# Section 1. 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 2007-2008 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 critical need areas 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 would be available to use in the construction of forms for future administrations.

*Test Specifications and Design*

The basic test design was pre-determined by MSDE and provided to ETS in the form of the content specific "Test Specs – Test Form Matrix" document presented in Tables 1.2 to 1.5. 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 the ETS and MSDE content specialists. Construction of the forms was based on test blueprints approved by MSDE.

*Item Type*

As noted in the introduction, there were four item types that were used in the MDHSA tests. Table 1.1 shows how these item types and associated points were distributed by content area. These item types include:

- selected response (SR) – multiple-choice format question 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) – a writing prompt for which the written response is not longer than a page (26 lines); each BCR item has a maximum score of 3
- extended constructed response (ECR) – a writing prompt for which the written response is no longer than two pages (52 lines); each ECR has a maximum score of 4.

Table 1.1 Number of Operational Items and Points by Item Type for each MDHSA Content Area

| Content Area | Operational Items by Item Type | | | | | Points by Item Type | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 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 |

*Item Writing*

Item writers were employed to develop high quality test items that were aligned with Core Learning Goals. At least 50% of the item writers were Maryland educators, though for most content areas the percentage was closer to 100. The item writers were selected on the basis of their depth of content knowledge and familiarity with the MDHSA testing program. Many were experienced item-writers with the MDHSA program.

The item writers were trained on general item writing techniques as well as writing guidelines that were specific to the MDHSA 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 their 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. The 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 edited according to standard rules developed in conjunction with MSDE.
- Items reviewed for accuracy, organization, comprehension, style, usage, consistency and fairness/sensitivity.
- Item content 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.

- Verification that copyright and/or trademark permissions had been obtained for any materials requiring permissions, for both field test and operational material.
- Internal reviews conducted and historical records 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 met to conduct a side-by-side review of the items. Any final edits to the items were made. The items were then prepared for review by the Content and Bias/Sensitivity Review Committees. All constructed response items were also submitted to Measurement Incorporated (MI) for review.

The final round of reviews involved the Content Review and Bias/Fairness Review Committees. These committees were diverse groups of Maryland educators who reviewed each item to ensure that the content in each item: 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 review, MSDE and ETS content specialists conducted another side-by-side meeting to evaluate reviews by MI, the Content Review Committee, and the Bias/Fairness Review Committee and to reconcile the results of the various groups. The ETS content specialist then made the requested edits to the items and/or revisions to the accompanying graphics. The items that survived this process were then eligible for placement in the field test sections of the test forms.

*Testing Accommodations*

A number of special test forms are available to HSA examinees, including Large-print, Braille, and Kurzweil versions of the HSA developed for each content area. All three special versions of the test are available at each administration. Data from the special forms are included in the psychometric analyses.

## Test Specifications

All of the 2008 test forms were constructed using items from the Maryland item bank. The pool of items available for use in the construction of the 2008 forms included all items that had been administered, calibrated and linked to the operational scale. For Algebra, Biology, and Government, the MDHSA operational scale was defined in 2003 and included items administered in 2002 and 2003. For English, the scale was re-defined

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 available item pool. Please refer to Section 5 for a more detailed account of these analyses and flagging criteria.

Each test form was constructed to meet specific test blueprints. Tables 1.2 through 1.5 indicate the distribution of items within each reporting category by item type and the number of score points associated with each item type. The 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. 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 1.2 Algebra Blueprint

| ALGEBRA | | | | | |
|---|---|---|---|---|---|
| **Reporting Category** | **Number of Items** | | | | **Total Points per Category** |
| | **SR** (1 pt) | **SPR** (1 pt) | **BCR** (3 pts) | **ECR** (4 pts) | |
| Expectation 1.1<br>The student will analyze a wide variety of patterns and functional relationships using the language of mathematics and appropriate technology | 8 | 1 | 0 | 1 | 13 |
| Expectation 1.2<br>The student will model and interpret real-world situations, using the language of mathematics and appropriate technology. | 10 | 3 | 0 | 1 | 17 |
| Expectation 3.1<br>The student will collect, organize, analyze, and present data. | 4 | 2 | 2 | 0 | 12 |
| Expectation 3.2<br>The student will apply the basic concepts of statistics and probability to predict possible outcomes of real-world situations. | 4 | 0 | 1 | 1 | 11 |
| **TOTAL** | 26 | 6 | 3 | 3 | 53 |

Note: Information on the referenced indicators can be found in the Maryland Core Learning Goals. The HSA Core Learning Goals documents can be found on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html

Table 1.3 Biology Blueprint

| BIOLOGY | | | | |
|---|---|---|---|---|
| **Reporting Category** | **Number of Items** | | | **Total Points per Category** |
| | **SR** (1 pt) | **BCR** (4 pts) | **ECR** (4 pt) | |
| Goal 1<br>Skills and Processes of Biology | 8 | 2 | 0 | 16 |
| Expectation 3.1<br>Structure and Function of Biological Molecules | 8 | 1 | 0 | 12 |
| Expectation 3.2<br>Structure and Function of Cells and Organisms | 9 | 1 | 0 | 13 |
| Expectation 3.3<br>Inheritance of Traits | 9 | 1 | 0 | 13 |
| Expectation 3.4<br>Mechanism of Evolutionary Change | 5 | 1 | 0 | 9 |
| Expectation 3.5<br>Interdependence of Organisms in the Biosphere | 9 | 1 | 0 | 13 |
| **TOTAL** | 48 | 7 | 0 | 76 |

Note: Information on the referenced indicators can be found in the Maryland Core Learning Goals. The HSA Core Learning Goals documents can be found on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html

Table 1.4 English Blueprint

| ENGLISH | | | | |
|---|---|---|---|---|
| **Reporting Category** | **Number of Items** | | | **Total Points per Category** |
| | **SR** (1 pt) | **BCR** (3 pts) | **ECR** (4 pt) | |
| 1: Reading and Literature: Comprehension and Interpretation (RC)<br><br>Includes the following indicators: 1.1.1; 1.1.2; 1.1.3; 1.2.1; 1.3.3; 3.2.2 | 13 | 1 | 0 | 16 |
| 2: Reading and Literature: Making Connections and Evaluation (RE)<br><br>Includes the following 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 |
| 3: Writing - Composing (WC)<br><br>Includes the following 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 |
| 4: Language Usage and Conventions (WL)<br><br>Includes the following indicators: 3.1.3; 3.1.4; 3.1.6; 3.1.8; 3.3.1; 3.3.2 | 14 | 0 | 0 | 14 |
| **TOTAL** | 46 | 2 | 2 | 60 |

Note: Information on the referenced indicators can be found in the Maryland Core Learning Goals for English. The HSA Core Learning Goals documents can be found on the Maryland School Improvement website at http://www.mdk12.org/assessments/standards/9-12.html.

