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, Yen & Fitzpatrick, 2006). (See IRT Calibration and Scaling in Section 7 for details about the three-parameter logistic model [3PL] used for the MD HSAs.) Students’ total test scores and subscores are scale scores derived using the 3PL model and item-pattern scoring procedures. There are essentially two ways of scoring a test: number-correct (NC) or item-pattern (IP) scoring. NC scoring considers how many test items a student answered correctly in determining that student’s total raw score. In contrast, the IP scoring method is based on an IRT model. IP scoring takes into account not only a student’s total raw score, but also which test items the student answered correctly, as well as the psychometric characteristics of these items.

Test items are not equal in their characteristics. For example, some items are better at discriminating between students that know the tested content and those who do not; some items are more difficult; and low-ability students are more likely to guess correctly on some test items than on others.

Two students with exactly the same total raw score will get the same test score in NC scoring. It is very possible; however, that even though they have the same total raw score, the actual items they answered correctly were different, and their different sets of correctly answered items could have different item characteristics. In such a case, the students will very likely get different reported test scores in IP scoring even though they have the same total raw score. This would be applicable to both total test scores and subscore category scores reported using IP scoring.

Item-pattern scoring has been found to produce smaller standard errors of measurement (SEM) than number-correct scoring. The smaller the SEM, the more confidence we have about the accuracy of the test results. In addition, test reliability is higher with IP scoring than with NC scoring (Yen & Candell, 1991), which means that fewer questions are needed in IP scoring than NC scoring for equivalent scoring accuracy. For these reasons, both the HSA total test scores and test subscores are reported using IP scoring.
Conditional Standard Errors of Measurement

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

\[
CSEM(\hat{\theta}) = \frac{1}{\sqrt{I(\hat{\theta})}}
\]

where \(\theta\) refers to the true location, 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.

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).

Cut Scores

MSDE established the cut scores associated with each of the performance levels in the content areas other than English in 2003. 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.

---

5 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.
Table 4.1  MD HSA Cut Scores by Content Area

<table>
<thead>
<tr>
<th>Content Area</th>
<th>Cut-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Proficient</td>
</tr>
<tr>
<td>Algebra</td>
<td>412</td>
</tr>
<tr>
<td>Biology</td>
<td>400</td>
</tr>
<tr>
<td>English</td>
<td>396</td>
</tr>
<tr>
<td>Government</td>
<td>394</td>
</tr>
</tbody>
</table>

With the reinstatement of Government tests in January 2013, students entering 9th grade in the 2012-2013 school year or in a prior year (including students who entered 9th grade in 2011-2012, 2010-2011, or 2009-2010) have the option of earning a combined score of 1602 on all four MD HSAs or a combined score of 1208 on three MD HSAs (excluding Government) to be granted a Maryland High School Diploma.6 Students entering 9th grade in 2013-14 and beyond will be required to pass all four HSAs, including Government, or obtain a combined score of 1602 to fulfill the graduation requirement.

Year-to-Year Scale Maintenance

The MD HSAs for Algebra, Biology, and Government have been preequated since 2004; English has been preequated since 2005. In the preequated 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 bank-obtained item parameter estimates, 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 items as anchors and the Stocking and Lord procedure (Stocking & Lord, 1983). Having all operational items serve as linking items ensures that the linking set is large enough to provide stable and reliable results. Item bank parameter estimates 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 (2004, p.313). These guidelines are:

1. Items should appear in the same contexts and positions as when the item parameter estimates were established. Operational items are placed as close as possible to the same position they were in when parameters were estimated 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

6 More information on the testing requirement for graduation is available on the Maryland State Department of Education website at http://mdk12.org/assessments/high_school/index.html.
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.