HSAs

Maryland High School Assessments

Maryland High School Assessment
Technical Report

Algebra and Data Analysis
Biology
English
Geometry
Government

Educational Testing Service
December 2005

Forward

The technical information included in this report is intended for use by those who evaluate tests, interpret scores, or use test results in making educational decisions. It is assumed that the reader has some technical knowledge of test construction and measurement procedures, as stated in Standards of Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, 1999).

Table of Contents

Introduction

The 2005 Maryland High School Assessments (MHSA) consisted of end-of-course tests in Algebra/Data Analysis, Biology, English, Geometry, and Government. The MHSA is referred to as “end-of-course” tests, because students took each test as they completed the appropriate coursework. In addition, results from the English and Geometry administrations were used as the High School English Language Arts and Mathematics components in the Maryland State Department of Education (MSDE) Adequate Yearly Progress reports as required under the No Child Left Behind (NCLB) act for the 2005 school year. In the 2006 school year, Algebra will replace the Geometry test as the NCLB reporting content and Geometry test will no longer be administered for the MHSA. A new English test administered at the 10th grade replaced the old English test which was administered at the 9th grade. The technical details of the new English test are described in Section 6.

MHSA consisted of selected-response (SR) items, which required students to choose between four short response options; brief constructed response (BCR) items, which required students to write a short response; and extended constructed response (ECR), which required students to write a longer response. The SR items were machine scored; the BCR and ECR items were scored by raters. In addition, Algebra/Data Analysis and Geometry included items based on student-produced response (SPR), which required students to grid in correct responses on the answer document. All items were based on content outlined in Maryland’s Core Learning Goals.

MHSA in the content areas of Algebra, Biology, Geometry, and Government were administered in January, May and July. The new English test was administered in May and July. In general, for January and May 2005 administrations, three operational test forms were constructed: one for the primary administration window, and one for each of two make-up administrations. There were two forms constructed for the Summer 2005 administration: one for the first week of testing and one for the second week of testing. Each test form for all content areas except English consisted of two types of items: operational and field test. Operational items were common across each of the operational forms and were used to produce student scores; field test items were not scored operationally, but were analyzed and placed into the item bank for future test form construction. The English forms consisted of all field test items and items selected for score reporting were determined after the test administration. Detailed information about how scoring items were selected is in Section 6. All English items were analyzed and placed into the item bank for future test form construction. For the other content areas, with the exception of items selected for public release, all operational items were returned to the item bank where they will remain unused for at least two years to minimize item exposure.

The underlying item response models used for MHSA were the three-parameter logistic (3PL) model and the two-parameter partial credit (2PPC) model, also known as the generalized partial credit model (GPCM; see Section 5). For each content area, both total test scores and subscores were calculated for students. The total test scores were reported to individual students and were based on item-pattern (IP) scoring (mean 400, standard deviation 40). Subscores were also reported based on associated item parameters, though these scores were obtained using number-correct (NC) to scale-score (SS) tables While subscores were not reported at individual student level, the subscores were aggregated at the classroom level to provide teachers and administrators with additional information about student performance in each of the reporting categories.

Beginning with the 2004 administration, a pre-equated design was implemented while scores from previous administrations were based on parameters that were estimated following the administration (post-equated[1]). In the pre-equated design, item parameters were not updated following an administration; instead existing bank parameters were used to produce student scores. Using this design, scores can be calculated and assigned to students immediately after the answer documents have been scored.

All technical support and analyses were carried out in accordance with both ETS Standards for Quality and Fairness and Standards for Educational and Psychological Testing, issued jointly by the American Educational Research Association, American Psychological Association, and National Council on Measurement in Education.

This report is divided into 5 sections: Section 1 describes test development, form construction and administration details; Section 2 discusses the validity and reliability of the MHSA; Section 3 describes the scoring procedures and score types; Section 4 provides statistical summary results for each of the test forms administered in 2005; Section 5 describes the analyses conducted using the field test data including classical item analyses, differential item functioning, and item response theory calibrations and equating; and Section 6 provides information regarding the English MHSA exams.

[1] In the post-equated design, anchor items representative of the content and difficulty of the test forms were used to equate the test forms using a Stocking and Lord procedure (CTB/McGraw-Hill, December, 2003).

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. 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, the remaining numbers of items to be developed (which is determined by the requirements set forth in the RFP) were distributed among the remaining indicators in a fashion that would best ensure sufficient depth of items from which to construct operational 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 1.6. 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 operational forms was based on test blueprints as approved by MSDE.

Item Type

There were four item types that were utilized by the MHSA exam. These item types were selected response (SR), student produced response (SPR), brief constructed response (BCR), and extended constructed response (ECR). The following table shows how these item types were used on operational forms.

Table 1.1 Number of Items on Operational MHSA Forms by Item Type

Content Area	SR	SPR	BCR	ECR
Algebra	26	6	3	3
Biology	48	-	7	-
English	46	-	2	2
Geometry	26	6	2	3
Government	50	-	7	1

Item Writing

Item writers, at least 50 percent of which were Maryland educators, were contracted to develop quality test items that were aligned with Core Learning Goals. Item writers were selected based on their depth of content knowledge and familiarity with MHSA testing program. The item writers were trained on general item writing techniques as well as writing parameters that were specific to the MHSA program. Approximately one month after the initial item writer training, writers were provided a follow-up training session geared to evaluate their writing skills developed up to that point and 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 started item review and revision process. Specific requirements of writing for the MHSA program can be found in the "Guidelines for Item Writers" document.

Item Review and Revision

All items developed for this program 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 and comprehension, style, usage, consistency and sensitivity.
- Item content reviewed so that each item measures intended Goal-Expectation-Indicator.
- Copyright and/or trademark permission has been obtained for any required materials.
- Internal reviews conducted and historical records will be maintained for all version changes.

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

MSDE performed a review of the items and provided feedback to ETS content specialists. These edits were incorporated into the items, then MSDE and ETS content specialists met and conducted a side-by-side review of the items. Any final edits to the items were made. The items were then prepared for Content Review Committee review. All constructed response items were also submitted to Measurement Incorporated (MI) for review.

The final round of reviews involved the Content Review Committee and Bias/Fairness Review Committee. These committees were diverse groups of Maryland educators who reviewed each item and ensured that content in each item accurately reflected what was taught in Maryland schools and that no individual or group would be unfairly favored or disadvantaged due to the content of the items.

Upon the completion of this final round of review, MSDE and ETS content specialists again conducted a side-by-side meeting to evaluate reviews by MI, Content Review Committee, and Bias/Fairness Review Committee. The ETS content specialist then made any necessary edits to the items. The items that survived this process were ready to be placed in field test sections of operational forms.

Test Specifications

All the 2005 operational test forms were constructed from items from the Maryland item bank. The pool of items available for use in the construction of the 2005 forms included all items that had been administered, calibrated and linked to the operational scale. The MHSA operational scale was defined in 2002 and included items administered in 2002 and 2003. Items administered prior to 2002 were not eligible for selection of the 2005 forms. In addition, items flagged for poor fit and items that had been flagged for severe differential item functioning (DIF) against one of the focal groups were excluded from the available item pool. Refer to Section 5 for a more detailed account of these analyses and flagging criteria.

Each test included a mixture of selected-response (SR), as well as brief and/or extended constructed-response (BCR, ECR) items. Algebra/Data Analysis and Geometry also included student produced response (SPR) items. Each test form consisted of two sections administered within a single sitting (the two sections were separated by a short break). SR and SPR items were worth one score point and were scored against specific keys. BCR and ECR items varied in number of score points by content area. In Algebra and Geometry BCR items were worth three points and ECR items were worth four points. English BCR items were worth three points and ECR items were worth four points. The BCR and ECR items for Government were both worth four points and Biology had only BCR items, which were worth four points. Rubrics for items can be found at the following locations:

Algebra and Geometry:	http://mdk12.org/rubrics/mathematics.
Biology	http://mdk12.org/rubrics/science
English	http://mdk12.org/rubrics/english
Government	http://mdk12.org/rubrics/socialstudies

In addition, each test form was constructed to meet specific test blueprints. Tables 1.2 to 1.6 indicate distribution of items within each reporting category by item type.

Table 1.2 Algebra Blueprint

ALGEBRA/DATA ANALYSIS					
Reporting Category	Item Type				
	SR	SPR	BCR	ECR	Percent of Points
	(4pts/ECR)	(3 pts/BCR)	(3 pts/BCR)	(4 pts/ECR)	
Totals	26	6	3	3	
Expectation 1.1 The student will analyze a wide variety of patterns and functional relationships using the language of mathematics and appropriate technology.					25%
Expectation 1.2 The student will analyze a wide variety of patterns and functional relationships using the language of mathematics and appropriate technology.					32%
Expectation 3.1 The student will collect, organize, analyze, and present data.					22%
Expectation 3.2 The student will apply the basic concepts of statistics and probability to predict possible outcomes of real-world situations.					21%

Table 1.3 Biology Blueprint

BIOLOGY			
Reporting Category	ITEM TYPE		Percent of Points
	SR	CR	
	(1 pt/SR)	(4 pts/CR)	
Totals	48	7	
Goal 1 Skills and Processes of Biology			21%
Expectation 3.1 Structure and Function of Biological Molecules			16%
Expectation 3.2 Structure and Function of Cells and Organisms			17%
Expectation 3.3 Inheritance of Traits			17%
Expecation 3.4 Mechanism of Evolutionary Change			12%
Expectation 3.5 Interdependence of Organisms in the Biosphere			17%

Table 1.4 English Blueprint

ENGLISH				
Reporting Category	ITEM TYPE			Percent of Points
	SR	BCR	ECR	
	(1pt/SR)	(3pt/BCR)	(4pt/ECR)	
TOTALS	46	2	2	
1: Reading and Literature: Comprehension and Interpretation (RC) Includes the following indicators: 1.1.1; 1.1.2; 1.1.3; 1.2.1; 1.3.3; 3.2.2				27%
2: Reading and Literature: Making Connections and Evaluation (RE) 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				23%
3: Writing – Composing (WC) 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				27%
4: Language usage and Conventions (WL) Includes the following indicators: 3.1.3; 3.1.4; 3.1.6; 3.1.8; 3.3.1; 3.3.2				23%

Table 1.5 Geometry Blueprint

GEOMETRY					
Reporting Category	ITEM TYPE				Percent of Points
	SR	SPR	BCR	ECR	
	(1pt/SR)	(1 pt/SPR)	(3 pt/BCR)	(4 pt/ECR)	
Totals	26	6	2	3	
Expectation 2.1 The student will represent and analyze two and three dimensional figures using tools and technology when appropriate.					32%
Expectation 2.2 The student will apply geometric properties and relationships to solve problems using tools and technology when appropriate.					34%
Expectation 2.3 The student will apply concepts of measurement using tools and technology when appropriate.					34%

Table 1.6 Government Blueprint

GOVERNMENT				
Reporting Category	ITEM TYPE			Percent of Points
	SR	BCR	ECR	
	(1 pt/SR)	(4 pt/BCR)	(4 pt/ECR)	
Totals	50	7	1	
Expectation 1.1 The student will demonstrate understanding of the structure and functions of government and politics in the United States				26-31%
Expectation 1.2 The student will evaluate how the United States government has maintained a balance between protecting rights and maintaining order.				23-28%
Goal 2 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.				15%
Goal 3 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.				13%
Goal 4 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.				18%

Item Selection and Form Design

In order to conserve the item pool, the operational set of items consisted of both a common set of items shared across forms within an administration and also a unique set of items. Approximately 60% of the total form was common across each of the operational test sections within each of the January and May forms. The balance of the forms consisted of different mixtures of items depending on the form. The guidelines used to construct the forms are listed in Tables 1.7 and 1.9. The exact composition of the forms varied slightly based on available items in the pool.

Table 1.7. Form Construction Specifications – January 05 Administration

Primary Week Form A	**Primary Week Form B**	**Make-Up #1 Form C**	**Make-Up #2 Form D**
Common set – 60%	Common set – 60%	Common set – 60%	Common set – 60%
Unique Items from the pool – 40% (same as items in Form B)	Unique Items from the pool – 40% (same as items in Form A)	Half of the items from primary week's 40% – 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 items	Field Test Section – unique items	Field Test Section – same as Form A	Field Test Section – same as Form A

Table 1.8. Form Construction Specifications – May 05 Administration

Primary Week Forms E -K	**Make-Up #1 Form X**	**Make-Up #2 Form Y**
Common Set –60%	Common Set –60%	Common Set – 60%
Items from the pool – 40% (the same for Forms E – *)	Half of items from primary week's40% items – 20%	Other half of items from primary week's40% items – 20%
	Unique items from the pool – 20%	Unique items from the pool – 20%
Field Test Section – unique sets of items for Forms E through K	Field Test Section – same as Form E	Field Test Section – same as Form E

Table 1.9. Form Construction Specifications – 2005 Summer Administration

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 May 05 forms	Field Test Section – items repeated from May 05 forms

In addition to the operational items, embedded field test items were included with each version of the test form, resulting in several versions of the operational form that differed only by the included field test items. These items consisted of either newly written items or previously administered items that had poor item statistics and/or had been revised. Items eligible for re-field testing included items from the 2000-2001 administration years. These items were judged to be acceptable from a content perspective, but had 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 one best answer, but with positive point-biserials on one or more distractors. For the administration, different versions of the forms were spiraled at the student level.

