Section 3. Scoring Procedures and Score Types

Scale Scores

The MHSA 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 MHSA).

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.

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

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. 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.

Table 3.1 MDHSA 2006 Cut-Scores by Content Area

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

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 large and reliable. 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 rules defined by Kolen & Brennan (1995). These rules are:

a) items should appear in contexts and positions as when the item parameters were established,
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, and
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.