Field Test Item Analysis and Calibration

Key Check Analysis of Field Test Data

Using preliminary data collected from the 2012 administration (a minimum of 200 responses were required for each form by mode of administration), Pearson computed Classical Test Theory statistics on all multiple choice items in order to screen for items with characteristics that could be associated with an item being scored with a wrong correct answer key (mis-keyed). These analyses were carried out in the same manner as those described for the operational key check analysis (see page 9). Any items identified during this process were presented to Pearson content specialists for review to ensure that items were keyed properly. No mis-keyed items were identified on either of the MSA Science tests.

Classical Item Analysis

The following classical item statistics that were calculated:

- P-value of SR items
- Mean of BCR items
- Point-Biserial Correlation
- Item Option Point-Biserial for SR items
- P-value by Item Option for SR items
- Item Score Distribution for BCR items

The results of the classical item analysis were banked for use during the construction of subsequent MSA Science tests. P-value and point-biserial statistics for the 2012 MSA field test items are reported in Appendix A.

Field Test Calibration

Field test items are embedded within each session of the MSA Science tests with unique items appearing in the same positions across the field test forms. A total of ten field test forms were created by embedding unique field test items into each operational form. Table 3 provides a graphical depiction of the field test design. This design ensured that one of two sets of operational test items were common to each field test form. This allows all field test item parameters to be estimated concurrently, thus placing all items on a common scale as is done with the two operational forms during operational equating. During this concurrent calibration all items (operational and field test) are freely estimated. As a result the item parameter estimated obtained for the field test items are not on the base scale. In order to place these parameter estimates on the base scale so that they may be use to construct equivalent operational test forms for subsequent administrations the Stocking and Lord procedure is used to calculate transformation constants with the anchor set being formed from all of the operational items (comparing the operational item parameters obtained during field test calibration to those banked following post-equating). This process was used to place all 2012 field test items on the base scale. The transformation constants derived and applied at each grade during this are shown in Table 12. The IRT parameters for grade 5 and 8 field test items are presented in Appendix A.

Table 12. Field Test Transformation Constants

	Grade 5		Grade 8	
	Slope	**Intercept**	**Slope**	**Intercept**
Field Test (12 FT items >> 12 OP items)	1.033233	0.266260	1.065505	0.290947

Differential Item Functioning (DIF) Analysis

One of the goals of the MSA Science test development is to assemble a set of items that provides a measure of a student's ability that is as fair and accurate as possible for all subgroups within the population. Differential item functioning (DIF) analysis refers to procedures that assess whether items are differentially difficult for different groups of examinees. DIF procedures typically control for overall between-group differences on a criterion, usually total test scores. Between-group performance on each item is then compared within sets of examinees having similar test scores. If the item is differentially more difficult for an identifiable subgroup when conditioned on ability, the item may be measuring something different from the intended construct. However, it is important to recognize that DIF-flagged items might be related to actual differences in relevant knowledge or skills or statistical Type 1 error. As a result, DIF statistics are used to identify potential sources of item bias. Subsequent review by content experts and bias committees are required to determine the source and meaning of performance differences. In the MSA Science DIF analysis, DIF statistics were estimated for all major subgroups of students with sufficient sample size: Black, Hispanic and Female[1]. Items with statistically significant differences in performance were flagged so that items could be carefully examined for possible biased or unfair content that was undetected in earlier fairness and bias content review meetings held prior to form construction.

Pearson used the Mantel-Haenszel (MH) chi-square approach to detect DIF in SR items. Pearson calculated the Mantel-Haenszel *delta* statistic (MH D-DIF, Holland & Thayer, 1988) to measure the degree and magnitude of DIF. The student group of interest is the *focal* group, and the group to which performance on the item is being compared is the *reference* group. The referent groups for this DIF analysis were White for ethnicity and male for gender. The focal groups were females and minority ethnicity groups.

Items were separated into one of three categories on the basis of DIF statistics (Holland & Thayer 1988; Dorans & Holland 1993): negligible DIF (category A), intermediate DIF (category B), and large DIF (category C). The items in category C, which exhibit significant DIF, are of primary concern.