Forms were constructed using the test construction software associated with the customer item bank. The goal was to match the conditional standard error curve (CSEM) and test characteristic curves (TCC) with the "target" form defined as the base form used to set the operational scale in 2002. The information function, standard error curve, and test characteristic curve were graphical displays based on the item parameters associated with the items selected and were inter-related – that is, changes to the set of items selected will result in changes in all three displays.

The following were general steps completed during the test construction process.

1. For each administration, all operational 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 test characteristic and standard error curves.
3. During the test construction procedure test developers were careful to ensure that the item selections met all content specifications, including matching items to the test blueprint, distribution of keys, removal of clueing or clang, etc.
4. After the operational forms were selected, the field test sets were constructed. Field test sets did not need to meet any psychometric criteria, but were selected such that the items could be completed within a 30-minute time frame. Field test sets consisted of a set of multiple choice items, a combination of brief constructed

response items and multiple choice items, or an extended constructed response item. The field test items were embedded throughout the set of operational items.

Section 2. Validity

Validity is one of the most important attributes of assessment quality. It refers to the degree to which evidence supports the interpretations of test scores by proposed users of tests and is one of the most fundamental considerations in developing and evaluating tests (AERA, APA, & NCME, 1999). Validity is not based on a single study or type of study, but should be considered an ongoing process of gathering evidence supporting the interpretation of the resulting test scores. This process begins with the test design and continues throughout the entire assessment process, including design, content specifications, item development, psychometric quality and inferences made from the results.

The development of test content for each MHSA was overseen by a content expert who has a depth of knowledge and teaching experience related to the course in which the MHSA was administered. The appropriate content leads that had similar qualifications reviewed the test development work of these individuals.

The test development process itself provided numerous opportunities for the client to review test content and make changes to ensure that the items, both individually and as collections within forms, were valid measures of the knowledge and skills of Maryland students according to course standards. Every item that was created is referenced to a particular instructional standard (goal, expectation, and indicator). At various points during the internal ETS development process, that specific reference was either confirmed or changed to reflect changes to the item. When the item went to a committee of Maryland educators for a content review, the members of the committee made individual judgments on the match of the item content with the standard it was intended to measure and the appropriateness for the typical age of students being tested. These judgments were tabulated and reviewed by the content experts who use the information to decide which items will advance to the field test stage of development.

The constructs measured by each MHSA were described in detail in the Maryland high school curriculum standards (Core Learning Goals). All ETS content staff working on item development had been trained in the Core Learning Goals. The test blueprint documents presented in Section 1 (see Tables 1.2 to 1.6) were created in collaboration with committees of Maryland educators and were directly derived from the Maryland goals, expectations, and indicators. These Learning Goals can be found on the MSDE website at http://www.mdk12.org.

Although all eligible students participated in the MHSA and information about student performance was provided to students, parents, teachers and other stakeholders, scores for all content areas had no consequences for individual students during this time. Geometry and English scores were also used for AYP as a component of the Maryland No Child Left Behind (NCLB) Accountability program. Information on the interpretation of scores was provided to students, parents, schools and other stakeholders via the MSDE website.

In addition to the validation documentation gathered and maintained by MSDE, reliability analyses were computed. The results are presented in Section 4. This report contains relevant empirical information in support of the Maryland HSA as follows.

- Section 3 provides detailed information concerning the scores that were reported for the MHSA, and the cut-scores for each content area.

- Section 4 provides demographic information for the population of students who were administered the MHSA. Summary statistics at the test level were reported for the student population and for subgroups. In addition, reliability analyses and two measures of decision consistency were provided for the student population.

- Section 5 includes documentation regarding the field test analyses. Descriptions of classical item analyses, differential item functioning, item response theory calibration and scaling are included. In addition, summary tables of item p-values and item-total correlations are provided.

- Section 6 provides information regarding the English test. The following are included: a description of the selection of operational items and scaling of items, factor analyses of forms, summary statistics of student achievement, and measures of classification consistency.

Section 3. Scoring Procedures and Score Types

Scale Scores

Scale scores based on maximum likelihood estimates (MLE) were reported for the total test score. All scores were reported on the operational reporting scale established in 2003. While the total test score was based on item-pattern (IP) scoring, the subscores were based on number-correct (NC) to scale score scoring tables.

With IP scoring, because the likelihood equation can have multiple maxima with the 3PL model, a numerical method was developed that found the scale score at the global maximum in the likelihood function. NC to scale score scoring tables were obtained by inversing the test characteristic curves (TCC) of items contributing to the associated subscores. The procedure produced what Yen (1984) called 'number correct trait estimates,' which is referred to as 'NC scale scores' in this report.

Conditional Standard Errors of Measurement

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

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

where,

$\mathrm{SEM}(\hat{\theta})$=standard error of measurement

$\mathrm{I}(\theta)$= test information function.

The test information function is the sum of corresponding information functions of the test items when optimal item weights are used, as in the MHSAs. 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 than selected response items, because they have more score points.

Lowest and Highest Obtainable Test Scores

Both the maximum likelihood procedure and NC scoring cannot produce scale score estimates for students with perfect scores or scores below the level expected by guessing. Also, while maximum likelihood estimates were available for students with extreme scores other than zero or perfect, occasionally these estimates have very large conditional SEMs, 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 either number-correct (NC) or item-pattern scoring. Table 3.1 lists the LOSS and HOSS values for each content area established at the beginning of the MHSA program. MSDE decided that the LOSS and HOSS values for the Summer of 2005 and subsequent administrations would be 240 and 650, respectively, for all content areas.

Table 3.1 LOSS and HOSS Values

Content	LOSS	HOSS
Algebra	240	625
Biology	260	650
English I	240	650
Geometry	275	575
Government	260	650

Cut-Scores

The cut-scores associated with each of the performance levels in the non-English content areas were established by MSDE in 2003 (refer to Table 3.2). The English cut-scores were established during the standard setting study held in October of 2005. One cutscore was established for all of the content areas except for Geometry and English. Because Geometry and English results are used as the High School Mathematics and English Language Arts components of the MD accountability plan under NCLB, two cut-scores were established.

Table 3.2 MHSA 2005 Cut-Scores

	Cutscore	
Content Area	Proficient	Advanced
Algebra	412	
Biology	400	
Geometry	411	447
Government	394	
English	396	429

Section 4. Test-Level Analyses

This chapter summarizes the test-level statistics obtained for the January and May 2005 administrations of the MHSA. The test-level analyses include demographic distributions, reliability analyses, summary statistics, and decision consistency.

Demographic Distributions

All eligible students completed the MHSA, though the scores were not used for individual accountability during this time. The demographic characteristics of the students were presented in Tables 4.1 to 4.4 for the January and May administrations of Algebra, Biology, Geometry, and Government, respectively. The number of students participating in the May administration was greater than the number of students participating in the January administration. As a result, only two field test versions were included in the January administration to ensure sufficient samples for the analyses of the field test items. Due to the small numbers of students participating in the July administration, the May field test sections were repeated to ensure that the test length was comparable.

Table 4.1. Demographic Information for Algebra

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-Up Forms	
		N	%	N	%	N	%	N	%
Overall		5275		554		64431		3828	
Gender									
	Male	2476	46.9	267	48.2	32353	50.2	1985	51.9
	Female	2785	52.8	280	50.5	32075	49.8	1842	48.1
	Missing	14	0.3	7	1.3	3	0.0	1	0.0
Special Education									
	Yes	67	1.3	8	1.4	6210	9.6	527	13.8
	No	5205	98.7	545	98.4	57230	88.8	3235	84.5
	504	3	0.1	1	0.2	991	1.5	66	1.7
Ethnicity									
	American Indian	27	0.5	5	0.9	248	0.4	20	0.5
	Asian/Pacific Islander	121	2.3	11	2.0	3472	5.4	108	2.8
	African American	2056	39.0	296	53.4	24339	37.8	1542	40.3
	White	2812	53.3	197	35.6	32308	50.1	1927	50.3
	Hispanic	206	3.9	23	4.2	4060	6.3	230	6.0
	Missing	53	1.0	22	4.0	4	0.0	1	0.0
Limited English Proficient									
	Yes	3	0.1	0	0.0	1818	2.8	100	2.6
	No	5272	99.9	554	100.0	62155	96.5	3704	96.8
	Exited	0	0.0	0	0.0	458	0.7	24	0.6

Table 4.2. Demographic Information for Biology

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-Up Forms	
		N	%	N	%	N	%	N	%
Overall		7711		609		47996		2161	
Gender									
	Male	3674	47.6	285	46.8	23820	49.6	1161	53.7
	Female	4027	52.2	320	52.5	24173	50.4	1000	46.3
	Missing	10	0.1	4	0.7	3	0.0	0	0.0
Special Education									
	Yes	85	1.1	8	1.3	4267	8.9	329	15.2
	No	7614	98.7	599	98.4	42945	89.5	1789	82.8
	504	12	0.2	2	0.3	784	1.6	43	2.0
Ethnicity									
	American Indian	18	0.2	3	0.5	179	0.4	12	0.6
	Asian/Pacific Islander	149	1.9	10	1.6	3145	6.6	54	2.5
	African American	2444	31.7	289	47.5	15456	32.2	958	44.3
	White	4827	62.6	267	43.8	26268	54.7	977	45.2
	Hispanic	230	3.0	27	4.4	2944	6.1	158	7.3
	Missing	43	0.6	13	2.1	4	0.0	2	0.1
Limited English Proficient									
	Yes	2	0.0	0	0.0	1344	2.8	50	2.3
	No	7707	99.9	609	100.0	46220	96.3	2095	96.9
	Exited	2	0.0	0	0.0	432	0.9	16	0.7

Table 4.3. Demographic Information for Geometry

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-Up Forms	
		N	%	N	%	N	%	N	%
Overall		6978		550		56631		2442	
Gender									
	Male	3163	45.3	253	46.0	27820	49.1	1296	53.1
	Female	3778	54.1	288	52.4	28810	50.9	1146	46.9
	Missing	37	0.5	9	1.6	1	0.0	0	0.0
Special Education									
	Yes	22	0.3	6	1.1	4759	8.4	307	12.6
	No	6945	99.5	544	98.9	50942	90.0	2093	85.7
	504	11	0.2	0	0.0	930	1.6	42	1.7
Ethnicity									
	American Indian	22	0.3	2	0.4	196	0.3	10	0.4
	Asian/Pacific Islander	175	2.5	14	2.5	3419	6.0	70	2.9
	African American	1844	26.4	236	42.9	19923	35.2	1136	46.5
	White	4660	66.8	250	45.5	30060	53.1	1069	43.8
	Hispanic	193	2.8	31	5.6	3032	5.4	157	6.4
	Missing	84	1.2	17	3.1	1	0.0	0	0.0
Limited English Proficient									
	Yes	2	0.0	0	0.0	1115	2.0	45	1.8
	No	6974	99.9	550	100.0	54968	97.1	2374	97.2
	Exited	2	0.0	0	0.0	548	1.0	23	0.9

Table 4.4. Demographic Information for Government

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-Up Forms	
		N	%	N	%	N	%	N	%
Overall		6783		758		46837		2089	
Gender									
	Male	3304	48.7	366	48.3	23327	49.8	1170	56.0
	Female	3470	51.2	383	50.5	23508	50.2	918	43.9
	Missing	9	0.1	9	1.2	2	0.0	1	0.0
Special Education									
	Yes	53	0.8	14	1.8	4267	9.1	331	15.8
	No	6719	99.1	741	97.8	41764	89.2	1719	82.3
	504	11	0.2	3	0.4	806	1.7	39	1.9
Ethnicity									
	American Indian	22	0.3	6	0.8	171	0.4	6	0.3
	Asian/Pacific Islander	132	1.9	18	2.4	3115	6.7	75	3.6
	African American	2230	32.9	366	48.3	13134	28.0	858	41.1
	White	4035	59.5	299	39.4	27587	58.9	981	47.0
	Hispanic	315	4.6	48	6.3	2829	6.0	167	8.0
	Missing	49	0.7	21	2.8	1	0.0	2	0.1
Limited English Proficient									
	Yes	7	0.1	2	0.3	1333	2.8	70	3.4
	No	6776	99.9	756	99.7	45020	96.1	2000	95.7
	Exited	0	0.0	0	0.0	484	1.0	19	0.9

Reliability

Reliability describes the extent to which differences in test scores reflect true differences in the knowledge, ability, or skill being tested rather than fluctuations due to chance or factors other than those which were being tested. The variance in the distributions of test scores (i.e., the differences among individuals) is partly due to real differences in the knowledge, skill, or ability being tested (true variance) and partly due to random errors in the measurement process (error variance). The number used to describe reliability is an estimate of the proportion of the total variance that is true variance. Several different ways of estimating this proportion exist. The estimates of reliability reported in this report were internal-consistency measures, which were derived from analysis of the consistency of the performance of individuals on items within a test (internal-consistency reliability). Therefore, the estimates apply only to the test form being analyzed. They do not take into account form-to-form variation due to equating limitations or lack of parallelism, nor were they responsive to day-to-day variation due to, for example, state of health or testing environment. Reliability coefficients may range from 0 to 1. The higher the reliability coefficient for a set of scores, the more likely individuals would be to obtain very similar scores if they took another form of the test. The formula for the internal consistency reliability as measured by Cronbach's Alpha (Cronbach, 1951) is reported below:

$$\alpha = \frac{n}{n-1}\left[1-\frac{\sum_{i=1}^{n}\sigma_i^{\;2}}{\sigma_x^2}\right]$$

where n is the number of items, σ_i^2 is the variance of scores on the *i*-th item, and σ_x^2 is the variance of the total score (sum of scores on the individual items).

