Section 2. Test Construction and Administration

Test Development

Planning

Planning for the test development process began with the creation of item development plans for each content area. ETS content leaders collaborated with their content counterparts at MSDE to create these plans. The item bank was reviewed to determine how well the available item pool matched the test form requirements set forth in the test form blueprint as defined by the Core Learning Goals and the 2008–2009 form construction templates provided by MSDE. Areas that contained low item counts were given priority when determining which indicators were to be addressed by the item writers. After these areas with critical need were defined and addressed, any remaining items to be developed (as determined by the requirements set forth in the RFP) were distributed among the indicators in a fashion that would best ensure sufficient numbers of items for use in the construction of forms for future administrations.

Test Specifications and Design

MSDE predetermined the basic test design and provided it to ETS in the form of the content-specific "Test Specifications—Test Form Matrix" document. This basic test design document provided information based on specified expectations and the distribution of the number of items by item type for each reporting category. How the specific items were placed throughout the forms was left to the collaborative efforts of ETS and MSDE content specialists. Construction of the forms was based on test blueprints approved by MSDE. Blueprints for each content area are presented in Tables 2.3 to 2.6.

Item Types

As noted in Section 1, four item types have been used in the MD HSA tests. Tables 2.1 and 2.2 show how these item types and associated points were distributed by content area. These item types include the following:

- Selected response (SR)—questions in multiple-choice format with four answer options; each SR item is worth one point.
- Student-produced response (SPR)—an item type used in Algebra only, for which the student works the problem and records the answer in an answer grid; each SPR item has a maximum score of 1.
- Brief constructed response (BCR)—writing prompts for which the written response is no longer than a page (26 lines); each BCR item has a maximum score of 3.
- Extended constructed response (ECR)—writing prompts for which the written response is no longer than two pages (52 lines); each ECR has a maximum score of 4.

Beginning with the May 2009 test forms, BCR and ECR items are no longer included on MD HSA forms. These constructed response (CR) items were replaced on a point-by-point basis with SR items and, in the case of Algebra, SPR items.

Table 2.1 Number of Operational Items and Points by Item Type for each MD HSA Content Area, October, January, and April Administrations

	Operational Items by Item Type					Points by Item Type				
Content Area	SR	SPR	BCR	ECR	Total	SR	SPR	BCR	ECR	Total
Algebra	26	6	3	3	38	26	6	9	12	53
Biology	48	-	7	-	55	48	-	28	-	76
English	46	-	2	2	50	46	-	6	8	60
Government	50	-	7	1	58	50	-	28	4	82

Table 2.2 Number of Operational Items and Points by Item Type for Each MD HSA Content Area After Removal of CRs, May and Summer Administrations

	Operational Items by Item Type			Points by Item Type		
Content Area	SR	SPR	Total	SR	SPR	Total
Algebra	43	10	53	43	10	53
Biology	76	-	76	76	-	76
English	60	-	60	60	-	60
Government	82	-	82	82	-	82

Item Writing

Item writers were employed to develop high-quality test items that were aligned with the Core Learning Goals. Nearly all the item writers were Maryland educators. Only a small portion of the total number of items written were developed by ETS content specialists. Item writers were selected on the basis of their depth of content knowledge and familiarity with the MD HSA program. Many were experienced MD HSA item writers.

Item writers were trained on general item writing techniques as well as writing guidelines that are specific to the MD HSA program. Approximately one month after the initial item writer training, a follow-up training session was provided. The session was designed to evaluate how well the item writers' writing skills had developed to that point, to facilitate peer review of items, and to provide constructive feedback to guide the rest of their writing assignment.

Upon completion of their writing assignment, item writers submitted their items to ETS. Items that were accepted proceeded to the item review and revision process.

Item Review and Revision

All items underwent a series of editorial reviews in accordance with the following procedures:

- Items were edited according to standard rules developed in conjunction with MSDE.
- Items were reviewed for accuracy, organization, comprehension, style, usage, consistency, and fairness/sensitivity.
- Item content was reviewed to establish whether the item measured the intended Goal-Expectation-Indicator-Assessment Limit, with the Goal being the broadest category and Assessment Limit being the narrowest parameter of content being assessed. Assessment Limit is defined as the maximum domain from which test questions will be developed.
- Copyright and/or trademark permissions were verified for any materials requiring permissions, for both field test and operational material.
- Internal reviews were conducted and historical records were established for all version changes.

After ETS performed the required internal reviews, items were submitted to MSDE for review. If the MSDE content specialist requested an original version of the item as submitted by the item writer, a copy was provided. Any associated stimulus material, graphic, and/or art was provided as well as information regarding the Goal-Expectation-Indicator-Assessment Limit that each question addressed.

MSDE content specialists performed a review of the items and provided feedback to ETS content specialists. The edits were incorporated into the items. MSDE and ETS content specialists then met to conduct a side-by-side review of the items. Any final edits to the items were made. Finally, the items were prepared for review by the Content and Bias/Sensitivity Review Committees.