Positive values of *delta* indicate that the item is easier for the *focal* group, suggesting that the item favors the *focal* group. A negative value of *delta* indicates that the item is more difficult for the *focal* group. The item classifications are based on the Mantel-Haenszel chi-square and the MH delta (Δ) value as follows:

- The item is classified as C category if the absolute value of the MH delta value (i.e., |Δ|) is significantly greater than 1 and also greater than or equal to 1.5.
- The item is classified as B category if the MH delta value (Δ) is significantly different from 0 and either the absolute value of the MH delta (|Δ|) is less than 1.5 or the absolute value of the MH delta (|Δ|) is not significantly different from 1.

[1] DIF analysis on the Asian students was not conducted due to small sample size.

- The item is classified as A category if the delta value (Δ) is not significantly different from 0 or the absolute value of delta (|Δ|) is less than or equal to 1.

The effect size of the standardized mean difference (SMD) was used to flag DIF for the BCR items. The SMD reflects the size of the differences in performance on CR items between student groups matched on the total score. The following equation defines SMD:

$$SMD = \sum_k w_{Fk} m_{Fk} - \sum_k w_{Fk} m_{Rk}$$

where $w_{Fk} = n_{F+k} / n_{F++}$ is the proportion of focal group members who are at the k th stratification variable, $m_{Fk} = (1/n_{F+k}) F_k$ is the mean item score for the focal group in the k th stratum, and $m_{Rk} = (1/n_{R+k}) R_k$ is the analogous value for the reference group. The SMD is the difference between the unweighted item mean of the focal group and the weighted item mean of the reference group. The weights applied to the reference group are applied so that the weighted number of reference group students is the same as in the focal group (within the same ability group). The SMD is divided by the total group item standard deviation to get a measure of the effect size for the SMD using the following equation:

$$\text{Effect Size} = \frac{\text{SMD}}{SD}$$

The SMD effect size allows each item to be placed into one of three categories: negligible DIF (AA), moderate DIF (BB), or large DIF (CC). The following rules are applied for the classification (Allen, Carlson & Zalanak, 1999). Only categories BB and CC were flagged in the results.

- The item is classified as CC category if the probability is <.05 and if |Effect Size| is >.25.
- The item is classified as BB category if the probability is < .05 and if .17<|Effect Size|≤.25.
- The item is classified as AA category if the probability is >.05 or |Effect Size| is ≤ .17.

Table 13 summarizes the results of the DIF analysis appearing in Appendix B for SR (B/C) and BCR (BB/CC) items. Items with a statistical indication of DIF were reviewed for bias by subject matter experts during data review. It should be noted that "Total" in Table 13 reflects total items flagged based on the largest DIF classification level. That is, items flagged at both the B and C would be counted as "C" in Table 13.

Table 13. DIF Flag Summaries from all MSA Science Field Test Items

	DIF Classification Level				
Grade	**B**	**BB**	**C**	**CC**	**Total**
5	12	4	1	2	19
8	5	3	0	4	12

Data Review of the Field Test Items

Background

Data review represents a critical step in the test development cycle. Pearson psychometricians provided a list of flagged items for the 2012 MSA Science field test data review based on the following criteria:

SR items will be flagged if:

- P-value < .10 or P-value > 0.90
- Point biserial correlation < 0.30
- Item omission > 5%
- Incorrect distractor p-value > 0.40
- Incorrect distractor point biserial correlation > 0.05
- 100% non-response to any distractor
- IRT *a* parameter < 0.50
- IRT *b* parameter < -4.00, or IRT *b* parameter > 4.00
- IRT *c* parameter > 0.50
- B or C level DIF

BCR items will be flagged if:

- BCR mean < 0.30 or BCR mean > 2.70
- Point biserial correlation < 0.30
- Any score point where 0% of students earn that score
- IRT *a* parameter < 0.50
- IRT *b* parameter < -4.00, or IRT *b* parameter > 4.00
- IRT step values (*d*) < -4.00, or IRT step value > 4.00
- BB or CC level DIF

The flagged items were reviewed by Pearson Content team and MSDE content experts. The final decision about the suppression of the flagged items was made in collaboration between MSDE and Pearson.

