Operational Item Analysis and Equating

Testing Population

Maryland Students in grade 5 and 8 took the Science operational test as part of the MSA program. Mode of testing (whether a test is administered by paper or via 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 4 and 5, respectively.

Table 4. Demographic Characteristics of Grade 5 and Grade 8 Sample for Overall, Online, and Paper

	Grade			
	5		8	
	N	%	N	%
Mode of Administration				
Online	32262	53.08	38540	63.08
Paper	28515	46.92	22560	36.92
Form				
1	5779	9.51	5809	9.51
2	5747	9.46	5886	9.63
3	5775	9.50	5745	9.40
4	5829	9.59	5812	9.51
5	5721	9.41	5907	9.67
6	8487	13.96	8636	14.13
7	5889	9.69	5776	9.45
8	5832	9.60	5779	9.46
9	5823	9.58	5842	9.56
10	5895	9.70	5908	9.67
Gender				
Female	29832	49.09	29918	48.98
Male	30941	50.91	31169	51.02
Unknown	4	**	13	**
Ethnicity				
Hispanic/ Latino	7214	11.87	6445	10.55
Non-Hispanic/ Latino	53559	88.13	54642	89.45
Race				
American Indian	177	0.33	188	0.34
Asian/Pacific Islander	3663	6.84	3526	6.45
African American	21358	39.88	21994	40.25
Native Hawaiian	64	0.12	78	0.14
White	25833	48.23	26775	49.00
Two or More Races	2464	4.60	2081	3.81
All	**60777**	**100**	**61100**	**100.00**

Note: differences in values reflect missing data
** less than 0.001

Distribution of Students across Forms

As described, MSA Science test forms are composed of a set of operational items and field test items. Ideally, each respective test form will be administered to randomly equivalent groups of students. This helps ensure that any item and test level statistics are more directly comparable. The administration of multiple test forms is commonly referred to as "spiraling." The MSA Science test forms were spiraled at the student level and within mode of administration so that there would be an even distribution of tests across forms. Table 5 presents this distribution of tests across forms by mode of administration at each grade. Within-form overages (i.e. Grade 5 online Form 6) reflect the inclusion of additional forms for special accommodations (i.e. read-aloud, audio presentation, etc.).

Table 5. Distribution of Forms by Grade

		Form									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Grade 5	Online	2958	2933	2977	2994	2893	5493	3051	2970	2970	3023
	Paper	2821	2814	2798	2835	2828	2994	2838	2862	2853	2872
	Overall	5779	5747	5775	5829	5721	8487	5889	5832	5823	5895
Grade 8	Online	3679	3709	3563	3597	3704	5811	3587	3560	3632	3698
	Paper	2130	2177	2182	2215	2203	2825	2189	2219	2210	2210
	Overall	5809	5886	5745	5812	5907	8636	5776	5779	5842	5908

Key Check Analysis of Operational Test Data

Using preliminary data collected from the 2012 operational test (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). Any items identified during this process were presented to Pearson content specialists for review to ensure that items were keyed properly. All operational MSA Science items were confirmed as correctly keyed and functioning sufficiently within the statistical parameters (described below) to conduct the classic and IRT analysis described in the next sections.

The key check analysis 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.

- **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).

- **P-Value by Response Option:** These data indicate the proportion of students who selected each response option.

The following criteria were used to designate items as potentially mis-keyed:

- P-value < 0.15
- Point-biserial < 0.20
- P-value for a single unkeyed response >= .40

Analysis

Following the complete processing of answer documents, student demographic and item response data were transmitted to Pearson's Psychometric and Research 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 analysis were conducted with SAS programs. Item response theory (IRT) analysis 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 analysis 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.

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 operational items are reported in Appendix A.

IRT Calibration

Pearson used a concurrent calibration IRT estimation procedure for placing all Form A and Form B operational MSA Science items on a common theta scale that was then equated to the original 2007 base scale (as described in the next section). The 3 parameter logistic (3-PL) model was used for SR items and the generalized partial credit (GPC) model was used for BCR items because of the mixed format of the test (i.e., multiple-choice and constructed response or polytomous items).

<u>Dichotomous Item Response Theory Model</u>

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

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

X_j is a random variable representing a response to item *j*, a_j is item discrimination, b_j is the item location parameter, and d_k, is a threshold or "step" difficulty for $k = 0, 1, 2, \ldots, m_j-1$ thresholds denoting the intersections of the respective m_j response functions.

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

- attempted at least one item on the test,
- attempted at least one BCR item, and
- had a student score that was not invalidated.

MULTILOG estimates parameters simultaneously for dichotomous and polytomous items via marginal maximum likelihood procedures. As mentioned in the test design section of this document, the MSA Science tests utilize two operational forms (Form A and Form B) per grade with a set of 20 items common to both forms. This set of 20 items was used to create an incomplete data matrix so that the unique items from each form could be calibrated concurrently, thus placing the parameters for all operational items administered at each grade on a common scale.

