Maryland High School Assessment
Technical Report

Algebra and Data Analysis
Biology
English I
Geometry
Government

Educational Testing Service
June 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

Tables

Figures

Introduction

The 2004 Maryland High School Assessments (MHSA) consisted of end-of-course tests in Algebra/Data Analysis, Biology, English I, Geometry, and Government. The HSA is referred to as "end-of-course" tests, because students took each test as they completed the appropriate coursework. In addition, results from the Geometry administration were used as a High School mathematics component in the Maryland State Department of Education (MSDE) adequate yearly progress reports as required under the No Child Left Behind (NCLB) act. HSA contained selected-response (SR) items, which required students to choose between three/four short response options and are machine scored; brief constructed response (BCR) items required students to write a short response and are scored by raters; extended constructed response (ECR) items required students to write a longer response and are also 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 to the answer document. All items were based on content outlined in Maryland's Core Learning Goals.

HSAs were administered in January, May and July. In general, for January and May 2004 administrations, three operational test forms were constructed: one for the main administration window, and one for each of two make-up administrations. In order to conserve the item pool, two May make-up forms were used for July main (May make-up form 1) and July make-up form (May make-up form 2). Each test form 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. In addition, with the exception of items selected for public release, all operational items were also returned to the item bank where they are to remain unused for at least two years to minimize item exposure.

The underlying item response models used for HSA 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 a total test score and subscores were reported to 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. A study was conducted to investigate the nature and extent of differences in subscores based IP scoring versus NC scoring (see Chapter 3). While subscores were not reported at individual student level, the subscores were aggregated at classroom level to provide teachers and administrators with additional information about student performance in each of the reporting categories. A special study was also conducted that involved reviewing and replicating English 2003 results using ETS programs. Results indicated that the ETS programs successfully replicated the English 2003 results reported by CTB (see Chapter 3).

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 HSAs; Section 3 describes the scoring procedures and score types; Section 4 provides statistical summary results for each of the test forms administered in 2004; and 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.

[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. Areas, as defined by the Core Learning Goals, 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 that future administrations have a sufficient depth of items from which to construct operational forms.

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 direction to session length, item number and type by session, and other form requirements. 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 Maryland HSA 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 HSA Forms by Item Type

Content Area	SR	SPR	BCR	ECR
Algebra	26	6	3	3
Biology	48	-	7	-
English	50	-	2	1
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 HSA testing program. The item writers were trained on general item writing techniques as well as writing parameters that were specific to the Maryland HSA 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. Many specific requirements of writing for Maryland HSA program can be found in "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 2004 operational test forms were constructed from items from the Maryland item bank. The pool of items available for use in the construction of the 2004 forms included all items that had been administered, calibrated and linked to the operational scale. For HSA 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 2004 forms[2]. 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 (see also Section 5 for more details about 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 I BCR items were worth four points and ECR items were worth six 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 I	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.

[2] Subsequent to the selection for the 2004 forms, a linking study was conducted to place some additional items onto the operational scale. The results of this study are located in Appendix 1.A.

Table 1.2 Algebra/Data Analysis 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 I Blueprint

ENGLISH				
Reporting Category	ITEM TYPE			Percent of Points
	SR	BCR	ECR	
	(1pt/SR)	(3pt/BCR)	(4pt/ECR)	
TOTALS	50	2	1	
Goal 1 The student will demonstrate the ability to respond to a text by employing personal experiences and critical analysis.				35%
Goal 2 The student will demonstrate the ability to compose in a variety of modes by developing content, employing specific forms, and selecting language appropriate for a particular audience and purpose.				31%
Goal 3 The student will demonstrate the ability to control language by applying the conventions of standard English in writing and speaking.				20%
Goal 4 The student will demonstrate the ability to evaluate the content, organization, and language of texts.				14%

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 30% 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 were listed in Tables 1.7 and 1.8. The exact composition of the forms varied slightly based on available items in the pool.

Table 1.7 January Administration

Primary Week	Make-Up #1	Make-Up #2[1]
January common set - 30%	January common set – 30%	January common set – 30%
Unique Items from the pool –70%	Items from January Operational - 35%[2]	Items from January Operational - 35%[2]
	Unique Items from the pool – 35%	Unique Items from the pool – 35%
Field Test Section – 2 versions	Field Test Section – same as 1st operational version	Field Test Section – same as 1st operational version

Notes. [1]For Government and Biology, the same make-up form was administered for both administrations.
[2]Items from the January Operational administration included in Make-up 1 and 2 must be different.

Table 1.8 May Administration

Primary Week	Make-Up #1	Make-Up #2
May Common Set - 30%	May Common Set - 30%	May Common Set - 30%
Unique Items from the pool –70%	Items from May Operational - 35%[1]	Items from May Operational - 35%[1]
	Unique Items from the pool – 35%	Unique Items from the pool – 35%
Field Test Section – 8 versions	Field Test Section – same as 1st operational version	Field Test Section – same as 1st operational version

Notes. [1]Items from the May Operational administration included in Make-up 1 and 2 must be different.

In addition to the operational items, an embedded field test section was included with each version of the test form, resulting in several versions of the operational form that

differed only by the set of 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 distracters. 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 forms were constructed simultaneously; in order 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, etc.
4. After the operational forms were selected, the field test sections were constructed. Field test sections 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 sections 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 section was included at the end of Session 2.

In each content area, TCCs and CSEMs for each of the test forms are plotted in figures 1.1 to 1.10. In general the TCCs and CSEMs closely matched the target. Where forms varied in difficulty, differences were minimized in the scale score region of the cut-scores and, in all cases the difference was less than 5% of the total raw score, i.e. the passing raw score difference of the two forms is less than 5%.

Figure 1.1 Test Characteristic Curve: Algebra

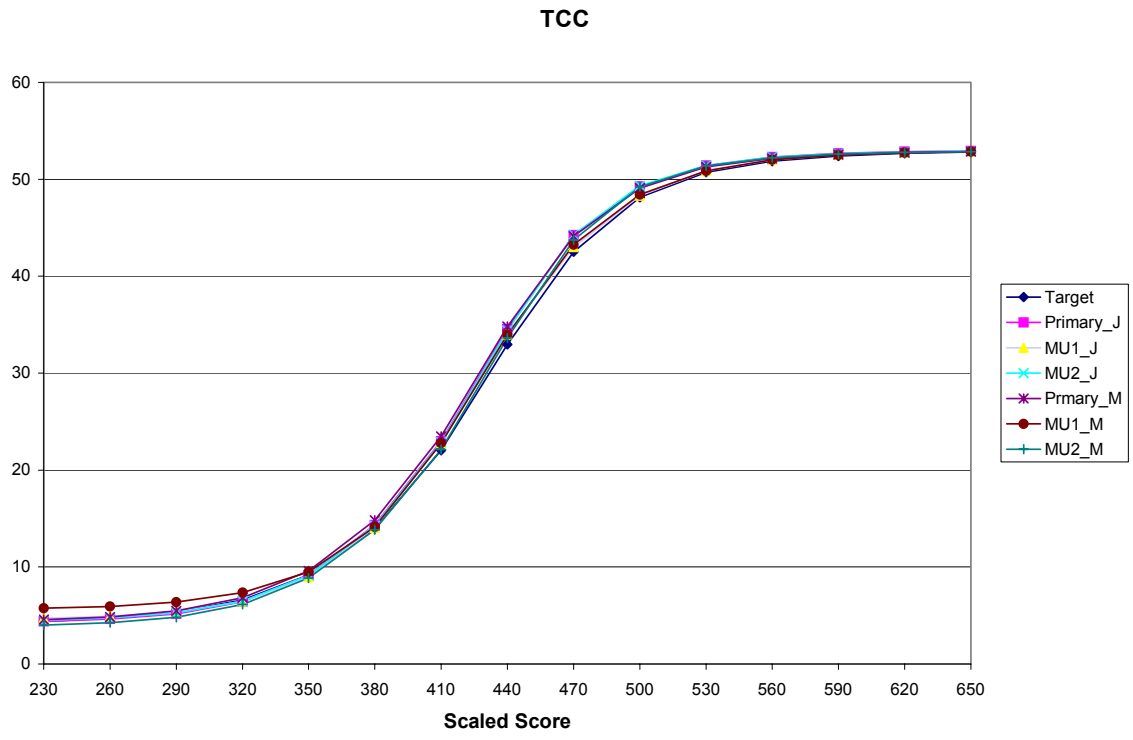


Figure 1.2. Conditional Standard Error of Measurement: Algebra

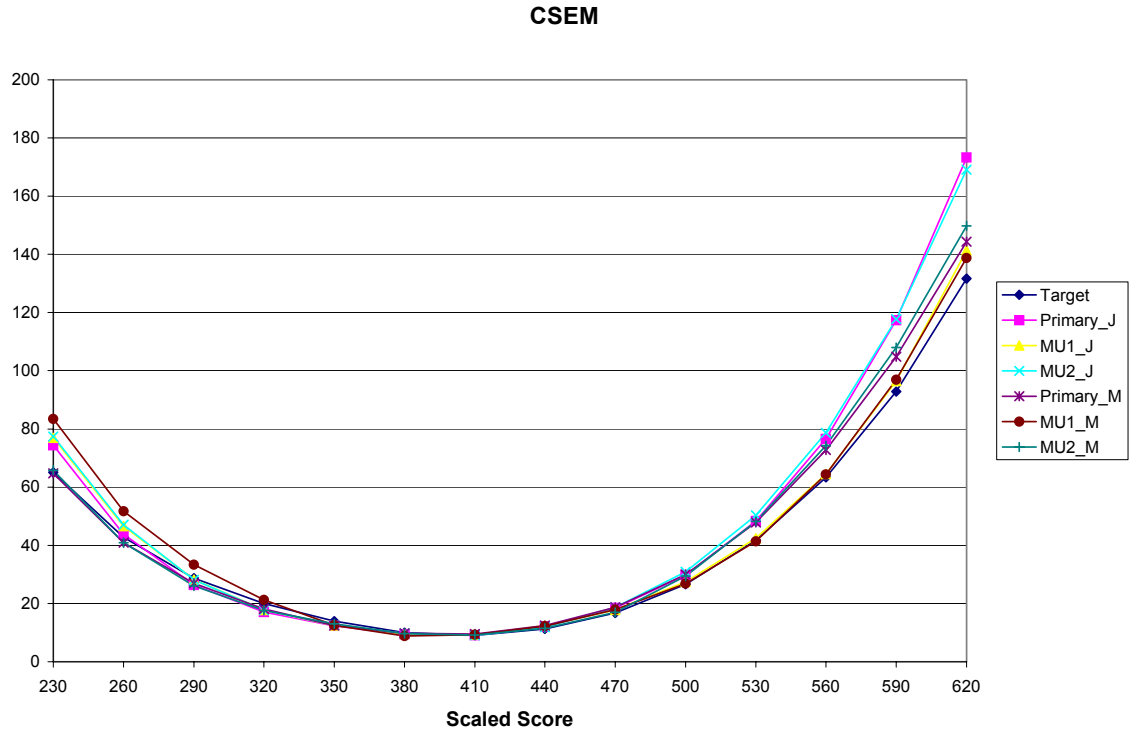


Figure 1.3 Test Characteristic Curve: Biology

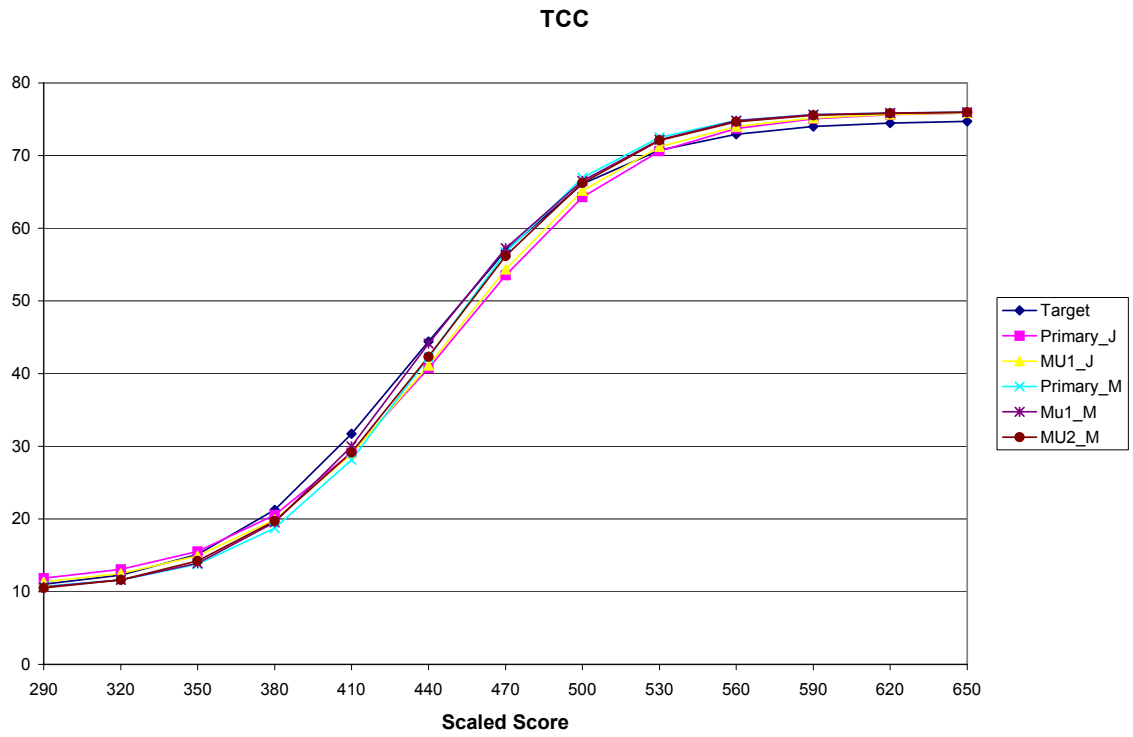


Figure 1.4 Conditional Standard Error of Measurement: Biology

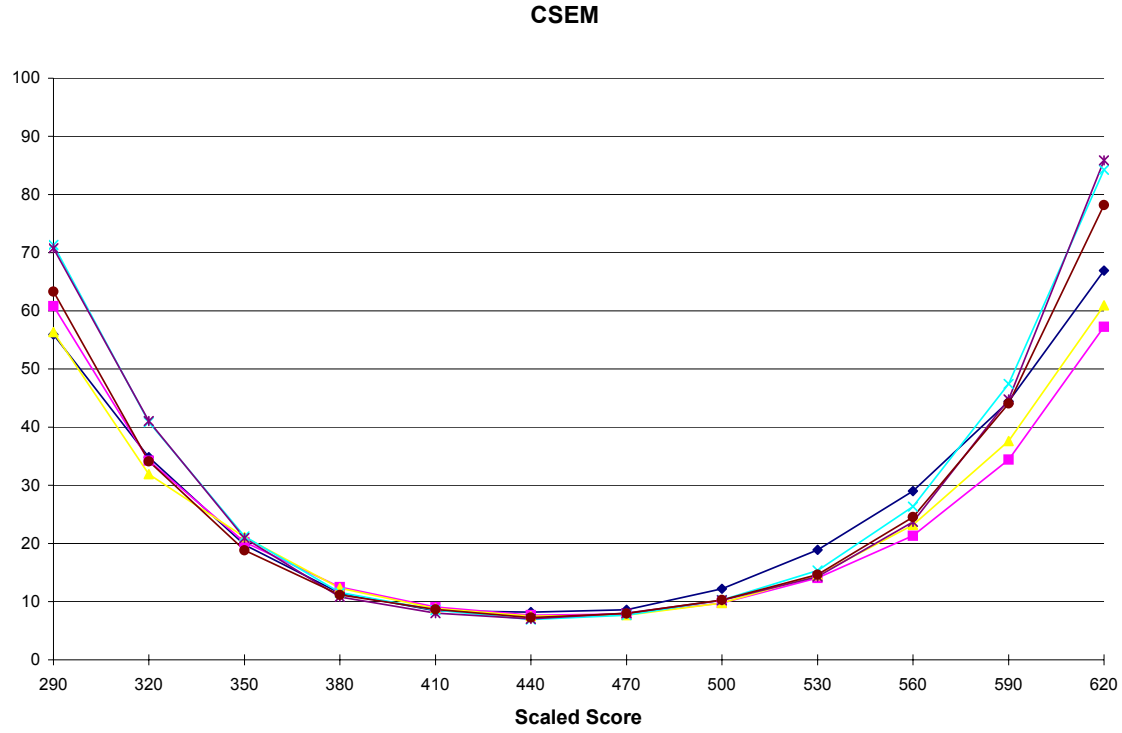


Figure 1.5. Test Characteristic Curve: English I

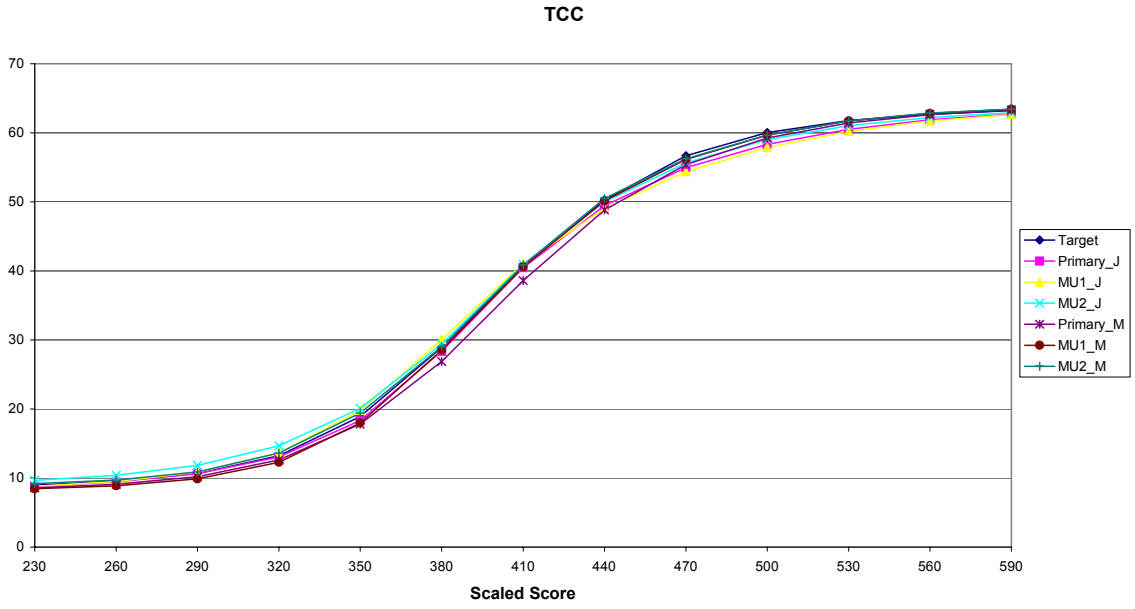


Figure 1.6. Conditional Standard Error of Measurement: English I

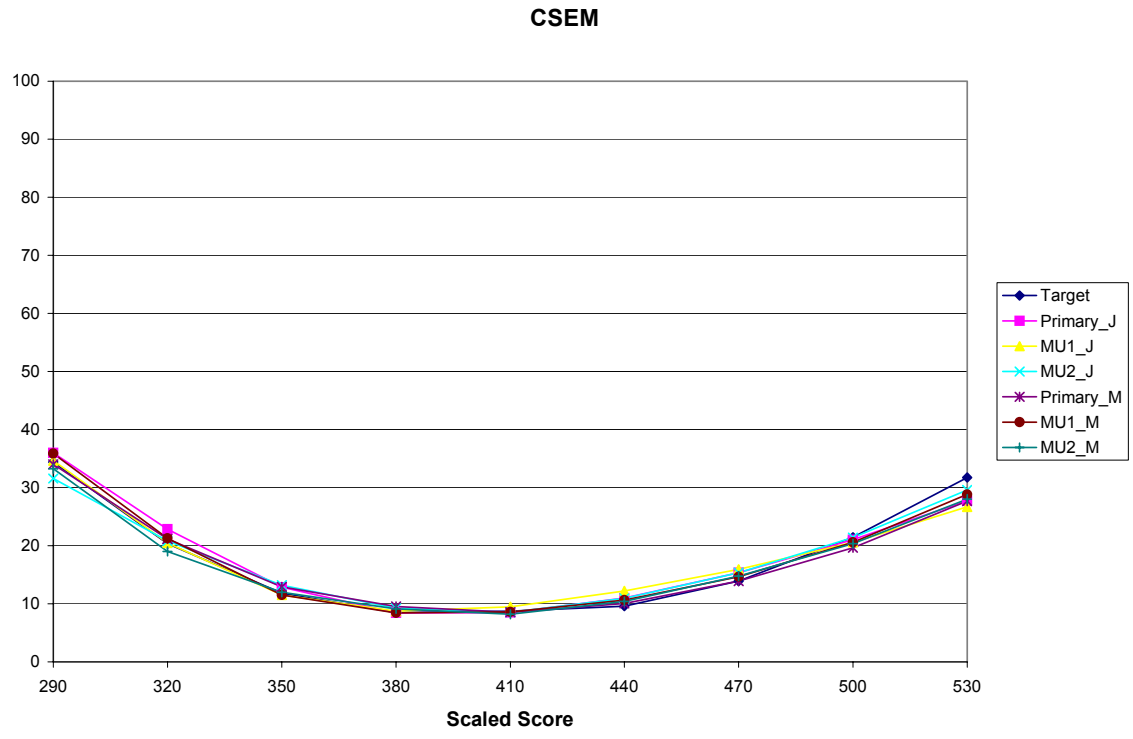


Figure 1.7. Test Characteristic Curve: Geometry

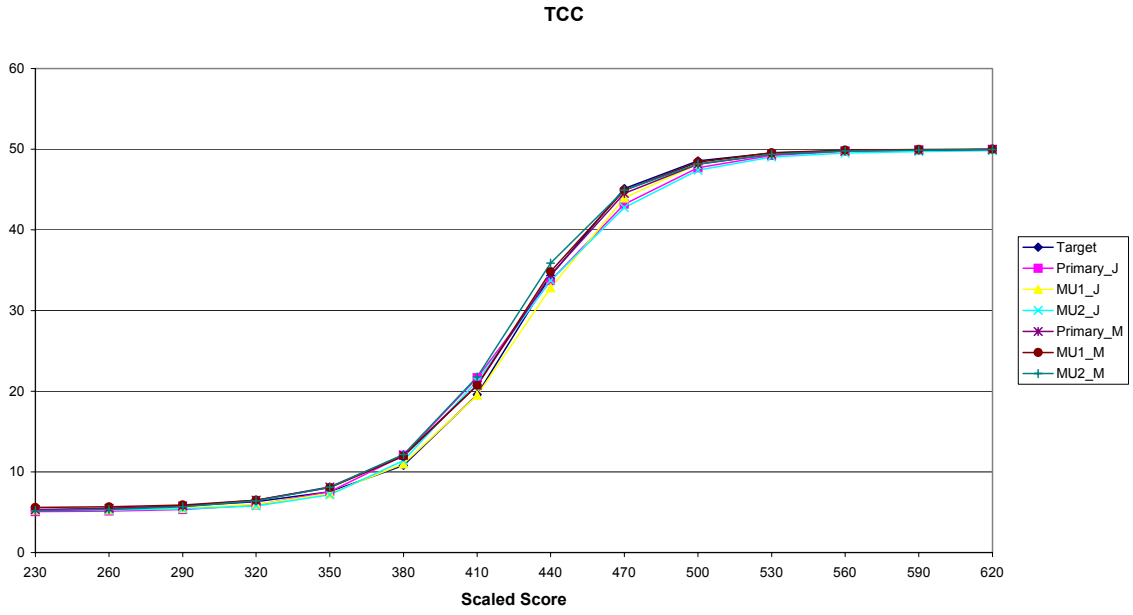


Figure 1.8. Conditional Standard Error of Measurement: Geometry

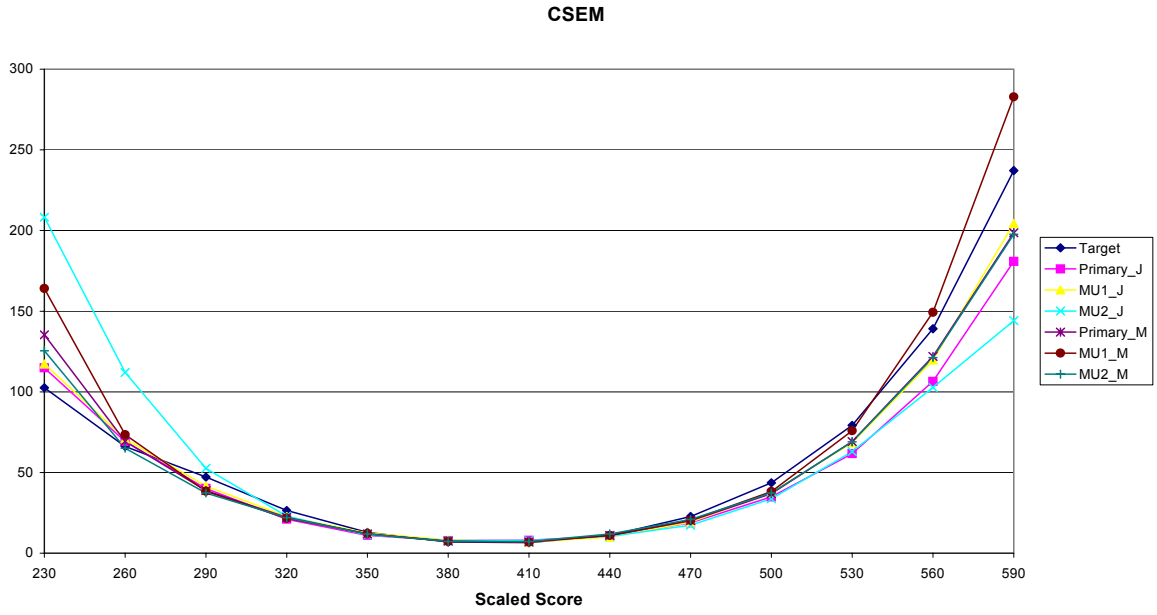


Figure 1.9. Test Characteristic Curve: Government

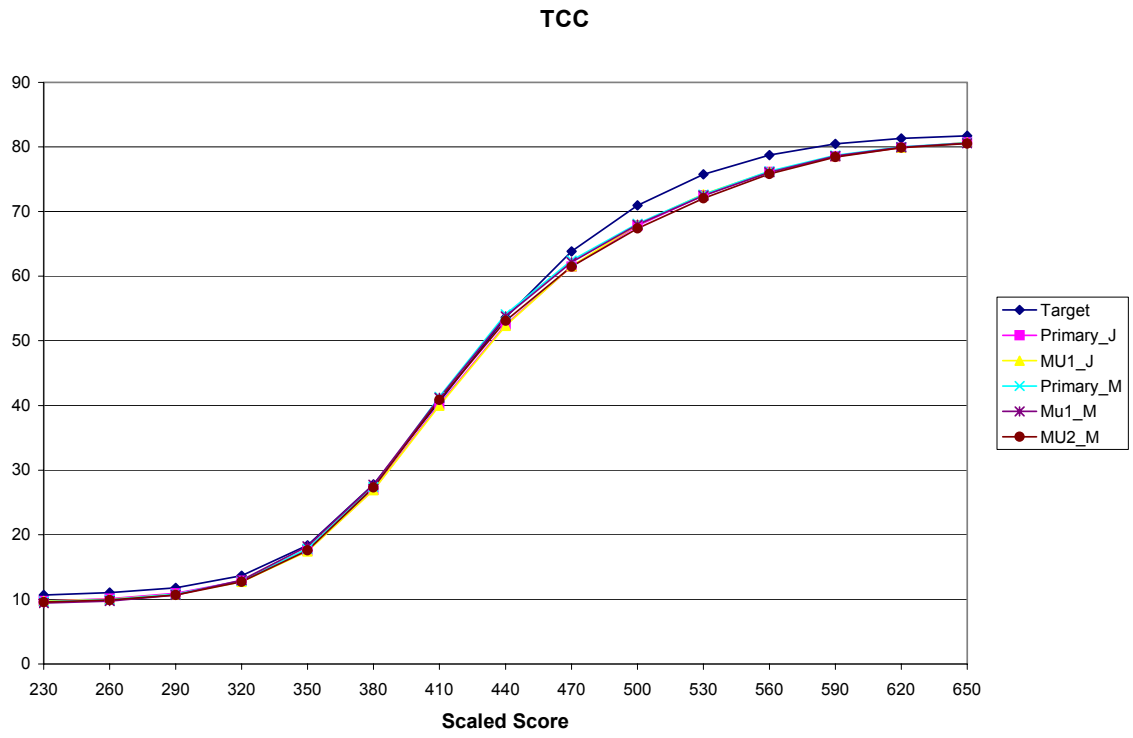
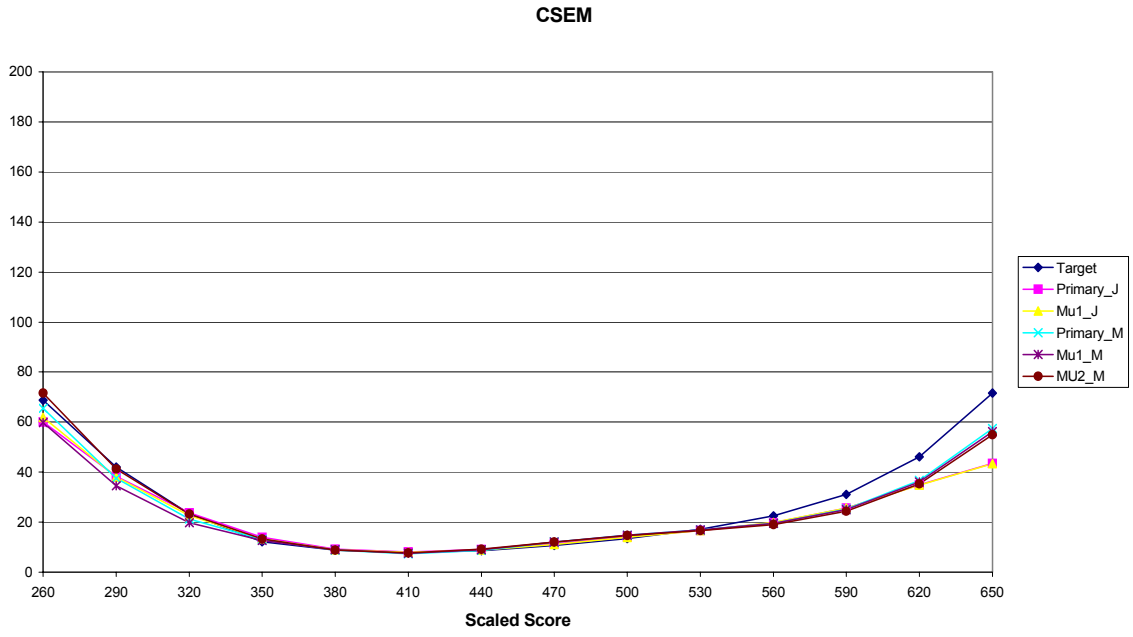


