analyses Undertaken Prior to the Creation of the Operational Forms

The 2011 Mod-MSA: Reading was administered as a single form, which included more than the required number of operational items. After administration of the form, operational items were selected during data review on the basis of their statistics and the number of items required for each strand of the operational test (see Table 2.3.1, above). All items not selected as operational were banked with their respective statistics as field test (FT) items. These items together with the 2009/2010 operational items could be used as common linking items in 2012 in order to place the 2012 examinations on the established 2009 (Grades 6-8) and 2010 (Grades 3-5) scales.

The statistical considerations for the selection of operational items were based on the following:

- Classical item analysis
- Differential item functioning (DIF) analyses
- IRT analyses

All analyses provided in this report, unless otherwise indicated, are based on the equating sample which consisted of the full population with exclusion criteria for the equating process.

<u>Classical Item Analyses</u>

Classical item analyses included the calculation of *p*-values, the point-biserials, distractor-to-total correlations, and distractor frequency analysis.

Items were flagged for further scrutiny if:

- An item distractor was not selected by any students (i.e., nonfunctional distractor), or selected by a large number of high proficiency students, with low selection from other proficiency groupings (i.e., ambiguous distractor).
- An item *p*-value was less than .10 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.
- Omit rate was flagged at > 5%.

All items required a careful decision for inclusion in the operational form. For example, an item that was flagged as having a point-biserial < 0.10 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 was made for including it in the operational test form, i.e., learning the content was deemed a necessary factor for an item’s inclusion in spite of it having poor statistics that were not related to miskeyed items.

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.

During the item development period, prior to statistical analysis of DIF, all items were subjected to the scrutiny of the Bias Review Committee. As explained in Section 2.2, the Bias Review Committee examined each reading item, looking for indications of bias that could impact the performance of an identifiable group of students. They discussed or rejected items based on gender, ethnicity, religious preferences, or geographical bias.

After items were scored, statistical item analysis pertaining to DIF was undertaken. In this analysis, the gender reference group was males, and the ethnicity reference group was white. The gender focal group was females and the ethnicity focal group was black (African Americans). The total score of each operational form was used as the matching variable.

Since the 2011 Mod-MSA: Reading was a single-format examination comprising only SR items, the DIF procedure used consisted of the Mantel-Haenszel contingency procedure (Mantel & Haenszel, 1959) together with an effect-size approach[2] based on the delta scale (Camilli & Shepard, 1994).

The Mantel Haenszel Chi-Square

The Mantel and Haenszel (1959) chi-square, which approximately follows a chi-square distribution with one degree of freedom, can be formulated as per the following (from Camilli & Shepard, 1994):

$$\text{MH}\,\chi^2 = \frac{\left\{\left|\sum_{j=1}^{S}\left[A_j - E\left(A_j\right)\right]\right| - 1/2\right\}^2}{\sum_{j=1}^{S} VAR\left(A_j\right)}, \text{ where}$$

[2] For a detailed discussion on Mantel-Haenszel chi-square, the delta scale and ETS categories, please refer to Camilli and Shepard (1994).

A_j and $E(A_{j)}$ are the observed number of correct responses and the expected number on the item, respectively for the reference group, while VAR(A_j) is the variance associated with the observed score.

The Delta Scale

The odds of a correct response are *P/Q* or *P/(1-P)*. The odds ratio, on the other hand, is simply the odds of a correct response of the reference group divided by the odds of a correct response of the focal group.

For a given item, the odds ratio is defined as follows:

$$\alpha_{M-H} = \frac{P_r / Q_r}{P_f / Qf}.$$

The corresponding null hypothesis is that the odds of getting the item correct are equal for the two groups. Thus, the odds ratio is equal to 1:

$$H_0\text{: } \alpha_{M-H} = \frac{P_r / Q_r}{P_f / Qf} = 1.$$

In order to calculate the delta scale, the Mantel and Haenszel (1959) log odds ratio was calculated using the following equation:

$$\alpha_{MH} = \frac{\sum_{j=1}^{S} A_j D_j / T_j}{\sum_{j=1}^{S} B_j C_j / T_j} \text{ , where}$$

the various variables in the equation are from the following 2 x 2 contingency table for the *jth* total score on the test (Camilli & Shepard, 1994, p. 106).

Score on studied item with general notation

		1	0	Total
Group	R	A_j	B_j	n_{Rj}
	F	C_j	D_j	nFj
		m_{1j}	m_{0j}	T_j

The log odds ratio is a transformation of the odds ratio with its range being in the interval $-\infty$ to $+\infty$. The simple natural logarithm transformation of this odds ratio is symmetrical around zero, in which zero has the interpretation of equal odds. The odds ratio is transformed into a log odds ratio as per the following: $\beta_{M-H} = \ln(\alpha_{M\text{-}H})$. β_{M-H}, also has the advantage of being transformed linearly to other interval scale metrics (Camilli & Shepard, 1994). This fact is utilized in creating the delta scale (*D*), which is defined as $D = -2.35\beta_{M-H}$.

DIF Classification

The *M-H* χ^2 is examined in conjunction with the delta scale (*D*) to obtain DIF classifications depicted in Table 2.4.1, below.

