8. TEST VALIDITY

8.1. Test Validity for the 2011 Mod-MSA: Reading

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 2011 Mod-MSA: Reading, content-related evidence, evidence from item development methods, bias review evidence during test development and for items that showed differential item functioning (DIF), and evidence from internal structure were collected. Also, a study comparing the mode of administration was undertaken by Pearson in 2009 and 2010 to validate the online administration of the test.

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).

Evidence regarding the alignment between the content in the 2011 Mod-MSA: Reading and the standards of achievement set by MSDE are provided in Appendix E that links each item to the specific standard(s) it measures. Information on the item composition of the operational test forms can be obtained from Section 2.6 *Items Selected for the 2011 Operational Tests.* The selected items are displayed in Appendix A with their UIN numbers.

Evidence from Item Development Methods

Test development for Mod-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 necessitated a great deal of time and energy. More specific information on item (test) development and review can be obtained in Section 2, *Test Design and Development of the 2011* Mod-MSA: Reading while the standards to

which the items were aligned can be obtained from the MSDE website at: http://mdk12.org/instruction/curriculum/index.html.

As explained in Section 2.4 to 2.6, once these items were scored, MSDE and Pearson conducted additional item analysis and content review to select items for the operational form, i.e., the form on which the student scores were reported. Any item 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.

Evidence Based on Excluding Bias Items Before and After DIF Analysis

One important consideration in evaluating the validity of a test is to examine the equity of each item performance between groups of interest. As explained in Section 2.2, all items went through a bias review committee to ascertain that items were not biased with respect to gender, ethnicity, geographical location, etc. Also, as explained in Section 2.4, after items were scored, DIF analysis was undertaken and those items that showed moderate or significant DIF were reviewed for bias with respect to gender, and ethnicity, which included white and black students. More information on DIF analyses can be obtained in Section 2.4, *Differential Item Functioning*.

Items that had moderate or extreme DIF are depicted in Table 8.1.1, below. These items for the Mod-MSA; Reading were checked for content bias, but did not show favoritism on the basis of gender or ethnicity (black vs. white students). DIF for all items across grades are provided in Appendix C.

Table 8.1.1. Category Classification of Items that Showed Moderate or Extreme DIF by Grades

			DIF Classification[1]	
Grade	Item Sequence No.	Item UIN No.	Gender	White/African-American
3	18	100000450388	A	<B
	55	100000260342	<B	A
4	19	100000479575	>B	A
5	22	100000269965	>B	A
	31	100000403023	>B	A
6	3	100000213665	>B	A
	8	100000213671	>B	A
	21	100000270691	A	>B
	33	100000273224	A	>B
	44	100000450957	A	<B
	51	100000257022	<B	A
7	4	100000213678	A	<B
	9	100000403118	>B	>B
	18	100000403125	<B	A
	20	100000403126	>B	A
	23	100000403134	>B	A
	45	100000257732	A	<B
8	25	100000270643	<B	A

Note: 1. '>' = in favor of the reference group, i.e. males and White Americans while '<' = in favor of the focal group. Extreme DIF = "C", Moderate DIF = "B", and No DIF is classified as an "A".

Evidence from Internal Structure of the Tests

As explained in Section 2.3, the 2011 Mod-MSA: Reading contains three reading strands: General Reading, Literary, and Information. Even though these are individual strands are "locally independent," they measure the same underlying reading trait. Therefore, the positive correlation among these strands is an indication of their relationship with each other in measuring the same underlying construct. To ascertain the homogeneity of the test, correlations were calculated to depict the relationship between each strand within a grade. Tables 8.1.2 through 8.1.7 show the correlations among the reading strands for Grades 3 through 8, respectively. The 2010 correlations are also provided.

Table 8.1.2. The 2010/2011 Mod-MSA: Reading Strand (Cluster) Correlations: Grade 3

	2010						2011					
Strand (Subscale)	**N**	**Mean**	**SD**	**GR**	**L**	**I**	**N**	**Mean**	**SD**	**GR**	**L**	**I**
General Reading (GR)	813	10.02	3.34	1.00			943	9.33	2.75	1.00		
Literary (L)	813	7.07	2.50	0.60	1.00		943	7.83	2.74	0.54	1.00	
Informational (I)	813	8.13	3.01	0.58	0.57	1.00	943	7.52	2.87	0.58	0.56	1.00

Note. The restriction of the range of scores on the strands could have resulted in the attenuation of the correlation coefficients between any two strands/modalities.