Figure 1.10. Conditional Standard Error of Measurement: Government



Appendix 1.A. Linking Study: 2000-2001 to the Operational Scale (2003)

Maryland High School Assessment

Linking Study

2000-2001 to the Operational Scale (2003)

March 17, 2004

Educational Testing Service

Appendix 1.A. Linking Study: 2000-2001 to the Operational Scale (2003)

Background

The Maryland High School Assessment (HSA) has been administered since 2000. While new items have been developed and there is a substantial item pool, not all items were on the operational/reporting scale - which is defined by the 2002 administration. Essentially there were two sets of items: 1.) items administered in 2000 or 2001 and not administered again in either 2002 or 2003 and 2.) items administered in 2002 or 2003 – these include both newly developed items and items previously administered in 2000 or 2001. The first set of items were on the "field test scale", the second set of items were on the operational or reporting scale (mean 400, sd 40).

The items from 2000 and 2001 have not been linked to this new scale. Rather the items remained on the previous "field test" scale, which was defined following the 2000 administration. The intention was to administer the 2000-2001 items again and recalibrate the items prior to use on future forms. However the items were transformed to be on a 400/40 scale without a linking study. This is explained in an email from R. Clymer, Program Manager, CTB (November 26, 2003),

> The psychometric council requested that all old field test items be recalibrated for the operational administration due to the quality of the field test items (e.g., high omit rates, motivation, etc.). The items were only transformed to 400/40 for the purpose of item selection in 2003.

In consideration of the large numbers of items that were on the 2000-2001 scale, MSDE requested the completion of a special linking study to help ascertain whether the items on the field test scale could be placed onto the operational scale without administering and recalibrating them again.

Method

To complete this study, items that could serve as a linking set were identified and included using a Stocking & Lord linking approach. Items that were administered first in 2000 or 2001 and again in 2003 were included. The numbers of items by administration that were included in the linking study were listed in Tables 1.A.1-1.A.5 below. The majority of the items were from the May administrations.

Table 1.A.1 Algebra (n=144)

	Jan-03	May-03
Jan-00	0	0
May-00	3	63
Jan-01	0	0
May-01	18	60

Table 1.A.2 Biology (n=185)

	Jan-03	May-03
Jan-00	0	0
May-00	26	36
Jan-01	6	7
May-01	43	67

Table 1.A.3 English I (n=156)

	Jan-03	May-03
Jan-00	4	2
May-00	30	19
Jan-01	0	0
May-01	15	86

Table 1.A.4 Geometry (n=126)

	Jan-03	May-03
Jan-00	0	0
May-00	23	67
Jan-01	0	0
May-01	14	22

Table 1.A.5 Government (n=138)

	Jan-03	May-03
Jan-00	1	0
May-00	26	52
Jan-01	0	10
May-01	21	28

Results

Results of the Stocking and Lord linking were presented in the tables and plots that follow. All available items included in the linking were retained for all content areas, except Geometry. In this content area, six items were identified as unstable in the expected p-values and B-value plots due to the difference in expected p-values whose values were greater than .20. The correlation of the reference (anchor) and linking items after the S/L procedure for expected p-values and the B-parameters were .84 and .86, respectively. After removing these items, the correlation improved to .92 and .90 (see Table 1.A.6). It was noted that these items appeared in vastly different regions of the test books (e.g., sequence #2 in 2000 and sequence #57 in 2003) and these differences may be related to context effects.

The results of the linking suggest that the items on the field test scale could be placed onto the operational scale. In all cases the correlation between the reference (anchor) and linking items after the S/L procedure were .90 or greater than for expected p-values and B-parameters (see Table 1.A.6). The correlation between the A-parameters is highest for government (.80) and lowest for geometry (.66). The correlation between the C-parameters is lowest for English I (.35) and Biology (.47).

Table 1.A.6 Correlations of Reference (Anchor) and Link Item Parameters

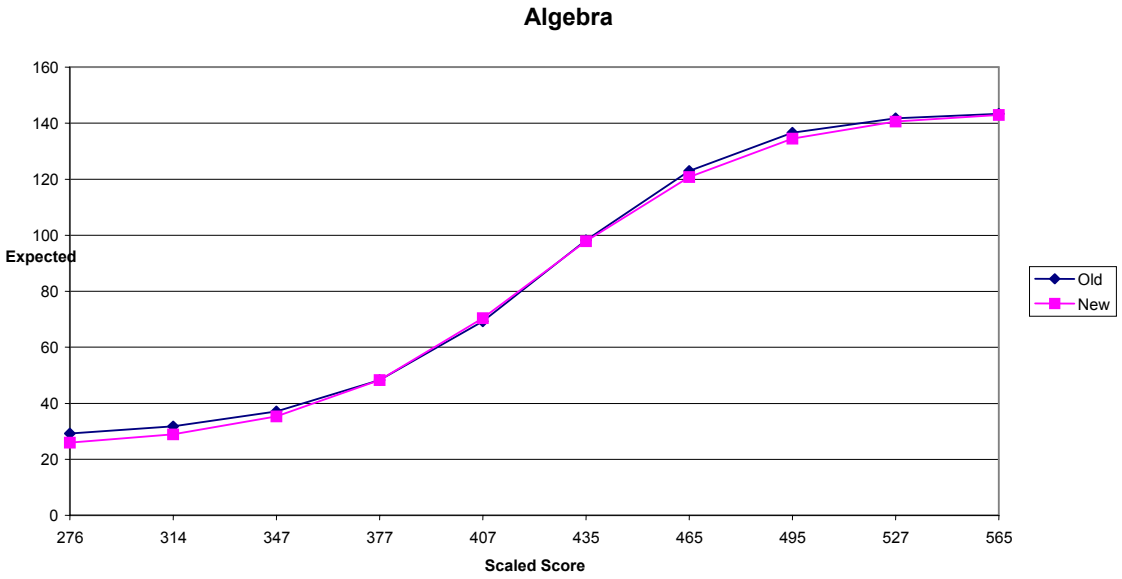
	expected p-value	B-parameter	A-parameter	C-parameter
Algebra	0.92	0.91	0.68	0.56
Biology	0.96	0.94	0.75	0.47
English I	0.94	0.90	0.71	0.35
Geometry	0.92	0.90	0.66	0.61
Government	0.92	0.90	0.80	0.55

For each content area, the following information is presented:

- Plot showing the alignment of the test characteristic curves based on the reference (anchor) and linking items after the S/L procedure.
- Transformation constants
- Bivariate plot showing the alignment of p-values estimated for the reference (anchor) and linking items after the S/L procedure. The correlation is noted in the second line of the title.
- Bivariate plot showing the alignment of the A-, B-, and C-parameters for the reference (anchor) and linking items before and after the S/L procedure. The correlation is noted in the second line of the title.
- A table of descriptive statistics (mean, sd, minimum, maximum) for the expected p-values, A-, B-, and C-parameters follows each plot.

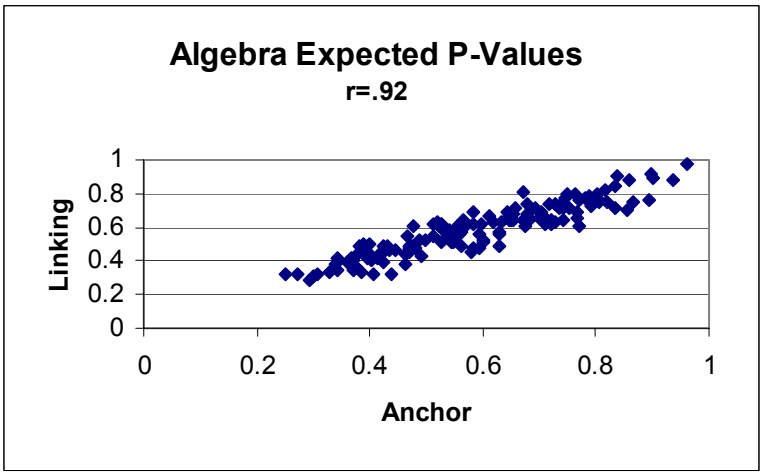
Appendix 1.A

Algebra



Transformation Constants

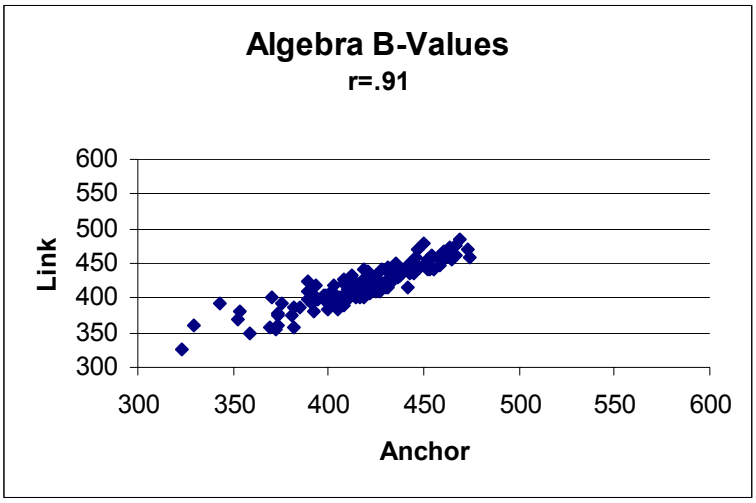
	Slope	Intercept
Algebra	0.93	50.67



Algebra Expected P-Values

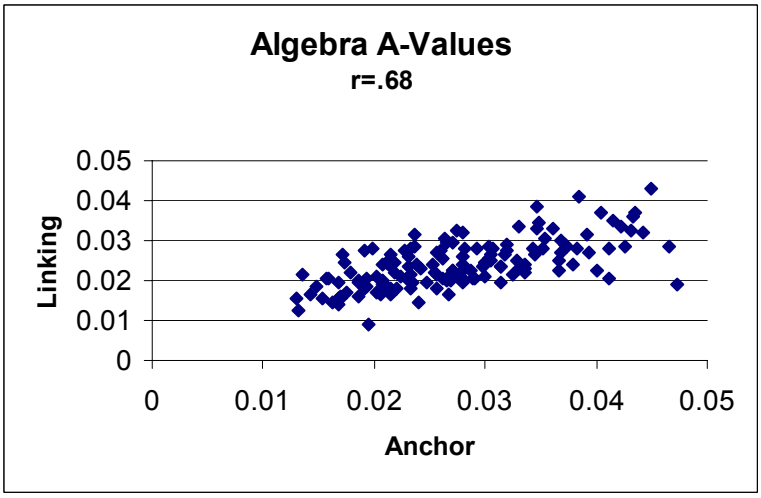
	Reference	Link
Mean	0.59	0.59
SD	0.16	0.15
Minimum	0.25	0.29
Maximum	0.96	0.97

Appendix 1.A



Algebra B-Values

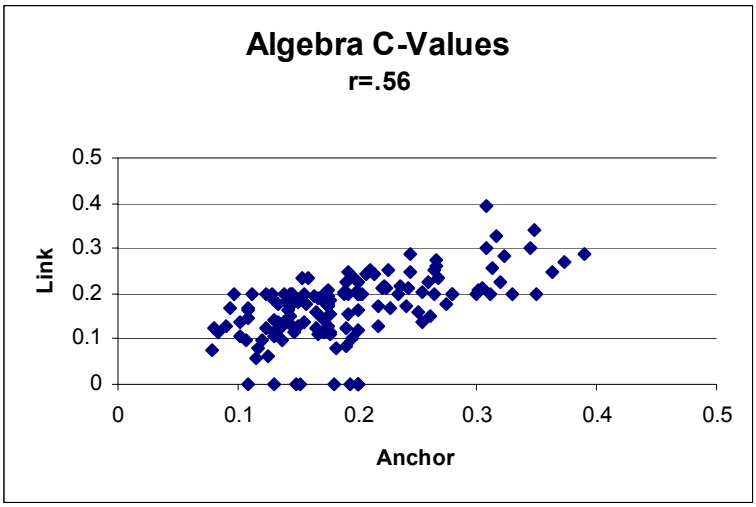
	Reference	Link
Mean	420.94	421.02
SD	29.74	29.14
Minimum	322.94	325.13
Maximum	474.44	483.68



Algebra A-Values

	Reference	Link
Mean	0.0280	0.0243
SD	0.0086	0.0061
Minimum	0.0130	0.0088
Maximum	0.0512	0.0429

Appendix 1.A

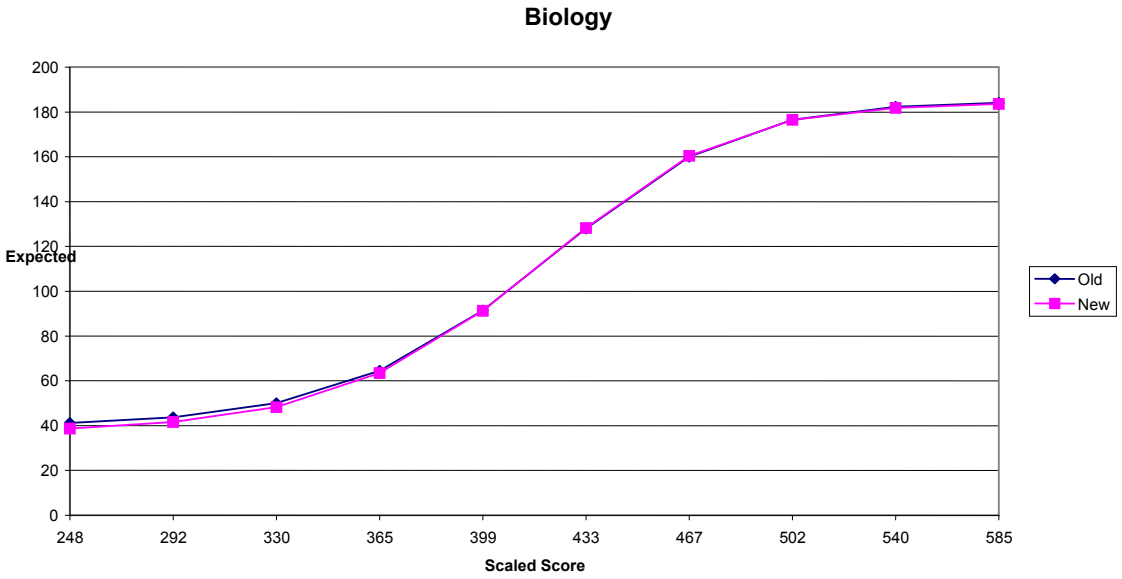


Algebra C-Values

	Reference	Link
Mean	0.19	0.17
SD	0.07	0.07
Minimum	0.08	0.00
Maximum	0.39	0.39

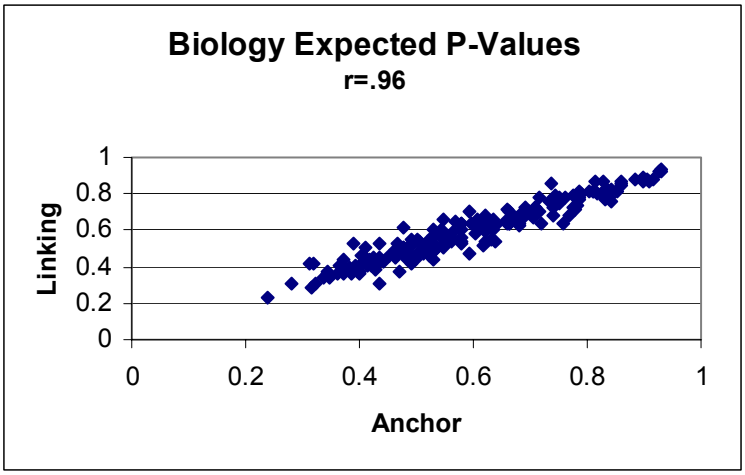
Appendix 1.A

Biology



Transformation Constants

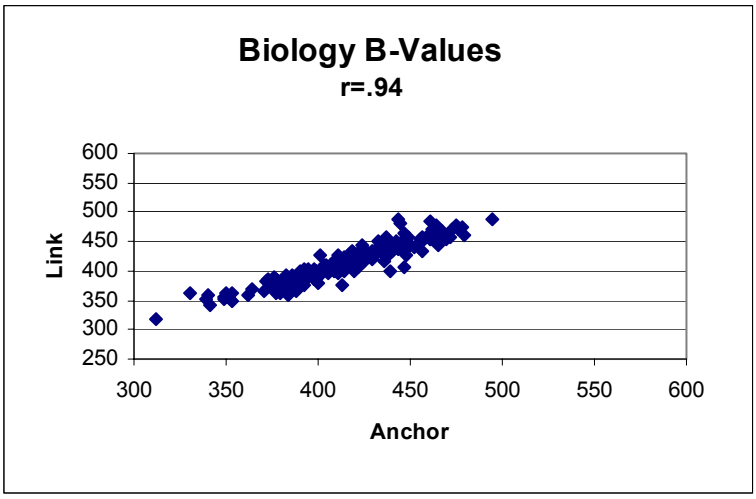
	Slope	Intercept
Biology	0.93	34.19



Biology Expected P-Values

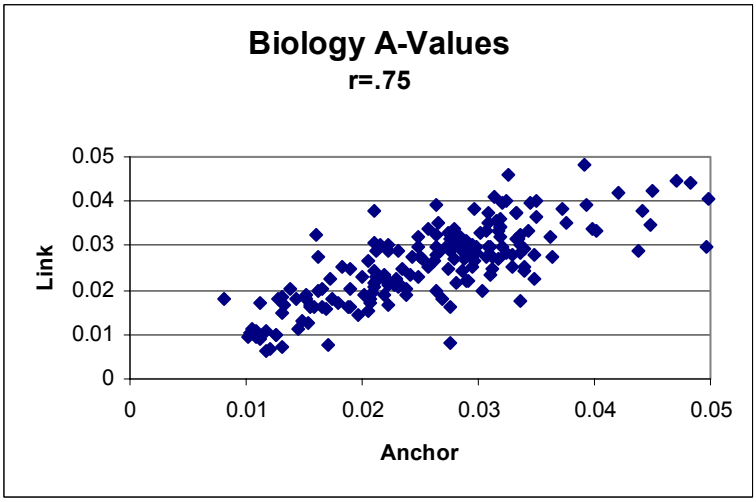
	Reference	Link
Mean	0.60	0.60
SD	0.16	0.16
Minimum	0.24	0.23
Maximum	0.93	0.94

Appendix 1.A



Biology B-Values

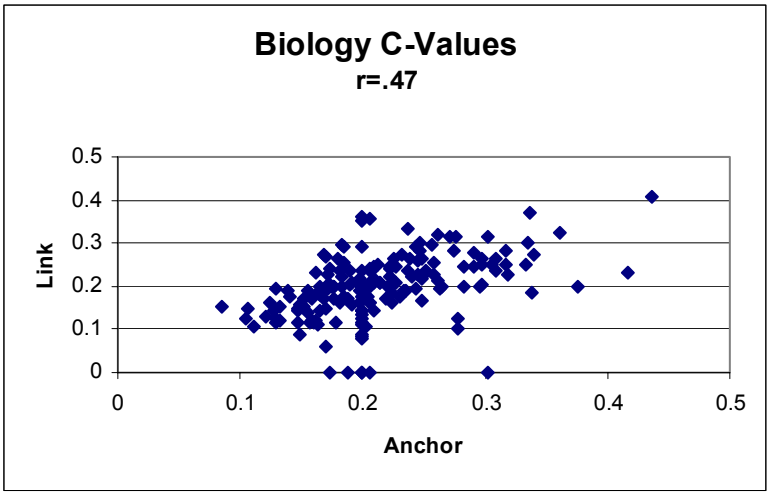
	Reference	Link
Mean	416.04	415.67
SD	34.76	35.08
Minimum	282.23	271.33
Maximum	495.07	487.84



Biology A-Values

	Reference	Link
Mean	0.026	0.026
SD	0.009	0.009
Minimum	0.008	0.006
Maximum	0.050	0.054

Appendix 1.A

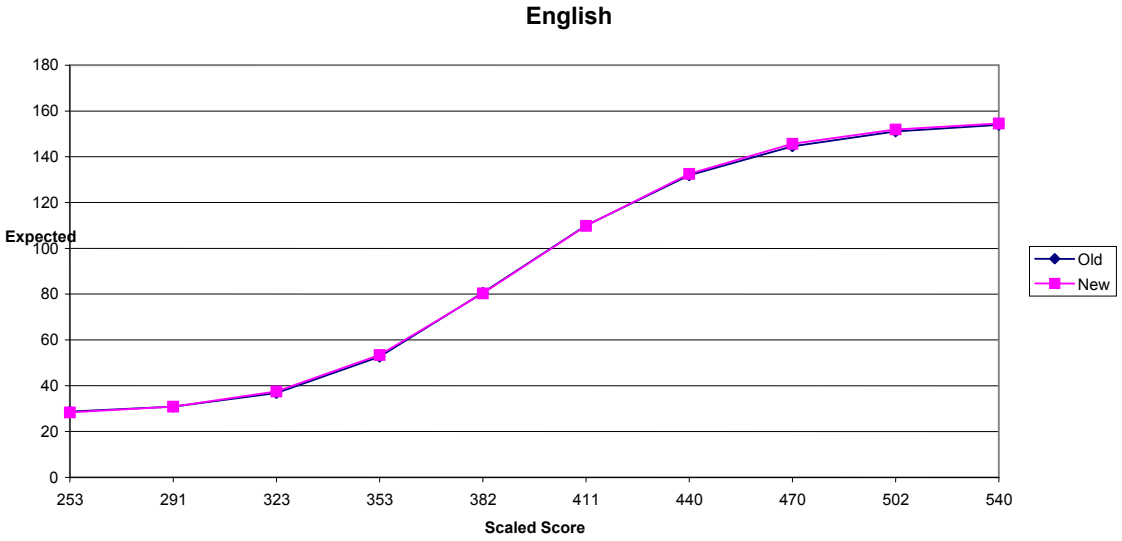


Biology C-Values

	Reference	Link
Mean	0.21	0.20
SD	0.06	0.07
Minimum	0.08	0.00
Maximum	0.44	0.41

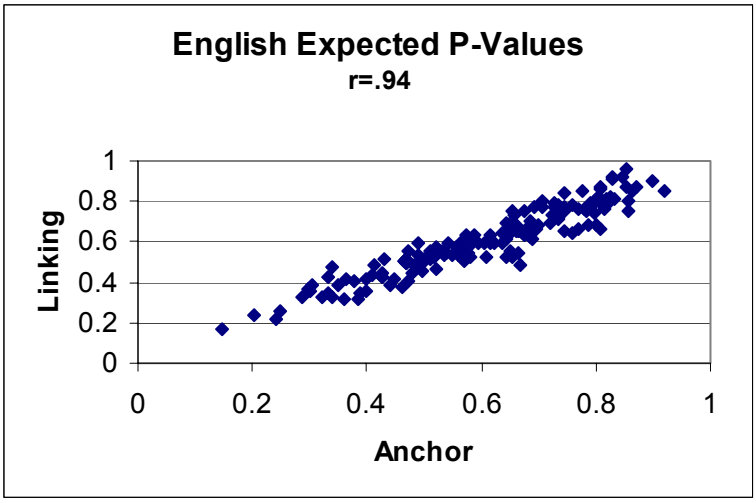
Appendix 1.A

English I



Transformation Constants

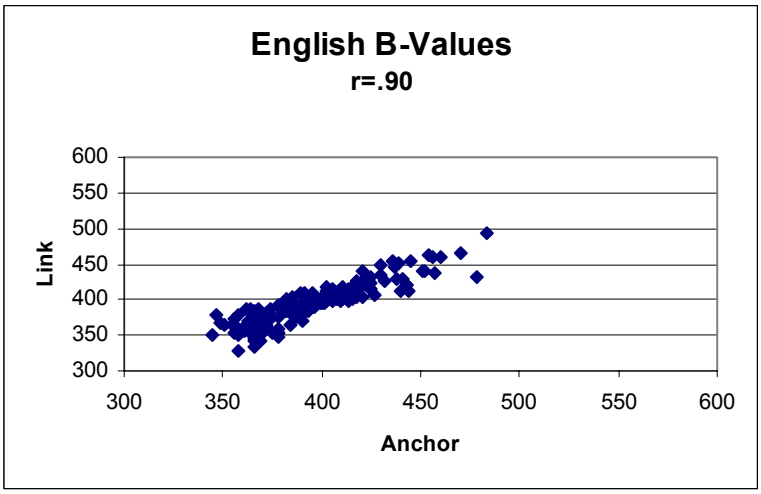
	Slope	Intercept
English I	0.80	74.89



English I Expected P-Values

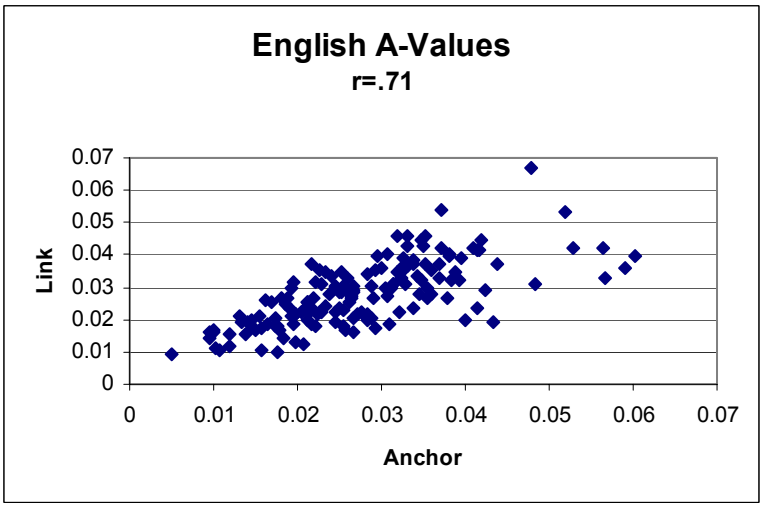
	Reference	Link
Mean	0.60	0.60
SD	0.17	0.17
Minimum	0.15	0.17
Maximum	0.92	0.96