Since all five MHSA have mix item type (both dichotomous and polytomous items), it is more appropriate to report stratified Alpha (Feldt & Brennan, 1989). The stratified Alpha is a weighted average of Cronbach's Alpha for item sets with different maximum score points or "strata." The formula for calculating the stratified Alpha is:

$$_{strata\alpha}\rho = 1-\frac{\sum\sigma_{X_j}^2(1-\alpha_j)}{\sigma_X^2}$$

where $\sigma_{X_j}^2$ is the variance for strata j of the test,

σ_X^2 is the total variance of the test, and

α_j is the Cronbach's Alpha for strata j of the test.

The results for the reliability analyses of the total test score are presented with the summary statistics in Tables 4.9 to 4.16. The reliability results indicate that all of the MHSA were highly reliable: reliabilities ranged from 0.91 to 0.95 for the primary forms, and from 0.85 to 0.94 for the make-up forms. In general, the make-up forms had slightly lower reliabilities than the primary forms. Because the make-up forms tended to have lower mean scale scores, the lower reliabilities may be related to a decrease in true-score variance.

Summary Statistics

Table 4.5 presents mean scale scores by content area for the January and May administrations. The mean scores for Algebra, Biology, and Government were higher for the May administration, whereas the mean score for Geometry was higher for the January administration. The difference between the January and May mean scores was less than 5 for all exams except Government, which yielded a difference of approximately 10.6.

Table 4.5. Mean Scores by Administration

	Jan-05			May-05		
	N	Mean	SD	N	Mean	SD
Algebra	5275	406.30	39.79	68259	411.27	50.58
Biology	7711	402.71	35.73	50157	406.70	42.48
Geometry	6978	411.61	31.02	59073	407.13	47.46
Government	6783	401.57	35.09	48926	412.14	42.44

Table 4.6 presents mean scale scores from 2003 to 2005 for each content area. The mean scores for Algebra were within 4 points, whereas the mean scores for Biology and Government were within 6 points. The largest change was evident for Geometry where a 7.6 point gain was observed from 2003 to 2005.

Table 4.6. Comparison of Mean Scores

	2003	2004	2005
Algebra	408.3	411.9	409.5
Biology	400.8	406.2	404.7
Geometry	398.8	405.2	406.4
Government	403.5	406.5	409.3

Table 4.7 presents the passing rates for Algebra, Biology and Government. As can be seen from the table, passing rates for Algebra and Biology improved approximately 6 and 8 percent, respectively, from 2003 to 2004, but declined slightly from 2004 to 2005. However, the passing rates for Government increased steadily from 2003 to 2005.

Table 4.7. Comparison of Passing Rates

	2003	2004	2005
Algebra	53.1	59.3	54.5
Biology	54.3	62.0	58.4
Government	39.8	54.6	67.1

Table 4.8 presents the percent of Geometry students classified as basic, proficient, and advanced from 2003 to 2005. Generally, there was a decline in the percent of students in the basic category, a slight increase in the percent of students in the proficient category, and an increase in the percent of students in the advanced category.

Table 4.8. Comparison of Classification Rates for Geometry

	2003	2004	2005
Basic	56.6	51.9	45.5
Proficient	33.2	36.1	36.8
Advanced	10.2	12.0	17.7

Summary statistics for all students and for subgroups based on gender, special education programs, ethnicity, and English language fluency are presented in Tables 4.9 through 4.16. The tables include number of students tested for whom valid scores were available, mean scale scores, and standard deviation of scale scores. In addition, test reliabilities are provided for the overall group of examinees. Information is presented for the primary forms of the content area, followed by the make-up forms. In all content areas, higher mean scores were noted for the primary forms compared to the make-up forms.

Table 4.9. Summary Statistics for Algebra Primary Forms

		January				May			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		406.30	39.79	5275	0.91	411.27	50.58	68259	0.93
Gender									
	Male	406.27	41.08	2476		408.22	55.33	34338	
	Female	406.55	38.51	2785		414.37	45.05	33917	
	Missing	*	*	14		*	*	4	
Special Education									
	Yes	376.31	40.36	67		357.51	57.52	6737	
	No	406.71	39.63	5205		417.37	45.98	60465	
	504	*	*	3		405.46	51.18	1057	
Ethnicity									
	American Indian	*	*	27		405.48	48.51	268	
	Asian/Pacific Islander	415.43	43.80	121		441.56	41.81	3580	
	African American	387.82	40.15	2056		385.44	48.65	25881	
	White	421.10	32.36	2812		429.34	43.32	34235	
	Hispanic	392.30	38.97	206		398.12	47.64	4290	
	Missing	366.77	37.77	53		*	*	5	
Limited English Proficient									
	Yes	*	*	3		382.97	52.63	1918	
	No	406.31	39.80	5272		412.10	50.32	65859	
	Exited			0		411.66	44.95	482	

* Statistics not reported for sample size less than 50 (N<50)

Table 4.10. Summary Statistics for Algebra Make Up Forms

		January Make-Up Form[a]				May Make-Up Forms							
		C				X				Y			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		387.72	39.39	551	0.89	386.33	60.76	3449	0.93	371.26	57.11	379	0.90
Gender													
	Male	387.11	39.64	266		379.99	65.83	1792		362.10	58.25	193	
	Female	388.27	38.89	278		393.21	53.96	1656		380.76	54.45	186	
	Missing	*	*	7		*	*	1				0	
Special Education													
	Yes	*	*	8		329.98	58.64	459		330.66	59.88	68	
	No	387.67	39.63	542		395.32	56.06	2930		380.58	52.08	305	
	504	*	*	1		378.33	65.48	60		*	*	6	
Ethnicity													
	American Indian	*	*	4		*	*	19		*	*	1	
	Asian/Pacific Islander	*	*	11		422.76	61.83	97		*	*	11	
	African American	378.42	37.51	296		359.31	55.46	1393		347.56	56.60	149	
	White	403.90	37.23	196		408.22	55.40	1724		387.38	50.84	203	
	Hispanic	*	*	22		367.90	58.37	215		*	*	15	
	Missing	*	*	22		*	*	1				0	
Limited English													
Proficient	Yes			0		349.59	53.90	90		*	*	10	
	No	387.72	39.39	551		387.32	60.71	3336		372.40	56.75	368	
	Exited			0		*	*	23		*	*	1	

[a] Form D is not summarized due to small sample (N = 3)

* Statistics not reported for sample size less than 50 (N<50)

Table 4.11. Summary Statistics for Biology Primary Forms

		January				May			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		402.71	35.73	7711	0.93	406.70	42.48	50157	0.94
Gender									
	Male	402.78	37.41	3674		402.96	46.27	24981	
	Female	402.73	34.13	4027		410.41	37.98	25173	
	Missing	*	*	10		*	*	3	
Special Education									
	Yes	367.06	26.82	85		365.16	45.25	4596	
	No	403.14	35.62	7614		411.07	39.78	44734	
	504	*	*	12		401.48	42.69	827	
Ethnicity									
	American Indian	*	*	18		396.29	45.99	191	
	Asian/Pacific Islander	416.62	35.32	149		428.17	39.34	3199	
	African American	378.65	29.10	2444		385.47	40.49	16414	
	White	415.37	32.06	4827		418.57	38.13	27245	
	Hispanic	388.13	32.23	230		393.38	41.37	3102	
	Missing	*	*	43		*	*	6	
Limited English Proficient									
	Yes	*	*	2		379.24	40.00	1394	
	No	402.72	35.74	7707		407.52	42.32	48315	
	Exited	*	*	2		404.27	38.62	448	

* Statistics not reported for sample size less than 50 (N<50)

Table 4.12. Summary Statistics for Biology Make Up Forms

		January Make-Up Forms								May Make-Up Forms							
		C				D				X				Y			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		381.81	33.54	383	0.90	374.54	26.48	226	0.85	374.23	49.60	1901	0.93	371.37	45.99	260	0.92
Gender																	
	Male	381.39	33.84	187		377.83	27.25	98		365.55	53.81	1020		362.21	48.14	141	
	Female	382.59	33.29	194		372.05	25.67	126		384.29	42.08	881		382.21	40.92	119	
	Missing	*	*	2		*	*	2				0				0	
Special Education																	
	Yes	*	*	6		*	*	2		341.11	49.88	288		*	*	41	
	No	382.15	33.72	375		374.64	26.33	224		380.12	47.07	1573		375.67	44.52	216	
	504	*	*	2		*	*	0		*	*	40		*	*	3	
Ethnicity																	
	American Indian	*	*	3				0		*	*	10		*	*	2	
	Asian/Pacific Island	*	*	6		*	*	4		*	*	44		*	*	10	
	African American	366.25	29.62	159		366.77	19.97	130		358.88	46.95	854		352.38	44.23	104	
	White	395.96	30.54	191		388.58	28.72	76		389.45	48.48	854		386.83	40.26	123	
	Hispanic	*	*	15		*	*	12		366.13	42.15	139		*	*	19	
	Missing	*	*	9		*	*	4				0		*	*	2	
Limited English																	
Proficient	Yes			0				0		*	*	45		*	*	5	
	No	381.81	33.54	383		374.54	26.48	226		374.67	49.87	1841		372.03	45.73	254	
	Exited			0				0		*	*	15		*	*	1	

* Statistics not reported for sample size less than 50 (N<50)

Table 4.13. Summary Statistics for Geometry Primary Forms

		January				May			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		411.61	31.02	6978	0.93	407.13	47.46	59073	0.94
Gender									
	Male	412.61	31.33	3163		405.94	50.80	29116	
	Female	410.98	30.68	3778		408.28	43.95	29956	
	Missing	*	*	37		*	*	1	
Special Education									
	Yes	*	*	22		363.66	50.65	5066	
	No	411.72	31.01	6945		411.27	45.10	53035	
	504	*	*	11		407.44	42.09	972	
Ethnicity									
	American Indian	*	*	22		396.52	45.23	206	
	Asian/Pacific Islander	426.30	29.57	175		438.33	41.27	3489	
	African American	387.84	28.65	1844		380.21	44.78	21059	
	White	421.41	26.13	4660		422.68	40.96	31129	
	Hispanic	401.65	27.63	193		399.58	42.97	3189	
	Missing	382.00	31.20	84		*	*	1	
Limited English Proficient									
	Yes	*	*	2		397.21	48.67	1160	
	No	411.62	31.02	6974		407.31	47.45	57342	
	Exited	*	*	2		409.18	44.22	571	

* Statistics not reported for sample size less than 50 (N<50)

Table 4.14. Summary Statistics for Geometry Make Up Forms

		January Make-Up Forms								May Make-Up Forms							
		C				D				X				Y			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		394.21	32.74	342	0.92	390.63	30.03	208	0.89	376.69	51.42	2154	0.92	374.14	49.36	288	0.92
Gender																	
	Male	392.64	33.87	159		391.99	31.19	94		373.48	55.51	1137		367.74	53.05	159	
	Female	396.07	31.45	177		390.37	28.45	111		380.28	46.19	1017		382.02	43.30	129	
	Missing	*	*	6		*	*	3				0				0	
Special Eductn																	
	Yes	*	*	5		*	*	1		343.61	49.38	266		*	*	41	
	No	394.45	32.73	337		390.58	30.09	207		381.42	50.00	1848		381.20	44.97	245	
	504			0				0		*	*	40		*	*	2	
Ethnicity																	
	American Indian	*	*	1		*	*	1		*	*	7		*	*	3	
	Asian/Pacific Island	*	*	9		*	*	5		406.90	41.29	63		*	*	7	
	African American	382.38	32.12	167		376.26	26.17	69		356.57	47.13	1019		352.41	49.35	117	
	White	406.25	26.46	143		402.79	27.36	107		397.68	47.74	921		389.30	44.19	148	
	Hispanic	*	*	14		383.82	31.52	17		370.65	49.01	144		*	*	13	
	Missing	*	*	8		*	*	9				0				0	
Limited English																	
Proficient	Yes			0				0		*	*	43		*	*	2	
	No	394.21	32.74	342		390.63	30.03	208		376.75	51.38	2091		373.90	49.49	283	
	Exited			0				0		*	*	20		*	*	3	

* Statistics not reported for sample size less than 50 (N<50)

Table 4.15. Summary Statistics for Government Primary Forms

		January				May			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		401.57	35.09	6783	0.94	412.14	42.44	48926	0.95
Gender									
	Male	399.01	36.11	3304		409.37	44.95	24497	
	Female	404.06	33.90	3470		414.91	39.56	24426	
	Missing	*	*	9		*	*	3	
Special Education									
	Yes	367.25	27.59	53		370.09	41.66	4598	
	No	401.84	35.01	6719		416.69	40.04	43483	
	504	*	*	11		406.75	40.34	845	
Ethnicity									
	American Indian	*	*	22		403.14	43.63	177	
	Asian/Pacific Islander	414.42	38.46	132		432.13	44.32	3190	
	African American	381.07	28.16	2230		392.81	38.35	13992	
	White	413.45	33.08	4035		420.60	40.57	28568	
	Hispanic	392.05	32.31	315		401.01	40.70	2996	
	Missing	*	*	49		*	*	3	
Limited English Proficient									
	Yes	*	*	7		385.24	37.33	1403	
	No	401.60	35.09	6776		413.00	42.38	47020	
	Exited			0		406.47	36.10	503	