Results of Data Review

A total of 71 items in grade 5 and 70 items in grade 8 were inspected during data review as a result of the item not meeting the statistical flagging criteria. Eleven of the 71 total flagged items were rejected from the grade 5 pool and eight of the 70 flagged items for grade 8 were rejected.

Validity

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

Messick (1989) defined validity as follows:

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

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

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

To investigate the validity evidence of the 2012 MSA-Science tests, content-related evidence, differential item functioning (DIF) analysis on gender and ethnicity, and evidence based on internal structure were collected.

Content-related Evidence

Content related 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).

As described in the Item Development and Review section, all MSA Science items were explicitly developed to measure the specific knowledge and skills described in the Maryland State Curriculum. As noted, the alignment of the items to the six Science standards was reviewed and verified independently by multiple content experts to include Pearson staff, MSDE staff, and Maryland educators.

The Test Overview and Design section details the connection between the MSA Science blueprint and the MSC. The 2012 MSA Science tests were constructed exclusively using items that met not only the statistical criteria described in this report, but also verified as aligning to the MSC by Maryland science content experts. As described, tests were constructed according to the test blueprints and as such, scores provided are reflective of overall Science ability as defined within the state standards.

Differential Item Functioning (DIF)

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 the use of language or context that might offer an advantage or disadvantage to particular subpopulations within 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 approaches incorporate both perspectives when reviewing test questions with respect to fairness. Bias and sensitivity committee review of all field tested items occurs each year as described in the Item Development and Review section.

DIF analyses are carried out on all MSA Science field test items according to the procedures in the Differential Item Functioning Analysis section. 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 analysis, the total raw score on the operational items is used as the ability-matching variable. Any items displaying DIF that are also judged to contain language or context favoring or disadvantaging a given subpopulation are removed from the pool of eligible items during data review. Because of this ongoing and thorough approach, the majority of items on the MSA Science operational tests exhibit no DIF or weak DIF, and no items judged to show bias are selected for operational use.

Inter-Correlations among Standards

There are six standards within the MSC frameworks for MSA Science that together contribute to the overall reported Science test score. Items are written to capture performance that not only reflects the overall construct of science as defined within the frameworks, but to capture content and skills by standard. To assess the extent to which items aligned with the standards are offering some unique characteristics based on each respective standard, while more strongly capturing an overall "science" construct, a correlation matrix was computed among the total scores of competencies. It should be noted that only overall scale scores and performance levels are reported for MSA Science.

Table 15 reports the correlations among the six standards based on scale scores. The standard-level (subtest) inter-correlations ranged from 0.53 to 0.85 where most are greater than .60. The standard subscores are moderately highly related to one another and more strongly related to the total test score. This suggests there is some uniqueness to items grouped by standard but that they are collectively measuring a dominant overall construct (science).