Table 8.1.3. The 2010/2011 Mod-MSA: Reading Strand (Cluster) Correlations: Grade 4

	2010						2011					
Strand (Subscale)	**N**	**Mean**	**SD**	**GR**	**L**	**I**	**N**	**Mean**	**SD**	**GR**	**L**	**I**
General Reading (GR)	967	10.10	3.22	1.00			1344	9.44	2.71	1.00		
Literary (L)	967	6.82	2.51	0.57	1.00		1344	7.92	2.58	0.48	1.00	
Informational (I)	967	7.48	2.85	0.60	0.51	1.00	1344	8.52	2.88	0.57	0.52	1.00

Note. The restriction of the range of scores on the strands could have resulted in the attenuation of the correlation coefficients between any two strands/modalities.

Table 8.1.4. The 2010/2011 Mod-MSA: Reading Strand (Cluster) Correlations: Grade 5

	2010						2011					
Strand (Subscale)	**N**	**Mean**	**SD**	**GR**	**L**	**I**	**N**	**Mean**	**SD**	**GR**	**L**	**I**
General Reading (GR)	1043	10.08	3.06	1.00			1534	10.21	3.01	1.00		
Literary (L)	1043	7.62	2.80	0.57	1.00		1534	7.46	2.75	0.54	1.00	
Informational (I)	1043	6.60	2.58	0.50	0.51	1.00	1534	7.35	2.70	0.52	0.53	1.00

Note. The restriction of the range of scores on the strands could have resulted in the attenuation of the correlation coefficients between any two strands/modalities.

Table 8.1.5. The 2010/2011 Mod-MSA: Reading Strand (Cluster) Correlations: Grade 6

	2010						2011					
Strand (Subscale)	**N**	**Mean**	**SD**	**GR**	**L**	**I**	**N**	**Mean**	**SD**	**GR**	**L**	**I**
General Reading (GR)	975	8.82	3.07	1.00			1518	9.15	2.83	1.00		
Literary (L)	975	8.12	2.80	0.59	1.00		1518	8.31	2.54	0.54	1.00	
Informational (I)	975	7.59	2.78	0.57	0.54	1.00	1518	7.77	2.66	0.51	0.46	1.00

Note. The restriction of the range of scores on the strands could have resulted in the attenuation of the correlation coefficients between any two strands/modalities.

Table 8.1.6. The 2010/2011 Mod-MSA: Reading Strand (Cluster) Correlations: Grade 7

	2010						2011					
Strand (Subscale)	**N**	**Mean**	**SD**	**GR**	**L**	**I**	**N**	**Mean**	**SD**	**GR**	**L**	**I**
General Reading (GR)	1158	9.50	2.79	1.00			1518	9.15	2.83	1.00		
Literary (L)	1158	8.77	2.83	0.62	1.00		1518	8.31	2.54	0.54	1.00	
Informational (I)	1158	7.01	2.80	0.50	0.49	1.00	1518	7.77	2.66	0.51	0.46	1.00

Note. The restriction of the range of scores on the strands could have resulted in the attenuation of the correlation coefficients between any two strands/modalities.

Table 8.1.7. The 2010/2011 Mod-MSA: Reading Strand (Cluster) Correlations: Grade 8

	2010						2011					
Strand (Subscale)	**N**	**Mean**	**SD**	**GR**	**L**	**I**	**N**	**Mean**	**SD**	**GR**	**L**	**I**
General Reading (GR)	1268	9.79	2.84	1.00			1803	10.29	2.90	1.00		
Literary (L)	1268	8.77	2.69	0.55	1.00		1803	8.05	2.55	0.55	1.00	
Informational (I)	1268	8.71	2.59	0.52	0.48	1.00	1803	8.38	2.26	0.49	0.41	1.00

Note. The restriction of the range of scores on the strands could have resulted in the attenuation of the correlation coefficients between any two strands/modalities.

8.2. Unidimensionality Analysis for the 2011 Mod-MSA: Reading

Measurement implies order and magnitude along a single dimension (Andrich, 1989). Consequently, in the case of scholastic achievement, a linear scale is required 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 proficiency measured by the test (Andrich, 1988; Hambleton et al., 1991; Ryan, 1983).

To check the unidimensionality of the 2011 Mod-MSA: Reading, correlation coefficients were computed with LISREL 8.5 (Jöreskog & Sörbom, 1993). Principal component analysis was then

applied to produce eigenvalues. The first and the second principal component eigenvalues were compared without rotation. Table 8.2.1 summarizes the results of the first and second principal component eigenvalues of the 2011 Mod-MSA: Reading. As shown in the table, the first factor extracted a much larger amount of eigenvalues across all grades.

Table 8.2.1. Eigenvalues between the First and Second Components of the 2011 Mod-MSA: Reading

Grade	First Eigenvalue	Second Eigenvalue
3	8.08	2.23
4	7.67	1.75
5	7.85	1.84
6	7.09	1.74
7	6.25	1.78
8	6.86	1.62