Section 3. Scoring Procedures and Score Types

Scale Scores

The MDHSA reporting scale ranges from 240 to 650. It was established in 2003 and defined so that the scores had a mean of 400 and a standard deviation 40. These scores represent ability estimates obtained using Item Response Theory (IRT). (See Section 5 IRT Calibration and Scaling for details on the 3-parameter logistic [3PL] and generalized partial-credit [GPCM] IRT models used for the MDHSAs).

Scale scores based on maximum likelihood estimates (MLE) are reported for the total test scores. While the total test score is based on item-pattern (IP) scoring, the subscores are based on raw score to scale score (RS-SS) scoring tables[5].

When IP scoring using 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. RS-SS scoring tables were obtained by taking the inverse of the TCC for items contributing to the associated subscores (Yen, 1984).

Conditional Standard Errors of Measurement

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

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

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

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.

[5] For operational scoring, omitted responses on the 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 the 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

The cut-scores associated with each of the performance levels in the non-English content areas were established by MSDE in 2003[6]. These values are given in Table 3.1. The English cut-scores were established during the standard setting meeting held in October of 2005. One cut-score was established for Biology and 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. Students who entered grade 9 in the 2005-2006 academic year and thereafter must pass all four HSAs or achieve an overall combined score of 1602. The Proficient cut-scores are used to determine Pass/Fail classifications.

Table 3.1 MDHSA 2007 Cut-Scores by Content Area

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

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

Year-to-Year Scale Maintenance

The Maryland HSA tests have been pre-equated since 2004. 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 MDHSA embeds field test items in the operational test forms. The field-test data for the January and May administrations are delivered in the fall, and the field-test items 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 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 & Brennan (1995). These guidelines are:

a) 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 when parameters were established and within the same 1/3 of the total test form.
b) 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; deviations must be approved by MSDE.
c) 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.

Special Analyses Carried Out to Support the January Biology Scoring

There were two issues with the January 2007 Biology test that required special consideration and analyses prior to operational scoring. The issues are described below. Courses of action were determined in consultation with MSDE.

Punnett Square Biology Item

For one brief constructed response (BCR) item, there was an inconsistency between the design of the answer space for this item when the banked item parameters were established in 2003 and the design of the answer space when the item was used in the January 2007 tests. Specifically, the answer space omitted a box and the instructions to "Draw the Punnett square in the box" which appeared on the answer document when the item was field-tested and scaled. The item (#42, MD ID 64715) appeared on all four operational January forms (A, B, C, D).

MSDE wanted to maintain the item in their bank as it appeared prior to the misprint; that is, no item information would be updated in the bank. The question posed by MSDE for the January 2007 administration was whether the item bank parameters could be used in operational scoring, and if not, whether new the item should be omitted in determining student scores. To this end, ETS recalibrated the item, compared the banked parameters with those obtained from the recalibration, and evaluated the effects of the new parameters on students' scores.

The results of the study indicated that 1) the differences in item parameters did not exceed differences that could be attributed to sampling, and 2) replacing the bank parameters with the January 2007 parameters had very little impact on student scores. MSDE decided to score students using the bank item parameters.

Misalignment of Form C Test Book and Answer Document

An additional issue for operational scoring involved a misalignment of the session break between the test book and the answer document in Form C. Specifically, the test book labeled Session 1 as items 1 to 34 and Session 2 as items 35 to 71; the answer document labeled Session 1 as items 1 to 36 and Session 2 as items 37 to 71. The answer document was correctly labeled.

Since it was not known whether student performance was affected by the mislabeling, MSDE offered Form C examinees the opportunity to retake the exam using Form D. However, not all Form C examinees retested with Form D. MSDE decided the following would determine how students were scored:

1) Students who took Form C only received the higher of the Form C full form and Form C Session 1 scores.
2) Students who took Forms C and D received the highest of the Form D full form, the Form C full form, and the Form C Session 1 scores.

Item pattern scores were generated using the items that appeared in Session 1 of Form C. This session contained 23 selected response items and 4 constructed response items. These scores had more measurement error than the full forms because they were based on fewer items, and the content on which they were based was not representative of the test blueprint. Students who received an operational score based on the Form C Session 1 did not receive subscores. Results of the Form C scoring options are provided in Table 3.2. Because of the special scoring, Form C students were excluded from field test item analysis samples.

Table 3.2 Summary Results of Form C Special Scoring Options

Forms	Scale Scores	N	Mean	SD	Min	Max
C Only	Form C Full	448	357.7	50.5	240	470
	Form C Session 1	448	362.1	50.5	240	472
C and D	Form C Full	163	403.8	35.9	240	487
	Form C Session 1	163	405.6	33.8	240	483
	Form D Full	163	405.7	36.0	240	497