Section 4. Scoring Procedures

Scale Scores

The MD HSA reporting scale ranges from 240 to 650. For Algebra, Biology, and Government, the scale was established in 2003 and defined so that the scale scores had a mean of 400 and a standard deviation 40. In 2005 a scale for English was established that had the same range, mean, and standard deviation.

These scores represent ability estimates obtained using Item Response Theory (IRT). (See IRT Calibration and Scaling in Section 7 for details about the three-parameter logistic model used for the MD HSAs.)

Students' total test scores are scale scores derived using the three-parameter logistic (3PL) model and item-pattern (IP) scoring procedures. In the October, January, and April administrations, students' subscores were based on raw score to scale score (RS–SS) conversion tables.[6] Since May 2009, students' subscores have been based on IP scoring.

When the 3PL model is used, the likelihood equation can have multiple maxima. Therefore, a numerical method was developed to find the scale score at the global maximum in the likelihood function. The RS–SS scoring tables were obtained by taking the inverse of the test characteristic curve (TCC) for items contributing to each subscore (Yen, 1984).

Conditional Standard Errors of Measurement

Corresponding conditional standard errors of measurement (CSEM) were produced for both types of scoring and were equal to the inverse of the square root of the test information function.

$$\mathrm{CSEM}(\hat{\theta}) = \frac{1}{\sqrt{I\left(\hat{\theta}\right)}},$$

where CSEM($\hat{\theta}$) refers to the conditional standard error of measurement and I($\hat{\theta}$) refers to the test information function for $\hat{\theta}$.

The test information function is the sum of corresponding information functions of the test items when optimal item weights are used. Item information functions depend on the item difficulty, discrimination, and conditional item score variance. Thus, while polytomous items often have lower discriminations than selected response items (Fitzpatrick et al., 1996), they may convey more information because they have more score points.

[6] For operational scoring, omitted responses on the MD HSA are assigned the lowest score—SR/SPR items are treated as incorrect and CR items are assigned an item score of 0.

Lowest and Highest Obtainable Test Scores

The maximum likelihood procedure under the 3PL model cannot produce reasonable scale score estimates for students with perfect scores or scores below the level expected by guessing. While maximum likelihood estimates are usually available for students with extreme scores other than zero or perfect, occasionally these estimates have very large CSEMs, and differences between these extreme values have little meaning. Therefore, scores were established for these students based on a rational procedure (refer to Appendix 3.C of the 2004 Technical Report). These values were called the lowest obtainable scale score (LOSS) and the highest obtainable scale score (HOSS). The same LOSS and HOSS values were used for RS-SS tables and for IP scoring. Starting with the summer 2005 administration, MSDE decided that the LOSS and HOSS values would be 240 and 650, respectively, for all content areas.

Cut-Scores

MSDE established the cut-scores associated with each of the performance levels in the content areas other than English in 2003.[7] The English cut-scores were established during the standard-setting meeting held in October 2005. One cut-score was established for Biology and one was established for Government. Because Algebra and English results are used as the high school mathematics and English/language arts components of the Maryland accountability plan under NCLB, two cut-scores were established for these content areas. To comply with NCLB requirements for secondary science, an Advanced cut score for Biology was established in 2008. These values are given in Table 4.1.

Table 4.1 MD HSA Cut-Scores by Content Area

	Cut-Score	
Content Area	Proficient	Advanced
Algebra	412	450
Biology	400	452
English	396	429
Government	394	

Beginning with the class of 2009, students must obtain either a passing score on all four MD HSAs or an overall combined score of 1602. Passing status is achieved when a student’s score meets or exceeds the Proficient cut score, as listed in Table 4.1. Students graduating prior to 2009 were not required to pass the MD HSAs but were required to take the tests.

[7] Technical documentation on the standard-setting method used to establish the MD HSA cut-scores is available on the Maryland State Department of Education website at http://www.marylandpublicschools.org/msde/divisions/planningresultstest/maryland+standard+setting+technical+reports.htm.

Year-to-Year Scale Maintenance

The MD HSAs for Algebra, Biology, and Government have been pre-equated since 2004; English has been pre-equated since 2005. In the pre-equated design, a pool of IRT-calibrated items expressed on the reporting scale exists for test form construction. The item parameter estimates for new forms are obtained from the bank and are used to build test forms that are parallel across administrations. Student scores are produced with the new form bank-obtained item parameters, thereby linking scores from one administration to the other.

