Student Performance

Score Interpretation

To help provide appropriate interpretation of the 2008 MSA Science operational test scores, two types of scores were created: scale scores and performance levels and descriptions.

Scale Scores

As explained in the proceeding section, the 2008 MSA Science yielded scale scores that ranged between 240 and 650. As a result of calibration, equating and scaling the scale scores yielded from the 2 base forms have the same meaning within the same grade; however, the scale 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 on the MSA Science test. Thus, performance levels and descriptions can give a specific interpretation other than a simple interpretation because they were developed to bring meaning to the scale scores.

Performance Levels and Descriptions

Performance levels and descriptions provide specific information about students' performance levels and help interpret the 2008 MSA Science scale scores. They describe what students at a particular level generally know and are able to do and can be applicable to all students within a grade level.

Performance standards for MSA Science were established in 2007. Details of the standard-setting process and outcomes are provided in MSA Science standard-setting technical report (Pearson, 2007). The State Board reviewed the performance standards recommended by the standard-setting committee and made a modification in the recommendation. The performance standards approved by the State Board are listed in Table 9. Students whose scale scores are lower than the Proficient cut score are classified as "Basic." The highest performance group whose scale score is equal or higher than Advanced cut score belongs to the "Advanced" group. The middle group is called "Proficient"

Table 9. Scale score cut scores for grades 5 and 8 MSA Science.

Grade	Proficient Cut score	Advanced Cut score
5	391	467
8	387	478

Tables 10 reports percentages of grade 5 students in three performance groups and the descriptive statistics for the selected subgroups (gender and ethnicity). The analysis was conducted for all students in grades 5 as well as by administration mode.

Table 10. Grade 5 performance level percentages and descriptive statistics

	Overall						Online Administration						Paper Administration					
	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N
	B	P	A				B	P	A				B	P	A			
Subgroup																		
All Students																		
All	36	56	9	405	45.7	60770	32	58	10	411	44.4	35017	42	52	7	398	46.5	25753
Gender																		
Female	37	56	7	404	44.3	29503	33	59	8	408	43.1	17007	42	52	6	398	45	12496
Male	35	55	10	407	47	31214	30	58	12	413	45.5	18008	41	51	7	399	47.8	13206
Ethnicity																		
Asian	20	64	16	424	44.4	3604	20	64	17	425	42.9	1994	21	65	14	422	46.2	1610
Black	54	44	2	385	41.4	22748	53	45	2	388	40.2	10666	56	43	2	383	42.4	12082
Hispanic	52	46	2	386	43.8	5625	48	48	4	391	44.5	2106	54	44	2	384	43.1	3519
Native American	36	58	7	403	46.2	224	34	59	7	406	43.1	149	39	55	7	397	51.6	75
White	20	66	14	423	41.2	28512	20	66	14	423	40.9	20098	21	66	13	422	41.6	8414
Note: Performance Levels, B=Basic, P=Proficient, A=Advanced																		

Tables 11 reports percentages of grade 8 students in three performance groups and the descriptive statistics for the selected subgroups (gender and ethnicity). The analysis was conducted for all students in grades 5 as well as by administration mode.

Table 11. Grade 8 performance level percentages and simple statistics

	Overall						Online Administration						Paper Administration					
	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N	Performance Levels			Mean	SD	N
	B	P	A				B	P	A				B	P	A			
Subgroup																		
All Students																		
All	39	58	4	397	51.7	63573	37	60	4	400	49.2	41583	43	54	4	392	55.6	21990
Gender																		
Female	39	58	3	397	48.8	30856	37	60	3	399	46.6	20377	42	55	3	392	52.6	10479
Male	39	57	5	398	54.1	32571	36	59	5	401	51.6	21200	43	52	4	391	58.1	11371
Ethnicity																		
Asian	18	71	11	427	46.6	3429	18	72	10	426	45.2	2079	18	70	12	428	48.6	1350
Black	61	38	0	370	47.4	24538	58	41	1	375	45	15477	66	33	0	362	50.3	9061
Hispanic	56	43	1	376	50	5154	55	44	1	379	48.4	2931	57	42	1	372	51.8	2223
Native American	39	57	4	397	51.2	255	38	60	2	399	45.1	184	42	51	7	392	64.5	71
White	20	74	7	420	42.6	30053	20	74	6	420	42.1	20905	20	74	7	420	43.8	9148
Note: Performance Levels, B=Basic, P=Proficient, A=Advanced																		

Validity

Pearson subscribes rigorously to the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 1999). The standards define validity as

> … the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating tests. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations.

Validity can be established through the collection of evidence to demonstrate the alignment of item content with the curriculum, compliance to the test specifications, test fairness, and valid uses and interpretations of test scores. This section describes various analyses to evaluate the validity and reliability evidence for the 2008 MSA Science test.