The Content and Bias/Sensitivity Review Committees are diverse groups of Maryland educators who reviewed each item to ensure that its content (a) accurately reflected what was taught in Maryland schools; (b) correctly matched the intended CLG indicator; and (c) did not unfairly favor or disadvantage an individual or group.

Upon completion of this final round of reviews, MSDE and ETS content specialists conducted another side-by-side meeting to evaluate the reviews and to reconcile the results of the various groups. The ETS content specialists then made the requested edits to the items and/or revisions to the accompanying graphics. The items that survived this process were eligible for placement in the field test sections of the test forms.

Testing Accommodations

A number of alternate test formats are available to MD HSA examinees, including large-print, Braille, online audio, and Kurzweil versions of the MD HSA developed for each content area. All four alternate test formats are available at each administration. Data from these alternate formats are included in the psychometric analyses.

Test Specifications

All 2009 test forms were constructed using items from the Maryland item bank. The pool of items available for use in the construction of the 2009 forms included all items that had been administered, calibrated, and linked to the operational scale. For Algebra, Biology, and Government, the MD HSA operational scale was defined in 2003 and included items administered in 2002 and 2003. For English, the scale was redefined in 2005 when the English test was updated to become an end-of-course assessment for English 2. Items flagged for poor fit and items flagged for substantial differential item functioning (DIF) against one of the focal groups were excluded from the item pool. (See Section 7 for a more detailed account of these analyses and flagging criteria.)

Each test form was constructed to meet specific test blueprints. Tables 2.3 through 2.6 indicate the distribution of items within each reporting category by item type and the number of score points associated with each item type. The October, January, and April forms for Algebra, Biology, and Government consisted of two sessions administered within a single sitting; the forms for English consisted of three sessions administered within a single sitting. In the May and Summer administrations, all content area tests consisted of three sessions administered within a single setting. Sessions were separated by a short break.

Rubrics for items can be found at the following locations:

http://www.mdk12.org/assessments/high_school/look_like/algebra/rubric.html
http://www.mdk12.org/assessments/high_school/look_like/biology/rubric.html
http://www.mdk12.org/assessments/high_school/look_like/english/rubric.html
http://www.mdk12.org/assessments/high_school/look_like/government/rubric.html

Table 2.3 MD HSA Algebra Blueprint

Reporting Category	Number of Items						Total Points per Category
	October, January, and April				May and Summer		
	SR (1 pt)	SPR (1 pt)	BCR (3 pts)	ECR (4 pts)	SR (1 pt)	SPR (1 pt)	
Expectation 1.1 Analyzing Pattern and Functions	8	1	0	1	11	2	13
Expectation 1.2 Modeling Real-World Situations	10	3	0	1	13	4	17
Expectation 3.1 Collecting, Organizing and Analyzing Data	4	2	2	0	8	4	12
Expectation 3.2 Using Data to Make Predictions	4	0	1	1	11	0	11
Total	26	6	3	3	43	10	53

Note: Information on the referenced expectations can be found in the Maryland Core Learning Goals for Algebra, available on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html.

Table 2.4 MD HSA Biology Blueprint

Reporting Category	Number of Items				Total Points per Category
	October, January, and April			May and Summer	
	SR (1 pt)	BCR (4 pts)	ECR (4 pts)	SR Only (1 pt)	
Goal 1 Skills and Processes of Biology	8	2	0	16	16
Expectation 3.1 Structure and Function of Biological Molecules	8	1	0	12	12
Expectation 3.2 Structure and Function of Cells and Organisms	9	1	0	13	13
Expectation 3.3 Inheritance of Traits	9	1	0	13	13
Expectation 3.4 Mechanism of Evolutionary Change	5	1	0	9	9
Expectation 3.5 Interdependence of Organisms in the Biosphere	9	1	0	13	13
Total	48	7	0	76	76

Note: Information on the referenced goal and expectations can be found in the Maryland Core Learning Goals for Biology, available on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html.

Table 2.5 MD HSA English Blueprint

Reporting Category	Number of Items				Total Points per Category
	October, January, and April			May and Summer	
	SR (1 pt)	BCR (3 pts)	ECR (4 pts)	SR Only (1 pt)	
1: Reading and Literature: Comprehension and Interpretation (RC) Includes indicators 1.1.1, 1.1.2, 1.1.3, 1.2.1, 1.3.3, 3.2.2	13	1	0	16	16
2: Reading and Literature: Making Connections and Evaluation (RE) Includes indicators 1.1.4, 1.2.2, 1.2.3, 1.2.4, 1.2.5, 1.3.5, 4.1.1*, 4.2.1	11	1	0	14	14
3: Writing: Composing (WC) Includes indicators 2.1.1*, 2.1.4*, 2.2.1, 2.2.2, 2.2.3, 2.2.5, 2.3.1, 2.3.3, 4.3.1	8	0	2	16	16
4: Writing: Language Usage and Conventions (WL) Includes indicators 3.1.3, 3.1.4, 3.1.6, 3.1.8, 3.3.1, 3.3.2	14	0	0	14	14
Total	46	2	2	60	60

Note: Information on the referenced indicators can be found in the Maryland Core Learning Goals for English, available on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html.
*This indicator not included in the May and Summer administrations.

