Calibration and Equating

MSA Science Linking, Calibration and Equating

The MSA Science operational test was calibrated using data obtained from the 2008 operational test. A concurrent calibration design was used to calibrate all core items in two base forms in a single calibration.

Dichotomously scored SR items were calibrated using the three-parameter logistic (3-PL) IRT model (Lord & Novick, 1968) and polytomously scored BCR items used the Generalized Partial Credit model (GPC; Muraki, 1992). The MSA Science assessments were calibrated using a concurrent calibration design. All concurrent calibrations were performed with students who tested without accommodations. All 3-PL/GPC model items were calibrated using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003), which can estimate parameters simultaneously for dichotomous and polytomous items via marginal maximum likelihood procedures. Due to the 20 items in common between the two base forms simultaneous calibration of the items automatically placed their parameter estimates on the same underlying measurement scale.

Analysis of Operational Test Data

Using the data collected from the 2008 operational test, Pearson computed Classical Test Theory statistics and performed a concurrent calibration of the incomplete data matrix. All analyses resulting from the operational test were then screened and flagged for undesirable psychometric properties. Flagged items were presented to MSDE and Pearson content specialists for review to ensure that items were keyed properly. No miss-keyed items were identified on either of the MSA Science tests.

Some of the results from the analyses included the following Classical Test Theory statistics:

- **P-Value:** proportion of students who answered the item correctly. An item's p-value shows how difficult the item was for the students who took the test.

- **Mean of BCR item:** This is a measure of the difficulty of the BCR items, in Classical Test Theory and is indicated by the average raw score for a BCR item across all students from the rubric ratings. MSA Science rubrics range from 1 to 3, with a 0 score indicating no response. As a result, the average item score for all MSA Science BCR items falls between 0 and 3.

- **Point-Biserial Correlation (Pt Bis):** describes the relationship between a student's performance on the item (correct or incorrect) and the student's performance on the subject area test form as a whole (number of correct items on the test form).

- **Item Option Point-Biserial:** provide information about the relationship of a particular response option and overall performance on the test. An expected pattern of item option biserials is that incorrect item options should have lower values (typically negative) than the correct item option.

- **Frequency Distribution of Item Options by Group:** These data provide information about how the lower third, middle third and upper third responded to items by response option. These distributions allow for comparisons among the different performance levels.

- **Mean Score by Response Option:** These data indicate the overall raw test score of students by response option.

- **Differential Item Functioning (DIF)**: The information will assist in examining differential item performance across the African American, Asian, White, and Hispanic groups and across the male and female groups. The Mantel-Haenszel Delta is a statistical approach to indicate possible differential item performance and was used here. The Mantel-Haenszel Delta statistic indicates a differential likelihood of similarly performing students from different ethnic or gender groups answering the item correctly.

The following IRT analyses were also completed:

- **Item Parameter Estimates.** Discrimination, difficulty, and guessing parameters for each SR item were computed based on the 3-Parameter Logistic IRT Model. The item characteristic curve for each item was plotted. Discrimination and difficulty parameters were computed for each BCR item using the Generalized Partial Credit model.
- **Standard Error Estimate.** The standard error of item parameter estimates was computed for each item parameter estimate.
- **Item Fit Estimate.** The extent to which the IRT model conforms to the data was estimated item by item.

The following criteria were used to designate items as potentially unsuitable:

- P-value < 0.25 or >.90
- Point-biserial < 0.15
- Point-biserial for distracter > 0.05
- DIF at B or C level
- IRT a parameter < 0.30
- IRT b parameter < -4.0 or >+4.0
- IRT c parameter > 0.30

Testing Population

Maryland Students in grade 5 and 8 took the Science operational test as part of the MSA program. Mode of testing (paper versus online administration) was determined by each school. The number of students per form, including demographic breakdowns and accommodations for grade 5 and grade 8 appear in Tables 3 and 4, respectively.