Appendix 1.A



English I B-Values

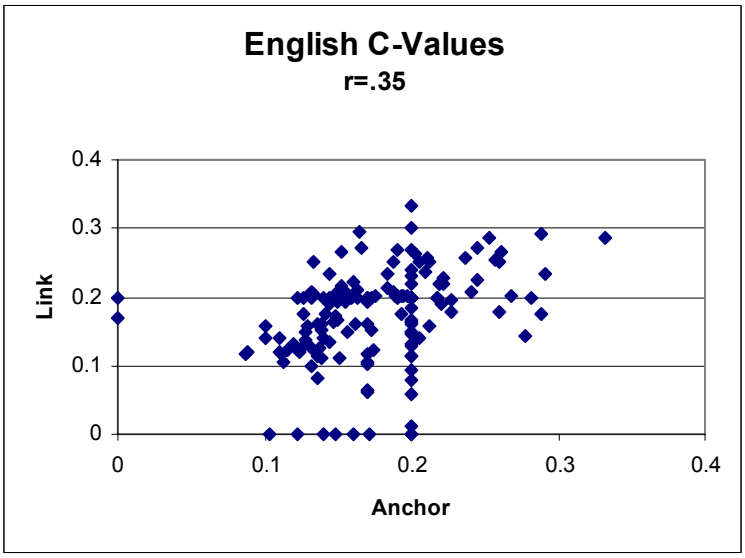
	Reference	Link
Mean	396.44	395.02
SD	29.48	29.79
Minimum	344.10	328.29
Maximum	483.00	492.34



English I A-Values

	Reference	Link
Mean	0.028	0.028
SD	0.010	0.010
Minimum	0.005	0.009
Maximum	0.060	0.067

Appendix 1.A

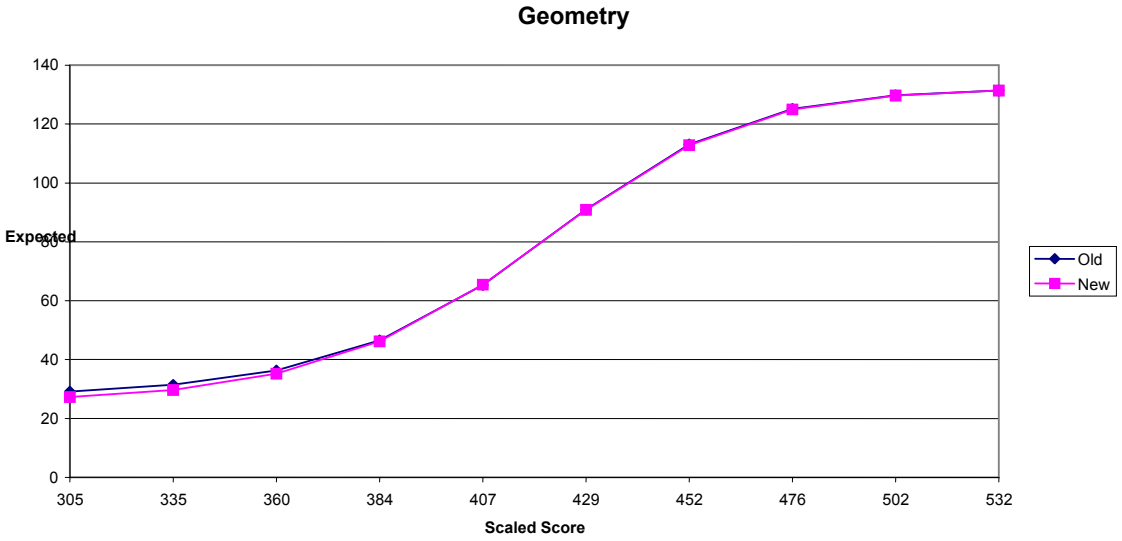


English I C-Values

	Reference	Link
Mean	0.18	0.17
SD	0.05	0.07
Minimum	0.00	0.00
Maximum	0.33	0.33

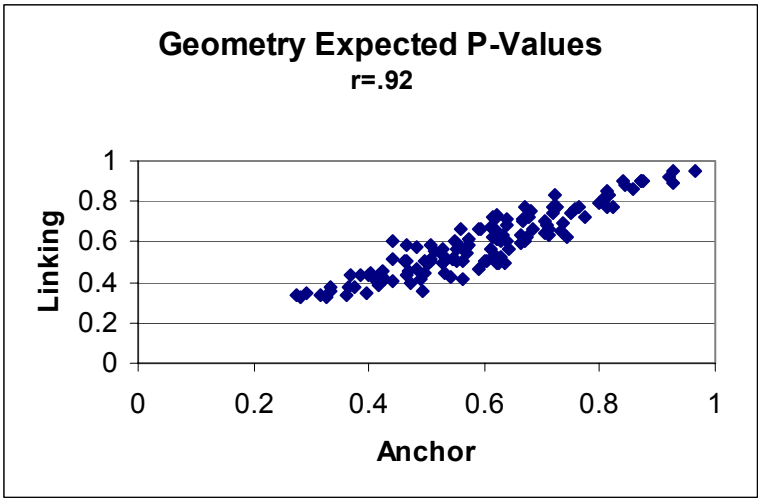
Appendix 1.A

Geometry



Transformation Constants

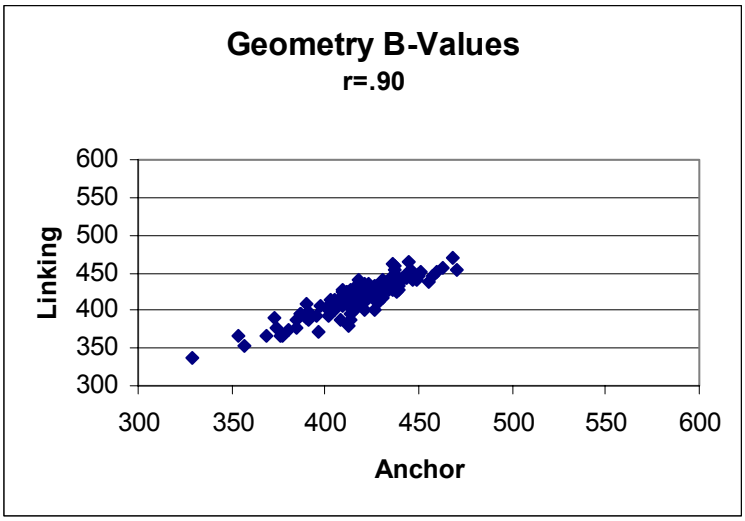
	Slope	Intercept
Geometry	0.72	113.28



Geometry Expected P-Values

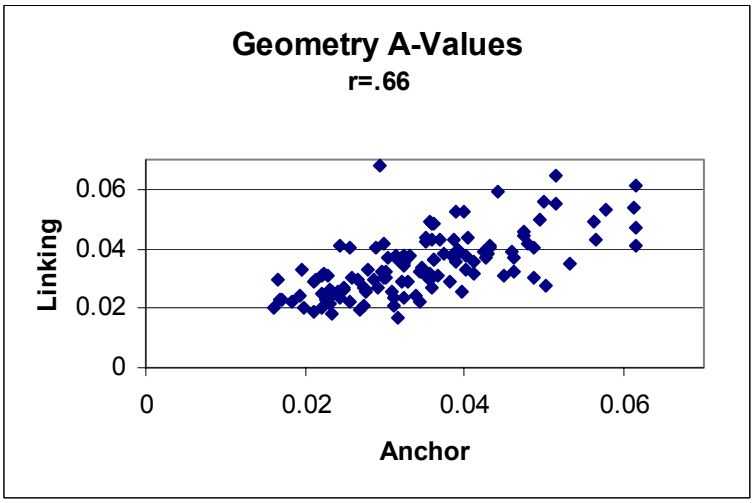
	Reference	Link
Mean	0.60	0.60
SD	0.15	0.16
Minimum	0.27	0.32
Maximum	0.96	0.95

Appendix 1.A



Geometry B-Values

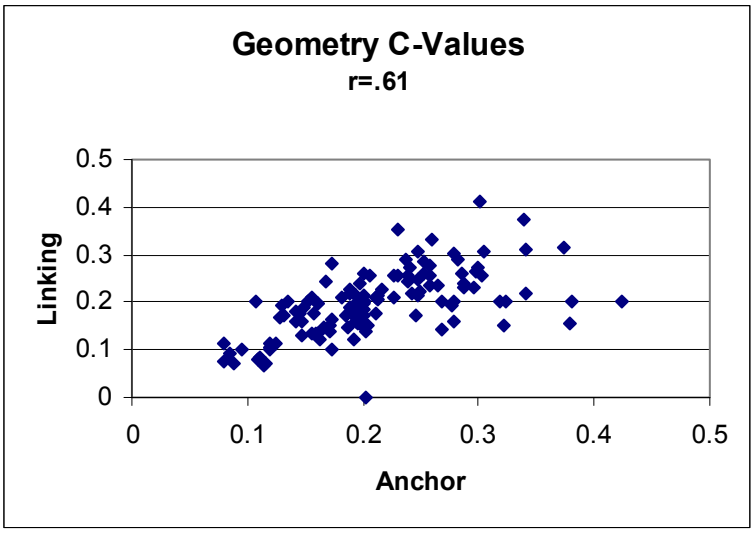
	Reference	Link
Mean	417.85	417.47
SD	23.32	24.33
Minimum	329.05	338.25
Maximum	469.84	469.20



Geometry A-Values

	Reference	Link
Mean	0.035	0.035
SD	0.011	0.010
Minimum	0.016	0.017
Maximum	0.062	0.068

Appendix 1.A

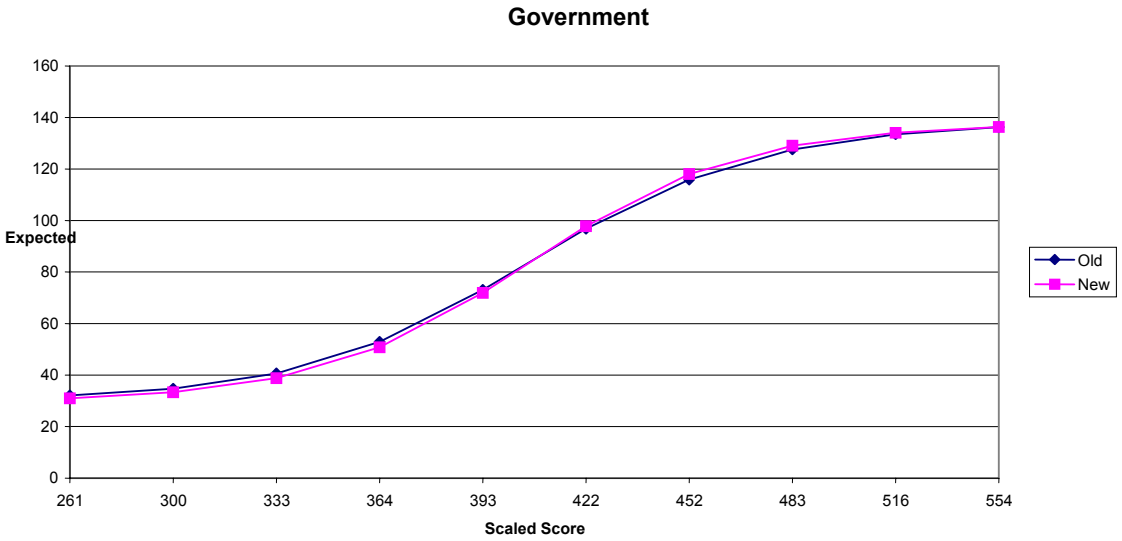


Geometry C-Values

	Reference	Link
Mean	0.21	0.20
SD	0.07	0.07
Minimum	0.08	0.00
Maximum	0.42	0.41

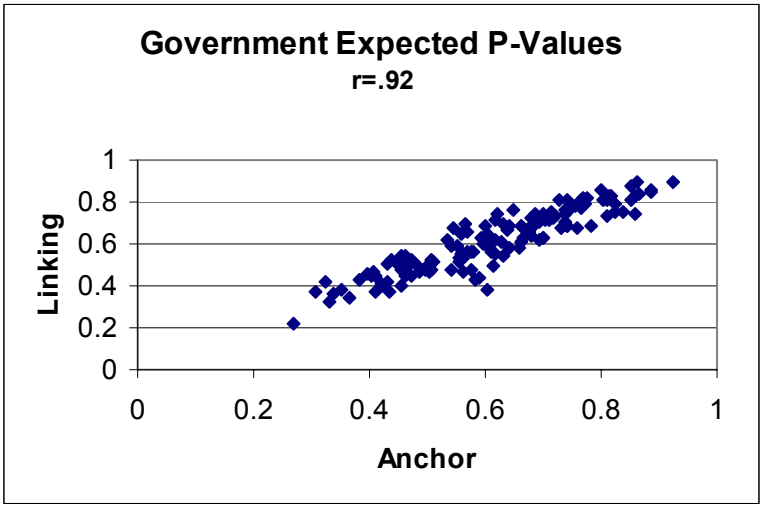
Appendix 1.A

Government



Transformation Constants

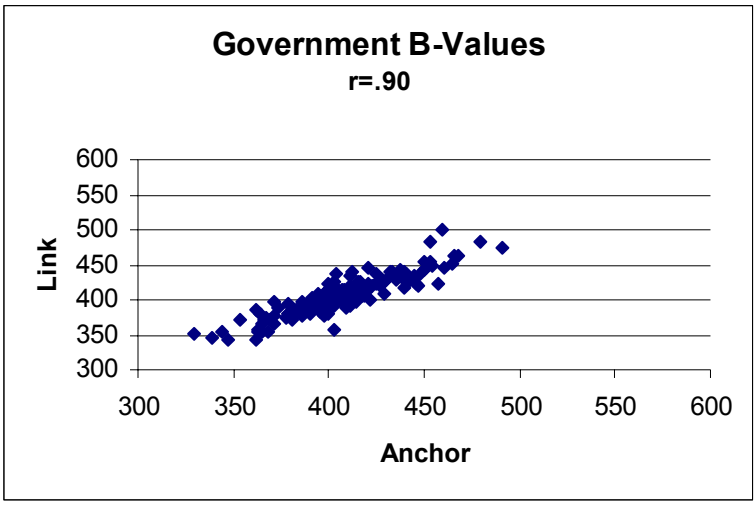
	Slope	Intercept
Government	0.99	0.34



Government Expected P-Values

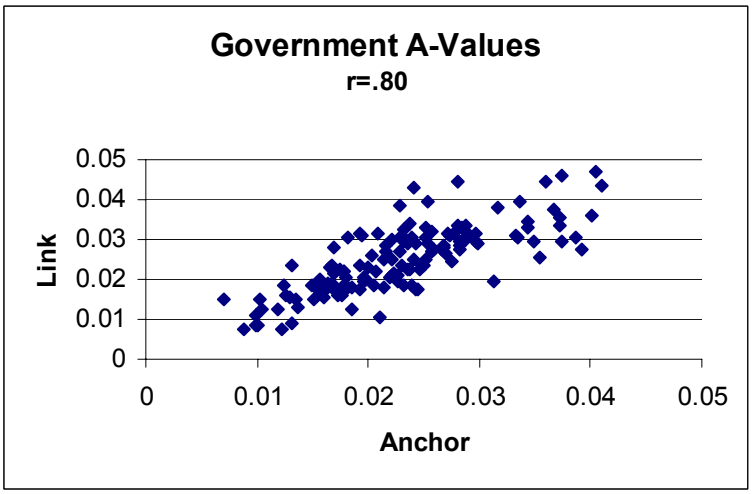
	Reference	Link
Mean	0.61	0.61
SD	0.15	0.15
Minimum	0.27	0.22
Maximum	0.92	0.90

Appendix 1.A



Government B-Values

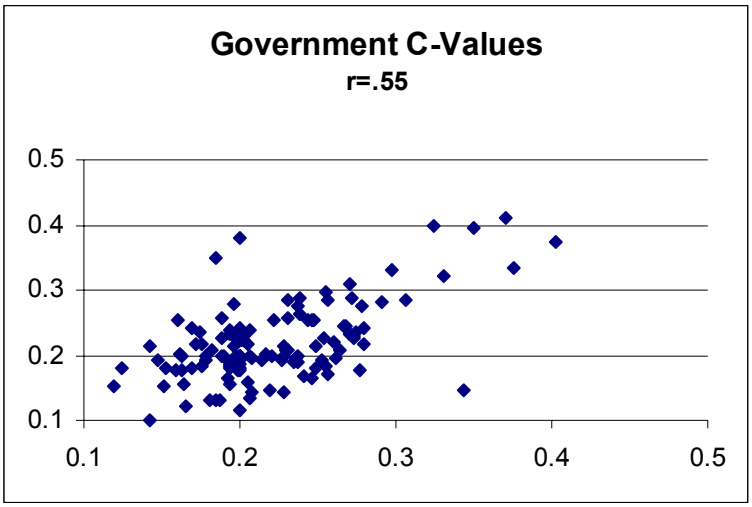
	Reference	Link
Mean	407.85	407.26
SD	30.16	30.12
Minimum	329.65	342.79
Maximum	491.29	500.15



Government A-Values

	Reference	Link
Mean	0.023	0.025
SD	0.008	0.009
Minimum	0.007	0.007
Maximum	0.049	0.050

Appendix 1.A



Government C-Values

	Reference	Link
Mean	0.22	0.21
SD	0.05	0.06
Minimum	0.12	0.00
Maximum	0.40	0.41

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 HSA was overseen by a content expert who has a depth of knowledge and teaching experience related to the course in which the HSA 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 HSA 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 in Section 1) 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 HSA 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 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, this report contains relevant empirical information in support of the Maryland HSA as follows.

- Section 3 provides detailed information concerning the particular scores that were reported for the Maryland HSA and includes an evaluation of different procedures for reporting subscore performance.

- Section 4 provides demographic information for the population of students who were administered the Maryland HSA as well as summaries of test level statistics. Summary statistics and reliability estimates were reported for the student population and by subgroups. Score distributions as well as the passing rates for all administrations and evidence that the tests were not speeded were also provided in this section.

- Section 5 includes documentation of the analysis procedures as well as distributions of item p-values and item-total correlations from the field testing activities. This section also includes an empirical evaluation of the impact of changing test directions for brief-constructed response items in the Government assessment.

- Appendix 2.A presents the results of factor analyses of the Maryland HSA item responses.

Appendix 2.A Factor Analysis Results

Maryland High School Assessment

Factor Analysis Study

March 17, 2004

Educational Testing Service

Appendix 2.A Factor Analysis Results

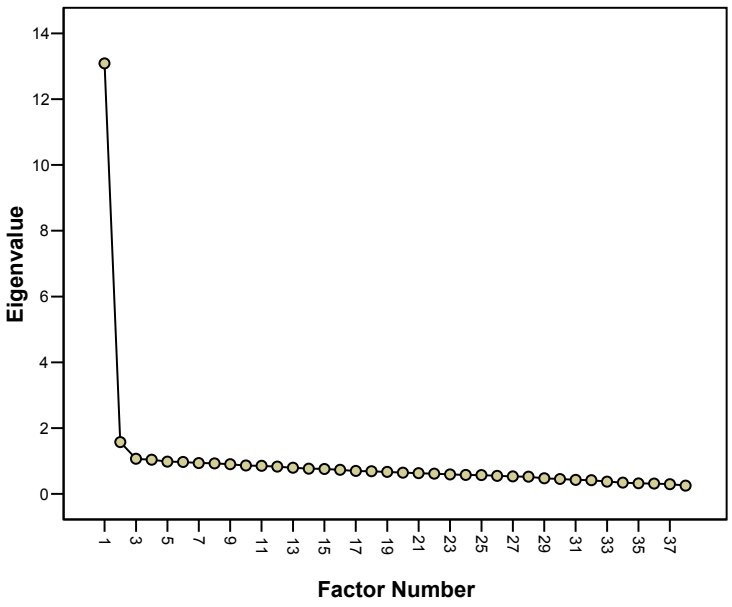
Factor analysis techniques were employed to investigate the dimensionality of the HSA content area tests. A random sample of 5000 students from the May 2004 administration was used for the analysis.

Given the ordinal nature of the item scores, matrices consisting of tetrachoric and polychoric correlations were produced for each subject area using PRELIS (Joreskog & Sorbom, 1993) and then analyzed within SPSS. The eigenvalues, percentage of variation accounted for, and the associated scree plots were provided.

Algebra

The Algebra factor analysis shows an initial eigenvalue of 13.091 for the first factor, which accounts for 34.45% of the variance. The next three factors have eigenvalues just slightly greater than one, for instance, the second factor's eigenvalue drops to 1.575, accounting for only 4.144% of the variance. The scree plot for this factor analysis is provided below; it appears as if one dominant factor is present.

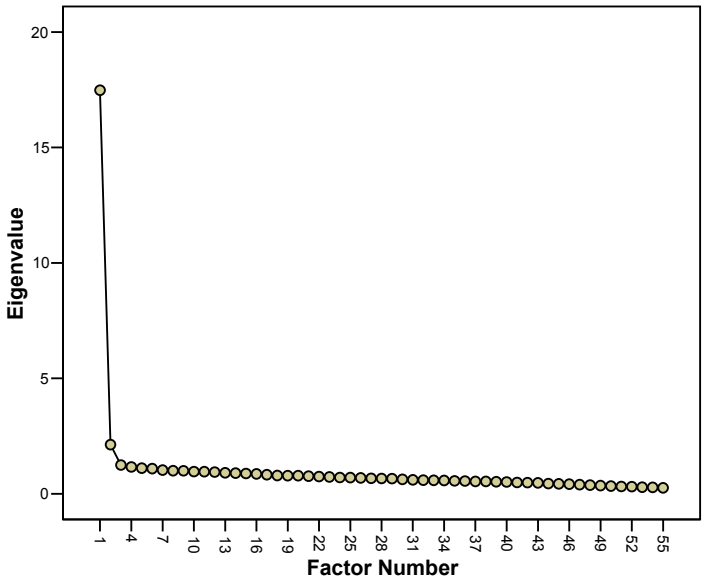
Figure 2.A.1 Algebra Scree Plot



Biology

The Biology factor analysis shows an initial eigenvalue of 17.480 for the first factor, which accounts for 31.783% of the variance. The next seven factors have eigenvalues greater than one. For instance, the second factor's eigenvalue drops to 2.130, accounting for only 3.872% of the variance while the third factor accounts for 2.26% of the variance with an eigenvalue of 1.245. The scree plot below gives a visual of this factor analysis; it appears as if one dominant factor is present.

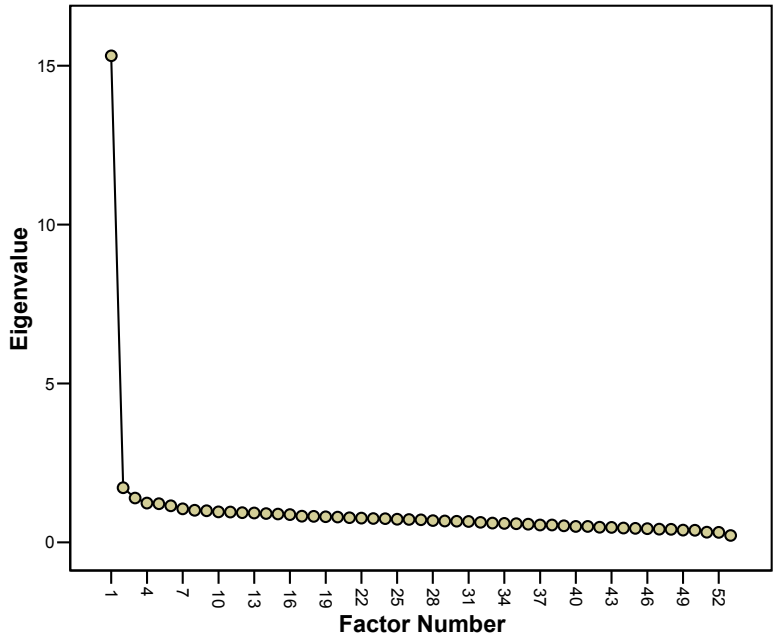
Figure 2.A.2 Biology Scree Plot



English

The English factor analysis shows an initial eigenvalue of 15.312 for the first factor, which accounts for 28.89% of the variance. The next seven factors have eigenvalues just slightly greater than one. For instance, the second factor's eigenvalue drops to 1.718, accounting for only 3.24% of the variance while the third factor's eigenvalue is 1.394, accounting for 2.63% of the variance. The scree plot below gives a visual of this factor analysis; it appears as if one dominant factor is present.