Table 2.6 MD HSA Government Blueprint

Reporting Category	Number of Items				Total Points per Category
	October, January, and April			May and Summer	
	SR (1 pt)	BCR (4 pts)	ECR[a] (4 pts)	SR Only (1 pt)	
Expectation1.1 U.S. Government Structure, Functions and Principles	13	2	1 (alt w/E2)	23	23
Expectation 1.2 Protecting Rights and Maintaining Order	11	2	1 (alt w/E1)	21	21
Goal 2 Systems of Government and U.S. Foreign Policy	8	1	0	12	12
Goal 3 Impact of Geography on Governmental Policy	7	1	0	11	11
Goal 4 Economic Principles, Institutions and Processes	11	1	0	15	15
Total	50	7	1	82	82

Note: Information on the referenced expectations and goals can be found in the Maryland Core Learning Goals for Government, available on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html.

[a] The ECR item on the Government assessment is aligned to and reported as either Expectation 1 or Expectation 2. Forms are developed to alternate between the two expectations across an assessment year.

Item Selection and Form Design

To conserve the item pool, each test form consisted of a common set of operational items shared across forms within an administration as well as a unique set of items. Within a given administration (i.e., October, January, April, May, Summer), approximately 60 percent of the operational items in each form were common across the test sections. The remaining items in the forms consisted of combinations of items that varied across forms. The guidelines used to construct the forms are provided in Tables 2.7 to 2.11. The exact composition of the forms varied slightly based on available items in the pool.

Table 2.7 Form Construction Specifications for the MD HSA October 2008 Administration

Primary Week Form R
Operational items—Unique items from the pool—100%
Field Test Section—Reuse of prior year field test set

Table 2.8 Form Construction Specifications for the MD HSA January 2009 Administration

Primary Week Form A	Makeup 1 Form B
Common set—60%	Common set—60%
Unique items from the pool—40%	Unique items from the pool—40%
Field Test Section—Unique items	Field Test Section—Reuse of field test set

Table 2.9 Form Construction Specifications for the MD HSA April 2009 Administration

Form S
Operational items—Reuse of prior year intact form
Field Test Section—Reuse of prior year field test set

Table 2.10 Form Construction Specifications for the MD HSA May 2009 Administration

Primary Week Forms C–N	Makeup 1 Form X	Makeup 2 Form Y
Common Set—60%	Common Set—60%	Common Set—60%
Unique Items from the pool—40%	Half of items from primary week's 40% unique items—20% Unique items from the pool—20%	Other half of items from primary week's 40% unique items—20% Unique items from the pool—20%
Field Test Section—unique sets of items for Forms C–N	Field Test Section—Reuse of one or a combination of the field test sets used in forms C–N, with a preference for form C; however, the actual selection of field test items was determined by the constraints imposed by the operational items	Field Test Section—Reuse of one or a combination of the field test sets used in forms C-N, with a preference of using the same set used for form X; however, the actual selection of field test items was determined by the constraints imposed by the operational items

Table 2.11 Form Construction Specifications for the MD HSA Summer 2009 Administration

Primary Week 1 Form P	Primary Week 2 Form Q
Common Set—60%	Common Set—60%
Unique items from the pool—40% Field Test Section— Reuse of prior year field test set	Unique items from the pool—40% Field Test Section— Reuse of prior year field test set

In addition to the operational items, embedded field test items were included with each version of the test form, resulting in multiple versions of a test form containing different sets of field test items. The percentage of field test items per form varied by content area and administration, as shown in Table 2.12.

Table 2.12 Number of Operational (OP) and Field Test (FT) Items by MD HSA Administration and Content Area

	October 2008, January 2009, and April 2009			May 2009 and Summer 2009		
Content Area	OP Items	FT Items	% FT Items	OP Items	FT Items	% FT Items
Algebra	38	10	21%	53	16	23%
Biology	55	16	23%	76	23	23%
English*	50	15–21	23–30%	60	31	34%
Government	58	6	9%	82	25	23%

* In the English test forms, the number of field test items differed slightly across administrations because the item sets varied in size.

The items being field tested were a combination of newly written items and/or previously administered items that had been revised due to content concerns or problematic item statistics. Items with problematic statistics were ones that were judged to be acceptable from a content perspective but had one or more of the following characteristics: p-values less than 0.10; item-total correlations of less than 0.15; collapsed score levels for constructed response items (i.e., very few responses in the top score levels); very high omit rates; or SR items with a positive point-biserial correlation for one or more distractors. For administrations in which there was more than one primary form (January and May), the forms were spiraled at the student level. Spiraling at the student level means that multiple forms of the test were packaged in order (e.g., D, E, F, etc.) and distributed to students according to this order. Spiraling at the student level helps ensure that all forms are randomly distributed throughout the state.