* Statistics not reported for sample size less than 50 (N<50)

Table 4.16. Summary Statistics for Government Make Up Forms

		January Make-Up Forms								May Make-Up Forms							
		C				D				X				Y			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		377.68	31.33	439	0.91	384.81	32.79	319	0.92	378.80	46.09	1787	0.94	373.35	41.06	302	0.93
Gender																	
	Male	374.88	32.38	213		381.65	35.83	153		371.81	48.49	1013		366.72	41.83	157	
	Female	380.47	30.16	223		387.66	28.81	160		387.94	41.01	774		380.92	38.96	144	
	Missing	*	*	3		*	*	6				0		*	*	1	
Special Eductn																	
	Yes	*	*	10		*	*	4		351.21	40.63	270		342.51	35.26	61	
	No	377.83	31.50	426		385.10	32.84	315		383.70	45.19	1486		380.47	38.61	233	
	504	*	*	3				0		*	*	31		*	*	8	
Ethnicity																	
	American Indian	*	*	5		*	*	1		*	*	6				0	
	Asian/Pacific Island	*	*	7		*	*	11		393.23	52.57	64		*	*	11	
	African American	367.22	27.64	231		375.44	27.12	135		367.17	42.73	726		364.70	39.85	132	
	White	393.52	30.24	168		395.50	35.45	131		389.09	46.26	838		381.32	39.47	143	
	Hispanic	*	*	21		*	*	27		371.01	43.79	152		*	*	15	
	Missing	*	*	7		*	*	14		*	*	1		*	*	1	
Limited English																	
Proficient	Yes	*	*	2				0		361.14	45.81	65		*	*	5	
	No	377.73	31.38	437		384.81	32.79	319		379.59	46.14	1704		373.53	40.83	296	
	Exited			0				0		*	*	18		*	*	1	

* Statistics not reported for sample size less than 50 (N<50)

Decision Consistency

The accuracy of decisions based on specified cut-scores was assessed for Reliability of Classification using the computer program RelClass, ETS proprietary software. RelClass provides two statistics that describe the reliability of classifications based on test scores (Livingston & Lewis, 1995). More specifically, information from an administration of one form is used to estimate the following:

1) Decision Accuracy describes the extent to which examinees are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy answers the question: How does the actual classification of test takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known.

2) Decision Consistency describes the extent to which examinees are classified in the same way as they would be on the basis of a single form of a test other than the one for which data are available. Decision consistency answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test.

The results are provided in Table 4.17 by administration and content area. The statistics are presented for the proficient cutscore for all exams, and the advanced cutscore for Geometry. High indices for decision accuracy and consistency were observed. All decision accuracy values estimated by this method were greater than 0.90. Therefore, the agreement between classifications based on an observable variable (scores on one form of a test) and classifications based on an unobservable variable (the test takers' true scores) was very good. Decision consistency values were greater than 0.87 for the proficient classifications, and greater than 0.91 for the advanced classifications. Since decision consistency statistics describe the agreement between classifications based on two variables (scores on the form students have taken and a parallel form of the same test that is not administered to the students), these values are within the acceptable range.

Table 4.17. Decision Accuracy and Consistency by Administration and Content Area

	Decision Accuracy		Decision Consistency	
	Proficient	Advanced	Proficient	Advanced
Jan, 2005				
Algebra	0.910		0.874	
Biology	0.918		0.885	
Geometry	0.920	0.938	0.889	0.913
Government	0.919		0.887	
May, 2005				
Algebra	0.916		0.883	
Biology	0.925		0.895	
Geometry	0.918	0.937	0.888	0.915
Government	0.914		0.917	

Section 5. Field Test Analyses

Following the receipt of the final scored file from Measurement Incorporated (MI), the field test analyses were completed. The analyses of the field test data consisted of four components: classical item analyses, differential item functioning (DIF), calibration, and scaling. All of the analyses were completed using Genasys, ETS proprietary software. The analysis procedures for each component are described in detail. Samples used for the analyses included all valid records available at the time of the analyses, including students classified as English as a second language, students with IEP or 504 plans, and students receiving accommodations. Only duplicate records, records invalidated by the test administrator, and records with five or fewer item responses were excluded from the analysis sample. The field test analyses presented in this section reflect only the January 2005 administrations. The May 2005 field test data were not available when the draft report was prepared. The May 2005 field test analyses will be presented in the final version of this report.

Classical Item Analyses

Classical item analyses involve computing a set of statistics based on classical test theory for every item in each form. The statistics provide key information about the quality of the items from an empirical perspective. The statistics estimated for the MHSA field test items are described below.

Classical item difficulty ("P-Value"):

This statistic indicates the percent of examinees in the sample that answered the item correctly. Desired p-values generally fall within the range of 0.25 to 0.90. Occasionally, items that fall outside this range can be justified for inclusion in an item bank based upon the quality and educational importance of the item content or the ability to measure students with very high or low achievement, especially if the students have not yet received instruction in the content or lack motivation to complete the field test items to the best of their ability.

The item-total correlation of the correct response option (for SR items) or the CR item score with the total test score:

This statistic describes the relationship between performance on the specific item and performance on the entire form. It is sometimes referred to as a discrimination index. Values less than 0.15 were flagged for a weaker than desired relationship and deserve careful consideration by ETS staff and MSDE before including them on future forms. Items with negative correlations can indicate serious problems with the item content (e.g., multiple correct answers, unusually complex content), an incorrect key, or students have not been taught the content.

The proportion of students choosing each response option (SR items):

> This statistic indicates the percent of examinees selecting each answer option. Item options not selected by any students or selected by a very low proportion of students indicate problems with plausibility of the option. Items that do not have all answer options functioning may be discarded or revised and field tested again.

The point-biserial correlation of incorrect response option (SR items) with the total score:

> These statistics describe the relationship between selecting an incorrect response option for a specific item and performance on the entire test. Typically, the correlation between an incorrect answer and total test performance is weak or negative. Values are typically compared and contrasted with the discrimination index. When the magnitude of these point-biserial correlations for the incorrect answer is stronger, relative to the correct answer, the item will be carefully reviewed for content-related problems. Alternatively, positive point-biserial correlations on incorrect option choices may indicate that students have not had sufficient opportunity to learn the material.

Percent of students omitting an item:

> This statistic is useful for identifying problems with test features such as testing time and item/test layout. Typically, we would expect that if students have an adequate amount of testing time, 95% of students should attempt to answer each question. When a pattern of omit percentages exceeds 5% for a series of items at the end of a timed section, this may indicate that there was insufficient time for students to complete all items. Alternatively, if the omit percentage is greater than 5% for a single item, this could be an indication of an item/test layout problem. For example, students might accidentally skip an item that follows a lengthy stem.

Frequency distribution of CR score points:

> Observation of the distribution of scores is useful to identify how well the item is functioning. If no students are assigned the top score point, this may indicate that the item is not functioning with respect to the rubric, there are problems with the item content, or students have not been taught the content.

Summaries of p-values by content area for the field test items administered in January are found in Table 5.1 for SR items and Table 5.2 for CR items. Summaries of item-total correlations by content area for the field test items administered in January are found in Table 5.3 for the SR items and Table 5.4 for the CR items. In addition, a series of flags was created to identify items with extreme values. Flagged items were subject to additional scrutiny prior to the inclusion of the items in the final calibrations. The following flagging criteria were applied to all items tested in the 2005 assessments:

- *Difficulty Flag*: P-values less than 0.25 or greater than 0.90.
- *Discrimination Flag*: Point-biserial correlation less than 0.15 for the correct answer.
- *Distractor Flag*: Point-biserial correlation positive for incorrect option.

- *Omit Flag*: Percentage omitted is greater than 0.05.
- *Collapsed Score Levels*: Items with no students obtaining the score point.

Following the classical item analyses, items with poor item statistics and items that were not scored were removed from further analyses. Refer to Table 5.5. These items have been identified for revision and possible re-field testing.

Differential Item Functioning (DIF)

Following the classical item analyses, DIF analyses were completed. One goal of test development is to assemble a set of items that provides an estimate of a student's ability that is as fair and accurate as possible for all groups within the population. DIF statistics are used to identify items whereby identifiable groups of students with the same underlying level of ability have different probabilities of answering correctly (e.g. females, African Americans, Hispanics). If the item is more difficult for an identifiable subgroup, the item may be measuring something different than the intended construct. However, it is important to recognize that DIF flagged items might be related to actual differences in relevant knowledge or skill (item impact) or statistical Type I error. As a result, DIF statistics are used to identify potential sources of item bias. Subsequent review by content experts and bias/sensitivity committees is required to determine the source and meaning of evident differences.

ETS used two DIF detection methods: the Mantel-Haenszel and standardization approaches. As part of the Mantel-Haenszel procedure, the statistic described by Holland & Thayer (1988), known as MH D-DIF, was used[2]. This statistic is expressed as the differences between the focal and reference group after conditioning on total test score. The statistic is reported on the ETS delta scale, which is a normalized transformation of item difficulty (proportion correct) with a mean of 12 and a standard deviation of 4.

[2] The formula for the estimate of constant odds ratio is:

$$\alpha_{MH} = \frac{\left(\sum_m \frac{R_{rm} W_{fm}}{N_m}\right)}{\left(\sum_m \frac{R_{fm} W_{rm}}{N_m}\right)},$$

where,

R_{rm} = number in reference group at ability level m answering the item right,
W_{fm} = number in focal group at ability level m, answering the item wrong,
R_{fm} = number in focal group at ability level m answering the item right,
W_{rm} = number in reference group at ability level m, answering the item wrong,
N_m = total group at ability level m.

This can then be used in the following formula (Holland & Thayer, 1985):
$MH\ D\text{-}DIF = -2.35\, ln[\alpha_{MH}]$.

Negative MH D-DIF statistics favor the reference group and positive values favor the focal group. The classification logic used for flagging items is based on a combination of absolute differences and significance testing. Items that are not statistically significantly different based on the MH D-DIF ($p>0.05$) are considered to have similar performance between the two studied groups; these items are considered to be functioning appropriately. For items where the statistical test indicates significant differences ($p < 0.05$), the effect size is used to determine the direction and severity of the DIF. For the ELA CR item, the Mantel-Haenszel procedure was executed where item categories are treated as integer scores and a chi-square test was carried out with one degree of freedom. The male and white groups were considered the reference groups for gender and ethnicity, respectively; the female and other ethnic groups were considered the focal groups.

Based on these DIF statistics, items are classified into one of three categories and assigned values of A, B or C. Category A items contain negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large DIF. Negative values imply that conditional on the matching variable, the focal group has a lower mean item score than the reference group. In contrast a positive value implies that, conditional on the matching variable, the reference group has lower mean item score than the focal group. For constructed-response items the MH D-DIF is not calculated, but analogous flagged rules based on the chi-square statistic have been developed resulting in classification into A, B, or C DIF categories.

There were 8 items flagged for C-level DIF against one of the identified focal groups (i.e., female, African American, American Indian, Asian, Hispanic) for two of the four content areas. For the government test, 2 items were flagged to have negative DIF against female (favor female), 3 items against African American (favor White) and 1 item against Hispanic (favor White). For the Biology test, 1 item was flagged to have DIF against African American (favor White) and another item was flagged to have negative DIF against African American (favor African American). These items are flagged in the bank, and will be reviewed for future use.

IRT Calibration and Scaling

The purpose of item calibration and scaling is to create a common scale for expressing the difficulty estimates of all the items across all versions of a test. The resulting scale has a mean score of 0 and a standard deviation of 1. It should be noted that this scale is often referred to as the “theta” metric and is not used for reporting purposes because the values typically range from –3 to +3. Therefore, the scale is usually transformed to a reporting scale (also know as a scale score), which can be more meaningfully interpreted by students, teachers, and other stakeholders.

The IRT models used to calibrate the MHSA test items were the 3-parameter logistic (3PL) model for SR items and the generalized partial credit model (GPCM) for CR items. Item response theory expresses the probability that a student will achieve a certain score on an item (such as correct or incorrect) as a function of the item's statistical properties and the ability level (or proficiency level) of the student.

The fundamental equation of the 3PL model relates the probability that a person with ability θ will respond correctly to item j:

$$P(U_j = 1 | \theta) = P_j(\theta) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j(\theta - b_j)}}$$

where:

U_j is the response to item j, 1 if correct and 0 if incorrect;

a_j is the slope parameter of item j, characterizing its discriminating power;

b_j is the threshold parameter of item j, characterizing its difficulty; and

c_j is the lower asymptote parameter of item j, reflecting the chance that students with very low proficiency will select the correct answer, sometimes called the "pseudo-guessing" level

The parameters estimated for the 3-PL model were discrimination (a), difficulty (b), and the pseudo-guessing level (c).