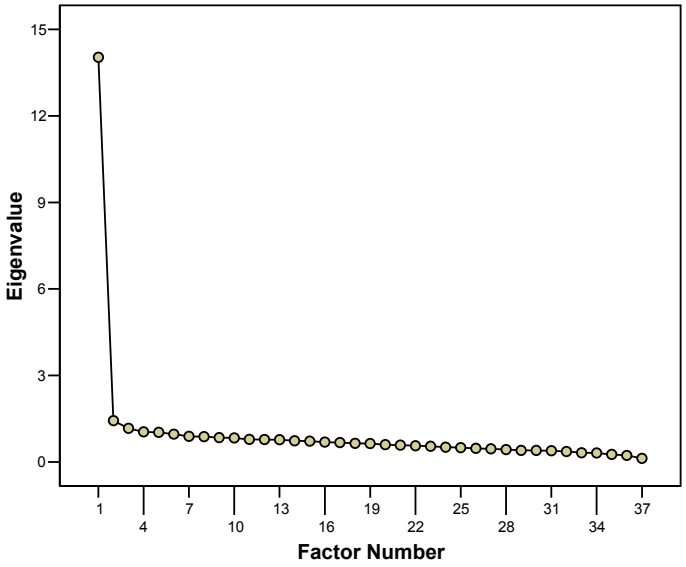
Figure 2.A.3 English I Scree Plot



Geometry

The Geometry factor analysis shows an initial eigenvalue of 14.032 for the first factor, which accounts for 37.924% of the variance. The next four factors have eigenvalues just slightly greater than one. For instance, the second factor's eigenvalue drops to 1.434, accounting for only 3.874% of the variance. The scree plot below gives a visual of this factor analysis; it appears as if one dominant factor is present.

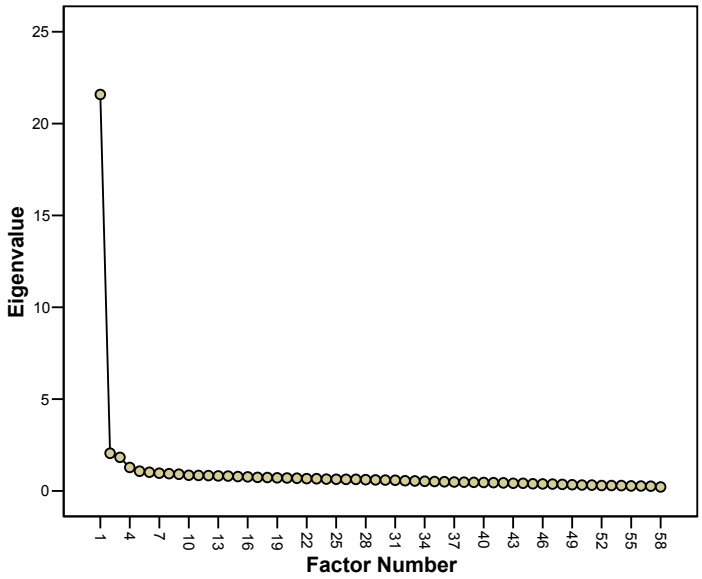
Figure 2.A.4 Geometry Scree Plot



Government

The government factor analysis shows an initial eigenvalue of 21.586 for the first factor, which accounts for 37.217% of the variance. The next five factors have eigenvalues greater than one. For instance, the second factor's eigenvalue drops to 2.053, accounting for only 3.54% of the variance, while the third factor accounts for 3.162% of the variance with an eigenvalue of 1.834. The scree plot below gives a visual of this factor analysis; it appears as if one dominant factor is present.

Figure 2.A.5 Government Scree Plot



Conclusions

All factor analyses indicated one dominant factor underlying the MD HSA data with the first factor accounting for a sizeable percent of the variance, followed by a few other factors accounting for considerably smaller percentage of the variance.

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 and this procedure produced what Yen (1984) called 'number correct trait estimates'. In this report, we call it 'NC scale scores'.

Prior to commencing with the 2004 scoring, MSDE had asked ETS to investigate and replicate the 2003 analyses for the English High School test completed by their previous vendor, CTB/McGraw-Hill. Using independent software, we were able to replicate the results, although small differences were noted in the parameter estimates, transformation constants, and mean scores. However, this is to be expected due to variations associated with inclusion/exclusion criteria for the calibration sample, and differences in the calibration software. Based on the results of this study, we also found no evidence of a systematic error or problem with the calibrations and linking studies completed by CTB/McGraw-Hill. The complete results of the study are presented in Appendix 3.A.

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.

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

where,

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

$I(\theta)$= 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 HSAs. 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.

The SEM curves for each test were presented in Section 1 (see Figures 1.2 for Algebra, Figure 1.4 for Biology, Figure 1.6 for English I, Figure 1.8 for Geometry and Figure 1.10 for Government). As can be observed in these figures, the SEMs vary across the scale. In all cases, extreme values were noted at the ends of the scale, but the SEM is minimized near the cut-scores for each content area, which were near the middle of the scale. This pattern is expected as 1) more items tend to be of middle difficulty; and 2) there were fewer items at the lower and upper ends of the scale. In all cases the SEM is less than 10 scale score points at the cut point.

Subscore Scoring

For the subscore scale scores, the NC to scale score scoring method (later called the NC scoring) was selected based on a special study that compared the two different scoring methods (see Appendix 3.B). At the classroom level, which is where these scores were used, the IP and NC methods produced nearly identical means for all subscores except the one with the fewest score points. This is consistent with other studies that have identified that while IP and NC ability estimates differ for individual examinees (i.e., for examinees with the same number-correct score, their item-pattern ability estimate may be higher or lower, depending on which items they got correct), these two ability estimates were tau-equivalent for groups of 30 or more examinees (Yen, 1984; Yen & Candell, 1991). While the benefit of using IP scoring is the reduced conditional SEMs relative to NC scoring, for the subscore with the fewest score points, IP scores had much higher conditional SEMs than NC scores through the lower part of the score scale. This occurred because a much larger number of scores were assigned the LOSS using IP scoring compared to NC scoring. The difference in results was caused by differential "interpretation" by the IP and NC methods of low scores that did/did not include score points earned on constructed response items. Essentially, IP scoring was not observed to be uniformly beneficial for subscores when there were a small number of score points that included both SR and CR items, and for subscores, the NC scoring method was subsequently recommended by the National Psychometric Committee (NPC).

Lowest and Highest Obtainable Test Scores

Both 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 (see Appendix 3.B; CTB/McGraw-Hill, December 2003). 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. In addition, the associated conditional SEMs were constrained to a maximum value of 80. Table 3.1 lists the LOSS and HOSS scores for each content area established following the first operational administration (CTB/McGraw-Hill, December, 2003).

Table 3.1 LOSS and HOSS Values

Content	LOSS	HOSS
Algebra	200	625
Biology	225	650
English I	200	625
Geometry	225	600
Government	225	650

Cut-Scores

The cut-scores associated with each of the performance levels in each of the content areas were established by MSDE in 2003 (see Table 3.2). One cut-score was established for all of the content areas except for Geometry. Because Geometry is used as the high school mathematics component of the MD accountability plan under NCLB, two cut-scores were established.

Table 3.2 HSA 2004 Cut-Scores

Content Area	Cut-score
Algebra	412
Biology	400
English I	407
Government	394
Geometry	Proficient – 411
	Advanced – 447

Appendix 3.A Review and Replication Analysis English 2003

Maryland High School Assessment

Review and Replication Analysis

English 2003

February 2, 2004

Educational Testing Service

Appendix 3.A Review and Replication Analysis English 2003

MSDE asked ETS to investigate and replicate the 2003 analyses for the English High School test completed by their previous vendor, CTB/McGraw-Hill. An estimated 6% drop in students classified as proficient in 2003 compared to 2002 at the state level prompted this request. The purpose of this study was to 1) review the technical documentation and steps completed by CTB/McGraw-Hill and note any suggested modifications; 2) replicate the study completed by CTB/McGraw-Hill; and 3) determine if a change in linking design would have made any important difference in the percent of students identified as proficient.

Summary of the Process Completed by CTB/McGraw-Hill

Based on the technical documentation, the analyses completed were consistent with high stakes assessment programs and involved item analyses, calibration, and equating. Item-pattern scoring was completed using the resulting item parameters. The completion of the work was within normal standard with the exception of the linking study design and outcome.

January Administration

For the January administration, four forms were administered, forms A, B, C, and W. Forms A, B, and C were built to match the test blueprint and consisted of items administered in 2002, as well as items field tested in 2000 and 2001, along with an embedded field test section. These forms also shared a common anchor set of 36 selected-response items. Form W was an exact duplicate of the 2002 Form W; all items were administered and calibrated in May 2002. This form did not match the test blueprint but did consist of a mix of selected-response (SR), brief constructed-response (BCR) and extended constructed-response (ECR) items (see Table 3.A.1). Including the embedded field test section on Forms A-C, all administered forms had very similar test lengths although Form W had 4 to 5 more SR items than the other forms.

Table 3.A.1. Number and Type of Item by Form

Form	Item Type	SR	BCR	ECR
A	FT	15	1	-
	OP	50	2	1
B	FT	16	1	-
	OP	50	2	1
C	FT	15	1	-
	OP	50	2	1
W	OP	70	3	1

Note. FT= field test. OP= operational.

Appendix 3.A

There were no common items between Form W and Forms A-C (see Table 3.A.2). The operational scale was defined by the 2002 administration, and the equating and linking design for the January 2003 forms was based on a mixed common item, randomly equivalent groups design. An intact form from 2002 (Form W) was spiraled along with the three new forms (A-C). Forms A-C shared a common anchor set, however there were no common items between these forms and Form W. The linking study design was to complete a concurrent calibration of Forms A-C with Form W, then link all of the forms to the 2002 scale through Form W. That is, the new, 2003 parameter estimates for Form W would be linked to the 2002 parameter estimates using a Stocking and Lord procedure. The resulting transformation constants would then be applied to Forms A-C. With the exception of four items that were removed from the calibration and anchor sets due to poor item performance[3], following the 2002 administration, all SR items on Form W were identified as the anchor set to place the 2003 forms onto the 2002 scale. Note, "X" represents a block of items.

Table 3.A.2. Composition of January Forms Relative to Previous Administrations

	2000 – 2001 Field Test Administrations		2002 Administration (0perational Scale)		Embedded Field Test
	Unique Items	Common Items	Unique Items	Common Items	Unique Items
F/T Pool	X	X			
2002 Pool			X	X	
	Items Contributing to Student Scores 2003				
2003 W			X		
2003 A		X			X
2003 B[1]		X			X
2003 C		X			X

[1]Note. Form B also included 2 items from the 2002 administration.

The intended linking study design was dependent on the assumption that the forms would be completed by randomly equivalent groups of students. To obtain randomly equivalent samples, the four forms were packaged and spiraled within each classroom. That is, the first student would receive Form A, the second Form B, the third Form C, the fourth form W, the fifth, Form A, and so on. The exception was the accommodation package – forms administered to students requiring specific accommodations. For these students,

[3] Items were not calibrated following the 2002 administration due to poor classical statistics. These items had very low or negative point-biserial correlations.

only Form A was included in the package; however, it was expected that only a small percentage of students required special accommodations to complete the form.

After the forms were administered and scored, it was determined that the forms were not administered to randomly equivalent groups of students. Based on the Draft 2003 Technical Report (CTB/McGraw-Hill, December, 2003) the forms were not administered to randomly equivalent groups of students because:

1. Large print and Braille forms were available for Form A only, resulting in disproportionate numbers of accommodated students receiving these forms.

2. Special Education students tended to be over represented [sic] on the first couple of forms within each content area. It appears that administrators tended to use the first one or two forms in each package for a disproportionate number of students who required special accommodations.

> Because of the requirement that these students be included in the calibration and equating, it was not possible to sample down in order to achieve comparable groups across test forms (p. 37).

As a result, a modification was made to the intended linking design. The steps completed were summarized below:

1. All forms were calibrated together in a single calibration run, then, using Form W, the forms were equated to the 2002 scale via a Stocking and Lord procedure, and the parameters for all forms adjusted with the resulting equating constants.

   Because Forms A, B, and C shared anchor items, this step placed these 3 forms on the same scale. However, W did not share items with A, B, and C, so this procedure did not place W on the same scale as A, B, and C via anchor items. Random equivalence of samples also did not place W on the same scale as the three other forms,[4] so an additional step was needed.

2. A second linking step was completed. This involved equating Form C to Form W using a linear approximation to equipercentile equating procedure. To complete this step Form W was scored with the 2002 item parameters and Form C was scored with the 2003 item parameters. The resulting equating constants were then applied to the items in forms A, B, and C. The rationale for this step was that the test scores and demographic characteristics of the students completing Form W were very similar to Form C. Due to the

[4] It appears that CTB/McGraw-Hill's parameter estimation software, Pardux, is not designed to automatically align parameters from non-overlapping, randomly equivalent samples. An external procedure, such as the linear equipercentile procedure, is needed. Because this external step is needed, the Stocking and Lord procedure used in Step 1 was not necessary and was over-ridden by Step 2.

disproportionately high representation of ESL and Special Education students, Form A had substantially lower test scores than the other forms.

For scoring purposes, the transformed parameters from step 2 above were used for Forms A, B, and C. Form W was scored with the item parameters estimated in 2002.

Appendix 3.A

May 2003 Administration

For the May administration, 11 forms were administered: Forms D, E, F, G, H, J, K, L, M, N, and P. All forms were similar with regard to the distribution of item type and test length (see Table 3.A.3).

Table 3.A.3. Number and Type of Item by Form

Form	Item Type	SR	BCR	ECR
D	ANC	33	-	-
	OP	17	2	1
	FT	19	1	-
E	ANC	33	-	-
	OP	17	2	1
	FT	16	1	-
F	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
G	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
H	ANC	33	-	-
	OP	17	2	1
	FT	16	1	-
J	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
K	ANC	33	-	-
	OP	17	2	1
	FT	17	1	-
L	ANC	33	-	-
	OP	17	2	1
	FT	15	1	-
M	OP	50	2	1
	FT	17	1	-
N	OP	49	3	1
	FT	18	-	-
P	OP	50	2	1
	FT	16	1	-

Note. FT= field test. OP= operational. ANC= anchor.

Forms D through L were built to match the test blueprint and consisted of items administered in 2002, as well as items field tested in 2000 and 2001, along with an embedded field test section. These forms shared a common anchor set of 36 selected-

response items with Forms A-C. Form M contained 28 SR items that were also administered in one of the forms administered in 2002, as well as newly developed items. Forms N and P were identified as "block field test books" and included only newly developed items. Forms M, N, and P had no items in common with either Form W or Forms A-L. Table 3.A.4 illustrates the composition of the 2003 forms relative to previous administration and new development. Note, "X" represents a block of items.

Table 3.A.4. Composition of 2003 Operational Forms Relative to Previous Administrations

	2000 – 2001 Field Test Administrations		2002 Administration (Operational Scale)		Field Test Items
	Common Items	Unique Items	Common Items	Unique Items	Unique Items
2002			X		
Items Contributing to Student Scores in 2003					
2003 W			X		
2003 A	X	X			X
2003 B[1]	X			X	X
2003 C	X				X
2003 D	X	X		X	X
⋮	⋮			⋮	…
2003 L	X	X		X	X
2003 M				X	X X
2003 N					X X
2003 P					X X

[1]Note. Form B also included 2 items from the 2002 administration.

Linking the May forms to the operational scale involved the following steps:

1. All forms were concurrently calibrated. This produced item parameters for each form approximately[5] relative to a true theta scale with distribution Normal (0,1).

2. Forms D-L were linked to the operational [scale score] scale via the set of common items shared with Forms A-C from the January administration in a

[5] Again it does not appear that Pardux is designed to precisely align parameters from simultaneous calibrations of forms with no over-lapping anchor items. Steps 2 and 3 provided the necessary link across forms.

Stocking and Lord procedure, and the item parameters were adjusted with the resulting equating constants.

3. Forms M, N, and P were placed onto the operational scale by equating each form to Form L using a linear approximation to equipercentile equating.

Item pattern scoring was completed using the resulting transformed item parameters.

The final resulting scale score means and standard deviations for each test form were listed in Table 3.A.5 below (CTB/McGraw-Hill Technical Report, pp. 40-41). The mean scale scores ranged from 390.3 to 399.4. The mean score was lowest for the first form of the spiral in both the January (Form A) and the May (Form D) administrations. In both cases these forms were also administered to the largest number of students within each of the calibration samples. While large print and Braille forms administered in May were the same forms administered in January (Form A), students with other types of accommodations were administered one of the May test forms. Of note, students completing a make-up form were excluded from the calibration samples.

Table 3.A.5. CTB/McGraw-Hill Summary Statistics English 2003

Form	N[1]	Mean	SD
January			
A	2370	390.8	38.1
B	2090	396.4	34.6
C	2019	395.5	34.6
W	1986	395.5	34.4
May			
D	5831	390.3	39.8
E	4797	397.7	34.5
F	4806	398.0	34.8
G	4772	397.1	34.3
H	4775	397.5	35.5
J	4720	397.9	35.5
K	4673	399.4	34.4
L	4600	398.8	36.2
M	4596	398.7	36.7
N	4508	398.7	36.7
P	4483	398.9	35.9

[1]Note. Based on calibration samples.

A summary of the results from the January and May administrations compared to the 2002 results were presented in Table 3.A.6; the 2002 results were taken from the CTB/McGraw-Hill Technical Report (p. 43) and included all students that participated in each administration. The results in Table 3.A.6 include a large number of students taking a make-up form: 933 students completed a make-up form in January and 3,543 students completed a make-up form in January. The make-up forms in both administrations were

the same. Form A was administered in the first make-up week; Form B was administered in the second make-up week. These students generally performed much poorer relative to the calibration sample. For example, students completing Form A in the first make-up week of the January administration had a mean scale score of 353 (sd 61.5).

The mean score for the January 2003 administration was 8.8 points lower. However, there was less than one score point difference between the May 2002 and May 2003 administrations. Unlike in 2002, in 2003 the scores for the January administration were 5.1 points lower than the scores for the May administration. Also noted was the difference in the test score variation in May 2003 compared to all other administrations.

Table 3.A.6. CTB/McGraw-Hill Summary Statistics by Administration and Year

Administration	N	Mean	SD
January 2002	9,339	398.3	41.0
May 2002	52,172	395.4	47.0
January 2003	9,488	389.5	42.2
May 2003	56,426	394.6	39.5

Study Methodology

The main purpose of this study was to replicate the results obtained by CTB/McGraw-Hill and to identify any design revisions that may produce different results. To replicate the results, all analyses steps, as described in the technical documentation supplied by CTB/McGraw-Hill were completed. Items were calibrated using Multilog (Scientific Software International, Inc.). This software allows for the estimation of item parameters for both selected response (SR) and constructed response (CR) items. ETS proprietary software was used to complete the Stocking and Lord and the linear approximation to equipercentile equating procedures.

Results

The percent of students included in the calibration sample overall and by form were presented in Table 3.A.7 for the January administration and Table 3.A.8 for the May administration. Specific information for the calibration sample was not included in the CTB/McGraw-Hill Technical Report. Therefore, the data contained in this column consists of all students, including students completing a Braille form or a make-up form.

As observed by CTB/McGraw-Hill, Form A had the largest case count - 283 more students completed this form compared to Form B. Moreover, the first form in the May administration (Form D) also had the largest case count – 5827 compared to 4799 for Form E. Regardless of the differences in case counts, the forms were spiraled to similar proportions of students for all of the demographic variables except Special Education students. In January 16.7% of the Form A sample were identified as Special Education students, compared to 9% for Forms B and C. In May, the differences were even more

pronounced: 21.8% of the Form D calibration sample were identified as Special Education students, compared to only 8.1 to 9.1% of the sample for the other forms.

Table 3.A.7. Characteristics of Calibration Samples by Form for January 2003

	CTB [1]	Replication				
	Total N=9488	Total N=8436	Form A N=2364	Form B N=2084	Form C N=2014	Form W N=1974
Female	49.7	50.3	48.8	51.3	50.7	50.7
Male	49.7	49.3	50.8	48.1	49.0	48.9
Gender Not Specified	0.4	0.4	0.5	0.5	0.3	0.4
African American	28.8	27.9	28.5	27.9	28.7	26.4
American Indian	0.3	0.3	0.3	0.1	0.4	0.3
Asian	2.5	2.3	2.3	2.4	2.1	2.5
Hispanic	2.2	1. 6	1.9	1.3	1.4	1.7
White	65.0	66.9	66.2	66.9	66.7	68.1
Other Ethnicity	1.1	0.9	0.8	1.3	0.7	0.9
Accommodated	-[2]	1.6	3.1	1.2	1.2	0.7
Eng Lang Learner	-[2]	0.3	0.3	0.2	0.2	0.5
Special Education	-[2]	10.9	16.7	9.0	9.0	7.9

Note. [1]. Data reported in this column is based on all students completing the January administration.
[2]. Information not included in the CTB/McGraw-Hill Technical Report

Table 3.A.8. Characteristics of Calibration Samples by Form for May 2003

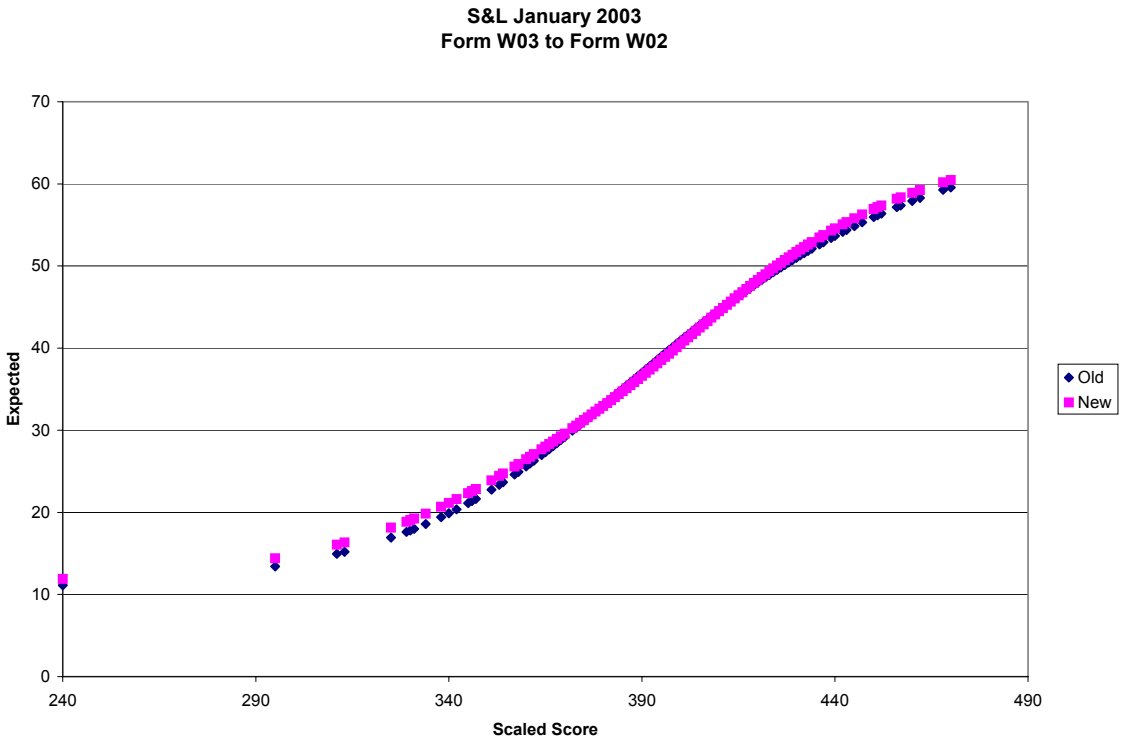
	CTB	Replication											
	Total N=56914	Total N=52549	Form D N=5827	Form E N=4799	Form F N=4807	Form G N=4773	Form H N=4770	Form J N=4712	Form K N=4672	Form L N=4599	Form M N=4600	Form N N=4507	Form P N=4483
Female	49.0	49.6	46.2	50.0	49.1	48.5	49.7	51.0	51.2	50.5	48.9	49.8	51.5
Male	50.1	50.0	53.1	49.6	50.5	51.1	50.0	48.4	48.5	49.2	50.6	49.8	48.0
Gender Unspecified	0.9	0.4	0.7	0.4	0.4	0.4	0.3	0.6	0.3	0.3	0.5	0.4	0.4
African American	35.7	35.0	36.2	35.2	34.8	35.8	35.5	34.6	35.2	34.8	34.5	34.5	34.0
American Indian	0.4	0.3	0.3	0.2	0.3	0.3	0.4	0.4	0.5	0.3	0.3	0.5	0.2
Asian	5.3	5.6	4.6	5.1	5.8	5.7	5.8	6.3	6.0	5.7	5.3	5.2	6.2
Hispanic	4.9	4.9	4.8	5.0	5.0	4.7	5.2	4.9	4.5	4.9	4.9	4.9	5.2
White	52.4	53.5	53.0	53.9	53.4	52.8	52.6	52.9	53.3	53.7	54.5	54.2	53.7
Other Ethnicity	1.2	0.7	1.0	0.7	0.7	0.7	0.6	0.8	0.5	0.6	0.6	0.7	0.6
Accom-modations	-[1]	1.6	3.6	1.1	1.5	1.6	1.3	1.4	1.4	1.2	1.2	1.3	1.5
Eng Lang Learner	-[1]	1.2	1.0	1.3	1.3	1.1	1.3	1.2	1.1	1.3	1.1	1.2	1.2
Special Education	-[1]	10.2	21.8	9.1	8.8	8.6	8.3	8.5	8.5	8.1	8.4	8.9	8.2

Note. [1] Data reported in this column is based on all students completing the January administration.
[2] Information not included in the CTB/McGraw-Hill Technical Report

January Results

As described earlier, a single calibration was completed for the January sample using Multilog. Forms A-C were then linked to the 2002 scale via the Form W item parameters in a Stocking and Lord procedure. As observed in Figure 3.A.1, differences in the test characteristic curves for Form W 2002 (old) and Form W 2003 (new) were noted at the lower end of the scale. This is related to differences between Pardux and Multilog in how the C-parameter is estimated. While many of the 2002 parameters had an estimated value of zero, non-zero estimates were obtained for the 2003 parameters using Multilog.

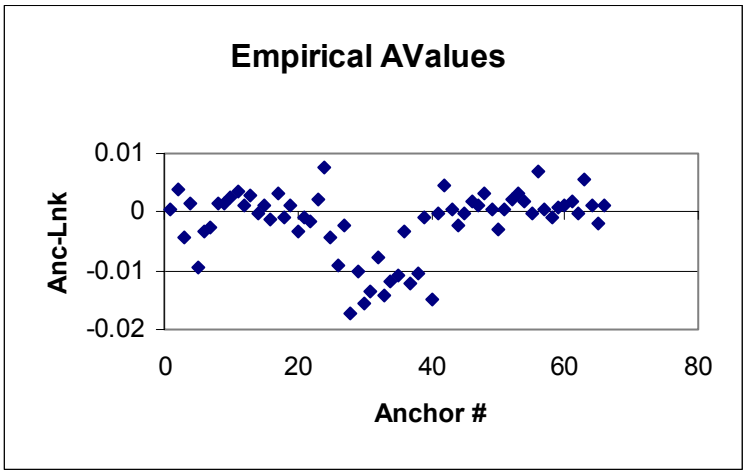
Figure 3.A.1.