Forms were constructed using the test construction software associated with the customer item bank. The goal was to match the test characteristic curves (TCCs) and the conditional standard error of measurement curves (CSEMs) with the "target" form defined as the base form used to set the operational scale. For Algebra, Biology, and Government, the base form was developed in

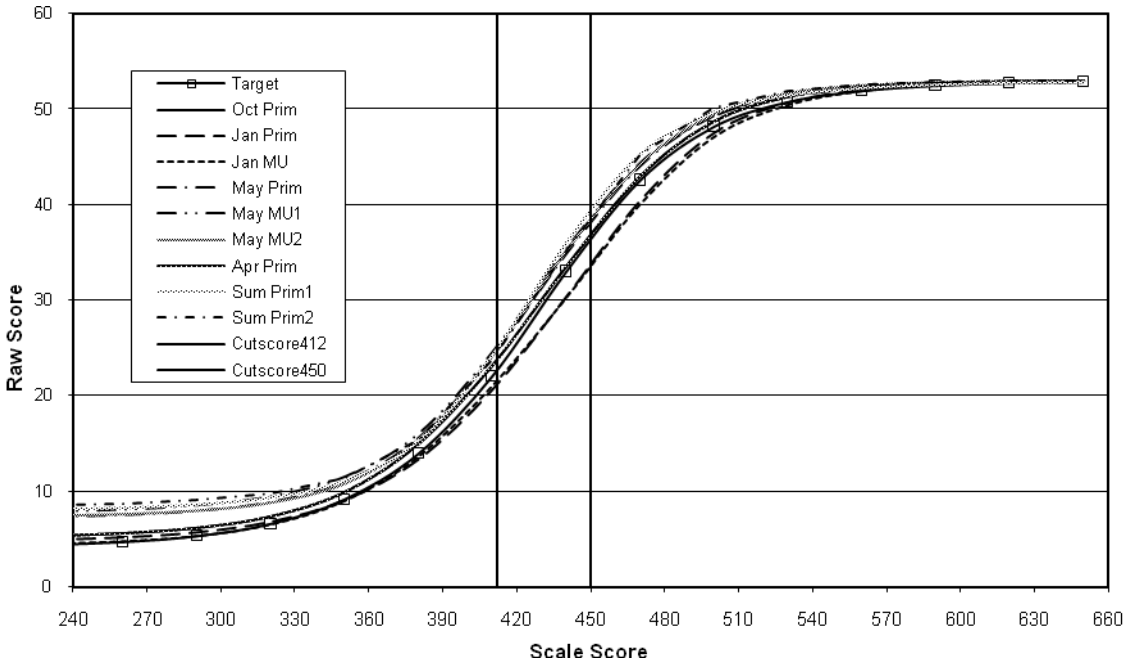
2003; for English the base form was developed in 2005. The TCCs and CSEMs were graphically displayed using item parameters associated with the operational items.
The following general steps were completed during the test construction process:

1. For each administration, all forms were constructed simultaneously in order to provide the best opportunity to construct parallel forms.
2. Items that matched the test blueprint were selected to match the target TCCs and CSEMs.
3. Test developers were careful to ensure that the item selections met all content specifications, including matching items to the test blueprint, distribution of keys, and avoidance of clueing[4] or clanging.[5]
4. After the operational items were selected for the test forms, the field test sets were constructed. Field test sets consisted of SR items in all content areas. While the field test sets were not constructed to meet any psychometric criteria, they were constructed to meet content criteria. For Algebra, Biology, and Government, the field test sets were estimated to be able to be completed by students in approximately 30 minutes for October, January, and April administrations and 35 minutes for May and Summer administrations. Due to the additional time required to read the passages and stimuli for English, the field test sets were estimated to be able to be completed by students in approximately 40 minutes for October, January, and April administrations and 50 minutes for May and Summer administrations. The field test items were embedded in the test according to a variety of content and template criteria, including, but not limited to, coverage of the reporting categories and assessment limits, cognitive balance, key balance/distribution, and clueing/clanging within the field test set and among the surrounding operational items.

Figures 2.1 to 2.8 show the plots of the TCCs and CSEMs for the target form and forms developed for each content area. It is important to note that the TCCs and CSEMs shown in the plots are based on pre-equated item parameters and therefore are theoretical curves calculated prior to administration of the tests. In general, the TCCs and CSEMs were similar to the target curves. The TCC plots indicate that all forms for each content area were within or very close to the acceptable range of the target curve for the full range of scale score values. Where forms varied in difficulty, differences between forms were typically less than 5 percent of the total raw score across the score range, especially in the range of the cut-scores. Where forms had differences slightly greater than 5 percent, these larger differences were typically seen at the very low end of the scale score range and at the high end of the scale. As expected, the CSEM plots indicate that the CSEMs for each content area were lowest in the middle range of scale scores, where the majority of student scores are located. (Please refer to figures 6.1 to 6.4 for histograms of student performance.)