The GPCM is given by

$$P_{jk}(\theta) = \frac{\exp\left[\sum_{v=1}^{k} Z_{jv}(\theta)\right]}{\sum_{c=1}^{mjj} \exp\left[\sum_{v=1}^{c} Z_{jv}(\theta)\right]}$$

where

$$Z_{jk}(\theta) = 1.7a_j(\theta - b_{jk}) = 1.7a_j(\theta - b_j + d_k)$$

$$\sum_{k=2}^{mj} d_k = 0$$

P_{jk} is the probability of responding in the kth category from m_j+1 categories for item j,

θ is the ability level,

a_j is the item parameter characterizing the discriminating power for item j,

b_{jk} is an item-category parameter for item j,

b_j is the item parameter characterizing the difficulty for item j,

d_k is the category parameter characterizing the relative difficulty of category k .

A proprietary version of the PARSCALE computer program (Muraki & Bock, 1995) was used for all item calibration work. This program estimates parameters for a generalized partial-credit model using procedures described by Muraki (1992). The resulting calibrations were then scaled to the bank estimates using the Stocking and Lord's (1983) test characteristic curve method using the operational items as the "anchor" set.

The calibration and equating process is outlined in the steps below:

1. For each test, calibrate all items using a sparse matrix design that places all items on a common scale. Essentially, this means that the data was analyzed using the following format. In the diagram below X's represent items, spaces indicating missing data. For example, items included on version 2 but not on version 1, 3, 4 or 5 were treated as "not reached" for the purposes of the analyses and were denoted as "missing" in the diagram below.

Common	Unique 1	Unique 2	Unique 3	Unique 4	Unique 5
XXXXXX	XXXXXXXXX				
XXXXXX		XXXXXXXXXX			
XXXXXX			XXXXXXXXX		
XXXXXX				XXXXXXXXX	
XXXXXX					XXXXXXXXX

2. Once the items have been calibrated, results are reviewed to determine if any items failed to calibrate.

3. After the final calibration parameters were obtained, the items were then linked to the bank scale using the test characteristic curve method. Specifically, the operational items were used to place the field test items onto the operational reporting scale.

Once the items were calibrated and placed onto the operational scale, the items were loaded into the item bank. Items were listed as unavailable based on the following criteria:

- Item-total correlation less than 0.1
- Item P-value less than 0.1
- Field test CR items that have fewer than 20 students achieving the highest score level
- Item not scored

Statistical Summary Tables

Table 5.1. Distribution of P-Values for the January Field Test SR Items

	Percentage and Number of Items							
P-Value	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.30	12.5	2	11.1	3	13.3	2	4.2	1
0.30 to 0.40	43.8	7	14.8	4	6.7	1	12.5	3
0.41 to 0.50	12.5	2	25.9	7	26.7	4	33.3	8
0.51 to 0.60	0	0	18.5	5	20.0	3	12.5	3
0.61 to 0.70	18.8	3	18.5	5	33.3	5	16.7	4
0.71 to 0.80	12.5	2	7.4	2	0	0	20.8	5
≥ 0.81	0	0	3.7	1	0	0	0	0
Descriptive Stats								
Number of Items	16		27		15		24	
Mean	0.44		0.51		0.50		0.54	
SD	0.18		0.16		0.17		0.17	
Min	0.11		0.22		0.12		0.26	
Max	0.75		0.91		0.70		0.80	

Table 5.2. Distribution of P-Values for the January Field Test CR Items

	Percentage of Items (N)							
P-Value	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.30	50.0	2	100	1	50.0	2	75.0	3
0.30 to 0.40	0	0	0	0	25.0	1	25.0	1
0.41 to 0.50	50.0	2	0	0	25.0	1	0	0
0.51 to 0.60	0	0	0	0	0	0	0	0
0.61 to 0.70	0	0	0	0	0	0	0	0
0.71 to 0.80	0	0	0	0	0	0	0	0
≥ 0.81	0	0	0	0	0	0	0	0
Descriptive Stats								
Number of Items	4		1		4		4	
Mean	0.30		0.17		0.29		0.20	
SD	0.16		0.00		0.10		0.11	
Min	0.11		0.17		0.19		0.10	
Max	0.44		0.17		0.41		0.33	

Table 5.3 Distribution of Item-Total Correlations for the January Field Test SR Items

	Percentage and Number of Items							
Correlation	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.15	0	0	7.4	2	0	0	12.5	3
0.15 to 0.24	18.8	3	14.8	4	6.7	1	16.7	4
0.25 to 0.34	37.5	6	22.2	6	6.7	1	8.3	2
0.35 to 0.44	18.8	3	37.0	10	46.7	7	50.0	12
0.45 to 0.54	25.0	4	18.5	5	20.0	3	12.5	3
≥ 0.55	0	0	0	0	20.0	3	0	0
Descriptive Stats								
Number of Items	16		27		15		24	
Mean	0.35		0.34		0.45		0.33	
SD	0.11		0.11		0.12		0.11	
Min	0.20		0.13		0.24		0.12	
Max	0.54		0.52		0.73		0.49	

Table 5.4 Distribution of Item-Total Correlations for January Field Test CR Items

	Percentage and Number of Items							
Correlation	Algebra		Biology		Geometry		Government	
	%	N	%	N	%	N	%	N
< 0.15	0	0	0	0	0	0	0	0
0.15 to 0.24	0	0	0	0	0	0	0	0
0.25 to 0.34	0	0	0	0	0	0	0	0
0.35 to 0.44	0	0	0	0	0	0	0	0
0.45 to 0.54	0	0	0	0	0	0	0	0
≥ 0.55	100	4	100	1	100	4	100	4
Descriptive Stats								
Number of Items	4		1		4		4	
Mean	0.63		0.70		0.76		0.63	
SD	0.02		0.00		0.05		0.04	
Min	0.62		0.70		0.71		0.57	
Max	0.66		0.70		0.81		0.66	

Table 5.5 January Field Test Items Excluded from Analyses

	Algebra		Biology		Geometry		Government	
	SR	CR	SR*	CR	SR	CR	SR	CR
Not Scored	0	0	0	1	0	0	0	0
Low/Neg Pt-Biserial/Flawed	0	0	1	0	1	0	0	0
No Response for Highest Score Level	0	0	0	2	0	0	0	2

* One additional item was excluded. Item was operational, and changed to FT due to key change.

The item was calibrated, but with too few cases. Not banked for future use.

Section 6. English Test

There are five parts to this section. The first part describes how the operational (scoring) items were selected for the new English test. The second part describes how the items were scaled, and the third part describes the factor analyses of the English forms. Summary statistics of student achievement and measures of classification consistency comprise the fourth and fifth parts.

Operational Item Selection

This section summarizes the procedures used for selecting ETS recommended English operational items. It reflects changes MSDE made to Form G, adding one extra SR item for subscore 2 in addition to the ETS recommendation. ETS recommended Form K to be the standard setting form and Form E or H be the secondary form depending on whether easier or harder items are needed for standard setting. Summary statistics (i.e., p-values, item-total correlations) for each of the recommended forms are presented in this document. Data files associated with item selection are posted on the MSDE DocuShare site. Three kinds of separate data files are posted.

1. Augmented form planners with item statistics and operational item designation are in Excel data files named: FP_English_0505_V5_Form***X***.D042005.xls where ***X*** stands for form code. Please note the operational item designation in the column heading "anchor_status." The value "O" indicates an operational item.
2. Item analyses by student ability group summary. This file, called "IA by category offload.xls," contains the item difficulties and item-total correlations for the high (H), medium (M) and low (L) ability groups.
3. Distractor analyses by student ability group. This is a flat text file called "IA by category Offload Distracter analyses.txt." This file contains the distractor analyses for all three ability groups. Please note that this is an extract file of the ETS item analyses output so there was no Maryland ID associated with each item. Instead, a form code and sequence number are added to the first column of the output for the reader to understand the statistics.

The processes that were used to select the operational items for the English forms are as follows:

1. Research conducted Item Analyses and DIF analyses and flagged items unavailable as operational items.
   Flagging criteria:
   P > 0.9 or P < 0.2
   Item-total correlation < 0.2
   Distractor item-total correlation > 0
   Omit rate (conditional code A or B) > 15%[3]
   Item with C-DIF

[3] Omit rates were considered not as crucial as the test blueprints according to correspondence with MSDE. After matching the test blueprint, CR items with omit rates as high as 23.32% were used.

2. Research conducted preliminary IRT calibration for all items except for items that were flagged for poor quality. Items that had poor fit were also excluded from item selection.
3. Research provided TD with augmented form planners with IA, DIF, and item fit flags.
4. Form K was thought to be the best form for standard setting. TD first selected operational items for the standard setting form – Form K.
5. TD used the test blueprint and reporting categories to select the operational items. One BCR item in form G was mislabeled as an ECR item so this item was not scored. TD was able to replace this BCR item with 2 SR items.
6. Research reviewed average item difficulties by form and suggested changes to item selection when necessary.

Table 6.1 shows the frequency distribution of item difficulty (P value) and item-total correlation (R_ITT) by SR and CR items for the proposed target form – Form K. Tables 6.2 to 6.10 show the same frequency distributions for the other forms. Table 6.11 presents the mean and standard deviation of the item difficulty and item-total correlation for each proposed operational form. Table 6.12 presents the number of items per subscore by form. Table 6.13 presents the number of items excluded from item selection by reason.

Table 6.1. Form K (Target)

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	0	0	0
0.3-0.4	1	11	0	0
0.4-0.5	5	21	2	0
0.5-0.6	4	13	1	1
0.6-0.7	15	1	1	1
0.7-0.8	17	0	0	2
0.8-0.9	4	0	0	0
0.9-1	0	0	0	0

Table 6.2. Form E

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	2	0	0
0.3-0.4	4	7	0	0
0.4-0.5	7	19	3	0
0.5-0.6	5	15	1	1
0.6-0.7	17	3	0	3
0.7-0.8	12	0	0	0
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.3. Form F

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	1	3	0	0
0.3-0.4	0	9	2	0
0.4-0.5	3	19	0	0
0.5-0.6	11	15	1	1
0.6-0.7	18	0	1	3
0.7-0.8	12	0	0	0
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.4. Form G

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	2	0	0
0.3-0.4	2	12	0	0
0.4-0.5	6	25	2	0
0.5-0.6	9	9	1	1
0.6-0.7	20	1	0	1
0.7-0.8	10	0	0	1
0.8-0.9	2	0	0	0
0.9-1	0	0	0	0

Table 6.5. Form H

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	4	0	0
0.3-0.4	2	10	0	0
0.4-0.5	7	22	2	0
0.5-0.6	5	10	1	1
0.6-0.7	12	0	1	3
0.7-0.8	14	0	0	0
0.8-0.9	6	0	0	0
0.9-1	0	0	0	0

Table 6.6. Form J

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	0	0
0.3-0.4	2	12	1	0
0.4-0.5	6	18	1	0
0.5-0.6	6	15	2	0
0.6-0.7	13	0	0	3
0.7-0.8	15	0	0	1
0.8-0.9	4	0	0	0
0.9-1	0	0	0	0

Table 6.7. Form L

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	0	0
0.3-0.4	1	9	0	0
0.4-0.5	5	24	3	0
0.5-0.6	12	10	1	1
0.6-0.7	15	2	0	1
0.7-0.8	12	0	0	2
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.8. Form M

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	0	0
0.3-0.4	0	6	0	0
0.4-0.5	2	17	2	0
0.5-0.6	10	22	1	1
0.6-0.7	13	0	1	2
0.7-0.8	19	0	0	1
0.8-0.9	2	0	0	0
0.9-1	0	0	0	0

Table 6.9. Form N

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	0	0	0
0.3-0.4	0	10	1	0
0.4-0.5	5	26	1	0
0.5-0.6	11	10	1	1
0.6-0.7	16	0	1	3
0.7-0.8	12	0	0	0
0.8-0.9	2	0	0	0
0.9-1	0	0	0	0

Table 6.10. Form P

	SR items		CR items	
P value/ R_ITT value interval	P value	R_ITT	P value	R_ITT
0.1-0.3	0	1	1	0
0.3-0.4	1	6	1	0
0.4-0.5	4	21	0	0
0.5-0.6	8	18	2	1
0.6-0.7	18	0	0	1
0.7-0.8	14	0	0	2
0.8-0.9	1	0	0	0
0.9-1	0	0	0	0

Table 6.11. Classical Item Statistics Summary by Form

Form code	K (Target)	E	F	G	H	J	L	M	N	P
P value										
mean	0.67	0.61	0.63	0.62	0.65	0.64	0.62	0.66	0.63	0.64
SD	0.12	0.14	0.12	0.11	0.13	0.14	0.11	0.11	0.11	0.12
R_ITT										
mean	0.48	0.48	0.47	0.45	0.45	0.47	0.48	0.49	0.46	0.48
SD	0.09	0.10	0.10	0.10	0.10	0.10	0.09	0.09	0.08	0.09

Table 6.12. Number of Items per Subscore Category by Form

Form code	K (Target)	E	F	G	H	J	L	M	N	P
Subscore 1	14	14	14	14	14	14	14	14	14	14
Subscore 2	12	12	12	14	12	12	12	12	12	12
Subscore 3	10	10	10	10	10	10	10	10	10	10
Subscore 4	14	14	14	14	14	14	14	14	14	14

Table 6.13. Number of Items Excluded from Selection by Reason

Analysis	Number of Items Flagged
Poor Content	4
High Omit Rate (SR)	8
High Omit Rate (CR)	4
DIF	16
Poor Fit	3

Calibration and Scaling

Items identified as operational items by both ETS and MSDE were calibrated using the Three Parameter Logistic (3PL) model for SR items and Generalized Partial Credit Model (GPCM) for CR items. There were 11 linking items shared by all 10 forms and additional linking items shared by adjacent forms. A concurrent calibration allowed us to put all item parameters on the same scale. The concurrent calibration converged successfully and item parameter estimates were obtained. Item fit statistics were examined and no item displayed poor fit. The maximum likelihood ability estimates (MLE) were obtained for all students in the calibration. For students with all correct or all incorrect responses, ability estimates were set to 4 and -4, respectively, on theta-scale. The mean and standard deviation of ability estimates were calculated and a set of transformation constants were derived such that the mean scale score was approximately[4] 400 and the standard deviation was 40. This set of transformation constants was applied to the item parameter estimates of the operational items in order to place the operational item parameters on the reporting scale.