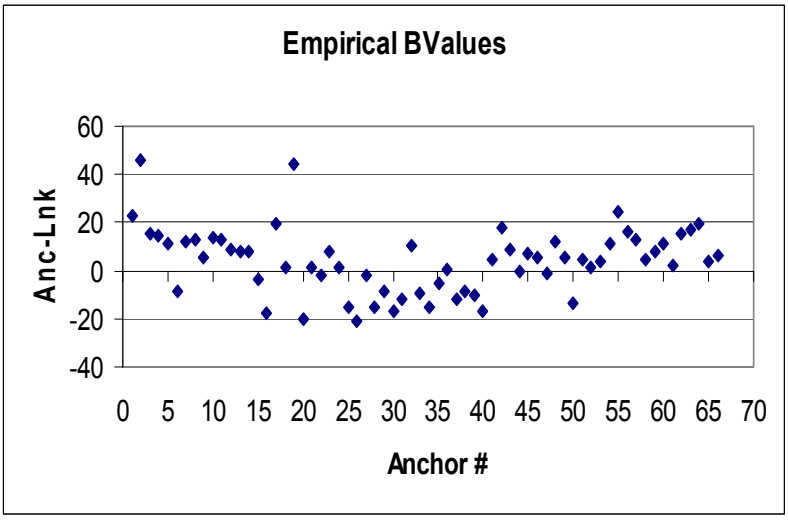


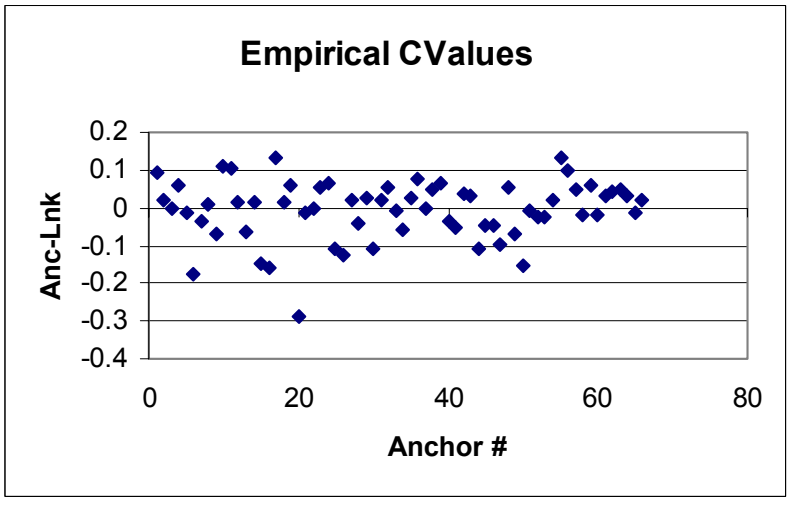
The differences in the 2003 A-, B- and C-parameters from our replication compared to 2002 were plotted in Figures 3.A.2 to 3.A.4.

Appendix 3.A

Figures 3.A.2 to 3.A.4. Differences in Item Parameter Values Compared to 2002.







Following CTB/McGraw-Hill's procedure, the Form W Stocking and Lord equating constants (slope=32.8196; intercept=393.2301) were then applied to all items in Forms A-C. Item-pattern scale scores were produced using the transformed parameters for Forms A-C and the 2002 parameters for Form W. Summary statistics are presented in Table 3.A.9.

Table 3.A.9. Descriptive Statistics January 2003 after Stocking and Lord

		CTB/McGraw-Hill		Replication		
Form	N	Mean	SD	N	Mean	SD
A	2370	380.8	45.8	2364	387.8	39.0
B	2090	387.6	41.7	2084	393.3	35.8
C	2019	386.5	41.6	2014	392.7	35.1
W	1986	395.5	34.4	1974	395.4	34.4

Following this transformation, a linear approximation to equipercentile equating was completed between Form C and Form W. The resulting transformation constants (slope=0.98302; intercept=10.5388) were then applied to Forms A, B, and C. Summary statistics are presented in Table 3.A.10. The additional transformation resulted in mean scores that were within one scale score point of the results reported in the Draft Technical Document (CTB/McGraw-Hill, December, 2003). In all cases, the replicated scores were higher, although the sample sizes were slightly different, which may account for the discrepancies.

Table 3.A.10. Descriptive Statistics January 2003 after Linear Equipercentile

		CTB/McGraw-Hill		Replication		
Form	N	Mean	SD		Mean	SD
A	2370	390.8	38.1	2364	391.8	38.3
B	2090	396.4	34.6	2084	397.2	35.2
C	2019	395.5	34.6	2014	396.5	34.5
W	1986	395.5	34.4	1974	395.4	34.4

After reviewing the design used by CTB/McGraw-Hill and noting that an extra step (Form W Stocking and Lord) had been used (see Footnote 2), we determined that the forms could have been placed onto the operational scale using only the linear approximation to equipercentile equating. As part of this study, we compare the results of the two-step linking to a single-step linking design. Not unexpectedly, the results were very similar (see Table 3.A.11).

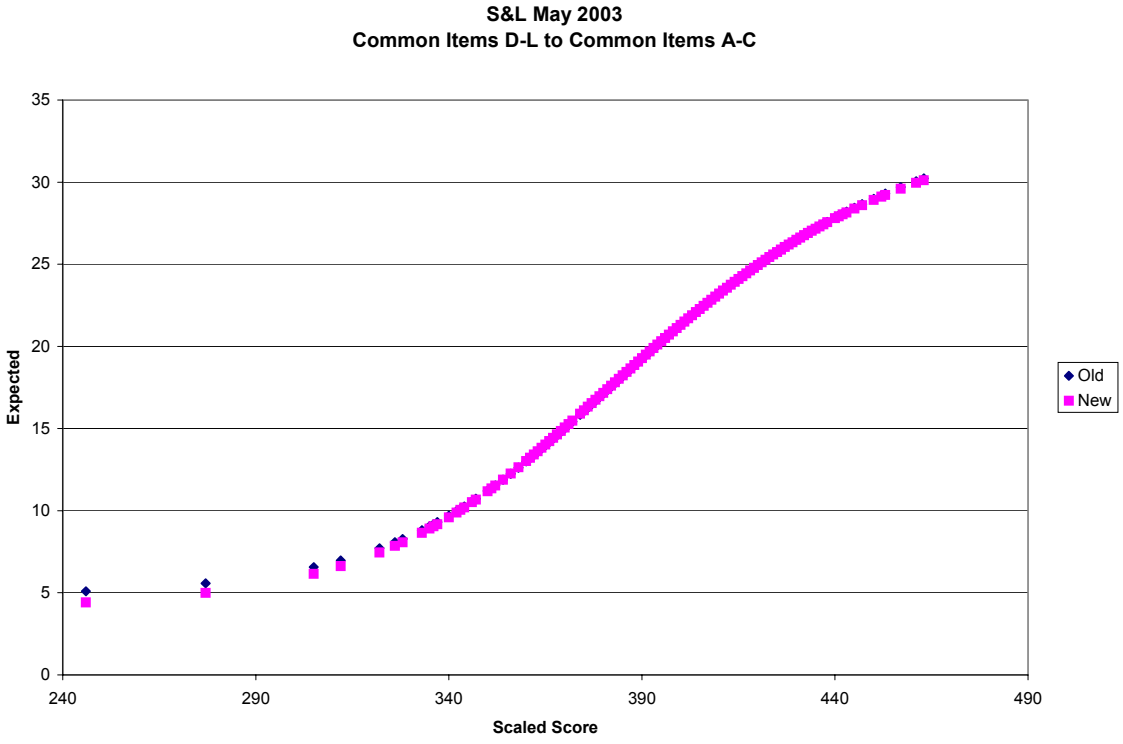
Table 3.A.11. Descriptive Statistics January 2003 Omitting Form W S&L Link

		CTB/McGraw-Hill		Replication Omitting Form W S&L Link		
Form	N	Mean	SD	N	Mean	SD
A	2370	390.8	38.1	2364	391.4	39.2
B	2090	396.4	34.6	2084	396.9	35.9
C	2019	395.5	34.6	2014	396.5	34.4
W	1986	395.5	34.4	1974	395.4	34.4

May Results

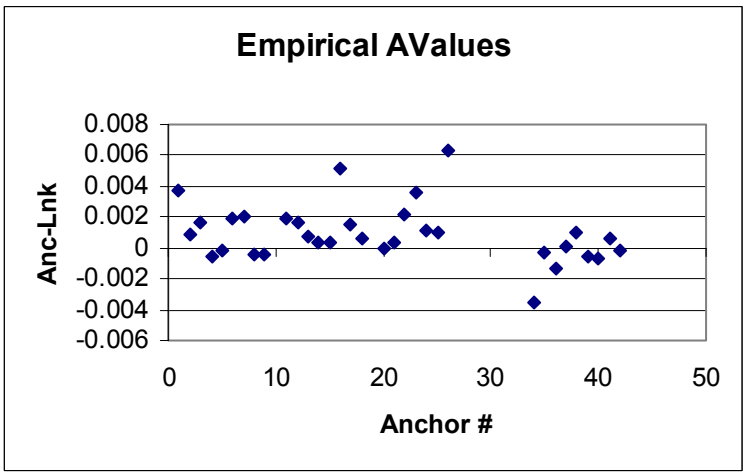
Following CTB/McGraw-Hill's procedure, the May 2003 forms were concurrently calibrated using Multilog. Forms D-L were placed onto the operational scale through the common item set shared between forms A-C (old) and Forms D-L (new) via a Stocking and Lord linking procedure. As observed in Figure 3.A.5, there were almost no differences in the test characteristic curves.

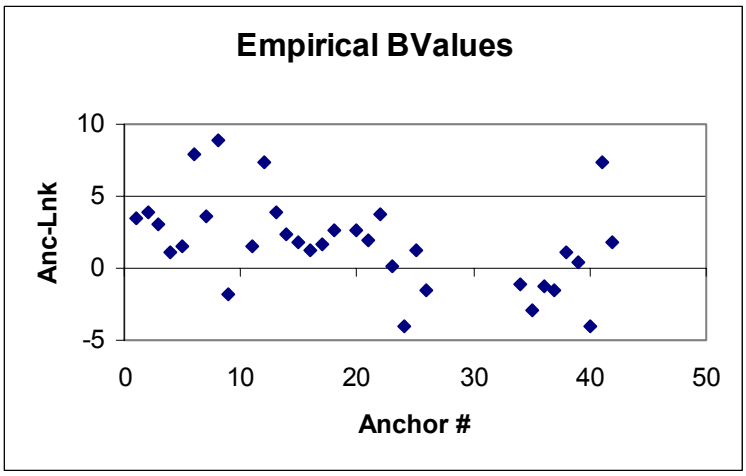
Figure 3.A.5.

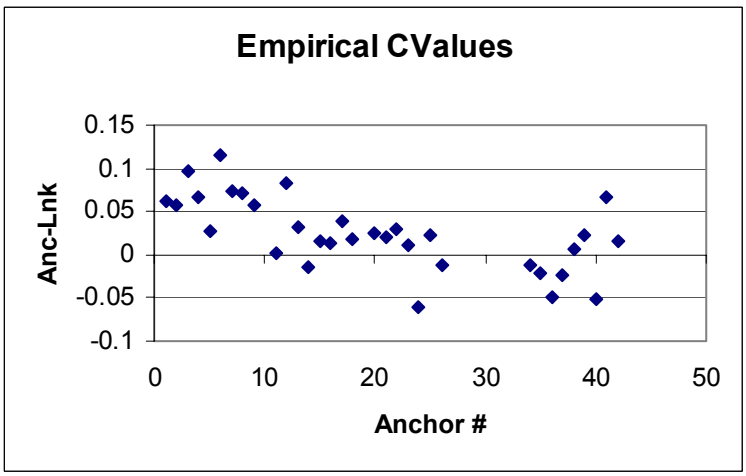


The differences in the January 2003 A-, B- and C-parameters compared to May 2003 are plotted in Figures 3.A.6 to 3.A.8.

Figures 3.A.6 – 3.A.8. Differences in Anchor Item Parameter Values: Forms A-C compared to Forms D-L.







The resulting equating constants (slope=33.13; intercept=399.5) were then applied to items in Forms D-P and item-pattern scale scores produced. Summary statistics are presented in Table 3.A.12. As observed in the January 2003 forms, the resulting means and standard deviations were very similar to the results reported in the Draft Technical Document (CTB/McGraw-Hill, December, 2003). In Forms D-L, the mean scores were approximately one scale score point higher.

Table 3.A.12. Summary Statistics May 2003 After Stocking and Lord

		CTB/McGraw-Hill		Replication		
Form	N	Mean	SD	N	Mean	SD
D	5831	390.3	39.8	5827	391.3	39.4
E	4797	397.7	34.5	4799	398.7	35.0
F	4806	398.0	34.8	4807	399.0	35.3
G	4772	397.1	34.3	4773	398.1	34.6
H	4775	397.5	35.5	4770	398.6	35.7
J	4720	397.9	35.5	4712	399.0	35.7
K	4673	399.4	34.4	4672	400.5	34.3
L	4600	398.8	36.2	4599	400.0	36.4
M	4596	388.1	38.7	4600	397.0	36.9
N	4508	393.0	36.9	4507	390.9	38.0
P	4483	395.3	37.7	4483	392.2	39.0

Because Forms M, N, and P shared no common items with Forms A-L or W, these forms were placed onto the operational scale using a linear approximation to equipercentile equating. Like the January analyses, the Stocking and Lord transformation constants were applied prior to completing the linear equipercentile equating. The descriptive statistics associated with these forms are presented in Table 3.A.13. The mean scores for these forms were very similar to Form L and slightly higher than the results obtained by CTB/McGraw-Hill.

Table 3.A.13. Descriptive Statistics January 2003 after Linear Equipercentile

	CTB/McGraw-Hill			Replication		
Form	N	Mean	SD	N	Mean	SD
M	4596	398.7	36.7	4600	401.3	34.9
N	4508	398.7	36.7	4507	402.6	34.4
P	4483	398.9	35.9	4483	402.8	33.9

Conclusions & Implications

Based on the results of this study, we found no evidence of a systematic error or problem with the calibrations and linking studies completed by CTB/McGraw-Hill. Using independent software, we were able to replicate the results. Small differences were noted in the parameter estimates, transformation constants, and mean scores; however, this is to be expected due to variations associated with inclusion/exclusion criteria for the calibration sample, and differences in the calibration software.

Several observations can be made. First, unless there were strict administration controls, it is very difficult to ensure that forms were be spiraled to randomly equivalent groups. For a variety of reasons, the spiral may have failed (e.g., seating assignments, re-ordering of forms by test administrators, etc.). In this study, the groups were very similar on all demographic variables except students classified as Special Education. A disproportionate number of these students were administered the first form in each administration. While the first of the May forms included relatively more special education students than the first January form, this did not affect the May equating. This is because the May forms were concurrently calibrated and linked using a common anchor set and the equating did not depend on the assumption of randomly equivalent groups[6]. If Forms A-C did not share a common item set, these forms could not have been placed onto the operational scale. Therefore, when randomly equivalent groups cannot be assured, it is prudent to always include common items across forms that can serve as an anchor set.

Second, the Stocking and Lord linking for Form W completed prior to the linear approximation to equipercentile equating was unnecessary. Completing two linking procedures only complicates the design. Essentially, this procedure would have produced similar final results to a single-step procedure. It appeared that the extra step was conducted by CTB/McGraw-Hill for January because it was not until after the Form W Stocking and Lord procedure was implemented that it was seen that this procedure was not sufficient. It is unclear why the two-step procedure was also implemented for the May forms.

Third, with a single cut-score near the middle of a score distribution, a relatively small difference in student scale scores can result in noticeable differences in percents of proficient students. Legitimate equating procedures can produce small variations in scale scores, which can make a noticeable difference in performance classifications.

[6] May forms without common items were linked using linear approximation to equipercentile equating.

Appendix 3.B Evaluating the Use of Item-Pattern and Number-Correct to Scale Score Scoring for Reporting Subscores

Maryland High School Assessment

Evaluating the Use of Item-Pattern and Number-Correct to Scale Score Scoring for Reporting Subscores

February 20, 2004

Educational Testing Service

Appendix 3.B Evaluating the Use of Item-Pattern and Number-Correct to Scale Score Scoring for Reporting Subscores

For the January 2004 administration of the Maryland High School Assessments, subscore scale scores were created using number-correct (NC) score to scale score conversion tables. However, the MSDE and the National Psychometric Council are interested in possibly reporting subscores based on item-pattern (IP) scoring, as will be used for reporting total test scores. While subscores will not be reported at the individual student level, the subscores will be aggregated at the classroom level to provide teachers and administrators with additional information about student performance by each of the reporting categories. To help determine the feasibility of implementing item-pattern scoring at the subscore level, this study investigates the nature and extent of differences in subscores based on item-pattern scoring versus number-correct scoring.

The results included in this report were based on the Algebra A04 form, which was administered this January. The distributions of items by type for each subscore (which were called Expectations) in Algebra A04 are listed in Table 3.B.1 below.

Table 3.B.1. Distribution of Items by Type for each Subscore

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

Because item responses were not yet available for the Algebra A04 form, item responses were simulated based on 5000 simulees with a mean scale score of 398.36, standard deviation 43.18, using the existing "pre-equated"[7] item parameters for this form from the item bank.

7 The items were administered in either 2002 or 2003 – these item parameters were on the operational scale.

Item-pattern scale scores based on these item response vectors were then estimated for each subscore and for the total test. NC scale scores based on NC to scale score conversion tables were also produced for each subscore and the total score (see Appendix 3.B.a). Thus, each item response vector yielded 10 scale scores: a NC scale score and an IP scale score, for each of the four Expectations and the total test.

Results

Individual Scores

The mean scale scores for both the IP and NC scale scores were lower than the mean true scale scores (see Table 3.B.2). Whereas the true score ranged from 254 to 557, both the NC and IP scale scores ranged from 240 to 625; this is due to the assignment of the lowest and highest obtainable score (LOSS; HOSS) for both the NC and IP estimated scores.[8]

Comparing the mean IP and NC scale scores, with the exception of Expectation 3.2, the NC means were very close to the IP means with less than a scale score difference. For Expectation 3.2, the NC scale score was higher by 11.02 scale score points; this result is examined in detail later in this section. The smallest difference between the mean scores was Expectation 1.1 with a difference of only 0.12 scale score points. All of the NC scale score means were slightly higher than the IP scale score means, except for the total scale core. See Appendix 3.B.b for the number, percent, mean, and standard deviation of NC and IP scale scores grouped at intervals of 10 true scale score points for each of the Expectations and the total scale core (i.e., a tabled true score of 405 includes results for all true scale scores from 400 to 409). The standard error associated with selected IP scale scores from each distribution of scores is listed in Appendix 3.B.c

Table 3.B.2. Summary Statistics

Scale Score	Total		Expectation 1.1		Expectation 1.2		Expectation 3.1		Expectation 3.2	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
True	398.36	43.18	-	-	-	-	-	-	-	-
NC	396.77	49.48	396.30	63.47	395.63	59.16	398.36	67.18	391.93	74.70
IP	397.11	48.78	396.18	63.08	395.46	59.50	398.13	67.63	380.91	89.54

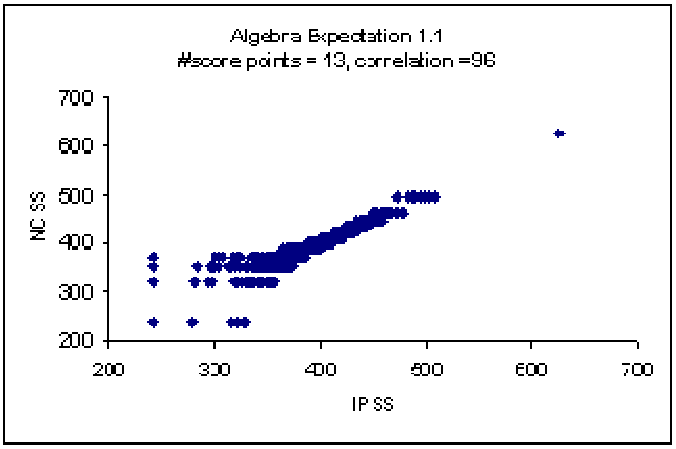
Not unexpectedly, the correlations between the IP and NC scale scores were high, ranging from .91 to .98 for the subscores and .98 for the total scale score. As noted by the bivariate plots (see Figures 3.B.1 –3.B.5) and the difference in standard deviation of NC and IP scale scores given the true scale score (see Figures 3.B.6-3.B.10), the largest differences in scores were noted at the lower end of the scale. This result is expected, given that the consideration

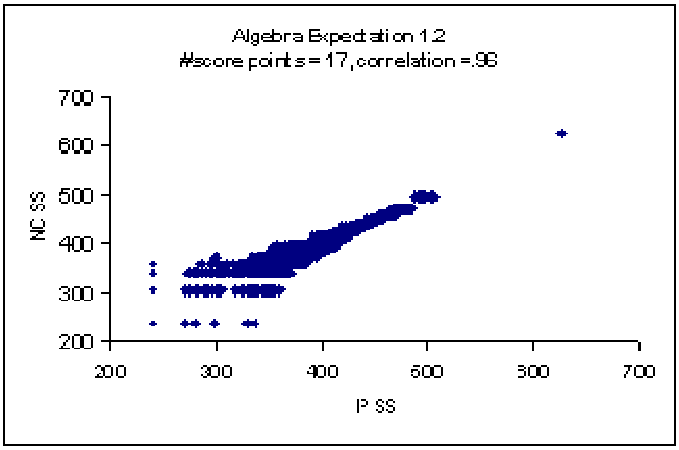
8 The LOSS and HOSS, which were assigned to extreme scores for which IRT does not provided maximum likelihood ability estimates, were set after examining the scale scores produced for the other scores.

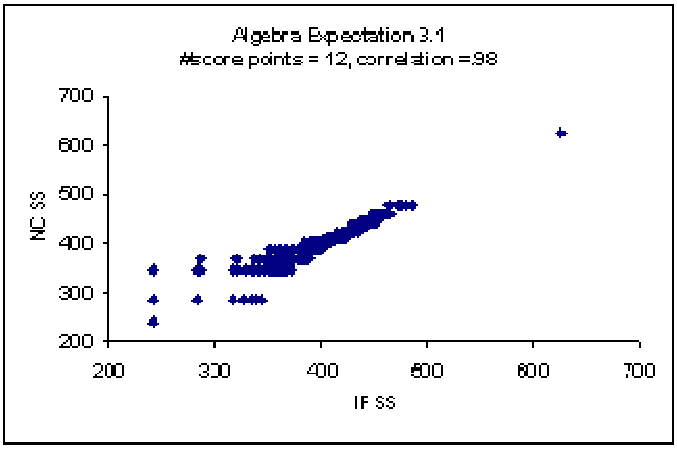
of the C-parameter ("guessing") has a greater effect among low-scoring examinees (Yen, 1984; Yen & Candell, 1991). The variation of scores was also greater at the lower end of scale for the total score, although the amount of variation was smaller than for the subscores. This result is also expected, given that as the number of score points increase, the influence of the uncertainty introduced by guessing decreases.

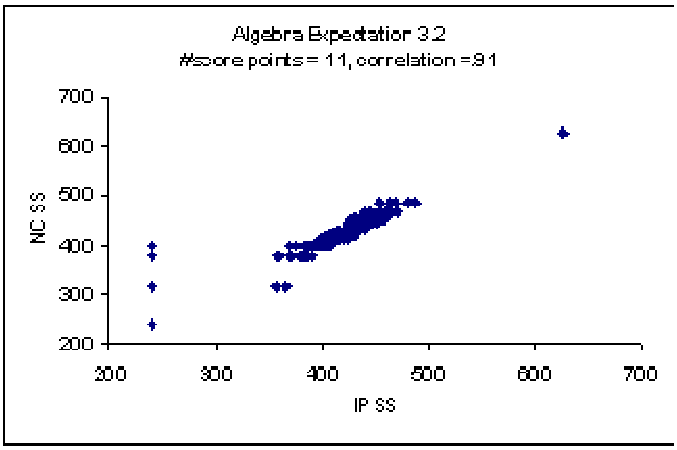
Appendix 3.B

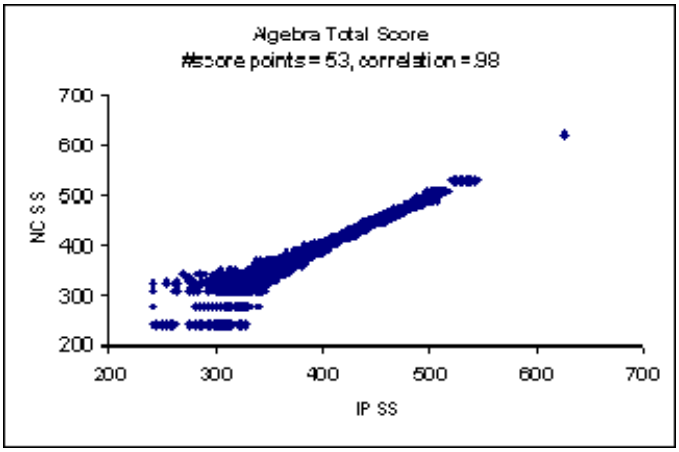
Figures 3.B.1 – 3.B.5 Bivariate Plots of NC and IP Scale Scores









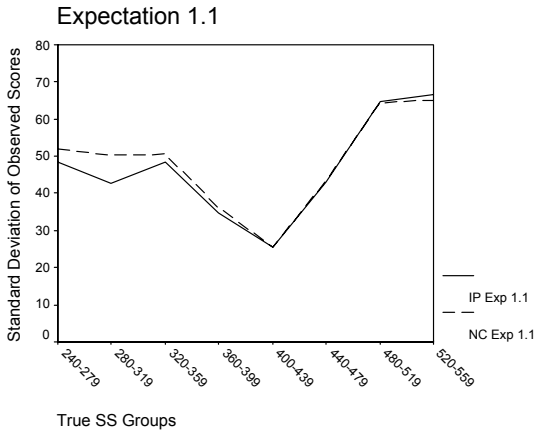


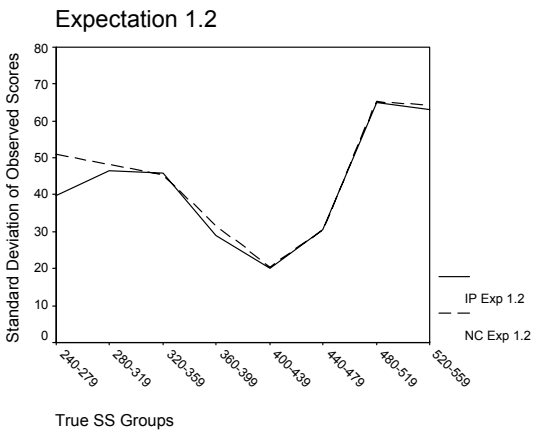
Following IRT principles, IP scale scores should have lower conditional standard errors of measurement than NC scale scores. This result is seen with the exception of Expectation 3.1

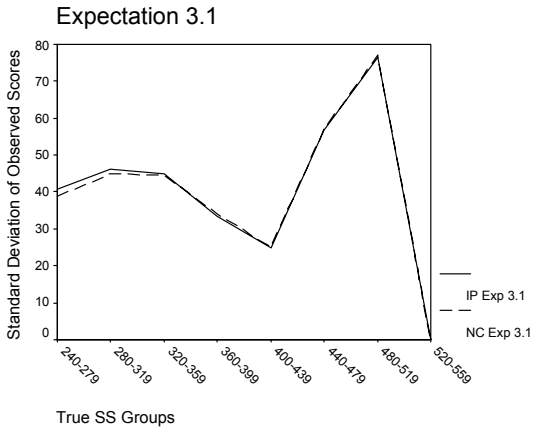
and 3.2 (see Figures 3.B.6-3.B.10). Both of these subscores have the fewest score points: Expectation 3.1 has 12 score points and Expectation 3.2 has only 11 score points. In both cases, the LOSS was assigned to more simulees using IP scoring compared to NC scoring (see Table 3.B.3). As noted in Table 2 of the two subscores, Expectation 3.2 has more variation and a larger difference in the average scale scores for the IP and NC scoring procedures. This is due to the large number of simulees that received the LOSS via IP scoring (n=1127) compared to the number of simulees that received the LOSS via NC scoring (384). In contrast, for Expectation 1.2, 119 simulees received the LOSS via IP scoring and 203 simulees received the LOSS via NC scoring. For this subscore, the NC scores were more variable than the IP scale scores and the difference in average scale scores was smaller.