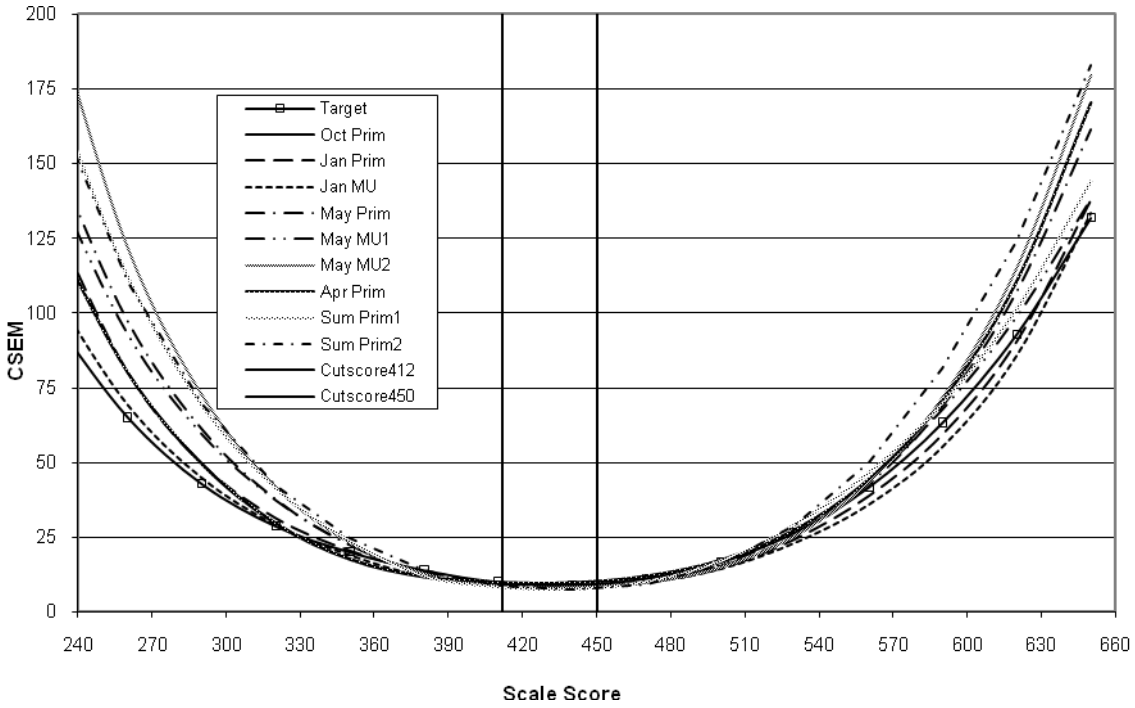
[4] Clueing refers to information within a passage, stimulus, item, graphic, or other test component that allows respondents to select/construct the correct answer to one or more items in an assessment without the knowledge and/or skill targeted by the item.
[5] Clanging occurs when an identical or resembling word(s) appears in both the item stem and one or more item distractors. Also, if two or more items that are near each other share common key words, even if the item content does not clue, the items are said to clang because the interpretation of the word in one item can affect the interpretation of another item.

Figure 2.1 Test Characteristic Curves for the MD HSA 2009 Algebra Forms



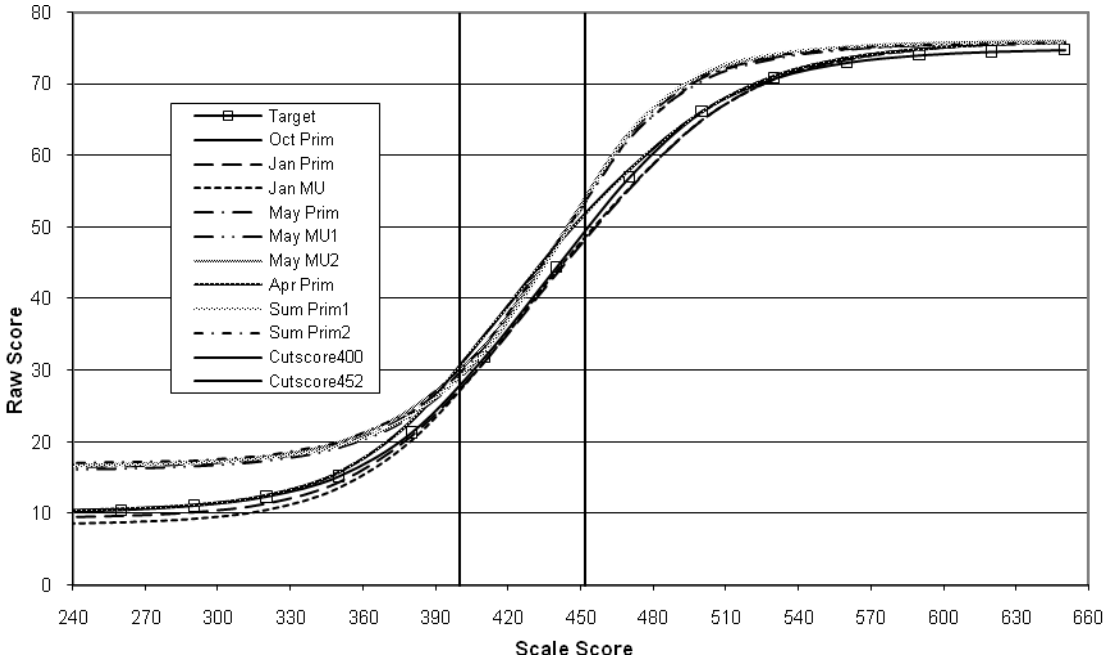
Note: Maximum possible raw score is 53.