Content-related Validity

All MSA Science items were explicitly developed to measure the specific knowledge and skills described in the Voluntary State Curriculum (VSC). In addition, the alignment of the items to the standards was reviewed and verified independently by multiple content reviewers and Maryland educators. The MSA Science core items were handed over to Pearson after the extensive reviews by the Mississippi educators and external reviewers.

Construct-related Validity

Construct validity refers to what test scores mean and what kinds of inferences they support. Construct validity is the central concept underlying the MSA Science test validation process. Evidence for construct validity is comprehensive and integrates evidence from both content- and criterion-related validity.

Construct-related validity evidence (internal consistency validity evidence) can come from many sources. The American Psychological Association provides the following list of possible sources (AERA, APA & NCME, 1999):

- high inter-correlations among assessment items or tasks attest that the items are measuring the same trait, such as a content objective, sub-domain or construct;
- substantial relationships between the assessment results and other measures of the same defined construct;
- little or no relationship between the assessment results and other measures which are clearly not of the defined construct;
- substantial relationships between different methods of measurement regarding the same defined construct;
- relationships to non-assessment measures of the same defined construct.

The collection of construct-related evidence is a continuous process, and at present substantial evidence is available representing internal structure (the first of the five bullets above). This section describes four sources of internal structure-based construct validity evidence for the MSA Science test: item-total/point-biserial correlations, inter-correlation among standards/subscales, unidimensionality, and DIF analysis.

Item-total Correlation

Item-total correlations provide another measure of the congruence between the way an item functions and our expectations. Typically students with high ability (i.e., those who perform well on the MSA Science overall) answer items correctly, and students with low ability (i.e., those who perform poorly on the MSA Science overall) answer items incorrectly. If these expectations are met, the point-biserial (i.e., item-total) correlation between the item and the total test score will be high and positive, indicating that the item is a good discriminator between high ability and low ability students. A correlation value above 0.20 is considered acceptable; values closer to 1.00 indicate greater discrimination. A test comprised of maximally discriminating items will maximize internal consistency reliability.

Assuming that the total test score represents the extent to which a student possesses the construct being measured by the test, high point-biserial correlations indicate that the tasks on the test require this construct to be answered correctly. Table 12 reports the mean, minimum, and maximum point-biserial correlation values for the MSA Science tests. The adjusted point-biserial removes the item score from the total score so that the index can be an unbiased estimate of the item with the test. As can be observed from this table, the average adjusted point-biserial ranged from 0.24 to 0.39 across the MSA Science tests for grades 5 and 8. Overall MSA Science core items in general seem to perform well in terms of differentiating students with high ability from low-performing students and measuring a common underlying construct. A portion of the field test items were somewhat less effective, which is to be expected.

Table 12. Summary of adjusted point-biserial

Subject	Grade	Status	Adjusted Point-biserial		
			Mean	Minimum	Maximum
SC	5	OP	0.36	0.10	0.56
SC	5	FT	0.29	-0.09	0.62
SC	8	OP	0.40	0.19	0.68
SC	8	FT	0.35	0.00	0.69

Inter-correlation among Standards

There are six standards within the VSC frameworks for MSA Science. Content judgment was made when classifying items into each of the standards, and the MSA Science subscales each represent one of these standards. To assess the extent to which items aligned with the standards are assessing the same underlying construct, a correlation matrix was computed among the total scores of competencies.

Table 13 reports the correlations among the six standards/subscales. The correlation ranged from 0.54 to 0.77 with majority of correlation around 0.65. The subscales are highly intercorrelated, indicating that a single overarching construct of Science is being measured.

Table 13. Correlation among MSA Science content standards

Grade 5 Form A		**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**
	Str1	1.0000					
	Str2	0.6180	1.0000				
	Str3	0.6960	0.6141	1.0000			
	Str4	0.6477	0.5880	0.6323	1.0000		
	Str5	0.6831	0.5985	0.6616	0.6299	1.0000	
	Str6	0.6580	0.5816	0.6512	0.6166	0.6261	1.0000
Grade 5 Form B		**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**
	Str1	1.0000					
	Str2	0.6138	1.0000				
	Str3	0.6615	0.6145	1.0000			
	Str4	0.6253	0.6042	0.6021	1.0000		
	Str5	0.6044	0.5791	0.5788	0.5846	1.0000	
	Str6	0.6773	0.6161	0.6688	0.6117	0.5825	1.0000
Grade 8 Form A		**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**
	Str1	1.0000					
	Str2	0.6896	1.0000				
	Str3	0.6769	0.7325	1.0000			
	Str4	0.6790	0.7231	0.7105	1.0000		
	Str5	0.5487	0.5897	0.5842	0.5969	1.0000	
	Str6	0.6723	0.6956	0.6921	0.6965	0.5775	1.0000
Grade 8 Form B		**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**
	Str1	1.0000					
	Str2	0.6510	1.0000				
	Str3	0.7138	0.6359	1.0000			
	Str4	0.6983	0.6099	0.6783	1.0000		
	Str5	0.6538	0.6141	0.6350	0.6160	1.0000	
	Str6	0.7657	0.6593	0.7164	0.6878	0.6597	1.0000