Table 1.5 Government Blueprint

| GOVERNMENT | | | | |
|---|---|---|---|---|
| **Reporting Category** | **Number of Items** | | | **Total Points per Category** |
| | **SR** (1 pts) | **BCR** (4 pts) | **ECR[a]** (4 pts) | |
| Expectation1.1<br>The student will demonstrate understanding of the structure and functions of government and politics in the United States | 13 | 2 | 1<br>(alt w/E2) | 21-25 |
| Expectation 1.2<br>The student will evaluate how the United States government has maintained a balance between protecting rights and maintaining order. | 11 | 2 | 1<br>(alt w/E1) | 19-23 |
| Goal 2<br>The student will demonstrate an understanding of the history, diversity, and commonality of the peoples of the nation and world, the reality of human interdependence, and the need for global cooperation, through a perspective that is both historical and multicultural. | 8 | 1 | 0 | 12 |
| Goal 3<br>The student will demonstrate an understanding of geographic concepts and processes to examine the role of culture, technology, and the environment in the location and distribution of human activities throughout history. | 7 | 1 | 0 | 11 |
| Goal 4<br>The student will demonstrate an understanding of the historical development and current status of economic principles, institutions, and processes needed to be effective citizens, consumers, and workers. | 11 | 1 | 0 | 15 |
| **TOTAL** | 50 | 7 | 1 | 82 |

Note: Information on the referenced indicators can be found in the Maryland Core Learning Goals. The HSA Core Learning Goals documents can be found 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

In order 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, May, Summer), approximately 60% 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 1.6 to 1.8. The exact composition of the forms varied slightly based on available items in the pool.

Table 1.6 Form Construction Specifications for the October 2007Administration

| **Primary Week<br>Form R** |
|---|
| Operational items – Re-use of intact Summer '06 Form (Alg, Bio, Gov), and a mix of '05-'06 forms (English) |
| Field Test Section – Reuse of '06 ('07 as needed for English) FT set(s) - no new development |

Table 1.7 Form Construction Specifications for the January 2008 Administration

| **Primary Week<br>Form A** | **Primary Week<br>Form B** | **Make-Up #1<br>Form C** |
|---|---|---|
| Common set – 80% | Common set – 80% | Common set – 80% |
| Unique Items from the pool – 20%<br>(same items as in Form B) | Unique Items from the pool – 20%<br>(same items as in Form A) | Unique items from the pool – 20% |
| Field Test Section<br>– unique items | Field Test Section<br>– unique items | Field Test Section – a reuse of one or a combination of the field test sets used in forms A-B, with a preference for form A; however, the actual selection of field test items was determined by the constraints imposed by the operational items |

Table 1.8 Form Construction Specifications for the May 2008 Administration

| **Primary Week Forms D -N** | **Make-Up #1 Form X** | **Make-Up #2 Form Y** |
|---|---|---|
| Common Set –60% | Common Set –60% | Common Set – 60% |
| Unique Items from the pool – 40% (the same items for all Forms D – N) | Half of items from primary week's 40% Unique items – 20% | Other half of items from primary week's 40% items – 20% |
| | Unique items from the pool – 20% | Unique items from the pool – 20% |
| Field Test Section – unique sets of items for Forms D through N | Field Test Section – a reuse of one or a combination of the field test sets used in forms D-N, with a preference for form D; however, the actual selection of field test items was determined by the constraints imposed by the operational items | Field Test Section – a reuse of one or a combination of the field test sets used in forms E-K, 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 1.9 Form Construction Specifications for the Summer 2008Administration

| **Primary Week #1 Form L** | **Primary Week #2 Form M** |
|---|---|
| Common Set –60% | Common Set –60% |
| Unique Items from the pool – 40% | Unique Items from the pool – 40% |
| Field Test Section – items repeated from Jan 08 forms | Field Test Section – items repeated from Jan 08 forms |

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. While the number of operational items remained the same for all administrations, the percentage of field test items per form was increased for the May and Summer 2008 administrations to increase the item bank. The percentage of field test items per form varied by content area and administration, as shown in Table 1.10.

Table 1.10 Number of Operational (OP) and Field Test (FT) Items by Administration and Content Area

| | October 2007 and January 2008 | | | May 2008 and Summer 2008 | | |
|---|---|---|---|---|---|---|
| Content Area | OP Items | FT Items | % FT Items | OP Items | FT Items | % FT Items |
| Algebra | 38 | 10 | 21% | 38 | 15 | 28% |
| Biology | 55 | 16 | 23% | 55 | 30 | 35% |
| English* | 50 | 15-21 | 23-29% | 50 | 31-32 | 38-39% |
| Government | 58 | 6 | 9% | 58 | 30 | 34% |

*In the English test forms the number of field test items differed slightly over forms 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. The items with problematic statistics were judged to be acceptable from a content perspective, but had one or more of the following characteristics: p-values less than 0.25; 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 were 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 (TCC) and the conditional standard error of measurement curves (CSEM) 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 TCC and CSEM curves were graphically displayed using item parameters associated with the operational items.

The general steps completed during the test construction process were:

1. For each administration, all forms were constructed simultaneously in order to provide the best opportunity to construct parallel forms.
2. First the common set of items was selected. Then items that matched the test blueprint were selected to match the target TCC and CSEM curves.