A second calibration was conducted to include all items (both operational and field test items) accepted from the MSDE review. Two items on Form P were considered to have poor fit. MSDE approved the removal of the two misfit items from calibration so a third calibration was conducted removing the two items. In a Stocking-Lord linking procedure, the operational items were used as linking items to bring the field test items on to the reporting scale.

Test Characteristic Curves and the Conditional Standard Error of Measurement (CSEM) plots were used to evaluate the extent to which the test forms were parallel. The ten forms appeared to be close to parallel forms. For example, the raw scores associated with a scale score of 410 for target Form K is 41.2 and for the most difficult form, Form N, it is 37.7. This translated to about 6% difference between the easiest and hardest forms. The CSEMs were minimized around scale scores of 350 to 440.

[4] Because of the boundary constraints of the MLE theta estimates (4 and -4), the actual scale score mean and standard deviation are not exactly 400 and 40.

Figure 6.1. Test Characteristic Curves for English Forms

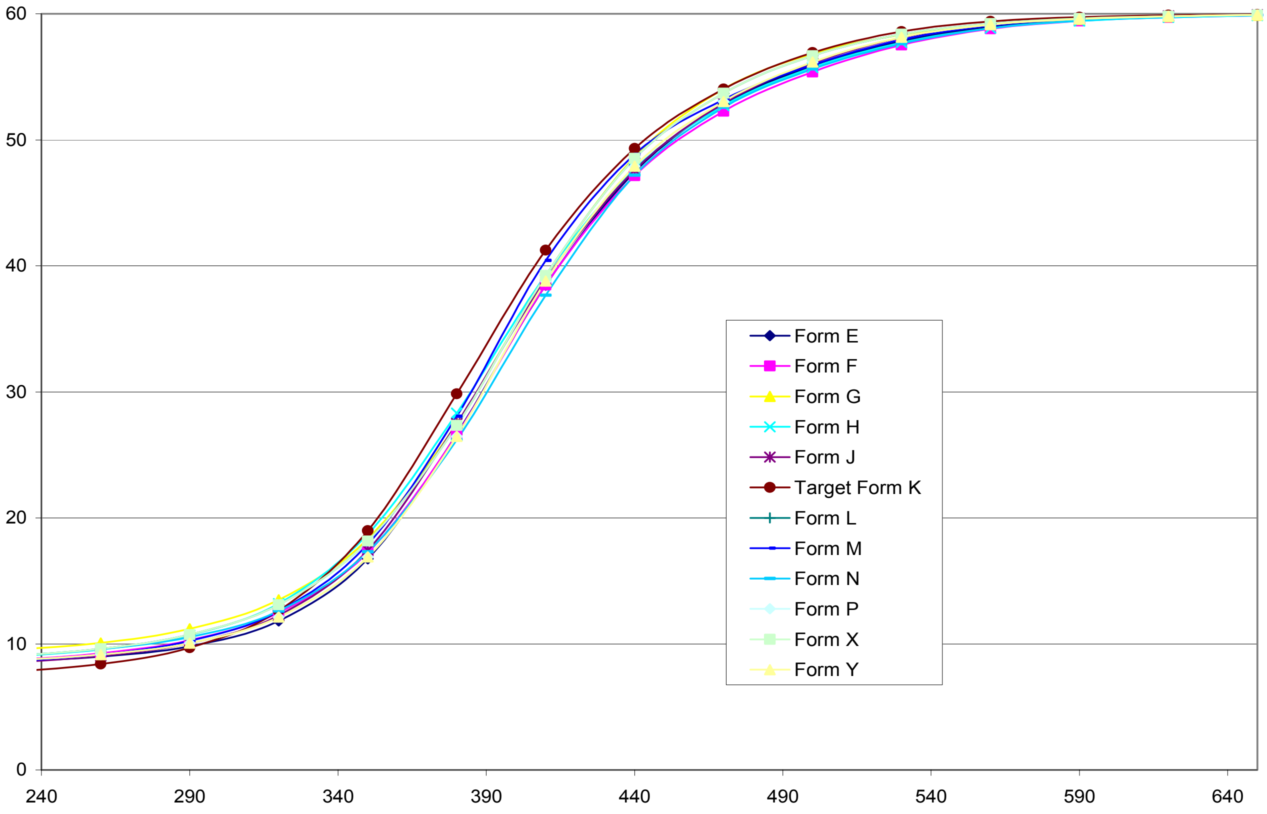
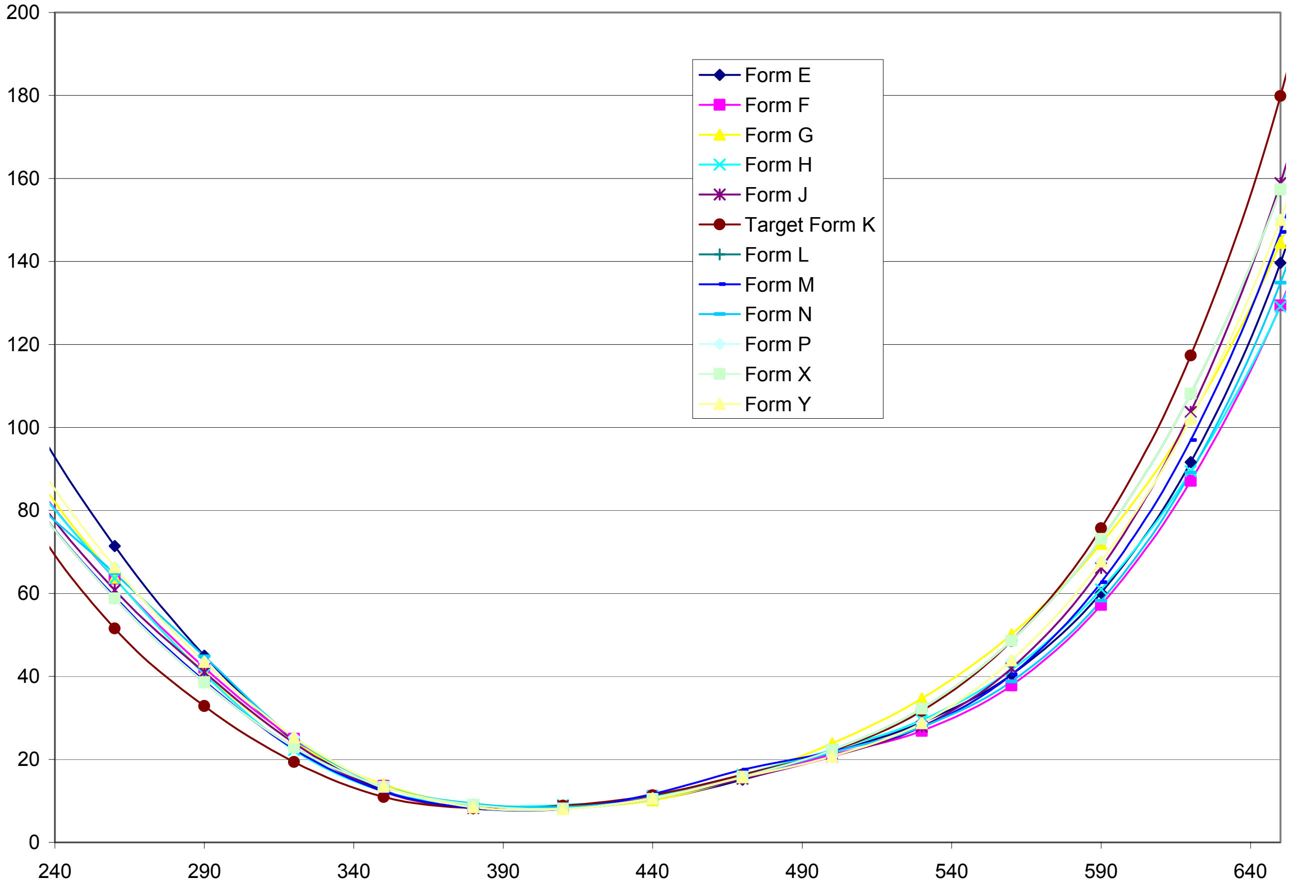


Figure 6.2. Conditional Standard Error of Measurement for English Forms



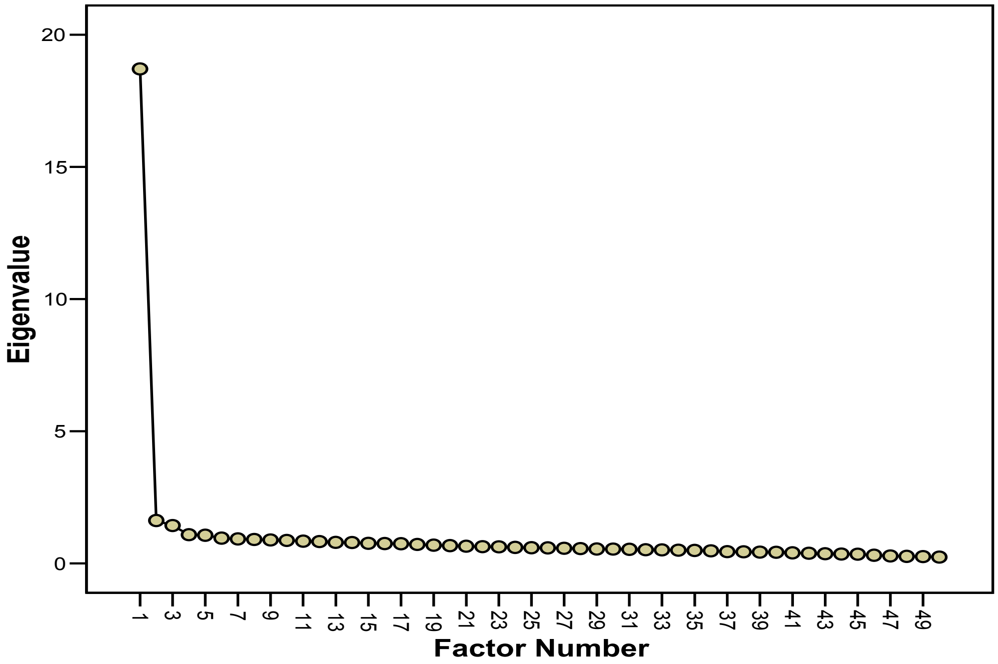
Factor Analysis Results

Factor analysis techniques were employed to investigate the dimensionality of the English MHSA primary forms. All students writing a particular form were used for the analyses. Given the ordinal nature of the item scores, matrices consisting of tetrachoric and polychoric correlations were produced for each form using PRELIS (Joreskog & Sorbom, 1993) and then analyzed using SPSS. The scree plots presented and discussed with respect to the eigenvalues and percentage of variation accounted for.

English Form E

The results of the factor analysis for Form E show an initial eigenvalue of 18.71 for the first factor, accounting for 37.42% of the variance. There were four other eigenvalues greater than one, ranging from 1.62, accounting for 3.25 % of the variance, to 1.07, accounting for 2.14 % of the variance. The scree plot for Form E illustrates one dominant factor.

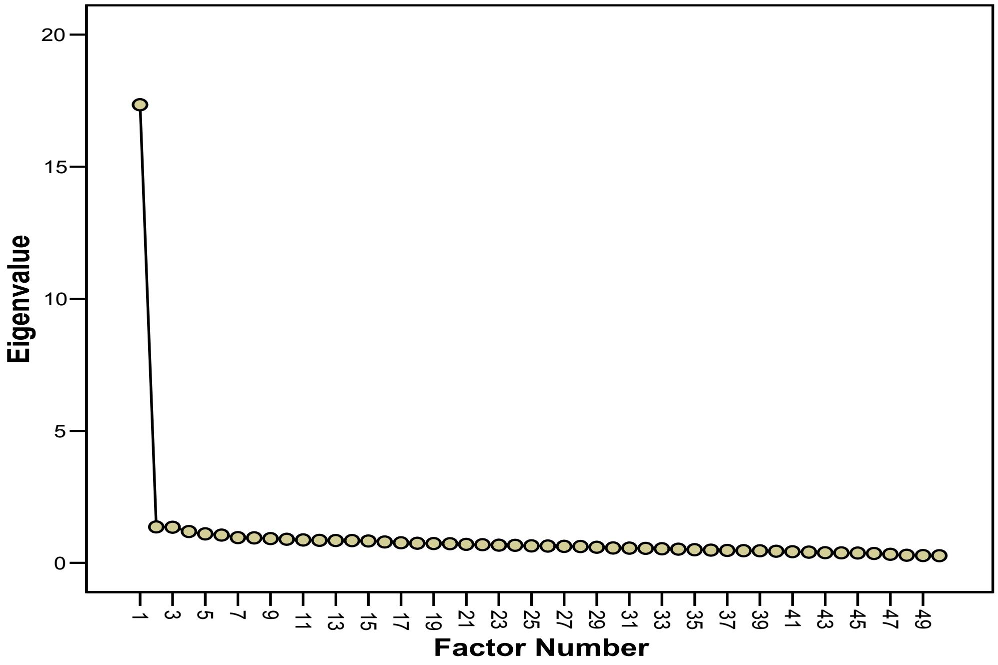
Figure 6.3 Form E Scree Plot



English Form F

The results of the factor analysis for Form F show an initial eigenvalue of 17.35, which accounts for 34.69% of the variance. There were six eigenvalues greater than one, although the remaining five eigenvalues were only slightly more than one, and accounted for less than 3% of the variance each. The scree plot for this factor analysis is provided below, indicating the presence of one dominant factor.