*Str: Standard

Unidimensionality

In addition to the processes and procedures Pearson employs during item and test form development to promote construct validity, a confirmatory factor analysis is also conducted to examine the construct validity of the 2008 MSA Science tests.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) was conducted to further examine the relationship between the subscales. CFA used SAS Proc Calis and the maximum likelihood estimation (MLE; Anderson & Gerbing, 1988) procedure. The model hypothesized that the subscale scores belong to a single latent trait. Model fit was tested through indices including adjusted goodness of fit (AGFI), and Root Mean Squared Error of Approximation (RMSEA). Values of the AGFI statistic which indicate good fit are higher than 0.90 (Tabachnick & Fidell, 2001). The RMSEA is a function of the estimated discrepancy between the population covariance matrix and the model-implied covariance matrix, with a value of less than or equal to .05 indicating close fit and a value between .05 and .08 indicating a "reasonable error of approximation" (Browne & Cudeck, 1993, p. 144). Hu and Bentler (1999) propose an RMSEA $\leq$ .06 as the guideline for close fit. Table 14 summarizes fit indicators estimated from the confirmatory factor analysis for the 2008 MSA Science tests. The confirmatory factor analysis results provide additional

evidence to support the validity of the MSA Science tests. For both grades, the lowest AGFI was 0.9809, and the highest RMSEA was 0.0518. The AGFI and RMSEA indicators supported the model fit.

Table 14. Fit indicators for confirmatory factor analysis on MSA Science

Grade/Form	AGFI	RMSEA
Grade 5 Form A	0.9974	0.0182
Grade 5 Form B	0.9849	0.0452
Grade 8 Form A	0.9845	0.0472
Grade 8 Form B	0.9809	0.0518

*AGFI: Adjusted Goodness of Fit; RMSEA: Root Mean Squared Error of Approximation

Validity Evidence for Scores from Accommodated Testing

Accommodations are offered to students with disabilities that preclude them from being fairly assessed by the tests as they are written (e.g., visually impaired students). In order to examine whether or not these accommodations are effective (i.e., result in valid test scores) the CFA conducted to examine the relationship between subscales was repeated using only students testing with accommodations and then again using only students testing without accommodations. The results of this analysis showed good model fit based on the data from both student populations (see Tables 15). This suggests that offering accommodations to disabled students preserves the internal structure of the test. One can infer from these results that the accommodations offered for the MSA Science tests are effective and produce scores that are as valid as those of students who are not in need of accommodation.

Table 15. Fit indicators for accommodations/non-accommodations based CFA

	Accommodations		No Accommodations	
Grade/Form	AGFI	RMSEA	AGFI	RMSEA
Grade 5 Form A	0.9968	0.0162	0.9937	0.0293
Grade 5 Form B	0.9935	0.0241	0.9869	0.0419
Grade 8 Form A	0.9915	0.0294	0.9910	0.0350
Grade 8 Form B	0.9967	0.0150	0.9891	0.0388

*AGFI: Adjusted Goodness of Fit; RMSEA: Root Mean Squared Error of Approximation

Validity Evidence for Different Populations

The primary evidence for the validity of the MSA Science lies in the content and construct being measured. The evidence of validity is sought from a statistical analysis to detect differential item functioning that could favor a particular sub-group over and beyond the difference in ability.

Since the test assesses the statewide content standards, which are required to be taught to all students, the test should not be more or less valid for use with one subpopulation of students relative to another. Great care has been taken to ensure that the MSA Science items are fair for students of various backgrounds. During the item development and review processes, efforts were made to avoid or detect possible bias toward or against any subpopulations in Maryland. Besides these content-based efforts that are put forth in the test development process, data-driven statistical procedures are also employed to identify items that behave differently for different populations. Statistical indices of Differential Item Functioning (DIF) are only a quantitative marker; bias is a qualitative condition that can only be determined by an examination of the content of the item. The MSA Science test development process approaches bias detection and

elimination from both viewpoints, at multiple steps in the process, and by multiple levels of reviews.

The DIF analysis was carried out on the data collected from the 2008 MSA Science administration. DIF statistics are used to identify items on which members of a focal group have different probability of getting the items correct from members of a reference group after members of both groups have been matched by the students' ability level on the test. In the DIF analyses, the total raw score on the operational items is used as the ability-matching variable. Details of the DIF analysis are provided in the DIF analysis section and the number of items displaying a significant level of DIF is summarized in Table 5. Because of the multi-layered approach to reducing or eliminating systematic bias, empirically the majority of items on the MSA Science operational tests exhibit no DIF or weak DIF, and the impact of DIF on the 2008 MSA Science scores can be considered negligible.