To increase the item pool, the MD HSA embeds field test items in the operational test forms. The field test data for the January and May administration are calibrated with the operational items at that time. The calibrations are linked to the reporting scale using all operational non-CR items as anchors and the Stocking and Lord procedure (Stocking & Lord, 1983). Having all operational non-CR items serve as linking items ensures that the linking set is both objectively scored and large enough to provide stable and reliable results. Item bank parameters are established at the time of the field test and are not updated following each administration.

To ensure that items behave the same way across administrations, construction of new forms follows guidelines defined by Kolen and Brennan (1995). These guidelines are:

1. Items should appear in the same contexts and positions as when the item parameters were established. Operational item are placed as close as possible to the same position they were in when parameters were established and within the same third of the total test form.
2. Operational items should appear in similar positions on the test. It may be problematic if an item is positioned in very different locations on the two forms, such as at the beginning of the test on one form and at the end of the test on another form. Operational items that appear in more than one form occupy consistent positions across forms; MSDE must approve any deviations.
3. The text is exactly the same in the old and new forms. Minor editorial changes and rearranging answer choices are discouraged; otherwise the items may function differently. All requests for minor editorial changes must undergo psychometric review to evaluate the implications for the response process.

Post-Test Calibration and Equating of the May 2009 Test Forms

As mentioned in the previous section, student scores on the MD HSAs typically have been generated using pre-equated item parameters. In May 2009 MSDE’s National Psychometric Council (NPC) advised that the item parameters used for scoring be estimated and equated using the May 2009 data. Given the replacement of the CR items in the MD HSA forms and the implementation of online testing, the NPC wanted to make sure that the item parameters and scores being reported were based on current data rather than the parameters in the item bank.

Accordingly the May 2009 operational items were calibrated using the 3PL model and the *PARSCALE* module of ETS’s proprietary software, *GENASYS*. Field test items were excluded from item calibration so that the parameters obtained were not influenced by field test item performance.

The number of students used to calibrate most test forms was large. The exceptions to this rule occurred for the Makeup Form Y in Biology, English, and Government, where the number of test takers was just under 1,000. As a consequence, in the first round of calibrations and equatings, the *c*-parameters for the items unique to these makeup forms were fixed to their reference values, at the NPC's suggestion (J. Bagsby, personal communication, April 21, 2009). For Biology, English, and Government, this meant that the *c*-parameters were fixed for fifteen, thirteen, and sixteen unique items, respectively.

All equatings between the post-test item calibrations and the pre-equated parameters for the operational items were carried out using the Stocking and Lord procedure as implemented within the *GENASYS* software. The pre-equated item parameters (parameters from the item bank) for the May 2009 operational items served as the reference parameters.

Evaluation of the equating results included comparing the reference parameters and post-test equated parameters in terms of means, standard deviations, correlations, item and test characteristic functions, and standard error curves. Scaled score results for May 2009 were also compared with historical trends.

Study of the first round of equating results indicated an excellent alignment of the reference and post-test parameters in Biology and very good alignment in Algebra, English, and Government. Within each content area, TCCs based on the reference and post-test equated parameters were very similar. Across the four content areas, the correlations between the reference and transformed *b*-parameters ranged from 0.86 to 0.96, and the correlations between the *a*-parameters ranged from 0.86 to 0.93. Finally, the correlations between *c*-parameters ranged from 0.64 to 0.74.

Further inspection of the results indicated that there were small, systematic differences between the reference and estimated *c*-parameters for the Biology, English, and Government items, which suggested that their equatings could be improved by fixing the *c*-parameters for all operational items to their reference values. This procedure was carried out, and final equated item parameters were delivered to Pearson for all content areas.

Based on a review of the May 2009 post-test equating and score results, the National Psychometric Council decided that

- scores based on the post-equated item parameters were approved for reporting May 2009 results for all content areas.
- the reference parameters for the May 2009 operational items will continue to be used as the statistics of record.
- field test items from the May 2009 administration will be scaled using reference parameters, following usual procedures.