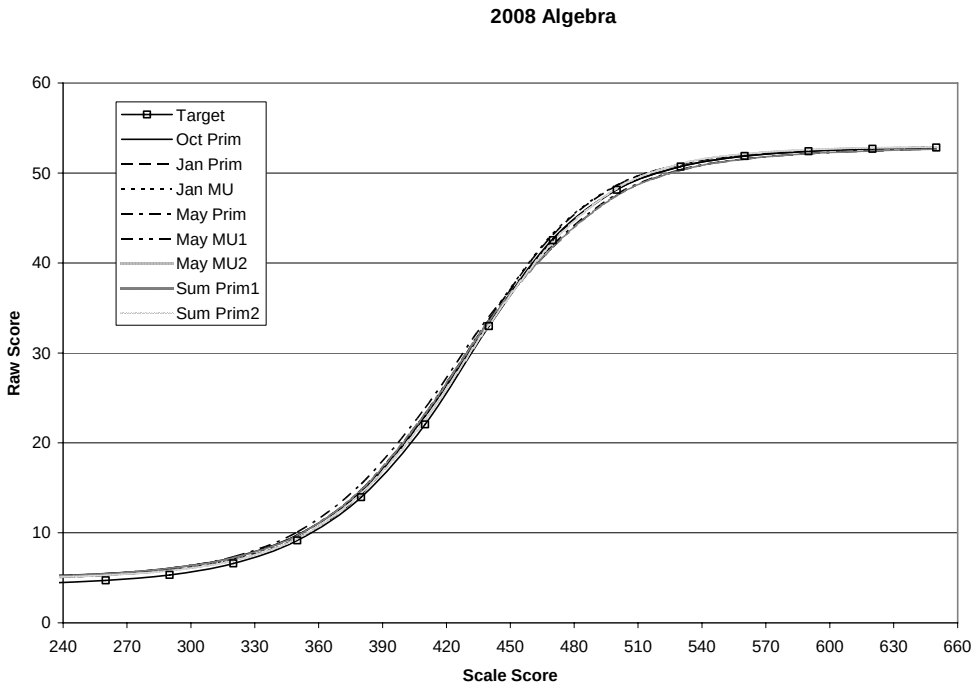
3. Test developers were careful to ensure that the item selections met all content specifications, including matching items to the test blueprint, distribution of keys, and avoidance of clueing[4] or clanging[5].
4. After the operational items were selected for the test forms, the field test sets were constructed. 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 completed by students in approximately 30 minutes. For English, the field test sets were estimated to be completed by students in approximately 40 minutes due to the additional time required to read the passages and stimuli. Field test sets consisted of SR items in all content areas. In addition, the sets consisted of SPR, BCR, and ECR in Algebra, BCR items in Biology, and either BCR or ECR items in English and Government. 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 1.1 to 1.8 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 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 closely matched the target curves. The TCC plots indicate that all forms for each content area were within the acceptable range of the target curve for the full range of scale score values. Where forms varied in difficulty, differences between forms were less than 5% of the total scale score, across the scale score range for Algebra, Biology and English. Four of the Government forms had differences slightly greater than 5% at the high end of the scale (500 to 530). As expected, the CSEM plots indicate that the CSEMs for each content area were lowest in the middle range of scale scores (375 to 475). Figures 1.9 through 1.12 contain the scale score distributions for the May 2008 administration by content area. These histograms have been included to aid interpretation of the CSEM plots by showing that the majority of students score in the middle range, where the CSEMs are smallest.

[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.

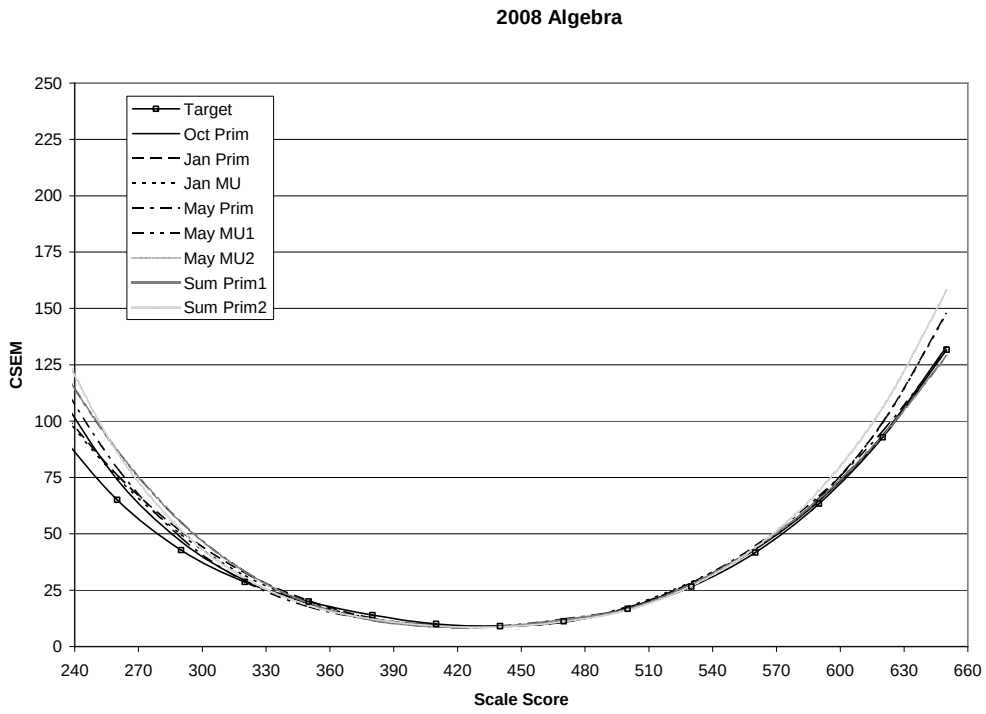
[5] Clanging occurs when an identical or resembling word(s) appears in both the item stem and one or more item distractors. Also, if two or more items that are near each other share common key words, even if the item content does not clue, the items are said to clang because the interpretation of the word in one item can affect the interpretation of another item.