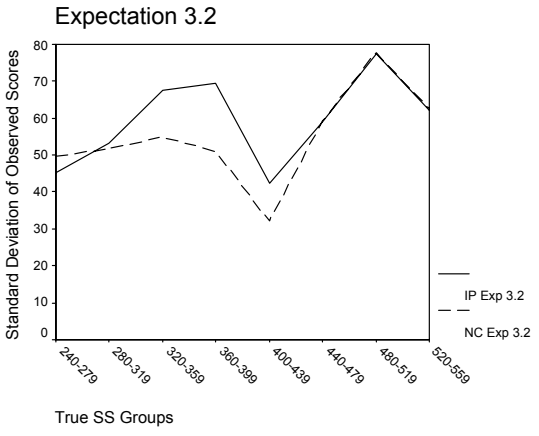
Appendix 3.B

Figures 3.B.6-3.B.10. Empirical Conditional Standard Errors of Scale Scores for Item Pattern (IP) and Number Correct (NC) Scoring Methods









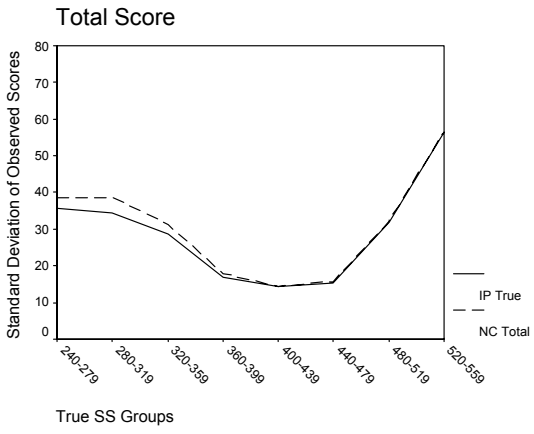


Table 3.B.3. Number and Percent of Simulees Assigned the LOSS by Subscore

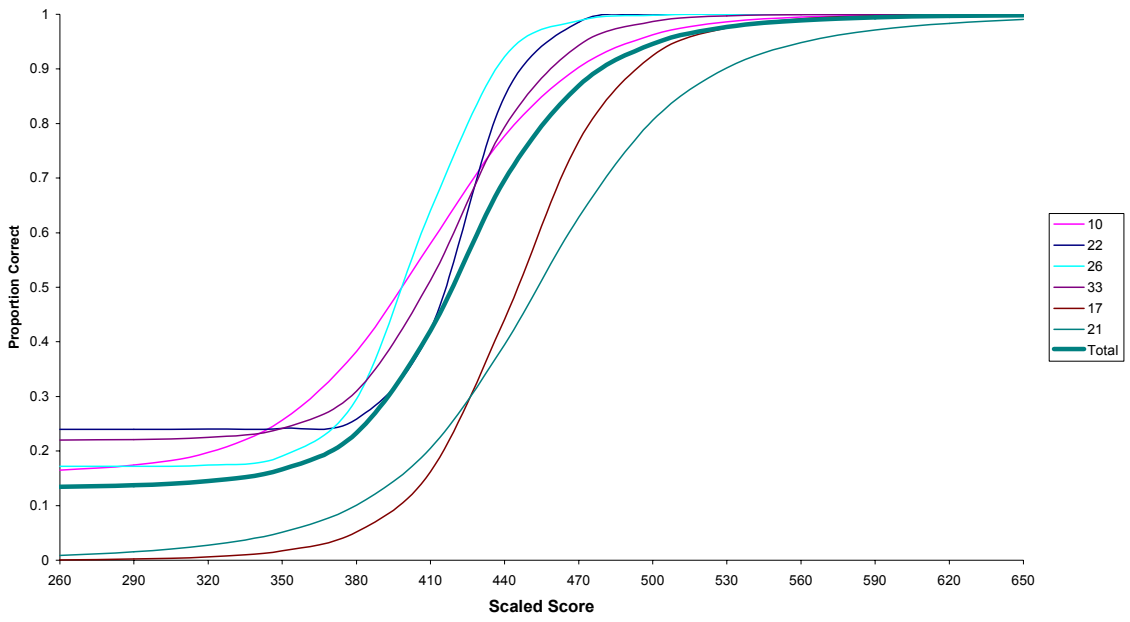
	IP		NC	
	N	%	N	%
Expectation 1.1	217	4.3	279	5.6
Expectation 1.2	119	4.0	203	4.1
Expectation 3.1	187	3.7	114	2.3
Expectation 3.2	1127	22.5	384	7.7
Total Score	66	1.3	80	1.6

Examining the IP and NC scale scores for Expectation 3.2 in more detail, it is noted that the differences in scores is related to characteristics of the items. This subscore included only 6 items: four selected response items (1 point each), one brief constructed response item (3 points) and one extended constructed response item (4 points). The SR items were moderately difficult, with B-values ranging between 406 and 424 (see Table 3.B.4) and have c values ranging from .16 to .24. In contrast, the BCR and ECR items were relatively more difficult, have 0 guessing, and contribute the most information (see Figure 3.B.11).

Table 3.B.4. Expectation 3.2 Item Parameters

		Parameters						
Item	Type	A	C	B	B-1	B-2	B-3	B-4
10		0.0212	0.16	410.04				
22	SR	0.0532	0.24	423.56				
26	SR	0.0393	0.17	406.02				
33	SR	0.0309	0.22	420.02				
17	BCR	0.0196			385.92	450.76	439.10	
21	ECR	0.0145			413.04	477.14	445.79	439.16

Figure 3.B.11. Expectation 3.2 Item Characteristic Curves and Expectation 3.2 Characteristic Curve



The effect of these item parameters on individual scores can be more clearly observed by examining the scores within the true score range of 320 to 359. In this score range, there were 33 possible IP scores compared to 6 possible NC scores (see Table 3.B.5).

Appendix 3.B

Table 3.B.5. Distribution of IP and NC Scale Scores for Expectation 3.2 within the True Score Grouping 320-359

Scale Score	IP		NC	
	N	%	N	N%
240	455	62.33	171	23.42
316			300	41.10
357	36	4.93		
358	11	1.51		
359	13	1.78		
366	43	5.89		
368	16	2.19		
370	10	1.37		
375	5	0.68		
377	6	0.82	180	24.66
379	14	1.92		
382	14	1.92		
383	9	1.23		
386	13	1.78		
387	9	1.23		
388	10	1.37		
390	6	0.82		
391	13	1.78		
394	5	0.68		
395	7	0.96		
398			65	8.90
400	3	0.41		
401	1	0.14		
402	3	0.41		
403	4	0.55		
404	1	0.14		
405	1	0.14		
406	7	0.96		
407	2	0.27		
408	2	0.27		
411			11	1.51
412	2	0.27		
414	2	0.27		
420	3	0.41		
421	2	0.27		
422			3	0.41
423	2	0.27		

Based on IP scoring, the LOSS (240) was assigned for all response patterns where only one, two, or three score points were obtained on the SR items. In contrast, one score point obtained on either the BCR or the ECR resulted in a much higher IP scale score: 366 and 357, respectively (see Table 3.B.6). This result is due to the item pattern scoring process: if a simulee gets 3 or less points from SR items, but 0 from the BCR or ECR items, the item pattern scoring process concludes that these points were likely to have come from guessing, and the IP scale score is at the LOSS. However, when a score point is obtained from a BCR or ECR item, the item pattern scoring process concludes that this score point was obtained via knowledge, not guessing, and the IP scale score is substantially higher than the LOSS.

Table 3.B.6. Expectation 3.2 Item Pattern Response Patterns and Associated IP and NC Scale Scores

			Items					
IP Scale Score	Raw Score	NC Scale Score	10 (SR)	22 (SR)	26 (SR)	33 (SR)	17 (BCR)	21 (ECR)
240	1	316	0	0	0	1	0	0
240	1	316	0	0	1	0	0	0
240	1	316	0	1	0	0	0	0
240	1	316	1	0	0	0	0	0
366	1	316	0	0	0	0	1	0
357	1	316	0	0	0	0	0	1
240	2	377	0	0	1	1	0	0
240	2	377	0	1	1	0	0	0
240	2	377	0	1	0	1	0	0
240	2	377	1	1	0	0	0	0
240	3	398	0	1	1	1	0	0

To shed further light on how IP scale scores were related to the NC scale scores for each subscore and the total score, the IP scale scores were grouped by the corresponding NC scale score and the following statistics were computed (see Tables 3.B.7 to 3.B.11):

1. Number of scores within the grouping (N)
2. Mean IP scale score (Mean)
3. Standard deviation IP scores (SD)
4. Number of IP scale scores within 5 Scale Scores of the NC scale score (N within 5 NC SS)
5. Percent of IP scale scores within 5 Scale Scores of the NC scale score (N within 5 NC SS)
6. Minimum obtained IP scale score (Low)
7. Maximum obtained IP scale score (High)
8. Mean IP scale score standard error (AveSE)

Table 3.B.7. Expectation 1.1

Raw Score	NC Scale Score	Pattern Scores							
		N	Mean	SD	N Within 5 NC SS	% Within 5 NC SS	Low	High	AveSE
0, 1	240	279	267.06	34.65	162	58.06%	240	328	95.57
2	320	388	314.67	34.31	52	13.40%	240	356	52.02
3	352	506	345.98	24.05	160	31.62%	240	372	34.40
4	372	535	367.99	14.42	256	47.85%	240	385	26.42
5	387	548	386.73	5.52	385	70.26%	364	399	21.65
6	400	503	399.62	4.59	375	74.55%	386	411	19.29
7	411	462	410.80	4.16	375	81.17%	401	421	18.00
8	422	405	422.18	4.28	332	81.98%	411	431	17.65
9	433	373	433.19	4.28	300	80.43%	422	444	18.36
10	446	380	445.74	5.16	246	64.74%	433	459	20.49
11	464	311	464.31	6.15	149	47.91%	449	478	25.66
12	494	228	491.48	8.48	165	72.37%	472	508	35.50
13	625	82	625.00	0.00	82	100.00%	625	625	206.86

Table 3.B.8. Expectation 1.2

Raw Score	NC Scale Score	Pattern Scores							
		N	Mean	SD	N Within 5 of NC SS	% Within 5 of NC SS	Low	High	AveSE
0,1	240	203	263.53	33.96	125	61.58%	240	336	162.30
2	306	284	307.40	39.46	7	2.46%	240	360	78.49
3	343	385	336.90	30.14	79	20.52%	240	369	40.86
4	361	448	359.20	17.24	148	33.04%	240	384	24.33
5	375	482	373.05	10.95	212	43.98%	300	394	19.68
6	386	419	385.46	8.88	200	47.73%	351	404	17.80
7	397	427	395.12	7.91	229	53.63%	357	412	16.92
8	406	395	405.99	6.11	226	57.22%	388	418	16.33
9	416	381	414.79	5.63	266	69.82%	392	424	16.09
10	425	337	424.34	4.65	276	81.90%	409	434	15.93
11	434	287	433.56	3.78	250	87.11%	419	441	15.81
12	443	234	442.10	2.98	216	92.31%	432	449	15.85
13	452	197	451.81	2.59	188	95.43%	442	458	16.39
14	462	181	462.15	2.44	177	97.79%	455	469	17.78
15	475	149	475.61	2.66	144	96.64%	467	485	20.80
16	496	127	496.48	4.14	107	84.25%	487	505	27.89

17	625	64	625.00	0.00	64	100.00%	625	625	286.66

Appendix 3.B

Table 3.B.9. Expectation 3.1

Raw Score	NC Scale Score	Pattern Scores							
		N	Mean	SD	N Within 5 of NC SS	% Within 5 of NC SS	Low	High	AveSE
0	240	114	240.00	0.00	114	100.00%	240	240	205.53
1	283	344	297.65	32.30	83	24.13%	240	344	83.36
2	344	606	338.94	24.92	176	29.04%	240	370	39.29
3	369	656	366.77	14.56	226	34.45%	286	387	26.19
4	387	627	385.33	8.86	327	52.15%	351	400	22.00
5	402	519	400.86	5.09	371	71.48%	382	410	19.39
6	413	462	413.66	3.30	418	90.48%	405	423	17.59
7	424	363	423.98	3.73	312	85.95%	416	433	16.64
8	434	362	434.06	3.92	290	80.11%	426	442	16.39
9	445	301	445.20	4.30	228	75.75%	437	453	17.17
10	458	248	457.46	4.92	171	68.95%	447	465	19.47
11	479	225	477.84	5.82	191	84.89%	463	484	26.80
12	625	173	625.00	0.00	173	100.00%	625	625	379.47

Table 3.B.10. Expectation 3.2

Raw Score	NC Scale Score	Pattern Scores							
		N	Mean	SD	N Within 5 of NC SS	% Within 5 of NC SS	Low	High	AveSE
0	240	384	240.00	0.00	384	100.00%	240	240	242.76
1	316	823	279.51	57.42	0	0.00%	240	366	178.03
2	377	832	350.16	55.80	212	25.48%	240	391	74.31
3	398	702	391.69	26.83	361	51.42%	240	407	28.92
4	411	530	412.86	6.27	290	54.72%	398	423	17.42
5	422	430	424.13	6.03	226	52.56%	406	432	15.84
6	432	345	433.90	6.30	106	30.72%	414	441	15.91
7	442	248	443.01	6.67	109	43.95%	424	450	16.83
8	453	197	452.52	7.36	80	40.61%	429	459	18.44
9	466	175	462.22	7.34	134	76.57%	440	470	20.80
10	484	164	481.70	7.76	148	90.24%	454	488	28.90
11	625	170	625.00	0.00	170	100.00%	625	625	328.74

Appendix 3.B

Table 3.B.11. Total Test

Raw Score	NC Scale Score	Pattern Scores							
		N	Mean	SD	N Within 5% NC SS	% Within 5 NC SS	Low	High	AveSE
0-4	240	80	271.51	30.78	35	43.75%	240	326	59.62
5	281	70	298.97	29.61	2	2.86%	240	339	38.29
6	309	99	303.64	35.00	14	14.14%	240	345	37.58
7	324	136	324.01	19.18	36	26.47%	240	349	23.55
8	335	135	334.87	15.60	36	26.67%	273	356	19.63
9	344	145	340.93	15.55	56	38.62%	269	363	18.08
10	351	159	350.57	10.80	65	40.88%	304	368	15.64
11	358	164	356.62	7.40	98	59.76%	336	370	14.44
12	363	184	362.86	6.36	123	66.85%	337	380	13.46
13	368	179	366.70	6.87	121	67.60%	337	380	12.94
14	373	183	372.50	5.50	139	75.96%	338	383	12.17
15	377	187	377.14	4.86	143	76.47%	362	388	11.62
16	382	143	380.78	4.47	115	80.42%	364	390	11.21
17	385	158	384.96	4.15	134	84.81%	372	393	10.77
18	389	158	388.32	3.32	139	87.97%	380	397	10.43
19	392	151	392.21	3.60	130	86.09%	381	400	10.07
20	396	129	395.45	3.32	120	93.02%	381	403	9.79
21	399	142	398.15	3.04	132	92.96%	389	405	9.56
22	402	125	401.66	2.87	118	94.40%	394	409	9.29
23	405	111	404.69	3.03	105	94.59%	396	412	9.08
24	408	112	408.00	2.41	112	100.00%	403	413	8.86
25	410	122	410.50	2.29	121	99.18%	405	416	8.71
26	413	111	413.10	2.30	111	100.00%	408	418	8.57
27	416	111	415.43	2.43	107	96.40%	410	422	8.46
28	418	107	418.37	2.71	103	96.26%	412	425	8.34
29	421	120	421.12	2.51	118	98.33%	414	426	8.26
30	423	108	423.51	2.45	106	98.15%	418	430	8.21
31	426	84	425.73	2.72	82	97.62%	420	432	8.18
32	428	97	428.47	2.56	93	95.88%	421	434	8.17
33	431	84	430.68	2.67	79	94.05%	424	435	8.18
34	433	83	433.66	2.30	82	98.80%	428	439	8.23
35	436	80	436.29	2.35	78	97.50%	429	441	8.29
36	439	84	438.63	2.25	83	98.81%	432	444	8.37
37	441	78	441.13	2.54	77	98.72%	435	446	8.48
38	444	65	444.03	2.49	64	98.46%	437	448	8.62

Raw Score	NC Scale Score	Pattern Scores							
		N	Mean	SD	N Within 5% NC SS	% Within 5 NC SS	Low	High	AveSE
39	447	64	447.19	2.56	61	95.31%	438	453	8.82
40	449	72	449.29	2.71	70	97.22%	440	454	8.97
41	452	63	452.29	2.88	60	95.24%	444	458	9.22
42	456	63	454.92	2.61	60	95.24%	449	461	9.46
43	459	69	459.29	2.43	67	97.10%	453	465	9.92
44	462	62	462.23	2.81	61	98.39%	456	467	10.28
45	466	48	466.31	2.49	47	97.92%	461	472	10.84
46	471	39	470.31	2.18	37	94.87%	465	475	11.45
47	476	44	474.84	2.57	42	95.45%	467	481	12.23
48	481	45	481.09	3.32	40	88.89%	475	488	13.46
49	488	45	488.20	2.64	45	100.00%	483	493	15.03
50	497	35	497.69	3.73	30	85.71%	489	507	17.51
51	510	37	508.41	4.18	29	78.38%	498	516	20.86
52	532	18	530.61	6.33	12	66.67%	521	544	30.23
53	625	12	625.00	0.00	12	100.00%	625	625	139.03

Note that regression effects affect these results: because simulees were grouped on the basis of an observed score (NC scale score), the dependent observed score (IP scale score) tends to be less extreme. Near the top and bottom of the scale, the means and standard deviations were also affected by the LOSS and HOSS.

Based on these tables, the mean IP scale score was similar to the NC scale score for the majority of the score groupings. As was observed in the true score groupings, the largest differences were noted at the lower end of the scale where the most variation of IP scale scores is also observed. In addition, the majority of the IP scale scores were within 5 scale score points of the NC scale score.

Aggregate Scores

As the primary purpose of the reported subscores will be to provide reports at the classroom level, aggregate scores were also simulated. To create these simulated results, 100 "classrooms" were simulated by randomly selecting 30 scores for each "classroom". These results are summarized in Table 3.B.12. The pattern of results is similar to the scores aggregated across the total sample (see Table 3.B.2). As with the total sample, the differences between the two types of scores were relatively small (less than one score point), with the exception of Expectation 3.2, where the NC scale scores were, on average, 10.24 points higher than the mean IP scale scores (see Table 3.B.13). The differences in IP and NC scale scores for each subscore are also observable in the bivariate plot (Figure 3.B.12).

Appendix 3.B

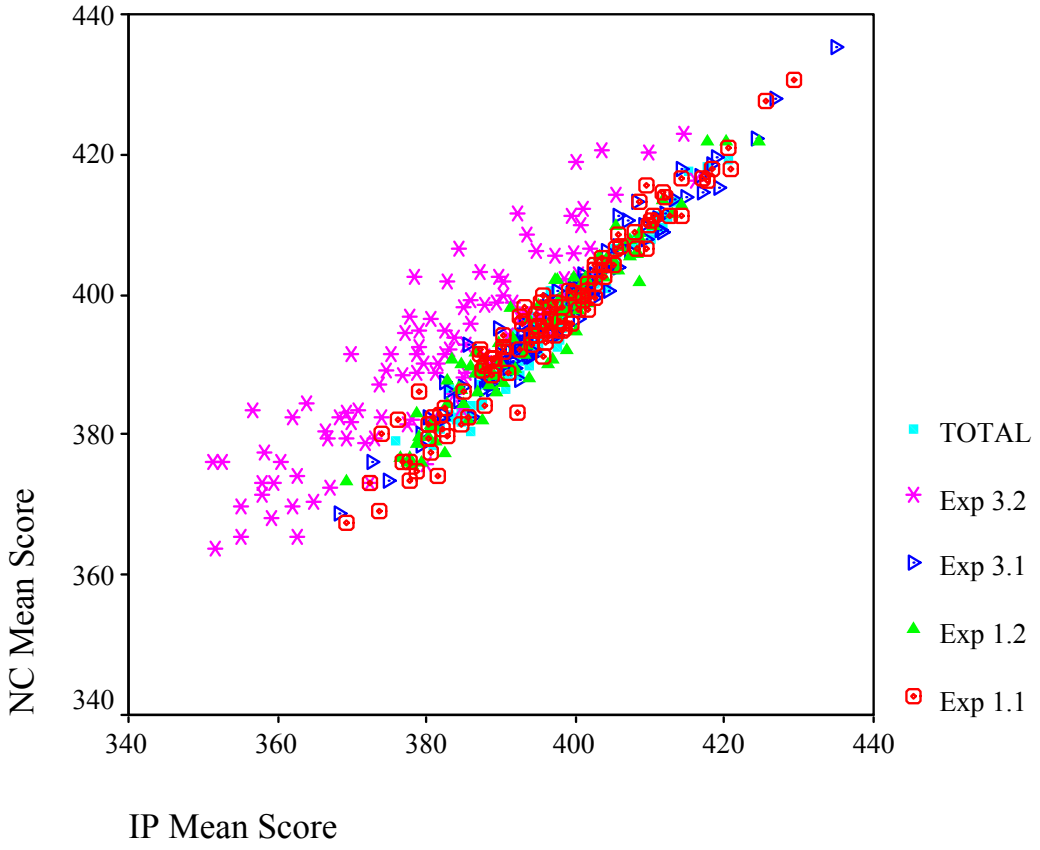
Table 3.B.12. Simulation of Aggregate Scores (n=30, 100 replications)

		Score Points	Mean	SD	Minimum	Maximum
Expectation 1.1	NC	13	396.23	13.01	367.57	430.47
	IP		396.16	12.61	369.40	429.27
Expectation 1.2	NC	17	394.67	10.77	373.40	422.10
	IP		394.63	10.85	369.10	424.70
Expectation 3.1	NC	12	397.96	11.64	368.83	435.17
	IP		398.00	11.73	368.27	435.03
Expectation 3.2	NC	11	391.37	13.54	363.70	422.87
	IP		381.13	14.94	351.47	415.97
Total Score	NC	53	396.04	9.26	378.60	419.37
	IP		396.78	8.92	375.90	420.57

Table 3.B.13. Differences between Mean IP and NC Scores (IP – NC)

	Mean	SD	Minimum	Maximum
Expectation 1.1	-0.07	2.88	-7.30	9.00
Expectation 1.2	-0.03	2.92	-7.54	6.64
Expectation 3.1	0.05	2.22	-7.20	4.60
Expectation 3.2	-10.24	6.53	-26.80	4.07
TOTAL	0.73	1.77	-3.94	5.30

Figure 3.B.12. Bivariate Plots IP and NC Mean Scores (n=30, 100 replications)



Summary and Conclusions

Based on the results of this study, the mean IP scale score was similar to the NC scale score for the total sample of the total score and all of the subscores except Expectation 3.2. For Expectation 3.2 the mean NC scale score was 11.02 scale score points higher than the mean IP scale score. For the samples of 30 scores, the mean IP and NC scores were similar across 100 replications except Expectation 3.2. In this case, the NC scale score was 10.24 points higher than the IP scale score.

The point of doing IP scoring is to benefit from a reduced conditional standard error of measurement relative to NC scoring. However, for the subscore with the fewest score points, Expectation 3.2, IP scale scores had much higher conditional SEMs than NC scores through the lower part of the score scale. This occurred because a much larger number of scores were assigned the LOSS using IP scoring compared to NC scoring. The difference in results was caused by differential "interpretation" by the IP and NC scoring methods of low scores that did/did not include score points earned on constructed response items. This study cannot determine the relative validity or meaningfulness of the scores produced by the IP and NC scoring methods, but only note that they can produce very different results when there are a small number of score points that include both SR and CR items.

It can also be noted that at the classroom level, which is where these scores are to be used, the IP and NC scoring methods produced nearly identical means—except for Expectation 3.2. Consistent IP and NC results at the group level reflect their tau-equivalence, which has been found in many other tests (Yen, 1984; Yen & Candell, 1991). In essence, the theoretical improvement in conditional SEM can be very useful for individual examinees, but is of no apparent value for groups of 30 or more students. The possibility exists that for small numbers of items with a mixed format, IP scoring will produce higher conditional SEMs and very different mean scores than NC scoring. Thus, IP scoring does not appear uniformly beneficial for subscores with small numbers of items with mixed formats.

This study demonstrates that conclusions about "areas of need" can be affected by the type of scoring used when there are small numbers of items with mixed formats contributing to a subscore. While Total scale scores are quite stable across IP and NC scoring, Expectation scores based on small numbers of items can be significantly affected by scoring procedure. For example, based on Table 3.B.2 results, the conclusion would be drawn that Expectation 3.2 is a serious area of need when IP scoring is used, but only a modest area of need when NC scoring is used. If IP scoring is used for subscores, then additional explanatory information will be needed so that scores are interpreted appropriately.

References

Yen, W. M. (1984). Obtaining maximum likelihood trait estimates from number-correct scores for the three-parameter logistic model. *Journal of Educational Measurement, 21*, 93-111.

Yen, W. M., & Candell, G. L. (1991). Increasing score reliability with item-pattern scoring: An empirical study in five score metrics. *Applied Measurement in Education, 4,* 209-228.

Appendix 3.B

Appendix 3.B.a

Number-Correct to Scale Score Scoring Tables

Expectation 1.1		
NC	Scale Score	SEM
0	240	80
1	240	80
2	320	42
3	352	30
4	372	25
5	387	22
6	400	19
7	411	18
8	422	18
9	433	18
10	446	20
11	464	25
12	494	36
13	625	80

Expectation 1.2		
NC	Scale Score	SEM
0	240	80
1	240	80
2	306	55
3	343	28
4	361	22
5	375	19
6	386	18
7	397	17
8	406	16
9	416	16
10	425	16
11	434	16
12	443	16
13	452	16
14	462	18
15	475	21
16	496	28
17	625	80

Expectation 3.1		
NC	Scale Score	SEM
0	240	80
1	283	80
2	344	32
3	369	25
4	387	22
5	402	19
6	413	18
7	424	17
8	434	16
9	445	17
10	458	19
11	479	27
12	625	80

Expectation 3.2		
NC	Scale Score	SEM
0	240	80
1	316	80
2	377	32
3	398	22
4	411	18
5	422	16
6	432	16
7	442	17
8	453	18
9	466	22
10	484	30
11	625	80

Appendix 3.B

Total Score		
NC	Scale Score	SEM
0	240	80
1	240	80
2	240	80
3	240	80
4	240	80
5	281	45
6	309	28
7	324	22
8	335	19
9	344	17
10	351	15
11	358	14
12	363	13
13	368	13
14	373	12
15	377	12
16	382	11
17	385	11
18	389	10
19	392	10
20	396	10
21	399	9
22	402	9
23	405	9
24	408	9
25	410	9
26	413	9
27	416	8
28	418	8
29	421	8
30	423	8
31	426	8
32	428	8
33	431	8
34	433	8

Total Score		
NC	Scale Score	SEM
36	439	8
37	441	8
38	444	9
39	447	9
40	449	9
41	452	9
42	456	10
43	459	10
44	462	10
45	466	11
46	471	12
47	476	12
48	481	13
49	488	15
50	497	17
51	510	21
52	532	31
53	625	80

Appendix 3.B.b
Grouped Frequency Distribution

The following tables list the number, percent, mean and standard deviation of NC and IP scores grouped at intervals of 10 true scale score points.