Figure 2.2 Conditional Standard Error of Measurement for the MD HSA 2009 Algebra Forms



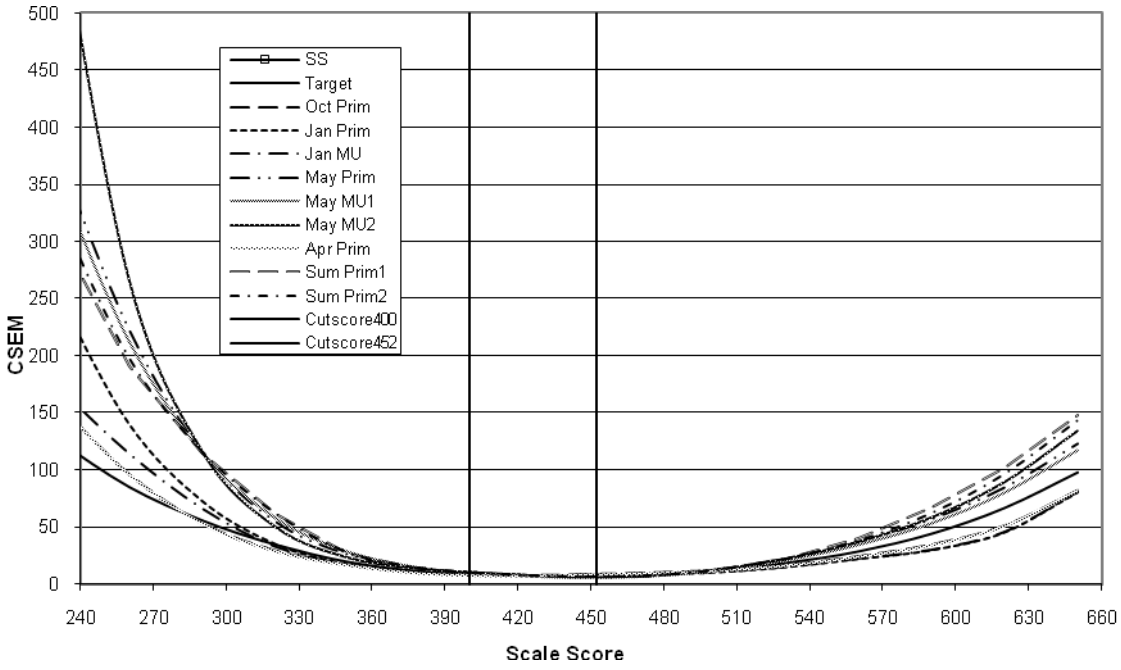
Note: Observed standard deviations for Algebra ranged from 27.1 to 38.1.

Figure 2.3 Test Characteristic Curves for the MD HSA 2009 Biology Forms



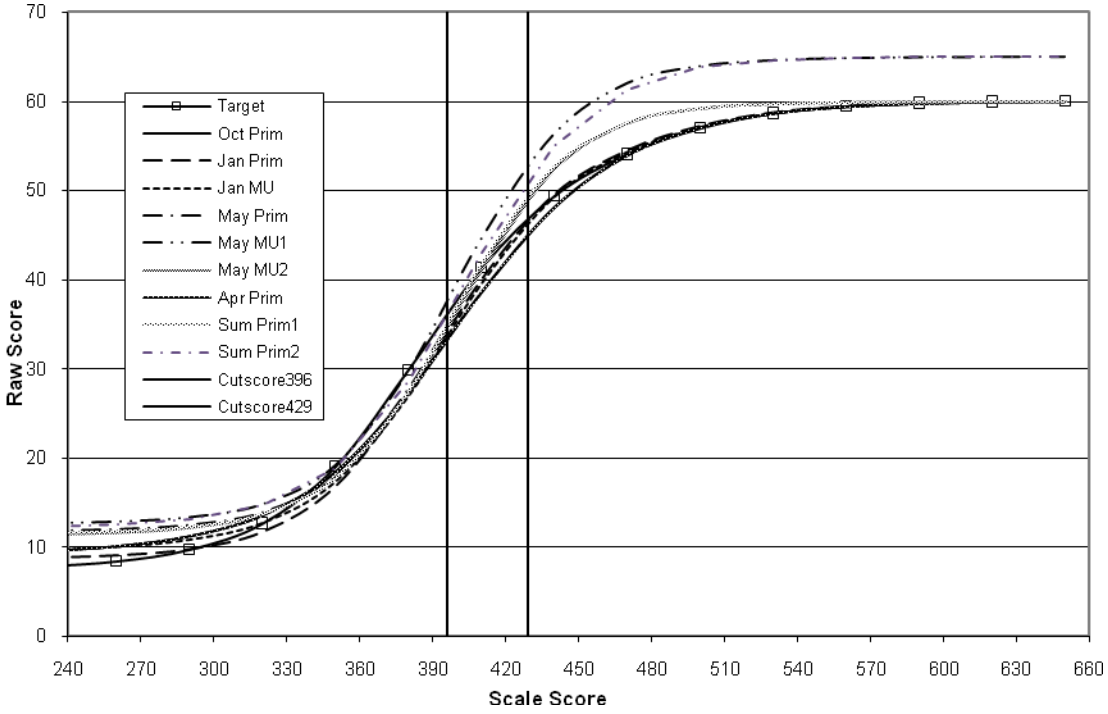
Note: Maximum possible raw score is 76.