Figure 1.1 Test Characteristic Curves for the 2008 Algebra Forms



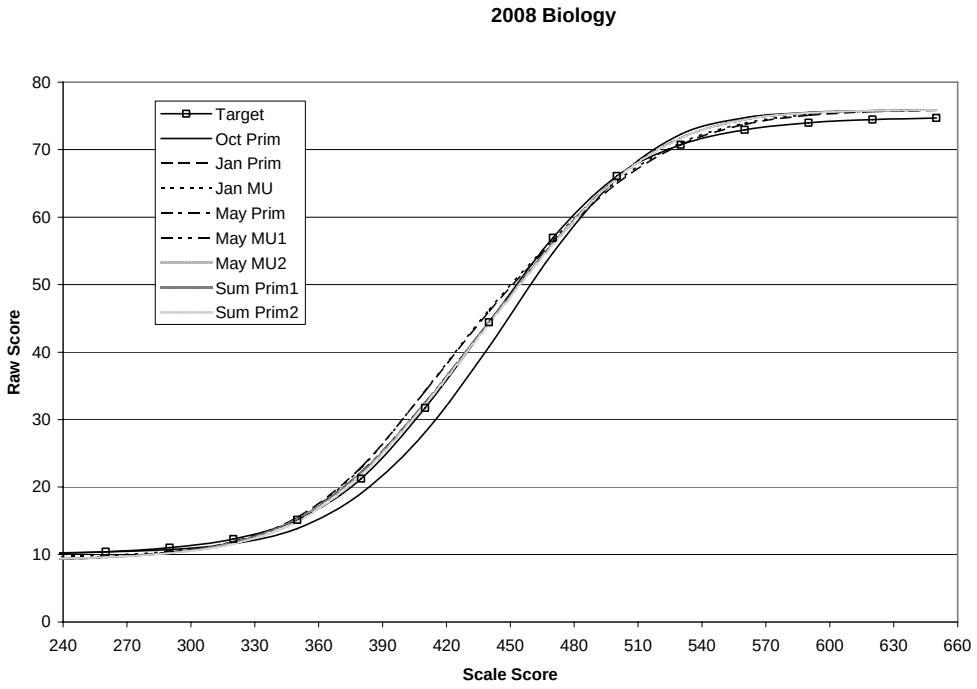

Note: Maximum possible raw score is 53.
Algebra Cut Scores: Proficient 412, Advanced 450

Figure1.2 Conditional Standard Error of Measurement for the 2008 Algebra Forms



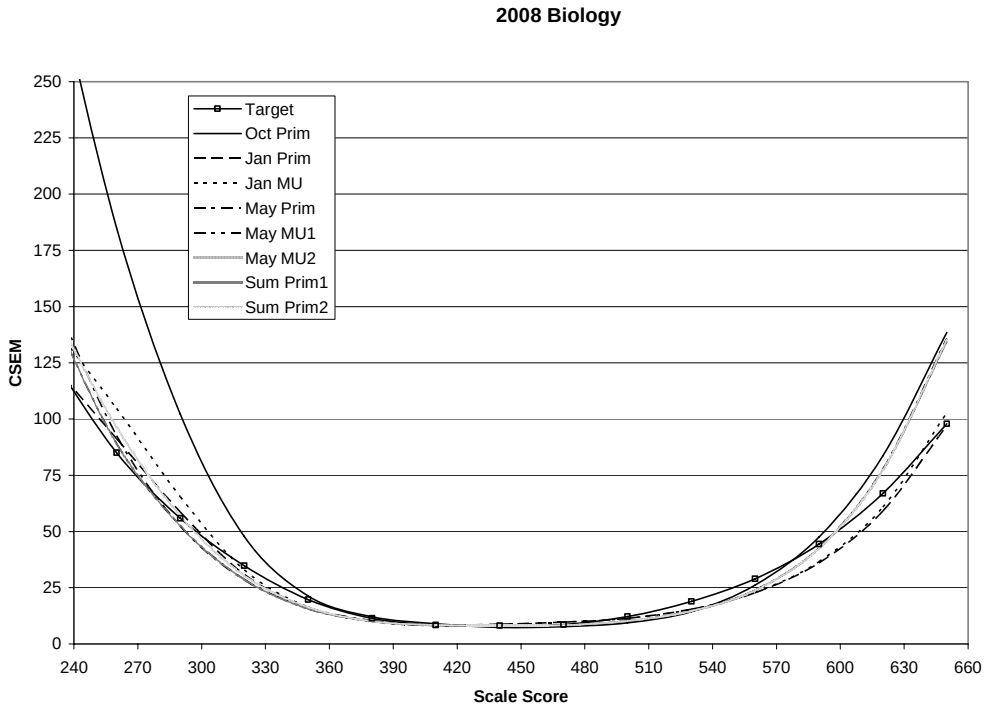

Note: Observed standard deviations for Algebra ranged from 27.1 to 38.1.
Algebra Cut Scores: Proficient 412, Advanced 450

Figure1.3 Test Characteristic Curves for the 2008 Biology Forms



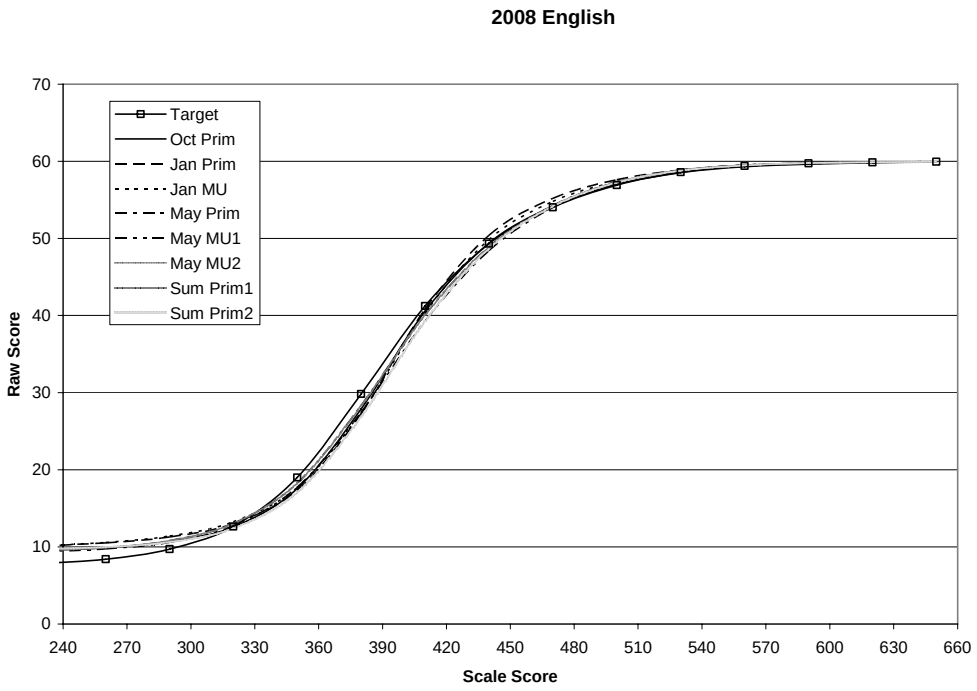

Note: Maximum possible raw score is 76.
Biology Cut Scores: Proficient 400, Advanced 452