Expectation 1.1

True Scale Score (midpoint)	N	%	NC		IP	
			Mean	SD	Mean	SD
255	4	0.08	288.00	56.94	240.00	0.00
265	3	0.06	277.33	64.66	284.33	47.82
275	10	0.20	307.60	49.90	303.00	48.95
285	13	0.26	273.23	44.52	284.69	45.66
295	34	0.68	288.12	44.79	276.00	40.95
305	45	0.90	301.96	54.29	293.04	44.75
315	61	1.22	282.80	50.00	283.85	41.24
325	100	1.98	303.70	51.21	299.25	47.66
335	135	2.70	321.84	51.14	319.74	49.74
345	203	4.06	330.13	49.40	329.96	47.87
355	292	5.84	342.38	46.88	343.79	42.43
365	363	7.26	352.78	41.97	354.56	36.37
375	424	8.48	368.89	35.87	367.92	36.90
385	453	9.06	378.63	30.04	379.21	29.52
395	429	8.58	391.90	25.26	392.11	24.95
405	442	8.84	403.05	21.60	402.70	22.12
415	401	8.02	415.05	21.73	415.20	21.71
425	401	8.02	426.97	21.85	427.19	21.75
435	336	6.72	436.74	24.37	436.45	23.48
445	265	5.30	450.86	29.87	451.12	29.34
455	194	3.88	464.02	43.37	464.29	42.74
465	143	2.86	475.57	42.64	475.57	42.69
475	86	1.72	491.14	60.32	490.93	60.02
485	74	1.48	501.72	54.24	502.42	54.32
495	43	0.86	522.26	70.53	521.23	71.24
505	19	0.38	529.05	68.00	530.63	66.98
515	13	0.26	562.23	71.01	561.85	71.68
525	5	0.10	546.40	71.75	545.40	72.78
535	4	0.08	592.25	65.50	591.00	68.00
545	4	0.08	592.25	65.50	591.00	68.00
555	1	0.02	625.00	0.0	625.00	0.0

Appendix 3.B

Expectation 1.2

True Scale Score (midpoint)	N	%	NC		IP	
			Mean	SD	Mean	SD
255	4	0.08	282.25	51.07	254.75	29.50
265	3	0.06	262.00	38.11	254.00	24.25
275	10	0.20	299.80	55.26	290.30	42.42
285	13	0.26	302.69	47.82	289.54	46.76
295	34	0.68	291.21	49.04	287.00	46.38
305	45	0.90	290.00	46.70	281.56	44.72
315	61	1.22	304.66	48.65	300.07	47.15
325	100	1.98	305.82	46.47	291.78	45.08
335	135	2.70	324.55	46.33	319.58	44.31
345	203	4.06	327.33	47.25	328.53	46.10
355	292	5.84	344.12	38.31	343.72	38.68
365	363	7.26	354.39	36.09	354.70	33.45
375	424	8.48	366.85	31.22	368.57	28.03
385	453	9.06	380.82	24.17	381.71	19.72
395	429	8.58	391.70	21.86	392.56	20.13
405	442	8.84	403.80	18.00	403.66	17.05
415	401	8.02	414.43	17.32	415.01	16.37
425	401	8.02	423.28	17.03	423.66	16.36
435	336	6.72	432.96	17.67	433.46	17.14
445	265	5.30	445.16	17.31	445.28	17.37
455	194	3.88	455.75	16.92	455.57	16.73
465	143	2.86	466.07	25.77	466.62	25.85
475	86	1.72	496.83	51.56	497.57	51.33
485	74	1.48	498.57	49.93	499.07	49.79
495	43	0.86	529.42	68.22	530.26	67.83
505	19	0.38	576.37	65.58	576.37	65.56
515	13	0.26	565.46	66.93	565.38	67.03
525	5	0.10	547.60	70.66	547.60	70.70
535	4	0.08	592.75	64.50	595.00	60.00
545	4	0.08	592.75	64.50	593.00	64.00
555	1	0.02	625.00	0.0	625.00	0.0

Appendix 3.B

Expectation 3.1

True Scale Score (midpoint)	N	%	NC		IP	
			Mean	SD	Mean	SD
255	4	0.08	313.50	35.22	298.50	39.03
265	3	0.06	303.33	35.22	297.33	50.46
275	10	0.20	307.40	44.57	291.60	43.45
285	13	0.26	282.46	44.80	261.31	37.10
295	34	0.68	296.09	42.31	285.79	44.02
305	45	0.90	292.58	49.26	286.84	46.67
315	61	1.22	299.30	43.46	300.25	46.23
325	100	1.98	308.20	47.05	305.27	46.37
335	135	2.70	323.70	42.63	321.33	43.27
345	203	4.06	328.80	45.34	325.30	47.31
355	292	5.84	344.58	39.55	344.77	37.57
365	363	7.26	354.13	36.72	355.59	34.24
375	424	8.48	365.77	35.00	366.67	34.53
385	453	9.06	378.43	28.64	378.16	29.01
395	429	8.58	391.95	24.04	392.24	23.99
405	442	8.84	402.10	22.08	402.27	22.22
415	401	8.02	413.48	20.23	413.51	20.22
425	401	8.02	424.22	18.96	424.09	18.68
435	336	6.72	435.86	25.84	436.13	25.38
445	265	5.30	448.58	31.98	448.61	31.73
455	194	3.88	469.22	48.32	469.24	48.09
465	143	2.86	492.04	65.87	492.40	65.56
475	86	1.72	520.55	75.97	520.67	75.85
485	74	1.48	545.07	78.85	545.88	77.97
495	43	0.86	555.63	75.45	555.47	75.69
505	19	0.38	569.00	75.56	569.84	74.65
515	13	0.26	567.23	76.25	567.92	75.33
525	5	0.10	625.00	0.00	625.00	0.00
535	4	0.08	625.00	0.00	625.00	0.00
545	4	0.08	625.00	0.00	625.00	0.00
555	1	0.02	625.00	0.0	625.00	0.0

Appendix 3.B

Expectation 3.2

True Scale Score	N	%	NC		IP	
			Mean	SD	Mean	SD
255	4	0.08	297.00	38.00	240.00	0.00
265	3	0.06	240.00	0.00	240.00	0.00
275	10	0.20	315.10	49.43	267.30	57.63
285	13	0.26	290.31	52.85	240.00	0.00
295	34	0.68	308.00	50.86	277.50	59.21
305	45	0.90	307.60	49.80	274.80	58.60
315	61	1.22	297.54	53.89	263.11	49.90
325	100	1.98	314.91	55.07	270.98	57.34
335	135	2.70	315.44	55.85	287.51	67.09
345	203	4.06	324.67	55.64	294.62	68.03
355	292	5.84	326.62	53.59	299.11	69.35
365	363	7.26	341.65	56.07	324.44	71.45
375	424	8.48	349.30	53.05	333.68	70.91
385	453	9.06	367.30	45.28	353.99	65.99
395	429	8.58	382.40	39.78	371.10	60.09
405	442	8.84	395.64	36.13	391.81	48.99
415	401	8.02	410.08	24.85	405.78	41.47
425	401	8.02	422.60	19.12	421.61	26.02
435	336	6.72	436.91	29.59	437.71	32.31
445	265	5.30	453.42	42.18	453.65	42.84
455	194	3.88	470.89	55.81	471.80	54.75
465	143	2.86	485.34	65.01	486.52	64.14
475	86	1.72	515.35	74.15	515.59	73.85
485	74	1.48	534.70	78.19	534.82	77.90
495	43	0.86	539.49	78.13	539.70	77.76
505	19	0.38	564.68	72.77	564.00	73.61
515	13	0.26	601.92	56.45	602.54	54.95
525	5	0.10	596.80	63.06	597.60	61.27
535	4	0.08	550.00	86.91	550.00	86.70
545	4	0.08	625.00	0.00	625.00	0.00
555	1	0.02	625.00	0.0	625.00	0.0

Appendix 3.B

Total Score

True Scale Score	N	%	NC		IP	
			Mean	SD	Mean	SD
255	4	0.08	284.75	32.62	240.00	0.00
265	3	0.06	240.00	0.00	275.67	31.56
275	10	0.20	314.20	28.46	301.30	28.85
285	13	0.26	279.15	41.19	284.15	33.75
295	34	0.68	288.53	36.84	290.71	34.44
305	45	0.90	296.00	39.66	291.69	34.07
315	61	1.22	294.79	38.51	299.48	35.12
325	100	1.98	310.48	36.91	306.20	35.13
335	135	2.70	330.35	32.49	329.18	30.91
345	203	4.06	337.35	28.33	340.85	19.98
355	292	5.84	349.97	22.80	352.24	17.80
365	363	7.26	361.12	16.52	361.97	14.00
375	424	8.48	372.21	13.90	372.70	13.29
385	453	9.06	382.72	12.31	382.85	11.42
395	429	8.58	393.85	11.15	393.89	10.82
405	442	8.84	404.14	10.15	404.12	9.73
415	401	8.02	414.60	9.36	414.87	8.94
425	401	8.02	424.19	8.65	424.26	8.50
435	336	6.72	433.84	9.34	434.13	8.94
445	265	5.30	445.14	9.93	445.17	9.40
455	194	3.88	455.11	10.35	455.20	9.77
465	143	2.86	463.84	12.28	464.12	11.44
475	86	1.72	478.17	15.32	478.28	15.21
485	74	1.48	486.54	17.28	486.82	16.72
495	43	0.86	495.63	25.99	495.28	25.50
505	19	0.38	517.16	33.26	517.53	33.04
515	13	0.26	542.69	58.19	542.46	58.16
525	5	0.10	537.40	49.89	537.20	50.66
535	4	0.08	546.50	54.87	548.00	52.62
545	4	0.08	596.25	57.50	595.75	58.50
555	1	0.02	625.00	0.0	625.00	0.0

Appendix 3.B.c

Pattern Scoring Standard Error of Measurement for Selected IP Scores

Expectation 1.1	
IP	IP SEM
240	126
280	69
290	59
300	54
320	42
330	38
340	34
350	31
360	28
370	26
380	23
390	21
400	19
410	18
420	18
430	18
440	19
450	21
460	24
470	27
480	31
489	34
493	36
499	38
503	40
625	207

Expectation 1.2	
IP	IP SEM
240	220
279	97
281	93
291	75
300	62
310	50
320	41
330	34
340	29
350	25
360	22
370	20
380	18
390	17
400	17
410	16
420	16
430	16
440	16
450	16
460	17
470	19
480	22
492	26
498	28
504	31
625	287

Expectation 3.1	
IP	IP SEM
240	206
283	87
286	82
320	44
340	33
351	29
360	27
370	25
380	23
390	21
400	20
410	18
420	17
430	16
440	17
450	18
460	20
471	24
484	30
625	379

Expectation 3.2	
IP	IP SEM
240	243
357	46
366	39
370	37
379	31
382	29
390	25
400	21
410	18
420	16
430	15
440	16
450	18
461	20
470	23
480	28
488	32
625	329

Total	
IP	IP SEM
240	90
251	75
260	64
269	55
273	51
280	45
290	38
300	32
310	27
320	23
330	20
340	18
350	15
360	14
370	12
380	11
390	10
400	9
410	9
420	8
430	8
440	8
450	9
460	10
470	11
480	13
490	15
500	18
510	21
521	26
530	30
540	35
625	139

Appendix 3.C Establishing the HOSS and LOSS

Maryland High School Assessment

March 17, 2004

Educational Testing Service

Appendix 3.C Establishing the HOSS and LOSS

Principles for determining the HOSS and LOSS in May 2002 were described in email correspondence from Diana Marr, Research Scientist at CTB (March 16, 2004). The text of the email is printed below.

> To determine the optimal HOSS and LOSS for each test form, we adopted the following principles (put forth by Wendy Yen in a 1991 memorandum):
>
> For HOSSes,
> 1. The HOSS must be greater than SS(n-1).
> 2. The HOSS must be high enough that it does not cause an unnecessary pileup of scale scores at the top of the scale.
> 3. The HOSS should be low enough that SE(HOSS) < 10*Min(SE).
> 4. The SE(HOSS) should change smoothly over levels. [The HOSS gaps should also change smoothly over levels, insofar as possible, but this is less important than maintaining smooth SE(HOSS) changes.]
> 5. The HOSS should be such that Number Correct SS and Item Pattern SS are tau equivalent.
> 6. The HOSS gap should be in the same ballpark as the penultimate HOSS gap.
>
> For LOSSes,
> 1. The LOSS should be low enough that it does not cause an unnecessary pileup of IP scale scores at the bottom of the scale.
> 2. The LOSS should be high enough that SE(LOSS)<15*MIN(SE); this criterion may be difficult to meet for some tests.
> 3. In general, the LOSS should be < SS(Sum c+1); however, if SS(Sum c+1) is poorly determined, causing violation of criterion b, then (Sum c+2) may be treated as the lowest determined scale score.
> 4. The SE(LOSS) should change smoothly over levels. [The LOSS gaps should also change smoothly over levels, insofar as possible, but this is less important than maintaining smooth SE(LOSS) changes.]
> 5. The LOSS should be such that Number Correct SS and Item Pattern SS are tau equivalent.
> 6. The LOSS gap should be in the same ballpark as the penultimate LOSS gap.
>
> After using these criteria to estimate the "optimal" HOSS and LOSS for each individual test form, results were then compared across all of the test forms within each content area to arrive at a single HOSS and LOSS for each content area. Because the 2002 test forms had been developed as field test forms, there was less consistency across forms than would be expected in a group of operational test forms. Thus, the optimal HOSS and LOSS varied considerably from form to form, and the selection of a single pair of values for each content area necessarily involved some compromises. For each content area, the final HOSS typically fell

somewhere between the lowest and highest individual test form HOSS, and the final LOSS typically fell somewhere between the lowest and highest individual test form LOSS.

References

Fitzpatrick, A. R., Link, V. B., Yen, W. M., Burket, G. R., Ito, K., & Sykes, R. C. (1996). Scaling performance assessments: A comparison of one-parameter and two-parameter partial credit models. *Journal of Educational Measurement, 33(3)*, 291-314.

Section 4. Test-Level Analyses

This chapter summarizes the test-level statistics obtained for the January, May and July 2004 administration of the HSAs. The test-level analyses included demographic distributions, scale score distributions, and reliability analyses.

Demographic Distributions

All eligible students completed the HSAs, 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 4.5 for the January and May administrations. The numbers of students completing the summer administrations ranged from 75 for Biology to 500 for Geometry. Algebra had 234 students, English I had 143 and Government had 95 students participate in the July administration. Due to the small sample sizes, results for this section were only included in overall results. The numbers of students participating in the May administration was greater than 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		4617	100	396	100	59398	100	2377	100
Gender									
	Male	2413	52.3	205	51.8	28749	48.4	1157	48.7
	Female	2204	47.7	191	48.2	30639	51.6	1219	51.3
	Missing	0	0	0	0	10	0	1	0
Special Education									
	Yes	459	9.9	65	16.4	4530	7.6	233	9.8
	No	4088	88.5	321	81.1	54236	91.3	2116	89
	504	72	1.6	10	2.5	632	1.1	28	1.2
Ethnicity									
	American Indian	14	0.3	2	0.5	227	0.4	11	0.5
	Asian/Pacific Islander	115	2.5	13	3.3	3461	5.8	75	3.2
	African American	1416	30.7	167	42.2	19970	33.7	969	40.9
	White	2950	63.9	158	39.9	32329	54.5	1190	50.2
	Hispanic	122	2.6	55	13.9	3332	5.6	125	5.3
	Missing	0	0	1	0.3	79	0.1	7	0.3
Limited English Proficient									
	Yes	68	1.5	30	7.6	1426	2.4	46	1.9
	No	4537	98.3	363	91.7	57449	96.7	2317	97.5
	Exited	12	0.3	3	0.8	523	0.9	14	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		7770	100	442	100	46550	100	1933	100
Gender									
	Male	3931	50.6	258	58.4	22433	48.2	938	48.6
	Female	3839	49.4	183	41.4	24106	51.8	992	51.4
	Missing	0	0	1	0.2	11	0	3	0.2
Special Education									
	Yes	799	10.3	97	21.9	3685	7.9	221	11.4
	No	6856	88.2	338	76.5	42404	91.1	1685	87.2
	504	115	1.5	7	1.6	461	1	27	1.4
Ethnicity									
	American Indian	21	0.3	0	0	163	0.4		
	Asian/Pacific Islander	161	2.1	20	4.5	2913	6.3	62	3.2
	African American	2206	28.4	213	48.2	15913	34.2	866	44.9
	White	5222	67.2	174	39.4	24925	53.6	874	45.3
	Hispanic	160	2.1	35	7.9	2609	5.6	115	6
	Missing	0	0	0	0	27	0.1	5	0.3
Limited English Proficient									
	Yes	73	0.9	12	2.7	1079	2.3	36	1.9
	No	7681	98.9	425	96.2	45039	96.8	1885	97.5
	Exited	1.6	0.2	5	1.1	432	0.9	12	0.6

Table 4.3. Demographic Information for English

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-up Forms	
		N	%	N	%	N	%	N	%
Overall		7193	100	392	100	55016	100	2271	100
Gender									
	Male	3704	51.5	253	64.5	27067	49.2	1092	48.2
	Female	3489	48.5	139	35.5	27939	50.8	1175	51.8
	Missing	0	0	4	0.2	10	0	4	0.2
Special Education									
	Yes	837	11.6	101	25.8	5376	9.8	311	13.7
	No	6255	87	283	72.2	49058	89.2	1935	85.2
	504	101	1.4	8	2	582	1.1	25	1.1
Ethnicity									
	American Indian	15	0.2	0	0	258	0.5	16	0.7
	Asian/Pacific Islander	179	2.5	18	4.6	3001	5.5	46	2
	African American	1625	22.6	192	49	19726	35.9	1171	51.7
	White	5202	72.3	147	37.5	28895	52.6	910	40.2
	Hispanic	172	2.4	33	8.4	3087	5.6	121	5.3
	Missing	0	0	2	0.5	49	0.1	7	0.3
Limited English Proficient									
	Yes	64	0.9	11	2.8	1072	1.9	48	2.1
	No	7119	99	376	95.9	53417	97.1	2206	97.1
	Exited	10	0.1	5	1.3	527	1	17	0.7

Table 4.4. Demographic Information for Geometry

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-up Forms	
		N	%	N	%	N	%	N	%
Overall		7113	100	588	100	45285	100	2512	100
Gender									
	Male	3423	48.1	317	53.9	21567	47.6	1245	49.6
	Female	3690	59.9	271	46.1	23713	52.4	1266	50.4
	Missing	0	0	0	0	5	0	1	0
Special Education									
	Yes	581	8.2	60	10.2	2655	5.9	205	8.2
	No	6417	90.2	520	88.4	42188	93.2	2272	90.4
	504	115	1.6	8	1.4	442	1	35	1.4
Ethnicity									
	American Indian	24	0.3	1	0.2	204	0.5	12	0.5
	Asian/Pacific Islander	190	2.7	127	21.6	2982	6.6	81	3.2
	African American	1745	24.5	200	34	14014	31	1032	41.1
	White	4985	70.1	189	32.1	25635	56.6	1234	49.2
	Hispanic	175	2.5	71	12.1	2426	5.4	149	5.9
	Missing	0	0	0	0	24	0.1	4	0
Limited English Proficient									
	Yes	45	0.1	36	6.1	918	2	33	1.3
	No	7036	98.9	521	88.6	43903	96.9	2463	98
	Exited	11	0.6	31	5.3	464	1	16	0.6

Table 4.5. Demographic Information for Government

		January Primary Forms		January Make-Up Forms		May Primary Forms		May Make-up Forms	
		N	%	N	%	N	%	N	%
Overall		8119	100	396	100	50408	100	2745	100
Gender									
	Male	4074	50.2	210	52.9	24708	49	1274	46.4
	Female	4045	49.8	186	46.9	25684	51	1470	53.6
	Missing	0	0	1	0.3	16	0	1	0
Special Education									
	Yes	787	9.7	61	15.4	4415	8.8	304	11.1
	No	7232	89.1	329	82.9	45460	90.2	2412	87.9
	504	100	1.2	7	1.8	533	1.1	29	1.1
Ethnicity									
	American Indian	14	0.3	1	0.3	298	0.6	19	0.7
	Asian/Pacific Islander	181	2.2	17	4.3	3098	6.2	80	2.9
	African American	2169	26.7	170	42.8	17549	34.9	1276	46.6
	White	5543	68.3	158	39.8	26574	52.8	1203	43.9
	Hispanic	202	2.5	50	12.6	2831	5.6	163	5.9
	Missing	0	0	1	0.3	58	0.1	4	0.1
Limited English Proficient									
	Yes	68	1.5	16	4	1114	2.2	44	1.6
	No	8063	99.3	375	94.5	48803	96.8	2679	97.6
	Exited	12	0.3	6	1.5	491	1	22	0.8

Score Distributions and Summary Statistics

Overall, comparisons of the combined mean scores for each administration are presented in Table 4.6. Scores for the May administration were higher than either the January or July administrations.

Summary statistics for all students and for subgroups based on grade, gender, ethnicity, language fluency, economic disadvantage and special education programs are presented in Tables 4.8 through 4.17. These tables include number of students tested for whom valid scores were available, mean scale scores, standard deviation of scale scores, as well as percentages of students in various proficiency levels. In all content areas, the mean scores were higher for the primary and make-up forms administered in May compared to the forms administered January. In addition, higher mean scores were noted for the primary week forms from both administrations compared to the make-up forms.

Table 4.6 Mean Scores by Administration

	Jan-04			May-04			July-04		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Algebra	5499	400.9	48.6	67595	412.8	49.5	234	400.84	32.05
Biology	8629	398.8	43.9	52116	407.5	42.5	75	391.23	34.61
English	8084	392.5	44.0	60768	407.2	39.6	143	377.38	31.49
Geometry	8375	401.7	42.1	53320	405.8	37.8	500	395.55	29.49
Government	9155	397.5	44.5	56626	408.0	42.0	95	392.69	39.63

The following figures graphically represent the distribution of scale scores for each of the content areas (see Figures 4.1 to 4.5). The data from the January and May administrations were overlaid to facilitate comparisons across two administrations.

Figure 4.1

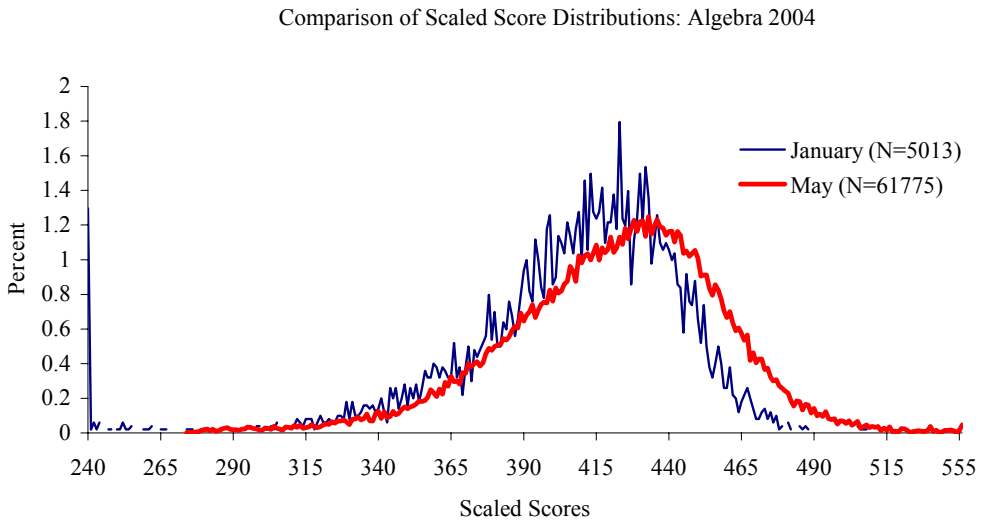


Figure 4.2

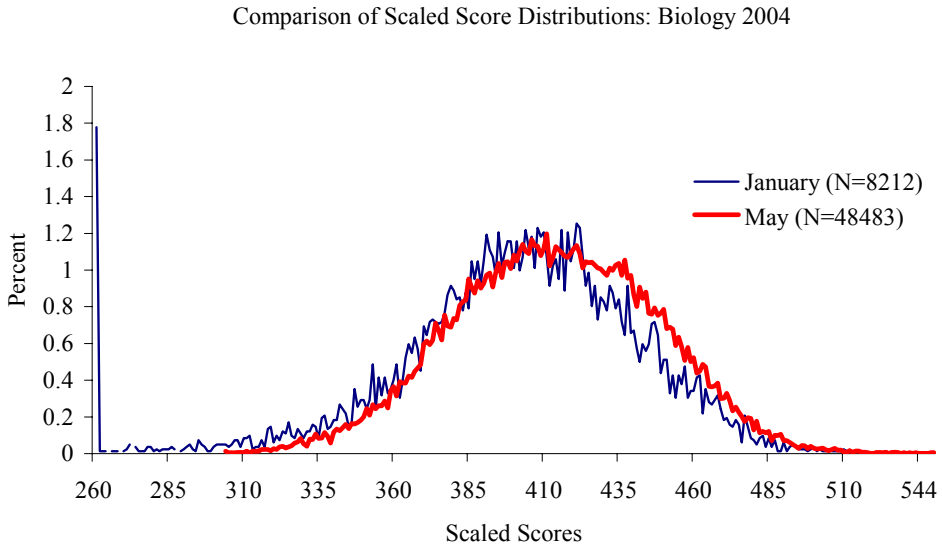


Figure 4.3

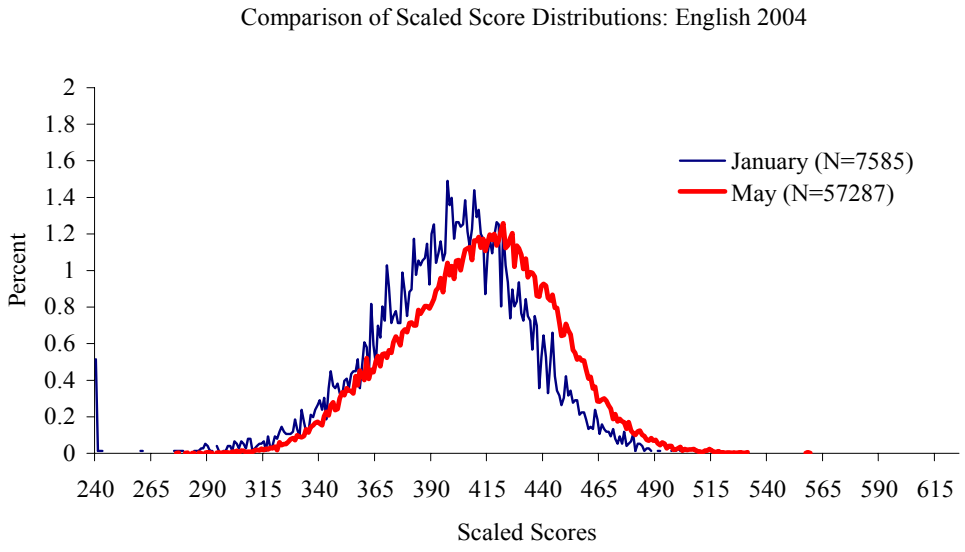


Figure 4.4

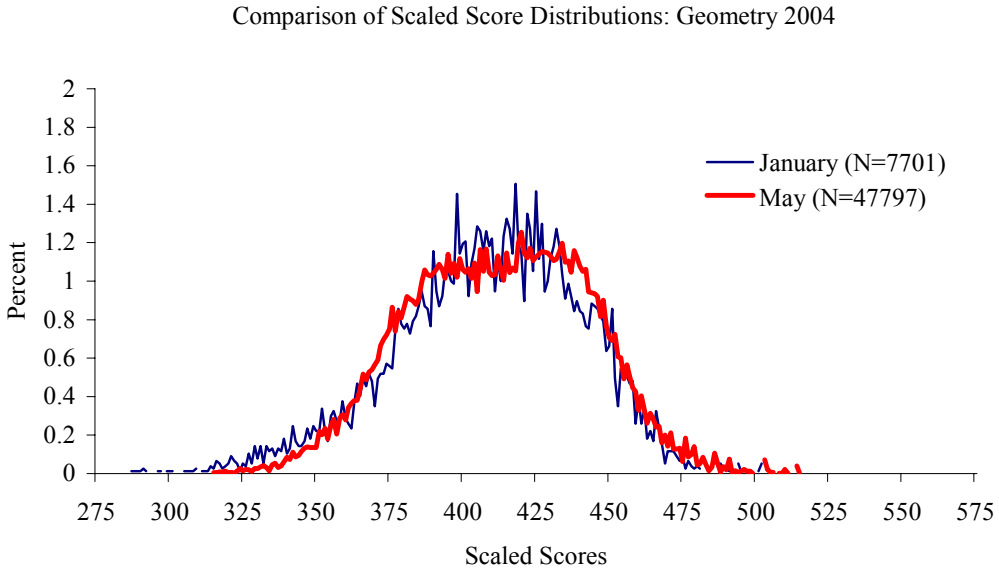
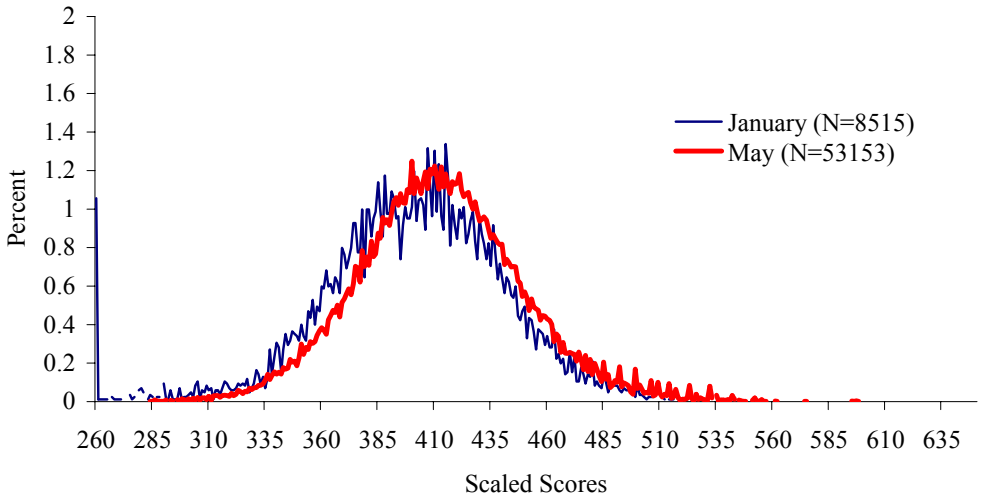