Figure 2.4 Conditional Standard Error Measurement for the MD HSA 2009 Biology Form



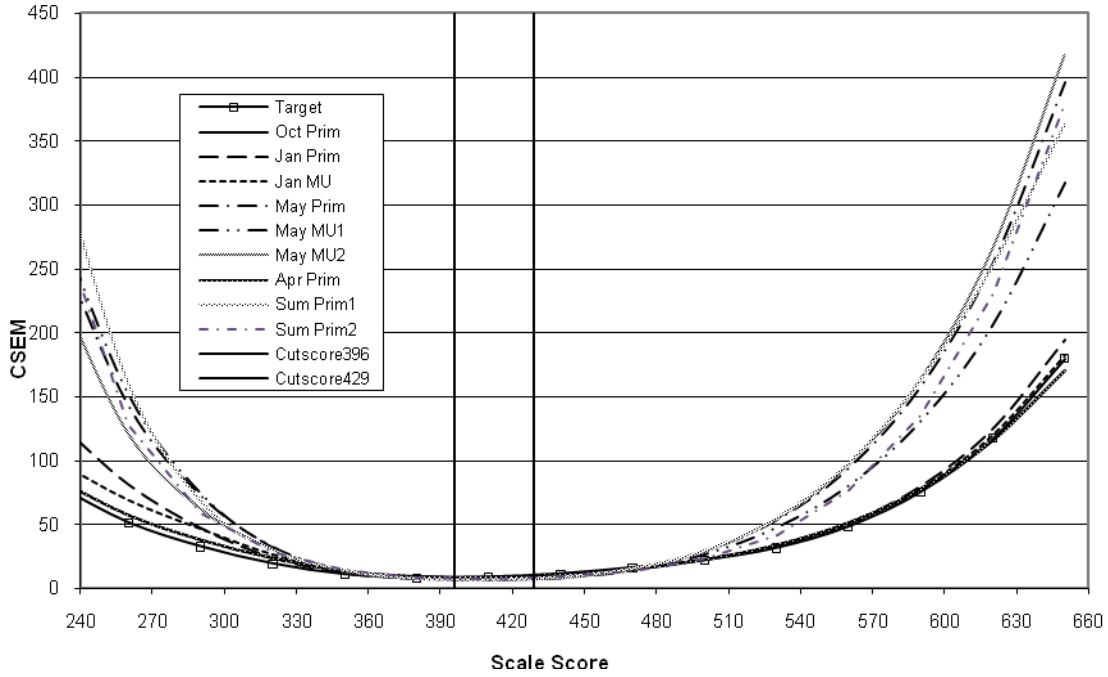
Note: Observed standard deviations for Biology ranged from 30.0 to 41.9.

Figure 2.5 Test Characteristic Curves for the MD HSA 2009 English Forms



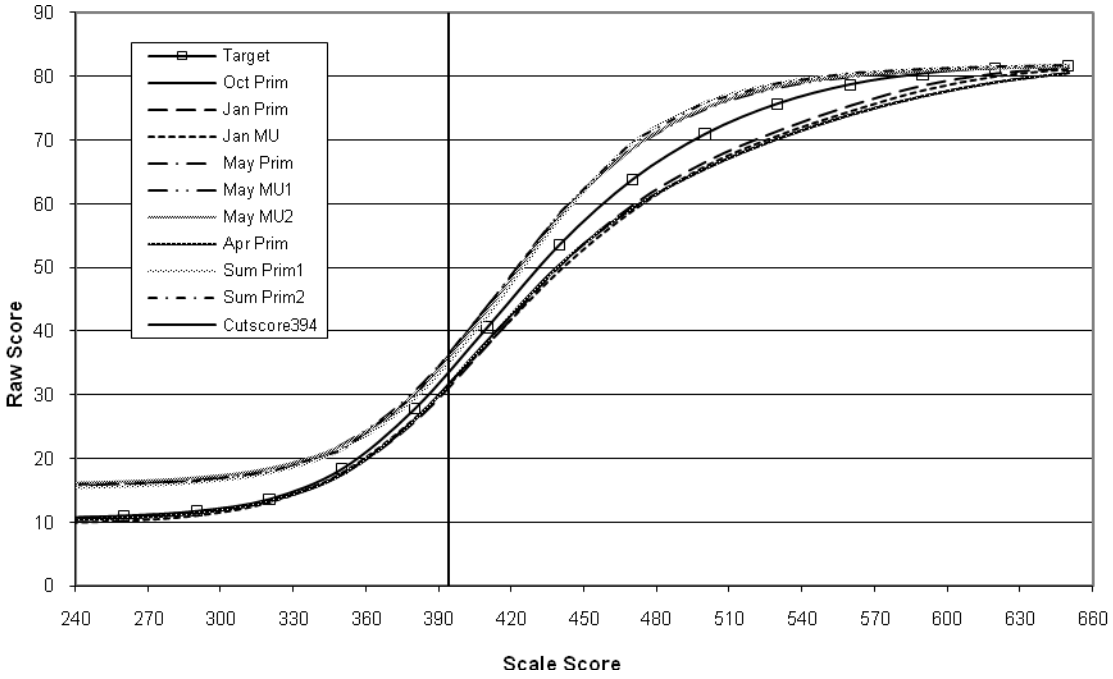
Note: Maximum possible raw score is 60.