Figure1.4 Conditional Standard Error Measurement for the 2008 Biology Form



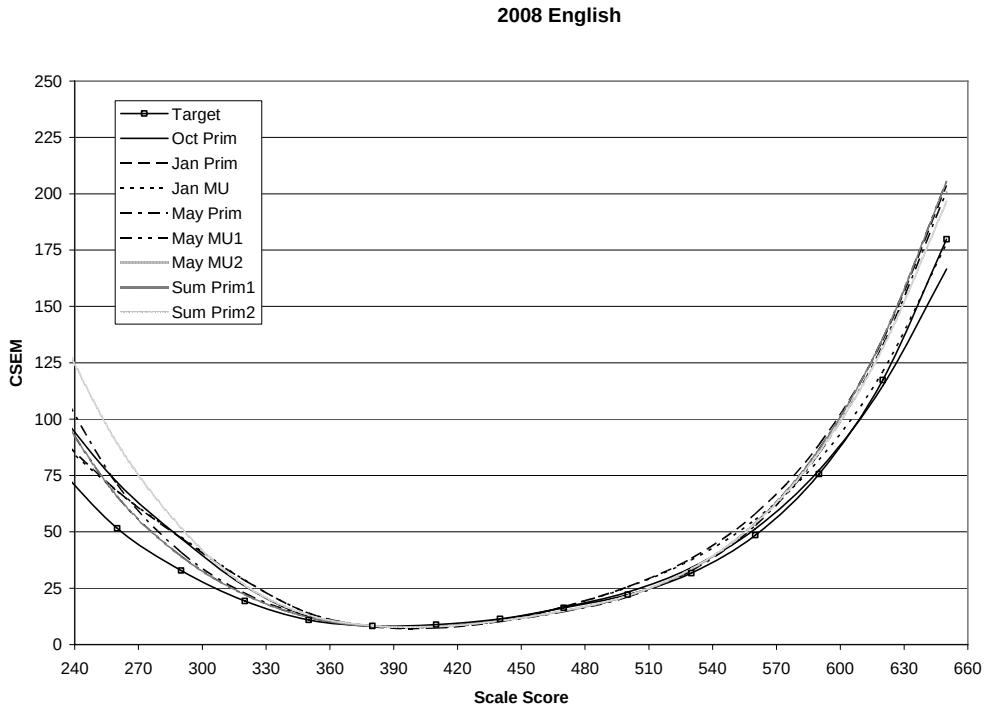

Note: Observed standard deviations for Biology ranged from 30.0 to 41.9.
Biology Cut Scores: Proficient 400, Advanced 452

Figure1.5 Test Characteristic Curves for the 2008 English Forms



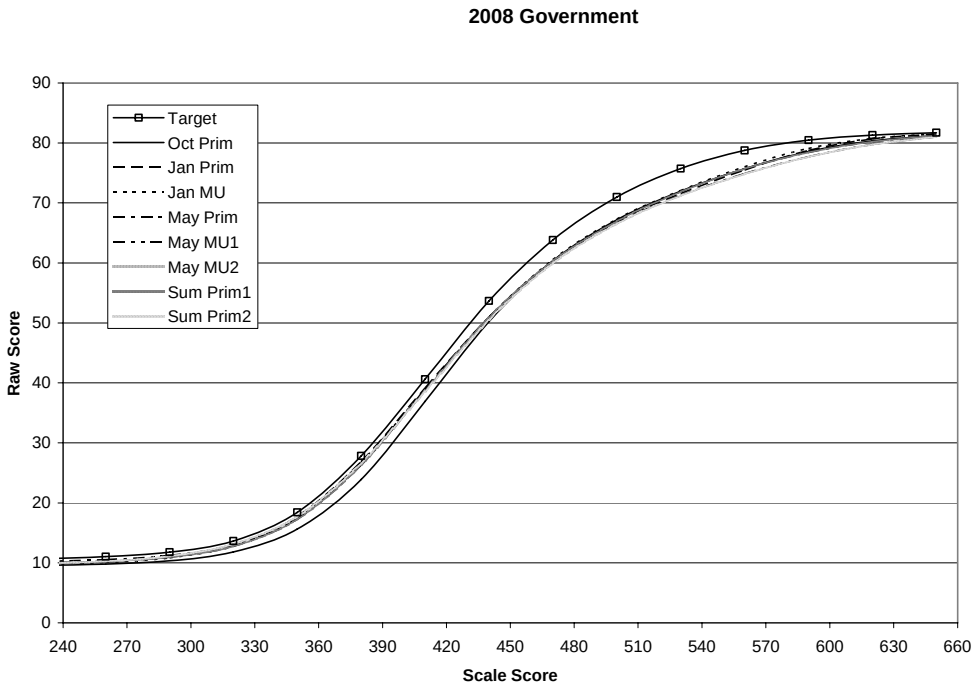

Note: Maximum possible raw score is 60.
English Cut Scores: Proficient 396, Advanced 429

Figure1.6 Conditional Standard Error of Measurement for the 2008 English Forms



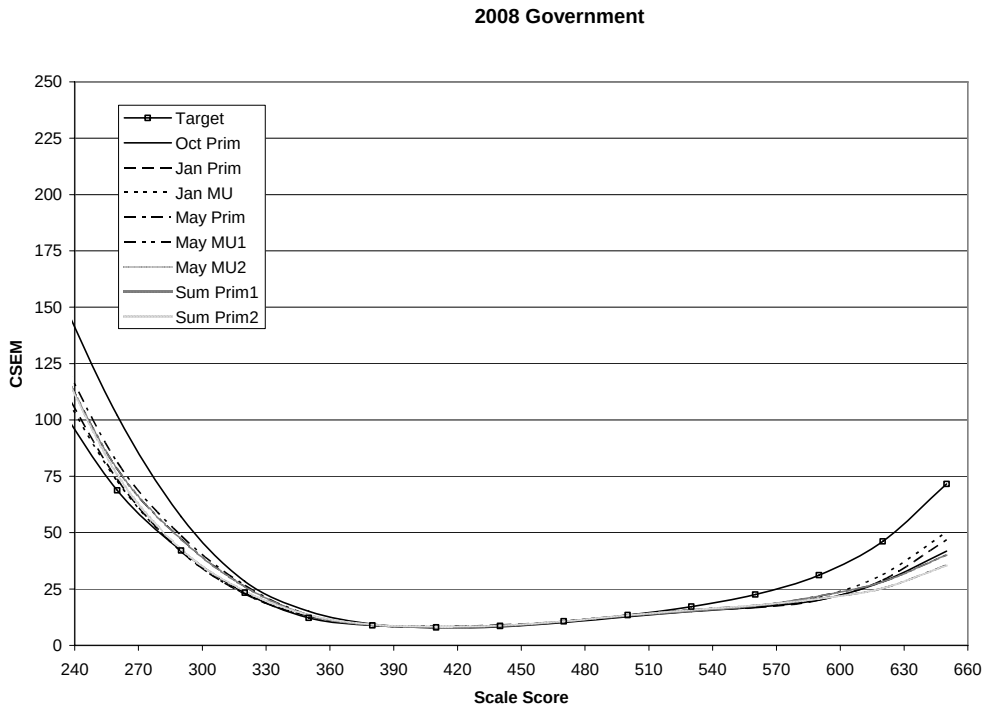

Note: Observed standard deviations for English ranged from 25.6 to 37.0.
English Cut Scores: Proficient 396, Advanced 429