Table 3. Demographic characteristics of grade 5 and grade 8 sample for overall, online, and paper

	Grade			
	5		**8**	
	N	**%**	**N**	**%**
Mode of Administration				
Online	25728	42	21886	35
Paper	35033	58	41585	65
Form				
1	6627	10.91	6096	9.60
2	5882	9.68	6180	9.74
3	7466	12.29	6189	9.75
4	5925	9.75	6018	9.48
5	5842	9.61	6099	9.61
6	5902	9.71	7294	11.49
7	5813	9.57	6168	9.72
8	5693	9.37	6223	9.80
9	5796	9.54	7108	11.20
10	5815	9.57	6096	9.60
Gender				
Female	29518	48.58	30858	48.62
Male	31226	51.39	32573	51.32
Ethnicity				
Unknown	21	.03	38	.06
Native	224	.37	255	.40
Asian	3605	5.93	3429	5.40
African American	22763	37.46	24540	38.66
White	28522	46.94	30055	47.35
Hispanic	5626	9.26	5154	8.12
All	**60761**	**100**	**63471**	**100**

* Differences in values reflect missing data

Distribution of Students Across Forms

Forms were spiraled at the student level. Forms were spiraled within mode of administration so that there would be an even distribution of forms.

Table 4. Distribution of forms by grade

		Form									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Grade 5	Online	3919	3290	4850	3362	3277	3333	3246	3216	3252	3288
	Paper	2708	2592	2616	2563	2565	2569	2567	2477	2544	2527
	Overall	6627	5882	7466	5925	5842	5902	5813	5693	5796	5815
Grade 8	Online	3931	4007	4048	3831	3949	4791	4009	4097	4973	3949
	Paper	2165	2173	2141	2187	2150	2503	2159	2126	2135	2147
	Overall	6096	6180	6189	6018	6099	7294	6168	6223	7108	6096

Analysis

Following the processing of answer documents, student demographic and item response data were transmitted to Pearson's psychometric services division. Pearson psychometric staff had primary responsibility for analyzing MSA Science data to ensure accuracy and validity of scoring. Most of the psychometric work was carried out using SAS Version 9.1 and MULTILOG 7.0, commercially available statistical analysis software. Traditional item analysis and data file QC analyses were conducted with SAS programs. Item response theory (IRT) analyses were conducted with the MUTLTILOG program (Thissen, Chen, & Bock, 2003). MULTILOG allows for estimation of IRT item parameters for dichotomously or polytomous scored items. It has been thoroughly tested and is currently utilized by several high-stakes testing programs administered by Pearson.

All technical support and analyses were carried out in accordance with both the *Standards* (AERA, APA, & NCME, 1999) and the Pearson Quality Assurance Program. Pearson staff verified the MSA Science data and analysis process at several steps in the procedure. This included verification of the SAS and MULTILOG programs prior to use on actual field data through review by a second member of the psychometric services staff and by using simulated data sets. Additionally, the output from the traditional and IRT item analysis programs were verified for out of range values and for consistent results across programs.

Pearson conducted extensive statistical analyses on all field items. These analyses showed which items were at an appropriate difficulty level for the testing population and screened for differential item functioning (DIF) for subgroups in the student population. The analysis of the test data is broken down into several components: 1) classical item analyses; 2) DIF analyses; 3) reliability analyses; and 4) calibration of items for bank values to be used in test construction. In the following sections, the analysis procedures for each component are described in detail. Tables summarizing the analyses are provided at the end of the chapter.

Classical Item Analyses

The following classical item statistics that were calculated:

- P-value of SR items
- Mean of BCR
- BCR Standard Deviation
- Point-Biserial Correlation
- Item Option Point-Biserial
- Frequency Distribution of Item Options by Group
- Mean Score by Response Option

The classical statistics for the 2008 MSA operational and field test items are reported in Appendix A.