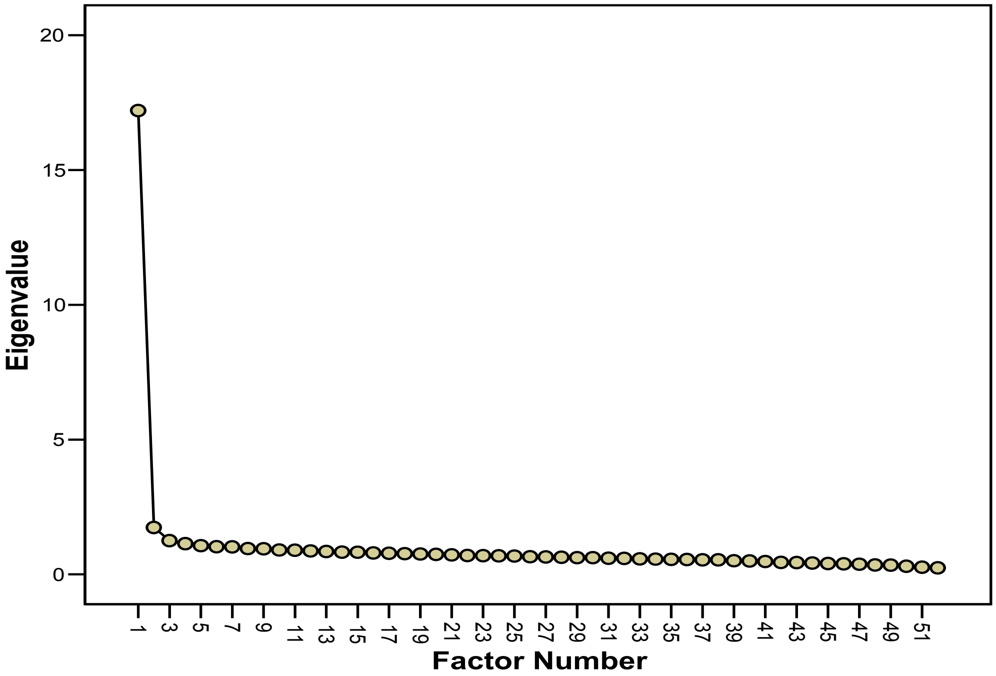
Figure 6.4. Form F Scree Plot



English Form G

The factor analysis results for Form G indicate an initial eigenvalue of 17.21 for the first factor, accounting for 33.09% of the variance. There were six other eigenvalues greater than one, ranging from 1.74 (3.34% of variance) to 1.02 (1.96% of variance). The scree plot for Form G indicates the presence of one dominant factor.

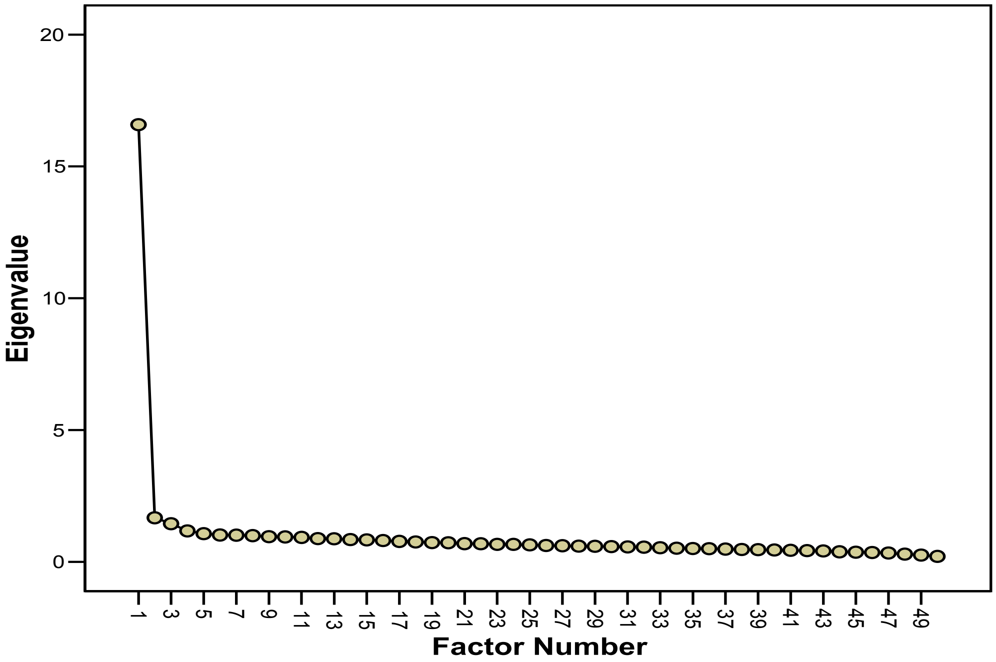
Figure 6.5. Form G Scree Plot



English Form H

The factor analysis results for Form H indicate an eigenvalue of 16.58 for the first factor, accounting for 33.17% of the variance. The remaining eigenvalues were less than two and accounted for less than 3.5% of the variance. The scree plot for Form H indicates the presence of one dominant factor.

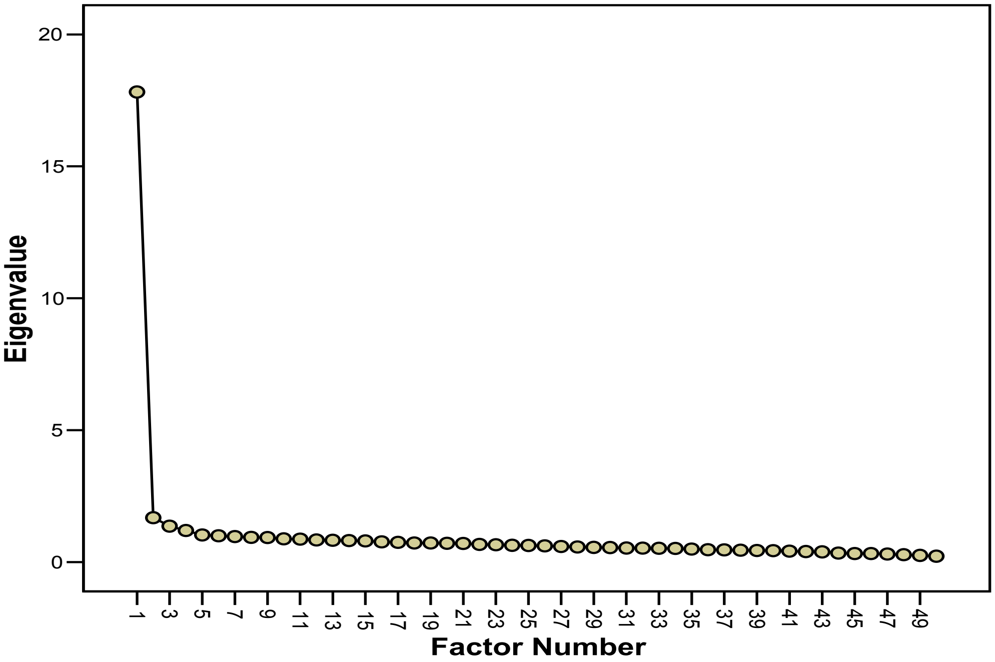
Figure 6.6. Form H Scree Plot



English Form J

The factor analysis results for Form J reveal an initial eigenvalue of 17.82 for the first factor, accounting for 35.64% of the variance. The remaining 4 eigenvalues with values greater than 1 ranged from 1.69, accounting for 3.37% of the variance, to 1.03, accounting for 2.06% of the variance. The scree plot for Form J illustrates one dominant factor.

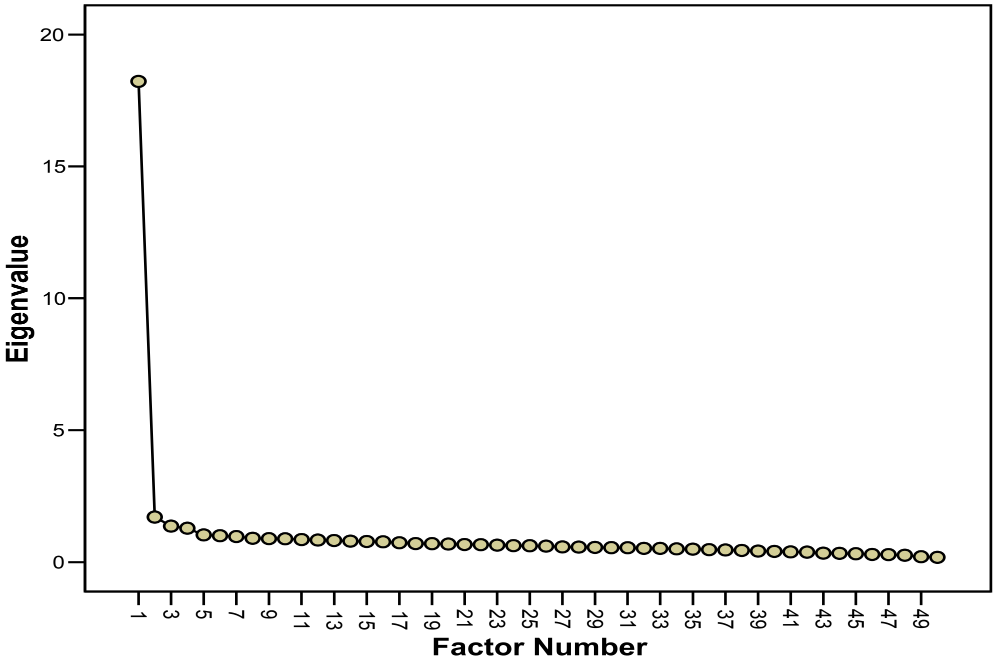
Figure 6.7. Form J Scree Plot



English Form K

The results of the factor analysis for Form K indicated an initial eigenvalue of 18.22 for the first factor, accounting for 36.45% of the variance. There were six eigenvalues greater than or equal to 1. The remaining 5 eigenvalues had values ranging from 1.71, accounting for 3.43% of the variance, to 1.00, accounting for 2.01% of the variance. The scree plot for Form K illustrates the dominance of one factor.

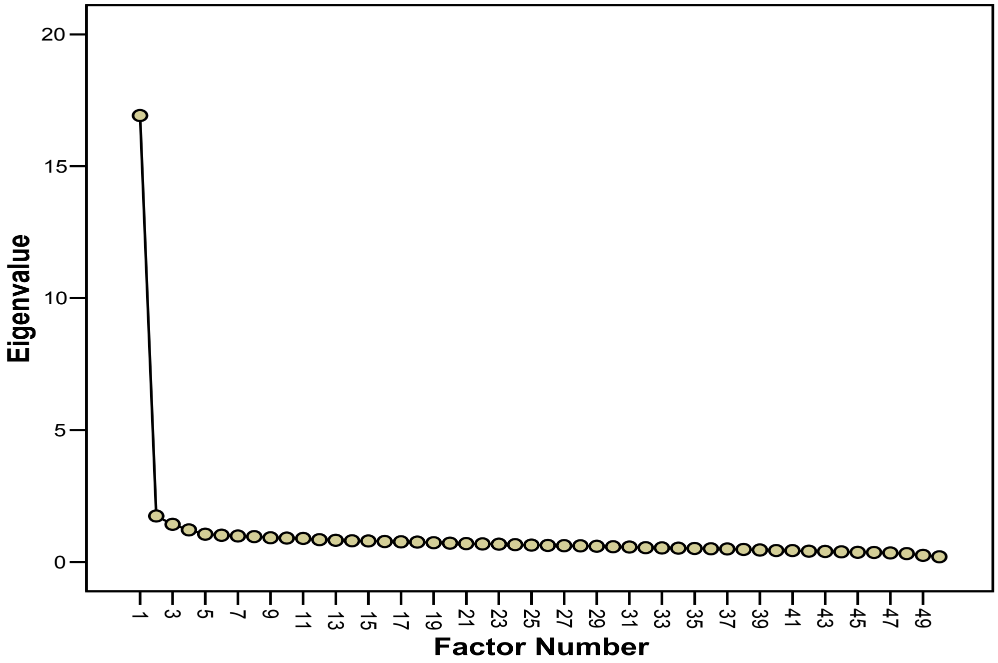
Figure 6.8. Form K Scree Plot



English Form L

The factor analysis results for Form L reveal an initial eigenvalue of 16.92, which accounts for 33.84% of the variance. There were six eigenvalues greater than one, although the remaining eigenvalues were less than 2, with variances ranging from 3.5 to 2%. The scree plot for this factor analysis is provided below, indicating the presence of one dominant factor.