Table 2.4.1. DIF Classification

Category	Description	Criterion
A	No DIF	Non-significant *M-H* χ^2 or $\|D\| < 1.0$
B	Weak DIF	Significant *M-H* χ^2 and $\|D\| < 1.5$ or Non-significant *M-H* χ^2 and $\|D\| \geq 1.0$
C	Strong DIF	Significant *M-H* χ^2 and $\|D\| \geq 1.5$

The groupings for the DIF analysis were based on matching students' scores on the Mod-MSA: Reading. Four proficiency groupings of the Mod-MSA students were formed at quarter intervals of the total Mod-MSA: Reading score. The performance on the Mod-MSA: Reading for the four proficiency-matched groups (gender, and ethnicity) was then compared for each item to evaluate potential differential performance by groups.

Items that were flagged as showing DIF (Category 'B', i.e., moderate DIF, and category 'C', i.e., extreme DIF) were subjected to further examination. For each of these items, experts judged whether the differential difficulty of the item was unfairly related to group membership based on the following guidelines:

- If the difficulty of the item was unfairly related to group membership, then the item should not be used at all.
- If the difficulty of the item was related to group membership, then the item should only be used if there was no other item matching the test alignment requirements presented in Appendix E.

All DIF results were stored in the Maryland item bank.

<u>Item Response Theory (IRT) Analyses</u>

Rasch fit statistics, infit and outfit (see Section 6.2) were used to examine model fit to the data. Items with fit indices < 0.5 or > 2.00 were flagged for misfit because, according to Linacre and Wright (1999), the inclusion of these items could be unproductive to the measurement system (< 0.5) or they could degrade the measurement system (> 2.0).

2.5. Items Flagged for Inspection Prior to the Creation of the Operational Forms

The following table provides content by grade summary with respect to the total number of items administered and the number of items that were flagged strictly on the basis of the statistics (classical, DIF and IRT) discussed above.

Table 2.5.1. Summary Stats Used in the Development of the 2011 Mod-MSA: Reading Operational Form

Grade	Total # of Items	DIF Flag B (for check only)	DIF Flag C	PB Flag <= 0.10 but > 0[1]	PB Flag < = 0 (Cannot be used)	Items Rejected (C DIF + PB <=0 Flag	Items Used for Operational Form Building Based on Statistical Criteria	Items Needed for Each Operational Form
3	57	2	0	3	1	1	56	45
4	57	1	0	4	1	1	56	45
5	57	2	0	4	2	2	55	45
6	57	6	0	2	2	2	55	45
7	53	6	0	5	2	2	51	45
8	53	1	0	4	1	1	52	45

Note: 1. Items in this column were generally not used unless a substitute could not be found for it.

As can be seen from the table, other than the point biserial (PB) and the DIF flags, all other statistical indices were well within the acceptable criteria. No items were flagged based on the fit analyses. For the PB we checked every item < 0.15 internally for the items being wrongly keyed. No such items were found across content and grade, even though some of the items had negative PBs.

2.6. Items Selected for the 2011 Operational Tests

As discussed above, the selection of items that were included in the final operational test forms of the 2011 Mod-MSA: Reading examination required a careful consideration based on test design, classical item analyses, DIF analyses, and IRT analyses. The general guidelines for the creation of the operational forms were as follows:

- Do not include items that are too easy or too hard.
- Do not include items with DIF classifications ‘C’ for the SR items *unless* they have been deemed acceptable by the external review of content experts.
- Finally, do not include items which have Rasch infit and outfit mean-squares higher than 2.0.

Appendix A provides a list of item UIN numbers of each item used to produce the operational form (the core items) from the total items administered in 2011.

Item level descriptive statistics (i.e., *p*-values and point biserials) are provided in Section 3.2 (Tables 3.2.1 to 3.2.6). Appendix B provides item analysis by distractors while differential item functioning (DIF) analysis is provided in Appendix C.

As shown in Tables 3.2.1 to 3.2.6, there were a few items that had negative point biserials. These items were examined by content specialists for key and content accuracy. Other than the Point Biserial (PB) and the DIF flags, all other statistical indices were well within the acceptable criteria. For the PB we checked every item < 0.15 internally for the items being wrongly keyed. No such items were found across content and grade, even though some of the items had negative PBs.

DIF analyses were conducted for gender and between White and African-Americans using the delta scale, D ($D = -2.35\log_e(\alpha_{MH})$, where $\log_e(\alpha_{MH})$ is the Mantel-Haenszel log odds ratio), in combination with the Mantel-Haenszel significant test of DIF detection (see Appendix C). Items with flags for moderate DIF (flag with ‘B’) were examined for bias. All items that were flagged as ‘C’ were not included in the operational form. However, there were no items with a DIF classification of ‘C’ across grades.

The MSDE and Pearson worked collaboratively to select items for 2011 operational scoring and to evaluate the psychometric properties of these operational item sets. In accordance with the NPC’s recommendation, no items with negative point biserial correlations were selected for operational scoring. However, in spite of our intention of abiding by the terms of rejection outlined above, some items that had PBs less than 0.10 (but not negative or zero PBs) were included as operational items because of not having corresponding substitute items to use. There were no item that had omits for greater than five percent of the students.

See Table 2.6.1 for the number of operational items with PBs less than 0.10 but greater than zero.

Table 2.6.1. Number of Items Included as Operational Items with 0 < PB < .10 by Grades

Grade	0 < PB < 0.10
3	1
4	2
5	2
6	2
7	3
8	2

2.7. Scoring Procedures of the 2011 Mod-MSA: Reading

Students’ responses were machine-scored. 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 were converted to alphanumeric format.

After students’ responses were scanned, 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 assigned a score of zero.