Figure 2.6 Conditional Standard Error of Measurement for the MD HSA 2009 English Forms



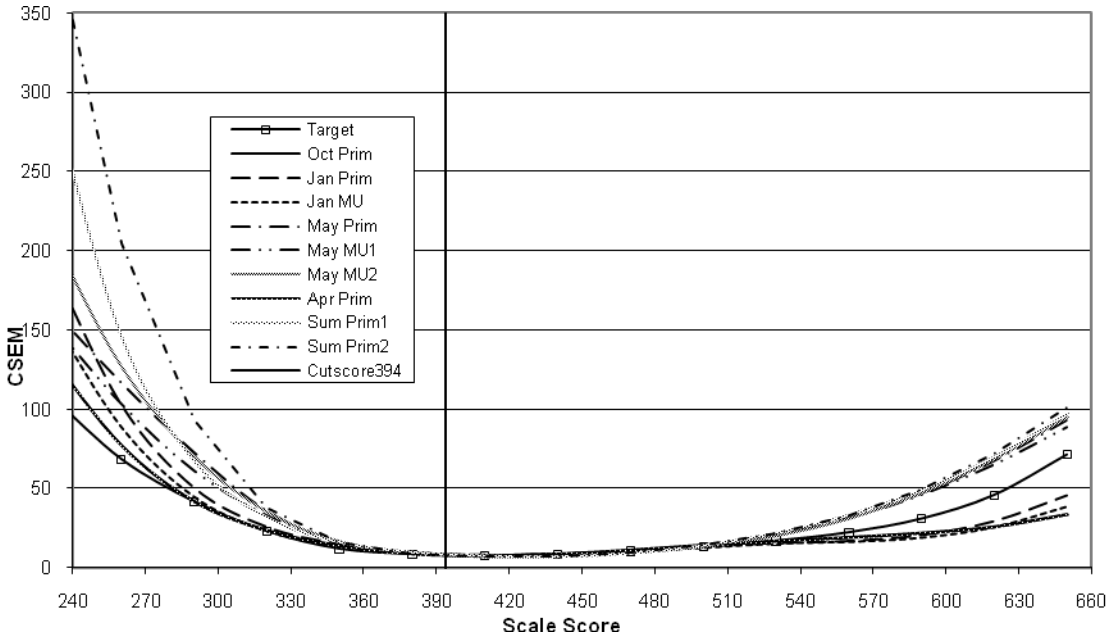
Note: Observed standard deviations for English ranged from 25.6 to 37.0.

Figure 2.7 Test Characteristic Curves for the MD HSA 2009 Government Forms



Note: Maximum possible raw score is 82.

Figure 2.8 Conditional Standard Error of Measurement for the MD HSA 2009 Government Forms



Note: Observed standard deviations for Government ranged from 23.9 to 40.9.

Test Administration

All MD HSA tests administered in October 2008, January 2009, and April 2009 were paper-and-pencil tests, except in the case of the Kurzweil form, which had an audio portion. In May 2009 the MD HSA online version was introduced. For May 2009 and Summer 2009, both paper-and-pencil and online versions of the MD HSA were available. A modality comparability study based on the May 2009 MD HSA data was completed in November 2009, and is included as Appendix C of this report.

For all administrations, paper-and-pencil primary forms were given during the first week of testing. For the January and May administrations, Makeup Form 1 was offered during the second week. For the May administration only, Makeup Form 2 was administered in the third week of testing. For the online versions in May and Summer 2009, all forms were spiraled equally throughout the testing window. In May the eleven primary and two makeup forms were spiraled over a three-week timeframe. In Summer the two primary forms were spiraled over the two-week timeframe.

All forms administered without extended time accommodations had timing limits indicated in Tables 2.13 and 2.14.

Table 2.13 Test Timing Schedule in Minutes by Content Area for the MD HSA October 2008 and January and April 2009 Administrations

Content Area	Session One	Break	Session Two	Break	Session Three
Algebra	75	5–15	75	NA	NA
Biology	80	5–15	70	NA	NA
English	60	5	60	5	50
Government	85	5–15	70	NA	NA

Table 2.14 Test Timing Schedule in Minutes by Content Area for the MD HSA May and Summer 2009 Administrations

Content Area	Session One	Break	Session Two	Break	Session Three
Algebra	50	5	50	5	50
Biology	45	5	45	5	45
English	50	5	50	5	50
Government	45	5	45	5	45