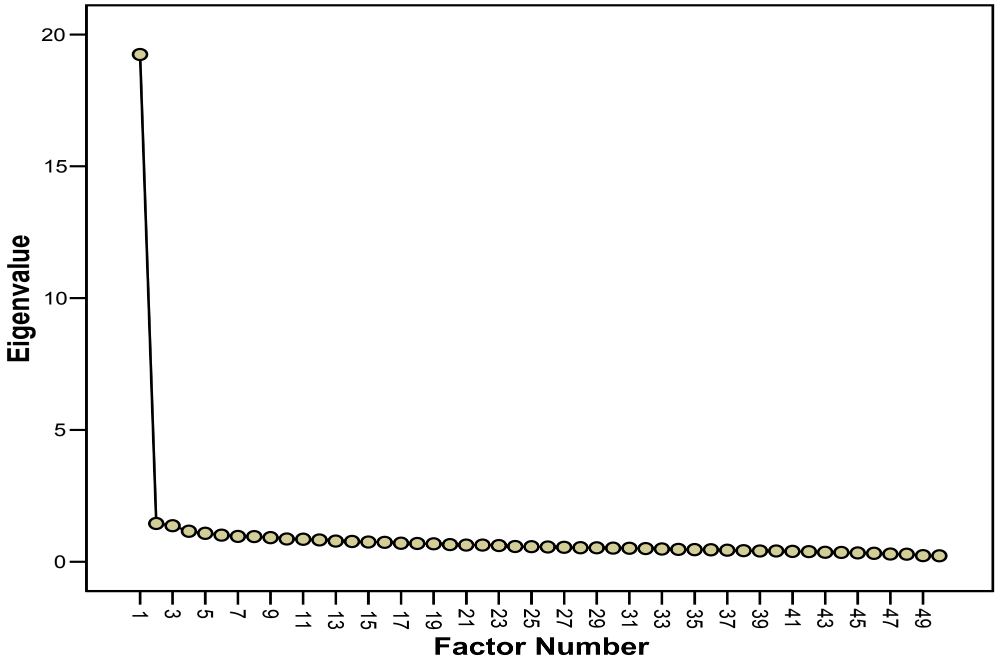
Figure 6.9. Form L Scree Plot



English Form M

The factor analysis results for Form M indicate an initial eigenvalue of 19.24, which accounts for 38.49% of the variance. Of the remaining 5 eigenvalues greater than 1, all were less than 1.5 and accounted for less than 3% of the variance. The scree plot for this factor analysis is provided below, indicating that one dominant factor is present.

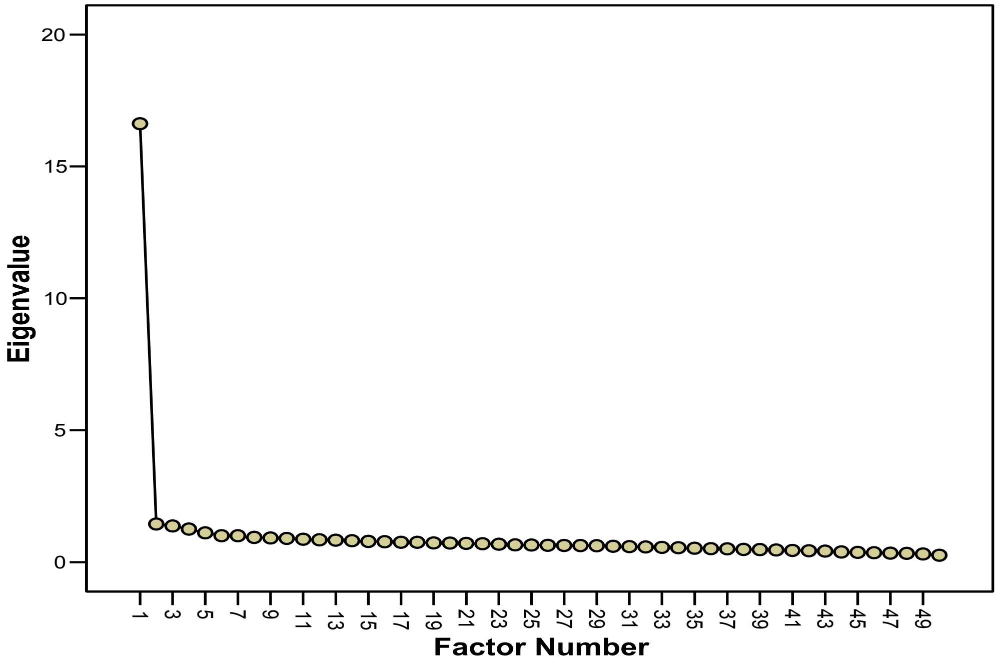
Figure 6.10. Form M Scree Plot



English Form N

The results of factor analysis for Form N shows an initial eigenvalue of 16.62, which accounts for 33.25% of the variance. There were 7 eigenvalues with values greater than 1. Of the remaining 6 eigenvalues, all were less than 1.5 and accounted for between 2 and 3% of the variance. The scree plot, provided below, demonstrates the presence of one dominant factor.

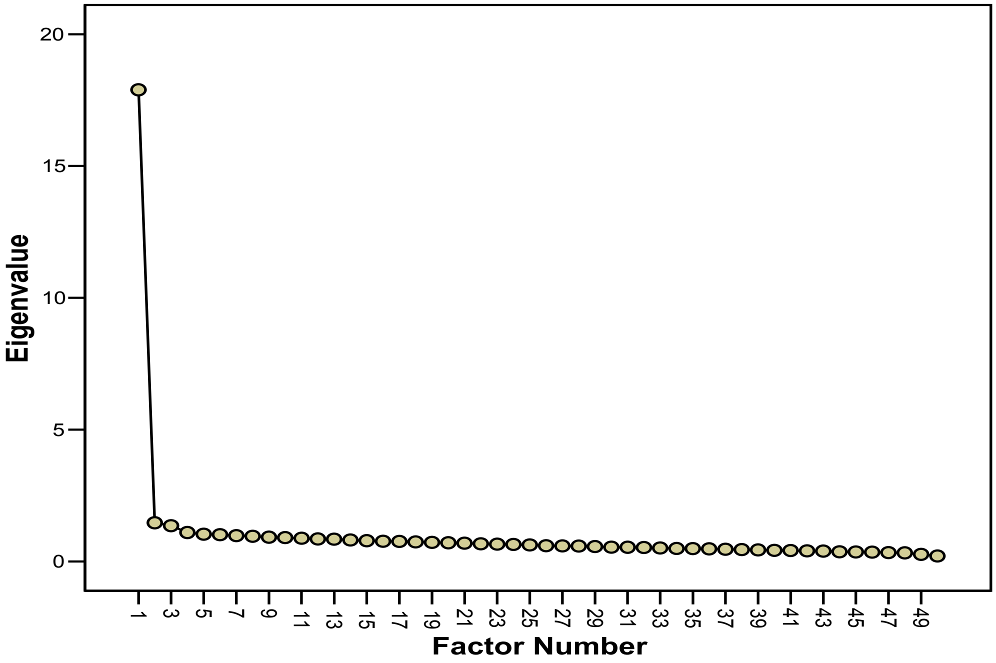
Figure 6.11. Form N Scree Plot



English Form P

The factor analysis results for Form P show an initial eigenvalue of 17.89, which accounts for 35.78% of the variance. Of the six eigenvalues that were greater than 1, the remaining five were less than 1.5 and accounted for between 2 and 3% of the variance. The scree plot for this factor analysis, provided below, illustrates one dominant factor.

Figure 6.12. Form P Scree Plot



Conclusion

The factor analyses results of the 10 primary forms indicate that one dominant factor underlies the MHSA English exams. In all cases, the first factor accounted for one-third or more of the total variance. The remaining factors accounted for considerably smaller percentage of the variance.

Summary Statistics of Student Achievement

This section summarizes the test-level statistics obtained for the English 2005 administration of the MHSA. The test-level analyses include demographic distributions, scale score information, and reliability analyses. The demographic characteristics of the students are presented in Table 6.14, whereas the scale score statistics and reliability analyses are presented in Table 6.15 for the primary forms and Table 6.16 for the make-up forms.

Table 6.14 Demographic Information for the English Exam

		May Primary Forms		May Make-Up Forms	
		N	%	N	%
Overall		54643		3250	
Gender					
	Male	27000	49.4	1771	54.5
	Female	27642	50.6	1478	45.5
	Missing	1	0.0	1	0.0
Special Education					
	Yes	5251	9.6	425	13.1
	No	48492	88.7	2765	85.1
	504	900	1.6	60	1.8
Ethnicity					
	American Indian	191	0.3	10	0.3
	Asian/Pacific Islander	3118	5.7	106	3.3
	African American	20546	37.6	1526	47.0
	White	27659	50.6	1396	43.0
	Hispanic	3128	5.7	211	6.5
	Missing	1	0.0	1	0.0
Limited English Proficient					
	Yes	920	1.7	61	1.9
	No	53256	97.5	3146	96.8
	Exited	467	0.9	43	1.3

Table 6.15. Summary Statistics for English Primary Forms

		May			
		Mean	SD	N	Alpha[a]
Overall		401.07	40.38	54643	0.93
Gender					
	Male	393.16	42.60	27000	
	Female	408.79	36.47	27642	
	Missing	*	*	1	
Special Education					
	Yes	359.42	40.47	5251	
	No	405.68	37.72	48492	
	504	395.49	38.61	900	
Ethnicity					
	American Indian	393.25	38.51	191	
	Asian/Pacific Islander	419.15	38.86	3118	
	African American	384.24	36.70	20546	
	White	412.80	38.58	27659	
	Hispanic	390.32	36.89	3128	
	Missing	*	*	1	
Limited English Proficient					
	Yes	369.25	31.02	920	
	No	401.73	40.37	53256	
	Exited	388.19	29.28	467	

* Statistics not reported for sample size less than 50 (N<50)

[a] alpha values ranged from 0.9239 to 0.9392 (M = 0.9300) across the 10 primary forms

Table 6.16. Summary Statistics for English Make-Up Forms

		May Make-Up Forms							
		X				Y			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		368.19	47.74	2782	0.93	366.57	42.10	468	0.92
Gender									
	Male	357.30	49.33	1501		355.85	45.11	270	
	Female	380.98	42.40	1280		381.18	32.41	198	
	Missing	*	*	1				0	
Special Education									
	Yes	335.09	46.32	352		338.47	42.39	73	
	No	372.88	46.03	2379		371.80	39.99	386	
	504	377.98	45.37	51		*	*	9	
Ethnicity									
	American Indian	*	*	9		*	*	1	
	Asian/Pacific Islander	379.94	46.49	90		*	*	16	
	African American	356.61	43.86	1314		360.92	40.80	212	
	White	380.29	49.71	1183		371.99	42.31	213	
	Hispanic	366.89	42.12	185		*	*	26	
	Missing	*	*	1				0	
Limited English									
Proficient	Yes	*	*	49		*	*	12	
	No	368.47	48.15	2694		367.00	42.29	452	
	Exited	*	*	39		*	*	4	

* Statistics not reported for sample size less than 50 (N<50)

Table 6.17 indicates the percent of students achieving the basic, proficient, and advanced levels. Results indicated that 56.3 percent of students achieved proficiency on the exam.

Table 6.17. Percent of Students by Classification

	2005
Basic	42.7
Proficient	34.7
Advanced	22.6

Decision Accuracy and Consistency

The accuracy of decisions based on specified cut-scores was assessed for Reliability of Classification using the computer program RelClass, ETS proprietary software. RelClass provides two statistics that describe the reliability of classifications based on test scores (Livingston & Lewis, 1995). More specifically, information from an administration of one form is used to estimate the following:

3) Decision Accuracy describes the extent to which examinees are classified in the same way as they would be on the basis of the average of all possible forms of a test. Decision accuracy answers the question: How does the actual classification of test takers, based on their single-form scores, agree with the classification that would be made on the basis of their true scores, if their true scores were somehow known.

4) Decision Consistency describes the extent to which examinees are classified in the same way as they would be on the basis of a single form of a test other than the one for which data are available. Decision consistency answers the question: What is the agreement between the classifications based on two non-overlapping, equally difficult forms of the test.

Table 6.18 provides the results for the decision classification of the proficient and advanced cut-scores. The decision accuracy indices were both greater than 0.90, indicating high agreement between classifications based on the observable variable (scores on one form of a test) and classifications based on an unobservable variable (the test takers' true scores). The decision consistency indices approached 0.90, which also indicate a high agreement between classifications based on two variables (scores on the form students have taken and score from a parallel form of the same test that is not administered to the students).

Table 6.18. Decision Statistics for the English Exam

	Decision Accuracy		Decision Consistency	
	Proficient	Advanced	Proficient	Advanced
English	0.914	0.920	0.884	0.886