Figure1.7 Test Characteristic Curves for the 2008 Government Forms




Note: Maximum possible raw score is 82.
Government Cut Score: Proficient 394

Figure1.8 Conditional Standard Error of Measurement for the 2008 Government Forms




Note: Observed standard deviations for Government ranged from 23.9 to 40.9.
Government Cut Score: Proficient 394

Figure1.9 Histogram of Total Scale Scores for May 2008 Algebra

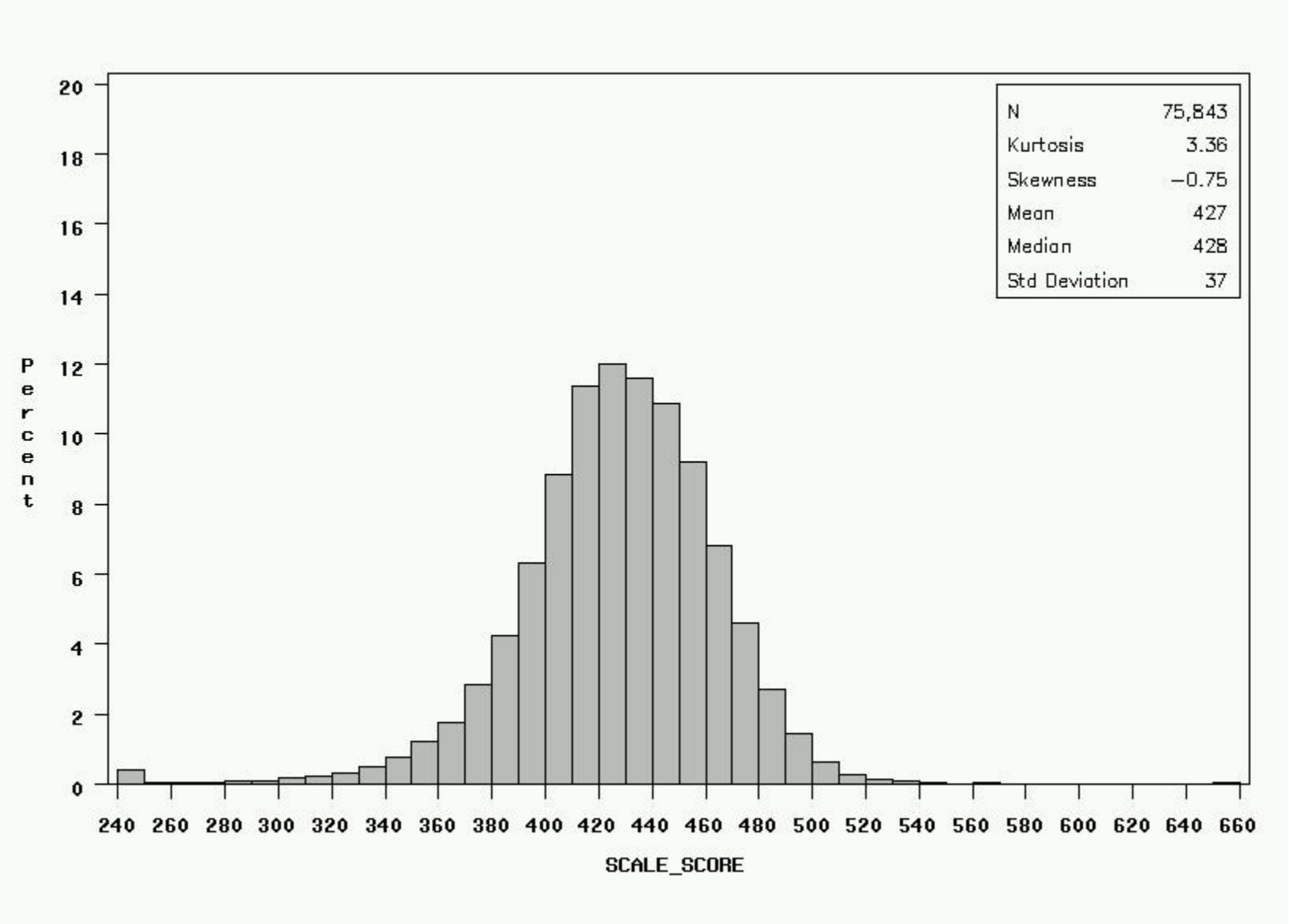


Figure1.10 Histogram of Total Scale Scores for May 2008 Biology

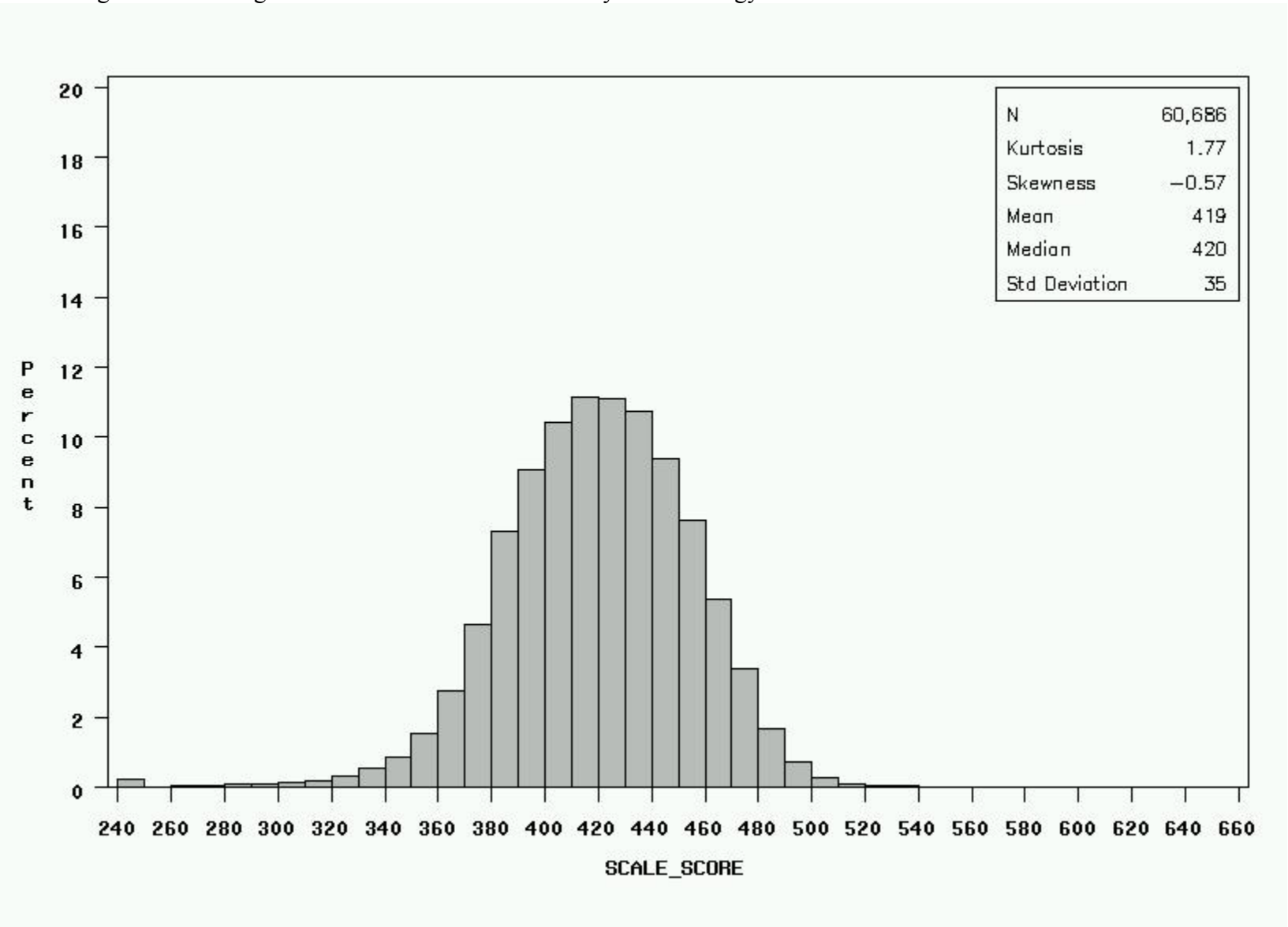


Figure1.11 Histogram of Total Scale Scores for May 2008 English

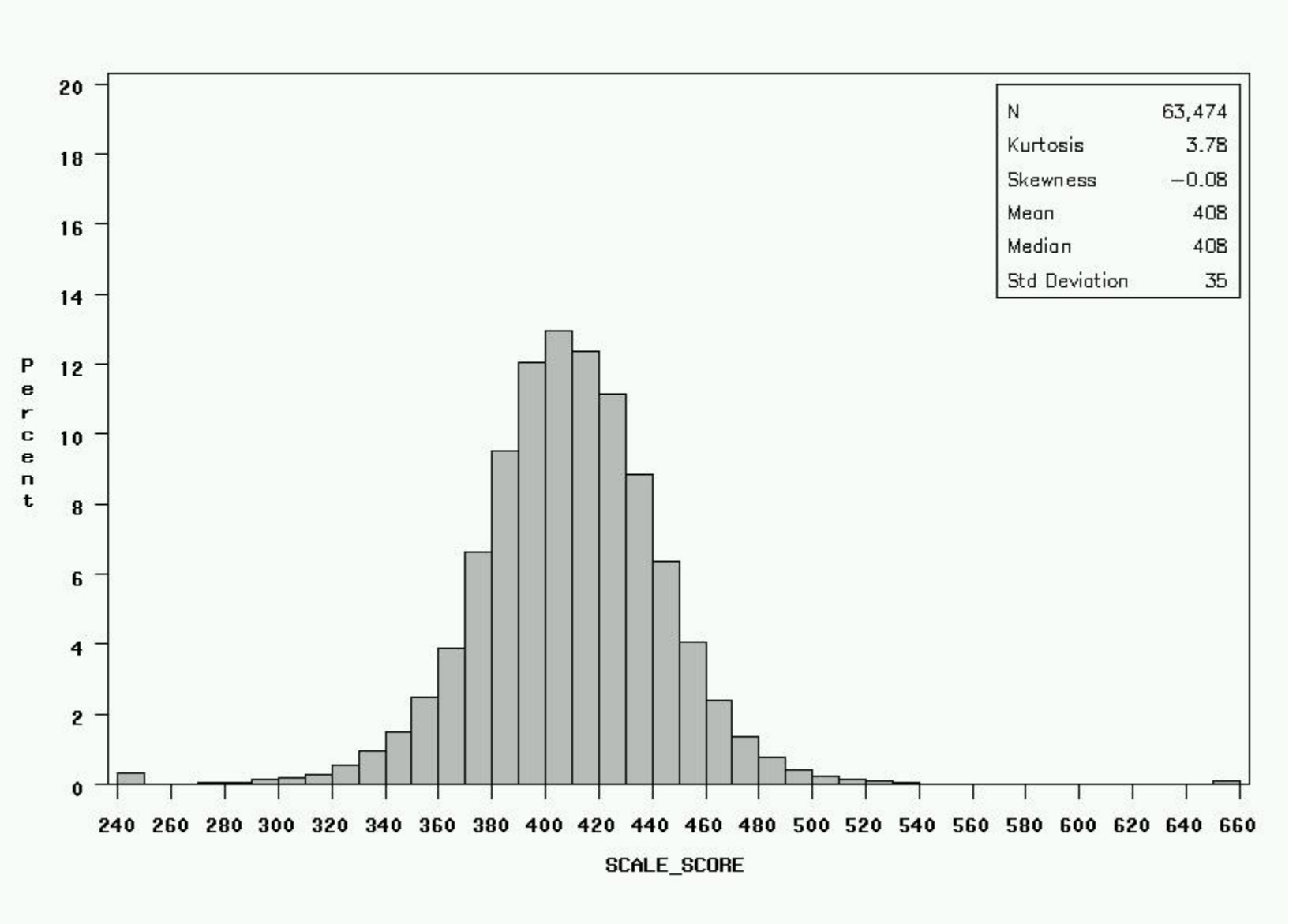
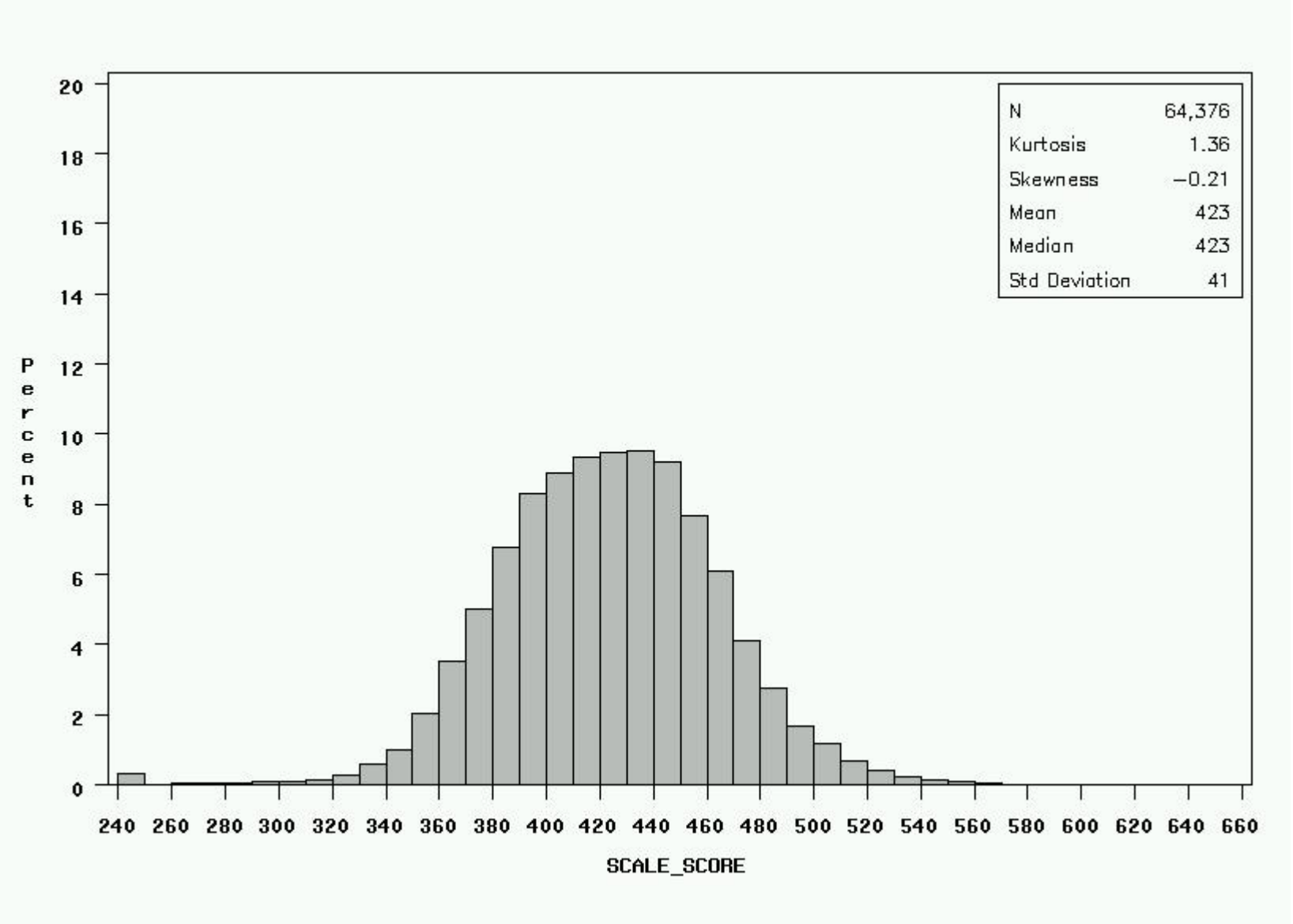


Figure1.12 Histogram of Total Scale Scores for May 2008 Government



## Test Administration

All HSA tests administered in October 2007, and in January, May, and Summer 2008 were paper-and-pencil tests, except in the case of the Kurzweil special form, which had an audio portion. Primary forms were administered during the first week of testing. For the January and May administrations, Make-up form 1 was