Differential Item Functioning (DIF) Analyses

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 the same total 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 analyses, 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.
- 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

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

$m_{Rk} = (1/n_{+Rk})R_k$ is the analogous value for the reference group. In words, 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:

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

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. 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.

The data in Table 5 summarize the number of field-test items in DIF categories for the grade 5 and 8 items and the full results of the DIF analyses appear in Appendix B. Items with a statistical indication of DIF were reviewed for bias by subject matter experts during data review.

Table 5. DIF flag incidence across all MSA Science field-test items

	Grade 5 (221 items total)		**Grade 8 (220 items total)**	
	DIF Level		**DIF Level**	
Group indicating DIF for FT Item	**B**	**C**	**B**	**C**
Black	0	0	2	0
Hispanic	1	0	2	1
Female	1	0	7	3
Total	2	0	11	4
Group indicating DIF for OP Item	**B**	**C**	**B**	**C**
Black	4	0	0	2
Hispanic	1	0	1	0
Female	3	0	4	0
Total	8	0	5	2

Test Score Reliability

The reliability of a test provides an estimate of the extent to which an assessment will yield the same results across subsequent administrations, provided the two administrations do not differ on relevant variables. Reliability coefficients are usually forms of correlation coefficients and must be interpreted within the context and design of the assessment and of the reliability study. The forms of reliability below measure different dimensions of reliability and thus any or all might be used in assessing the reliability of MSA Science.

The estimates of reliability reported in this report are internal consistency measures, which are derived from analysis of the consistency of the performance of individuals on items within a test (internal consistency reliability). Therefore, they apply only to the test form being analyzed. They do not take into account form-to-form variation due to equating limitations or lack of parallelism, nor are they responsive to day-to-day variation due, for example, to state of health or testing environment.

This is the formula for the most common index of reliability, namely, Cronbach's coefficient *alpha* (α; Cronbach, 1951). In this formula, the s_i^2's denote the variances for the k individual items; s_{sum}^2 denotes the variance for the sum of all items.

$$\alpha = (k/(k-1)) * [1- \Sigma(s^2_i)/s^2_{sum}]$$

Because of the mixed item types on the MSA Science test (i.e., MC and BCR), a stratified alpha (Feldt & Brennan, 1989) was computed. A stratified alpha is based on a weighted average of Cronbach's alpha for each item set. These results are in Table 6.

Table 6 Reliability estimate by form

Form	Grade 5	Grade 8
Base Form A	.91	.94
Base Form B	.91	.93

The coefficient alpha estimates for all forms meet conventional guidelines and legal benchmarks for applied test reliability (i.e., α > .85).

IRT Analysis

Pearson estimated IRT parameters for all MSA Science items to establish the underlying theta scale. These parameter estimates will serve to calibrate students who tested without accommodations and test items onto the same underlying scale. The 3-PL model SR items and the GPC model for BCR items were selected because of the mixed format (i.e., multiple-choice and constructed response or polytomous items) of the test.

Dichotomous Item Response Theory Model

For the SR items, or dichotomously scored items, calibration was done using Birnbaum's 3-PL item response theory (IRT) model (Lord & Novick, 1968). The formulation of the 3-PL model is presented below:

$$P_i(\theta) = c_i + (1 - c_i)\frac{1}{1 + e^{-Da_i(\theta - b_i)}} \qquad (1)$$

where θ (theta) is the student proficiency parameter, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the lower asymptote parameter and D is a scaling constant. The scaling constant is traditionally 1.7. With multiple-choice items it is assumed that, due to guessing, examinees with minimal proficiency have a probability greater than zero of responding correctly to an item. This probability is represented in the 3-PL model by the c_i parameter.