Equating

The purpose of equating is to maintain a common scale (theta) for expressing the item parameter estimates across versions (i.e., annual administrations) of a test. The theta distribution is commonly scaled to have the mean set to 0 and the standard deviation set to 1. Once the 2012 MSA Science tests were concurrently calibrated, it was necessary to place each respective scale (Grade 5 and Grade 8) onto the originating 2007 base scale. This was carried out using what is referred to as a common item, non-equivalent groups design (CINEG; Kolen & Brennan, 2004). In this case, the common item sets from the operational forms consisted of *all* operational SR items. That is, all operational items aside from BCRs served as linking items back to the base scale. For the item parameter estimates reflecting the base form, the most current parameter estimates were used, whether from the 2007 through 2011 field test calibrations or from the 2008 through 2011 operational administrations.

When conducting equating with nonequivalent groups, the parameters from different forms (Form X and Form Y) need to be placed on the same IRT scale. This can be accommodated under the IRT framework, because when the IRT model holds, the parameter estimates from different groups are on linearly related theta scales (Lord, 1980). Thus, a linear equation can be used to place IRT parameter estimates onto an existing (base) scale. A publicly available equating program, STUIRT (Kim & Kolen, 2004), was used to calculate transformation constants from the Stocking and Lord Procedure. In the Stocking and Lord approach (Stocking & Lord, 1983), the difference between two test characteristic curves is first squared for a fixed theta value:

$$\boldsymbol{SLdiff}(\theta_i) = \left[\sum_{j:V} \boldsymbol{P}_{ij}(\theta_{Yi}; \hat{\boldsymbol{a}}_{Yj}, \hat{\boldsymbol{b}}_{Yj}, \hat{\boldsymbol{c}}_{Yj}) - \sum_{j:V} \boldsymbol{P}_{ij}(\theta_{Yi}; \frac{\hat{\boldsymbol{a}}_{Xj}}{\boldsymbol{A}}, \boldsymbol{A}\hat{\boldsymbol{b}}_{Xj} + \boldsymbol{B}, \hat{\boldsymbol{c}}_{Xj}) \right]^2 .$$

The estimation proceeds by finding the combination of *A* and *B* minimizing the following criterion:

$$\boldsymbol{SLcrit} = \sum_{i} \boldsymbol{SLdiff}(\theta_i) ,$$

where the summation is over examinees. An iterative approach needs to be used to solve for *A* and *B* in the above equations.

Stability Check Procedure

Dramatic changes in item parameter values can result in systematic errors in equating results (Kolen & Brennan, 2004). It is customary to track changes in item parameters and to evaluate how those changes affect the results of equating. Thus, it was necessary to examine the stability of the MSA Science anchor item parameters after equating. Specifically, Pearson evaluated stability in the operational linking item parameters by examining differences in the originating (base) and transformed item characteristic curves. All items used for linking the 2012 MSA Science tests to the base scales were included in this stability check.

Pearson used an iterative anchor stability check approach that is analogous to examining differential item functioning. The steps of this process are as follows:

1) Place the current item parameters for all anchor items on the base-year scale by computing Stocking & Lord (SL) transformation constants using STUIRT (Kim & Kolen, 2004) and all anchor items.
2) For each linking item, calculate the weighted sum of the squared deviation (d^2) between the Item Characteristic Curves (ICC) using a theoretical weighted posterior theta distribution with 40 quadrature points:
    a) Apply the SL constants to the thetas associated with the standard normal theta distribution used to generate the SL constants.
    b) For each anchor item calculate a weighted sum of the squared deviation between the ICCs based on old (x) and new (y) parameters at each point in this theta distribution.

$$d_i^{\ 2} = \sum^{k} \left[P_{ix}(\theta_k) - P_{iy}(\theta_k) \right]^2 \bullet g(\theta_k)$$

c) Compute the mean and standard deviation of the d^2 values, and flag any item with a d^2 more than two standard deviations above the mean.
d) Review and sort the items in a descending (largest to smallest) fashion according to the d^2 value.
e) Step 2d) results in an item with the largest area between pre- and post-equated ICCs at the top of the list of anchor items:
   i) Drop the largest d^2 item from the anchor set.
   ii) Repeat steps 1 through 2d – omitting 2c (use the original mean and standard deviation) until no more items are flagged or more than 20% of the operational items appearing across the two OP forms will be dropped.
f) Review all dropped items with a d^2 flag to determine at what point in the process no more items should be dropped. Items not flagged in this process should not be dropped, but a flag alone is not the sole criteria for removing an item from the linking set. In other words, the flag is a necessary, but not sufficient criterion for dropping an anchor item.

Flagged items were further reviewed through examination of the classical item analysis, IRT estimates, item characteristic curves, fit statistics, item sequence change (change from location of the most recent administration), and impact on the test blueprint representation. Any item considered for removal was evaluated by a Pearson Content Specialist to determine of the content of the item or an event in the item's development history might explain the change in item performance. Decisions about whether to keep or remove an item were evaluated on a per item basis. When an item (note, only one item can be removed at a time) was removed from the anchor set, then this process (beginning with the computation of transformation constants) was repeated until there were no further items to be removed.

This process resulted in three items removed from grade 5 and four items removed from the grade 8 common item sets. The final transformation constants for each grade following this procedure are listed in Table 6.

Table 6. Operational Transformation Constants

	Grade 5		**Grade 8**	
	Slope	**Intercept**	**Slope**	**Intercept**
Operational (12 OP items >> 07 base scale)	1.035406	0.265267	1.069871	0.284065

The transformation constants were applied to the 2012 item parameters so that all items in the MSA Science pool can be put onto the original base scales. The equated IRT parameters for grade 5 and 8 items are presented in Appendix A.