administered during the second week. For the May administration only, Make-up Form 2 was administered in the third week of testing. All forms administered without accommodations have the following timing limits:

Table 1.11 Testing Timing Schedule in Minutes by Content

| Content | 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 |

If more than five percent of students omit an SR or SPR item, or more than 15% of students omit a CR item, the item is flagged as having a high omit rate. Table 1.12 shows omit rates for each content area by administration and item type. Relatively few SR items were flagged for omit rate. Most of the items flagged for omit rate were SPR and CR items, which tend to have higher omit rates in general, because students have to generate a response rather than choose one from available options. The tendency for SPR and CR items to have higher omit rates is consistent with findings from previous test years.

An exception to this was the Summer 2008 administration of the English test, where eleven SR items had omit rates higher than 5%. There were two Primary forms administered in Summer 2008. Form P was administered to 1324 students; Form Q was administered to 74 students. All eleven items flagged for high omit rate were on Form Q. None of the eleven items were located at the end of a test session. Ten of the eleven items flagged had an omit rate of 5.4%, which means 4 students omitted each item. The remaining item had an omit rate of 6.8%, which means that 5 students omitted the item. Appendices A and B include omit rates for each operational item and field test item, respectively, administered in October 2007, January, May and Summer 2008.

Table 1.12 Number of Operational Items Flagged for High Omit Rate

| | October | | | January | | | May | | | Summer | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Item Types | | | Item Types | | | Item Types | | | Item Types | | |
| | SR | SPR | CR | SR | SPR | CR | SR | SPR | CR | SR | SPR | CR |
| Algebra | 0 | 4 | 3 | 3 | 6 | 3 | 0 | 4 | 1 | 2 | 9 | 4 |
| Biology | 0 | -- | 5 | 0 | -- | 2 | 0 | -- | 0 | 0 | -- | 2 |
| English | 0 | -- | 0 | 0 | -- | 0 | 0 | -- | 0 | 11 | -- | 2 |
| Government | 0 | -- | 1 | 0 | -- | 1 | 0 | -- | 1 | 0 | -- | 5 |