Polytomous Item Response Theory Model

For the BCR items, or polytomously scored items, calibration was done using the GPC model (Muraki, 1992). For an item j with m_j possible scores (0, 1, . . . , m_j-1), the GPC model gives the probability of response r as a function of latent variable θ as

$$\Pr(X_j = r \mid \theta) = \frac{e^{z_{jr}}}{1 + \sum_{k=0}^{m_j-1} e^{z_{jk}}}, \tag{2}$$

where

$$z_{ji} = \sum_{k=0}^{i} a_j(\theta - b_j + c_k), \tag{3}$$

X_j is a random variable representing a response to item j and a_j, b_j and c_k, k = 0, 1, 2, ... , m_j-1 are item parameters.

Calibration of the mixed test format (3PL/GPC model) was conducted using MULTILOG 7.0 (Thissen, Chen, & Bock, 2003) and included only the students in the population who:

- Tested without accommodations,
- attempted at least one item on the test,
- attempted at least one BCR item, and
- the student's score was not invalidated.

MULTILOG estimates parameters simultaneously for dichotomous and polytomous items via marginal maximum likelihood procedures.

Item Calibration and Equating

The purpose of item calibration and equating is to create a common scale (theta) for expressing the item parameter estimates across versions of a test. The theta distribution is commonly scaled to have the mean set to 0 and the standard deviation set to 1. This scale is not often used for reporting because of interpretation issues arising from a scale with values typically ranging from -4.0 to +4.0. Therefore, following calibration and equating, the scale was transformed to a reporting scale which can be meaningfully interpreted by students, teachers and other stakeholders.

The following IRT analyses were completed for all items and are reported in Appendix A.

- The *a* parameter estimation for both SR and BCR items.
- The *b* parameter estimation for both SR and BCR items.
- The *c* parameter estimation for SR items only.
- Category step values (*d*) for BCR items only.
- The mean total-test theta estimate for all students earning a given score point for each category for BCR items only.

The item parameter estimates for the 2008 core items were equated to the base scale (established during the calibration of the 2007 census field test) using the 2007 item parameters from the census field test. Since the entire set of 2008 core items were field tested in 2007, two item parameter sets for the core items were available - one from the 2007 field test calibration and another from the 2008 calibration. A publicly available equating program, STUIRT (Kim & Kolen, 2004), was used to calculate equating constants using the Stocking and Lord Procedure. In order to place the 2008 field test items on the base scale the 2008 operational items were calibrated concurrently with the field test items. These new operational parameters were then used, along with the equated 2008 operational parameters, to calculate equating constants with the Stocking and Lord Procedure using STUIRT. The equating constants are listed in Table 7.

Table 7. Equating constants for operational and field test

	Grade 5		**Grade 8**	
	Slope	**Intercept**	**Slope**	**Intercept**
Operational (08 OP items -> 07 FT items)	1.015867	0.216272	1.08793	0.208599
Field Test (08 FT items -> 08 OP items)	1.006532	0.215406	1.065945	0.205523

The equating constants were applied to the 2008 item parameters so that all items in the MSA Science pool can be put onto the same theta metric. The complete IRT estimates for students in grade 5 and 8 who met the criteria for inclusion in the equating sample are in Appendix A.

Data Review of the Field Test Items

Background

Data review represents a critical step in the test development cycle. The 2008 MSA Science field test data review procedure was different from that of the 2007 field test. Instead of formal data review meeting, Pearson Psychometric team provided the list of flagged items for the criteria based on the following criteria:

For SRs:

1. Omit rate > 5%
2. 0.10 > p value < 0.90
3. Point biserial < 0.10
4. Non-responses to any one of the distractors
5. DIF flag with C

For BCRs:

1. Omit rate > 20%
2. Non-response to any of the rubric score point
3. p value <= 0.10 or >= 0.90
4. Item total correlation < 0.10
5. DIF indicator with CC

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 19 items in grade 5 and 12 items in grade 8 were inspected during data review as a result of the item not meeting the statistical flagging criteria for the classical item analyses and DIF. Three of the 19 total flagged were rejected from the grade 5 pool and one of the 12 flagged items for grade 8 was rejected.