Table 15. Correlation among MSA Science content standards

Grade 5 Form A	Mean	SD		Str1	Str2	Str3	Str4	Str5	Str6	Total
	414.80	62.60	**Str1**	1.000						
	422.20	79.25	**Str2**	0.577	1.000					
	413.40	62.05	**Str3**	0.645	0.580	1.000				
	413.41	62.08	**Str4**	0.638	0.568	0.621	1.000			
	429.23	83.79	**Str5**	0.561	0.526	0.530	0.548	1.000		
	414.89	74.69	**Str6**	0.618	0.578	0.607	0.607	0.547	1.000	
	412.39	46.91	**Total**	0.836	0.769	0.819	0.816	0.729	0.806	1.000
Grade 5 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	415.24	61.05	**Str1**	1.000						
	413.56	59.80	**Str2**	0.642	1.000					
	408.57	81.62	**Str3**	0.580	0.592	1.000				
	414.79	73.49	**Str4**	0.613	0.616	0.575	1.000			
	419.68	81.18	**Str5**	0.551	0.563	0.530	0.572	1.000		
	414.17	67.51	**Str6**	0.635	0.638	0.577	0.605	0.568	1.000	
	410.79	46.56	**Total**	0.820	0.830	0.762	0.809	0.750	0.817	1.000
Grade 8 Form A				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	419.46	72.13	**Str1**	1.000						
	418.79	81.37	**Str2**	0.578	1.000					
	414.83	64.17	**Str3**	0.648	0.611	1.000				
	408.40	72.93	**Str4**	0.599	0.567	0.633	1.000			
	409.31	83.91	**Str5**	0.572	0.571	0.598	0.571	1.000		
	415.47	65.88	**Str6**	0.646	0.602	0.663	0.613	0.588	1.000	
	412.61	48.04	**Total**	0.804	0.780	0.842	0.787	0.764	0.838	1.000
Grade 8 Form B				**Str1**	**Str2**	**Str3**	**Str4**	**Str5**	**Str6**	**Total**
	411.95	68.40	**Str1**	1.000						
	401.14	77.08	**Str2**	0.628	1.000					
	409.51	66.75	**Str3**	0.667	0.656	1.000				
	406.82	78.45	**Str4**	0.585	0.563	0.595	1.000			
	409.41	79.49	**Str5**	0.650	0.642	0.669	0.595	1.000		
	435.99	106.21	**Str6**	0.596	0.591	0.606	0.550	0.599	1.000	
	408.72	50.50	**Total**	0.824	0.817	0.850	0.748	0.825	0.762	1.000

*Str1=Skills and Processes; Str2=Earth/Space Science; Str3=Life Science; Str4=Chemistry; Str5=Physics; Str6=Environmental

Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was conducted for the 2012 MSA Science tests to examine the relationship between the subtest scores relative the total test score. Subtest raw scores were used for this analysis. CFA used SAS Proc Calis and the maximum likelihood estimation (MLE; Anderson & Gerbing, 1988) procedure. The model hypothesized that the subtest scores belong to a single latent trait. Model fit was tested through indices including adjusted goodness of fit (AGFI), and Root Mean Square Error of Approximation (RMSEA). Values of the AGFI statistic that 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 ≤ .06 as the guideline for close fit. Table 16 summarizes fit indicators estimated from the confirmatory factor analysis for the 2012 MSA Science tests. The confirmatory factor analysis results provide additional evidence to support the conclusion that scores from the MSA Science tests reflect a single latent trait (Science). For both grades, the lowest AGFI was 0.9873, and the highest RMSEA was 0.0419. The AGFI and RMSEA indicators supported the model fit.

Table 16. Fit indicators for confirmatory factor analysis on MSA Science

Grade/Form	AGFI	RMSEA
Grade 5 Form A	0. 9946	0. 0273
Grade 5 Form B	0. 9985	0. 0138
Grade 8 Form A	0. 9887	0. 0396
Grade 8 Form B	0. 9873	0. 0419

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

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 standards was repeated using only students testing with accommodations and then again using only students testing without accommodations. The results of this analysis showed comparable levels of model fit based on the two groups (see Table 17). This suggests that the accommodations offered to disabled students are effective at preserving the underlying latent structure of the MSA Science tests in comparison to that standard (non-accommodated) administration. By extension, MSA Science scores for accommodated and non-accommodated students are comparable.

Table 17. Fit indicators for accommodations/non-accommodations based CFA

	Accommodations		No Accommodations	
Grade/Form	AGFI	RMSEA	AGFI	RMSEA
Grade 5 Form A	0. 9900	0. 0332	0. 9948	0. 0269
Grade 5 Form B	0. 9993	0. 0000	0. 9982	0. 0149
Grade 8 Form A	0. 9837	0. 0439	0. 9896	0. 0379
Grade 8 Form B	0. 9921	0. 0307	0. 9878	0. 0409

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