Figure 4.5

Comparison of Scaled Score Distributions: Government 2004



Combined across all three administrations, the mean scores were higher compared to the mean scores from 2002 and 2003, as reported in the 2003 Technical Report (CTB/McGraw-Hill, 2003; see Table 4.7).

Table 4.7 Comparisons of Mean Scores from 2002, 2003, and 2004

	2002	2003	2004
Algebra	405.1	408.3	411.9
Biology	399.3	400.8	406.2
English	395.8	393.9	405.4
Geometry	398.3	398.8	405.2
Government	397.8	403.5	406.5

Passing rates were also higher in 2004 compared to the previous two years (see Table 4.8). The most notable increase was in English, – 11% and 14.8% more students were classified as passing in 2004 in comparison with 2002 and 2003, respectively. Smaller increases were noted in 2004 compared to 2003 for Algebra, Biology and Government (6.1%, 7.7% and 7.0%, respectively) and 2002 (7.2%, 7.5%, and 9.9%, respectively. In Geometry, more students were classified as Proficient and Advanced in 2004 compared to 2003 (2.9% and 1.8%, respectively) and 2002 (1.5% and 1.6%, respectively; see Table 4.9).

Table 4.8 Comparisons of Passing Rates from 2002, 2003, and 2004

	2002	2003	2004
Algebra	52.1	53.2	59.3
Biology	54.5	54.3	62.0
English	43.6	39.8	54.6
Government	57.3	60.2	67.2

Table 4.9 Comparisons of Geometry Passing Rates from 2002, 2003, and 2004

	2002	2003	2004
Basic	55.0	56.6	51.9
Proficient	34.6	33.2	36.1
Advanced	10.4	10.2	12.0

Speededness

The HSAs were untimed tests, therefore students had sufficient time to complete all items. Extensive timing studies have been conducted in previous years and the number and type of items adjusted for each of the content areas. As a verification that the tests were not speeded, the percentage of students who responded to the last items in each of the test sections can help identify any speededness issues. Tables 4.10 and 4.11 display the proportion of students who did not respond to the last 5 operational items in the first test section for each of the content areas for the January and May primary forms, respectively. Since the last 5 items in the end of the second section were all field test items, we have only presented the omit rates for the first section. For Biology and Government, the omission rates for each of the last 5 items were small and consistent with one another, suggesting that students had sufficient time to complete the entire assessment. Omission rates for the other three content areas were higher and were tending to increase toward the end of the session. This is particularly noticeable in the January Algebra and English forms. This may be related to insufficient time to complete the section. Alternatively students may not be motivated to complete the test, especially for the CR items as shown in Table 4.10 and 4.11 for Algebra and Geometry. MSDE planed to change the placement of the field test items (i.e. embed filed test items within operational items so that item statistics obtained for the field test items will be closer to operational settings. A more detailed omit analyses will be conducted for the 2005 administrations to discern whether the high omit rates were due to speededness or student motivation.

Table 4.10 Proportion of Students Omitting the Last 5 Items in the First Session:

January					
Algebra		Biology		English	
Item Number	%	Item Number	%	Item Number	%
19	3.3	32	1.9	32	10.4
20	3.1	33	2.3	33	10.8
21(CR)*[1,2]	35.2	34	2.2	34	11.3
22	9.9	35	2.5	35	11.6
23	10.6	36		36	12.0
Geometry		Government			
Item Number	%	Item Number	%		
21(CR)*[1,2]	23.3	30	1.9		
22	5.2	31	2.3		
23	5.0	32	1.9		
24	5.3	33	2.3		
25	5.1	34	2.2		

Table 4.11 Proportion of Students Omitting the Last 5 Items in the First Session: May

May					
Algebra		Biology		English	
Item Number	%	Item Number	%	Item Number	%
20	1.5	31	1.1	31	3.6
22(CR)	15.2	32	3.6	32	3.9
23	4.8	33	2.2	33	4.0
24	5.0	34	2.3	34	4.4
25	5.4	35	2.5	35	4.6
Geometry		Government			
Item Number	%	Item Number	%		
21(CR)	10.8	30	1.4		
22	2.9	31	1.5		
23	3.8	32	1.5		
24	3.0	33	1.6		
25	4.0	34	1.5		

*[1,2] CR – Constructed response items
*[1,2] CR Omit rates were defined as percent of student receiving condition code 'A' or 'B'.

Reliability

Reliability focuses on 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, they 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, for example, to 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 HSAs have mix item type (both dichotomous and polytomous items), it is more appropriate to report stratified Alpha (Feldt and Brennan, 1989). The stratified Alpha is a weighted average of Cronbach's Alpha for item sets with different maximum score points, i.e. "strata". The formula for calculating the stratified Alpha is:

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

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 score is presented in Tables 4.12 to 4.21 The results in these tables indicate that all of the HSAs were highly reliable with overall reliabilities ranging from 0.85 to 0.95. The lowest reliabilities were observed in Algebra. In general, the make-up forms had slightly lower reliabilities than the primary forms. Reliability estimates for the some of the tests were lowest for the make-up forms, which also have lower mean scale scores. This suggests that these lower reliabilities may be related to a decrease in true-score variance.

Table 4.12. Summary Statistics for Algebra Primary Forms

		January				May			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		408.6	39.83	4617	0.92	422.1	36.66	59398	0.91
Gender									
	Male	406.2	41.63	2413		422.3	37.58	28749	
	Female	411.3	36.57	2204		421.9	35.78	30639	
Special Education									
	Yes	374.5	50.15	459		385.5	40.14	4530	
	No	412.4	36.15	4086		425.2	34.64	54236	
	504 Only	409.6	35.17	72		415.2	35.64	632	
Ethnicity									
	American Indian	*	*	*		415.5	36.07	227	
	Asian/Pacific Islander	412.9	38.34	115		441.9	34.92	3461	
	African American	391.5	43.07	1416		402	33.86	19970	
	White	416.9	34.81	2950		433.3	32.62	32329	
	Hispanic	401.8	37.23	122		412.8	35.24	3332	
Limited English Proficient									
	Yes	384.3	39.22	68		402.8	37.43	1426	
	No	409	44.17	4537		422.6	36.49	57449	
	Exited	*	*	*		420	37.89	523	

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

Table 4.13. Summary Statistics for Algebra Make-Up Form

		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		398.8	43.15	366	0.91	385.8	32.12	296	0.85	404.4	38.74	1672	0.92	396.3	31.4	418	0.89
Gender																	
	Male	397.6	43.13	194		380.9	34.03	153		402.3	40.4	817		395.4	34.8	198	
	Female	400.3	43.26	172		391.1	29.16	143		406.3	36.99	855		397.1	28.05	220	
	Missing			*				*				*				*	
Special Education																	
	Yes	*	*	*		*	*	*		370	38.57	162		*	*	*	
	No	403.6	40.17	316		387	31.58	250		408.2	36.91	1491		398.5	29.94	377	
	504 Only	*	*	*		*	*	*		*	*	*		*	*	*	
Ethnicity																	
	American Indian	*	*	*				8		*	*	*		*	*	*	
	Asian/Pacific Islander	*	*	*		*	*	*		414.9	34.48	50		*	*	*	
	African American	370.1	49.16	93		375.9	36.87	113		386.8	35.32	643		387.6	28.24	238	
	White	410.1	34.54	260		395.4	27.72	110		417.2	36.55	875		407.5	32.24	146	
	Hispanic	*	*	*		383.5	25.87	62		399.9	35.82	100		412.5	26.62	*	
	Missing			*				*				*				*	
Limited English Proficient																	
	Yes	*	*	*		*	*	*		*	*	*		*	*	*	
	No	399.1	43.15	363		387.7	31.87	259		404.8	38.7	1625		396.2	31.18	412	
	Exited	*	*	*		*	*	*		*	*	*		*	*	*	

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

Table 4.14 Summary Statistics for Biology Primary Forms

		January					May			
		Mean	SD	N	Alpha		Mean	SD	N	Alpha
Overall		403.3	39.37	7770	0.92		414.5	33.53	46550	0.93
Gender										
	Male	400.5	43.2	3931			413.8	34.51	22433	
	Female	406.2	34.8	3839			415	32.56	24106	
Special Education										
	Yes	365.2	43.88	799			385	31.08	3685	
	No	407.8	36.33	6856			417.1	32.51	42404	
	504 Only	400	38.82	115			408.6	30.56	461	
Ethnicity										
	American Indian	*	*	*			410.2	30.87	163	
	Asian/Pacific Islander	422.1	40.35	161			432.8	34.53	2913	
	African American	381.3	37.55	2206			396.2	29.05	15913	
	White	412.5	35.74	5222			425	30.6	24925	
	Hispanic	386	47.27	160			403.5	30.98	2609	
Limited English Proficient										
	Yes	363.5	42.88	73			387.5	29.55	1079	
	No	403.7	39.14	7681			415.2	33.35	45039	
	Exited	*	*	*			405.1	31.92	432	

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

Table 4.15. 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
Overall		372.6	48.44	709	0.92	392.9	31.55	1356	0.91	380	29.12	342	0.86
Gender													
	Male	364.5	51.83	393		391.3	32.87	640		378.4	29.96	178	
	Female	382.84	41.72	316		394.2	30.27	716		381.9	28.16	164	
	Missing			*				*				*	
Special Education													
	Yes	333.4	46.27	131		372.9	29.32	149		*	*	*	
	No	381.5	44.6	568		395.2	30.96	1186		381.9	28.09	302	
	504 Only	*	*	*		*	*	*		*	*	*	
Ethnicity													
	American Indian	*	*	*		*	*	*		*	*	*	
	Asian/Pacific Islander	429.4	38.08	21		*	*	*		*	*	*	
	African American	359.6	46.31	314		380.6	27.12	570		376	28.04	218	
	White	383.4	46.18	329		403.8	30.62	661		388.3	30.2	101	
	Hispanic	358.3	43.9			381.9	30.13	82		*	*	*	
	Missing			*				*				*	
Limited English Proficient													
	Yes	*	*	*		*	*	*		*	*	*	
	No	373.1	48.68	691		393.2	31.59	1325		380.2	29.21	337	
	Exited	*	*	*		*	*	*		*	*	*	

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

**Make-Up forms for this administration contain the same set of operational items I couldn't find the ** in the table. Perhaps this sentences should be a "Note:" rather than appearing with

Table 4.16. Summary Statistics for English Primary Forms

		January				May			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		399.8	33.6	7193	0.90	411.6	34.35	55016	0.92
Gender									
	Male	393.3	34.65	3704		405.4	34.68	27067	
	Female	406.7	30.97	3489		417.6	32.92	27939	
Special Education									
	Yes	362.9	32.64	837		374.4	30.12	5376	
	No	404.8	30.65	6255		415.8	32.26	49058	
	504 Only	395.9	25.12	101		400.3	31.47	582	
Ethnicity									
	American Indian	*	*	*		403	31.55	258	
	Asian/Pacific Islander	411.3	33.7	179		425.8	34.16	3001	
	African American	380.1	31.55	1625		397.1	30.77	19726	
	White	405.9	31.9	5202		421.3	33.06	28895	
	Hispanic	391	27.93	172		401.2	30.97	3087	
Limited English Proficient									
	Yes	373.4	27.14	64		382.2	25.24	1072	
	No	400.1	33.57	7119		412.4	34.28	53417	
	Exited	*	*	*		397.2	26.98	527	
* Statistics not reported for sample size less than 50 (N<50)									

Table 4.17 Summary Statistics for English 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		368.4	38.87	307	0.90	366.3	48.9	308	0.91	386.6	32.45	1510	0.92	387.2	31.17	528	0.91
Gender																	
	Male	361.5	39.91	196		359.5	50.89	184		379	32.26	736		378.7	29.89	237	
	Female	380.5	33.86	111		376.4	44.09	124		393.8	31	773		394.3	30.54	290	
	Missing	*	*	*		*	*	*		*	*	*		*	*	*	
Special Education																	
	Yes	340.6	43.48	73		331.3	46.05	67		364.7	28.29	211		*	*	*	
	No	377.5	33.18	227		376.2	45.53	235		390.3	31.59	1282		389.9	29.59	482	
	504 Only	*	*	*		*	*	*		*	*	*		*	*	*	
Ethnicity																	
	American Indian	*	*	*		*	*	*		*	*	*		*	*	*	
	Asian/Pacific Islander	*	*	*		*	*	*		*	*	*		*	*	*	
	African American	359.3	34.85	135		350.6	49.66	118		379	29.19	678		386.6	29.82	384	
	White	375.9	40.77	161		378.3	47.07	138		394.3	33.89	701		388.4	35.79	117	
	Hispanic	*	*	*		*	*	*		383.7	30.18	81		*	*	*	
	Missing	*	*	*		*	*	*		*	*	*		*	*	*	
Limited English Proficient																	
	Yes	*	*	*		*	*	*		*	*	*		*	*	*	
	No	368.7	39.08	302		366.4	49.86	292		387	32.54	1465		387.4	31.29	516	
	Exited	*	*	*	*	*	*	*		*	*	*		*	*	*	

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

Table 4.18. Summary Statistics for Geometry Primary Forms

		January Primary Forms				May Primary Forms			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		407.2	35.95	7113	0.92	413.3	30.61	45280	0.93
Gender									
	Male	406	37.76	3423		414.7	30.95	21567	
	Female	408.3	34.16	3690		411.9	30.24	23713	
Special Education									
	Yes	373.8	41.65	581		389.2	28	2655	
	No	410.3	33.86	6417		414.9	30.15	42188	
	504 Only	403.8	32.34	115		405.1	27.09	442	
Ethnicity									
	American Indian	*	*	*		398.6	27.73	204	
	Asian/Pacific Islander	427.6	27.29	190		432.1	30.35	2982	
	African American	383	37.43	1745		398.9	25.6	14014	
	White	415.3	31.78	4985		422.6	27.69	25635	
	Hispanic	403.8	29.56	175		404.2	28.04	2426	
Limited English Proficient									
	Yes	408.2	30.25	61		398.1	30.29	918	
	No	407.2	36.02	7036		413.6	30.5	43903	
	Exited	*	*	*		405.65	33.28	464	

Table 4.19 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		385.8	34.93	332	0.87	401.3	41.8	549	0.93	395.1	27.26	1684	0.89	390.8	27.25	549	0.90
Gender																	
	Male	381.6	38.02	172		398.3	43.41	285		395.6	28.16	851		391.5	27.96	264	
	Female	390.4	31.75	160		404.6	39.83	264		394.7	26.31	833		390.2	26.61	285	
	Missing	*	*	*		*	*	*		*	*	*		*	*	*	
Special Education																	
	Yes	*	*	*		357.4	47.3	53		376.5	24.4	139		*	*	*	
	No	389	32.83	296		406.4	38.29	488		396.9	26.92	1519		392.1	26.91	501	
	504 Only	*	*	*		*	*	*		*	*	*		*	*	*	
Ethnicity																	
	American Indian	*	*	*		*	*	*		*	*	*		*	*	*	
	Asian/Pacific Islander	*	*	*		437.2	29.29	127		412.7	27.54	57		*	*	*	
	African American	374.6	33.97	132		375.5	33.76	157		384	23.42	678		380.2	24.97	271	
	White	394.4	32.64	187		402.8	39.47	192		403.7	26.95	843		403.7	24.49	233	
	Hispanic	*	*	*		390.6	36.62	73		389.7	23.15	99		*	*	*	
	Missing	*	*	*		*	*	*		*	*	*		*	*	*	
Limited English Proficient																	
	Yes	*	*	*		*	*	*		*	*	*		*	*	*	
	No	386	35.03	329		400.5	41.88	482		395.3	27.27	1644		391.1	27.19	544	
	Exited	*	*	*		*	*	*		*	*	*		*	*	*	

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

Table 4.20 Summary Statistics for Government Primary Forms

		January Primary Forms				May Primary Forms			
		Mean	SD	N	Alpha	Mean	SD	N	Alpha
Overall		402.5	39.76	8119	0.95	413.5	36.17	50408	0.95
Gender									
	Male	398.8	42.69	4074		411.9	37.37	24708	
	Female	406.3	36.19	4045		415.2	34.9	25684	
Special Education									
	Yes	358.2	41.83	787		381.1	33.19	4415	
	No	407.5	36.47	7232		416.8	34.89	45460	
	504 Only	395.2	33.98	100		406.8	33.38	533	
Ethnicity									
	American Indian	*	*	*	*	383.8	30.39	298	
	Asian/Pacific Islander	420.7	39.06	181		402.3	39.17	3098	
	African American	379.8	35.76	2169		432.2	30.46	17549	
	White	411.1	37.63	5543		397.5	35.31	26574	
	Hispanic	394.2	39.29	202		42.31	32.99	2831	
Limited English Proficient									
	Yes	*	*	*	*	388.9	29.75	1114	
	No	402.7	39.71	8063		414.2	36.14	48803	
	Exited	*	*	*	*	403.7	32.02	491	

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

Table 4.21 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
Overall		378.7	40.99	748	0.95	391.8	36.95	1692	0.95	390	35.39	721	0.95
Gender													
	Male	374.1	43.59	414		388.1	37.86	790		387.5	38.16	337	
	Female	384.2	36.82	334		395.1	35.84	902		392.2	32.66	384	
	Missing	*	*	*		*	*	*		*	*	*	
Special Education													
	Yes	354.3	35.88	119		362.6	31.7	156		361.8	26.49	90	
	No	383.5	40.61	614		395	36.21	1519		394.2	34.61	622	
	504 Only	*	*	*		*	*	*		*	*	*	
Ethnicity													
	American Indian	*	*	*		*	*	*		*	*	*	
	Asian/Pacific Islander	*	*	*		415.3	46.8	52		*	*	*	
	African American	364.3	38.11	249		379.5	30.93	794		382.8	31.42	356	
	White	386.9	41.29	421		406	37.35	717		396.3	36.7	325	
	Hispanic	376.3	33.7	58		379.2	31.65	122		*	*	*	
	Missing	*	*	*		*	*	*		*	*	*	
Limited English Proficient													
	Yes	*	*	*		*	*	*		*	*	*	
	No	378.9	41.32	724		392.4	37.08	1639		390.1	35.52	713	
	Exited	*	*	*		*	*	*		*	*	*	

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

** Make-Up forms for this administration contain the same set of operational items

References

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. Psychometrika, 16, 292-334.

Feldt, L & Brennan, R. (1989). Reliability. In Linn, R. L. (Ed.), Educational Measurement (3rd. Ed., pp. 117-118), NY: Macmillan

Section 5. Field Test Analyses

Following the receipt of the final scored file from Measurement Incorporated (MI), the field test analyses were completed. The analysis of the field test data can be broken down in four componentsThere are four types of analyses conducted for the field test data: 1) classical item analyses; 2) differential item functioning (DIF) analyses; and 3) calibration and scaling. All of the analyses were completed using Genasys, ETS proprietary software. The analysis procedures for each component are described in detail. The samples used for all analyses included all valid records available at the time of the analyses. Students classified as English as a second language, students with IEP or 504 plans and those receiving accommodations were included in all analyses. Only duplicate records, records invalidated by the test administrator and those with five or fewer item responses were excluded from the analysis sample.

Classical Item Analyses

Classical item analyses involve computing, for every item in each form, a set of statistics based on classical test theory. Each statistic is designed to provide some key information about the quality of each item from an empirical perspective. The statistics estimated for the HSA 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 to better measure students with very high or low achievement, especially if the students have not yet received instruction in the content or if they lack motivation to complete the field test items to the best of their ability.

The item-total correlation of the correct response option (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., incorrect key, multiple correct answers or unusually complex content), or can indicate that students have not been taught the content.

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

These statistics indicate the percent of examinees that select each of the available answer options and the percent of examinees that omitted the

item. Item options not selected by any students indicate problems with plausibility of the option. Items that do not have all answer options functioning should 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 of this correlation 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 can also 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 that 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 in identifying how well the item is functioning. If no students are assigned the top score point, this indicates that the item may not be functioning with respect to the rubric and/or that the is with no students can indicate serious problems with the item content or can indicate that students have not been taught the content.

Summaries of the items administered based on p-values are listed in and item-total correlations are listed in Tables 5.1-5.8 for each content area. In addition, a series of flags were created in order 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 was applied to all items tested in the 2003-2004 assessments:

- Difficulty Flag: P-values less than 0.25 or greater than 0.90.
- Discrimination Flag: Point- biserial correlation less than 0.15 for correct answer.

- Distractor Flag: Point-biserial correlation is 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 (see Tables 5.9 and 5.10). These items have been identified for revision and possible future re-field testing.

Differential Item Functioning (DIF)

Following the classical item analyses, DIF studies were completed. One of the goals 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 those items that identifiable groups of students (e.g. females, African Americans, Hispanics) with the same underlying level of ability have different probabilities of answering correctly. If the item is differentially more difficult for an identifiable subgroup, the item may be measuring something different from 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 are required to determine the source and meaning of any differences that are seen.

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[9]. This statistic is expressed as the

[9] 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):

$$MH\ D\text{-}DIF = -2.35\ ln[\alpha_{MH}] .$$

differences between the focal and reference group after conditioning on total test score. This 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. 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 as reference groups and the female and other ethnic groups are categorized as 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 contains negligible DIF, Category B items exhibit slight or moderate DIF, and Category C items have moderate to large values of 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.

No items were flagged for C-level DIF against one of the identified focal groups (female, African American, American Indian, Asian, and Hispanic) for both January and May administrations.

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 versions within 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 HSA test items were the 3-parameter logistic (3PL) model for selected response items and the generalized partial credit model (GPCM) for constructed response 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:

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

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

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

where

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

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

P_{jk} is the probability of responding in the k^{th} 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
XXXXXXXXXXXXXXX
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. In some cases, there may be several iterations of calibrations whereby items that do not converge are removed from analysis. No items were omitted from the final calibrations.

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
- Collapsed score level
- Item not scored

Government Constructed Response Study

In the evolution of the item writing process, the directional statements associated with the Government brief and extended constructed response items were modified to be more specific, beginning with the May, 2004 administration. In reviewing the item bank, there were several items that could be used on future forms, however, these items included the previous directional statements and formatting. As a result, available items have two different formats and future test forms could include items with both types of formatting. While changing all of the items to the "new" format would be desirable, MSDE was concerned that this change could impact item performance. To obtain new item parameters, the items would need to be re-field tested, which would decrease the numbers of items available for form construction in the short term, would delay the field testing of newly written items, and would increase the development costs associated with these existing items. A study completed during the May 2004 administration that involved printing two items in both the old and new formats found that there were virtually no differences between the two sets of item parameters. Therefore the change in the directions does not appear to have an important or systematic effect on item performance (see Appendix 5.A).

Statistical Summary Tables

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

	Percentage of items (N)				
P-Values	Algebra	Biology	English I	Geometry	Government
< 0.30	62.5 (10)	17.9 (5)	2.9 (1)	75.0 (12)	16.7 (4)
0.30 to 0.40	25.0 (4)	21.4 (6)	20.6 (7)	18.8 (3)	20.8 (5)
0.41 to 0.50	6.3 (1)	28.6 (8)	2.9 (1)	6.3 (1)	12.5 (3)
0.51 to 0.60	0	10.7 (3)	26.5 (9)	0	20.8 (5)
0.61 to 0.70	0	7.1 (2)	23.5 (8)	0	20.8 (5)
0.71 to 0.80	6.3 (1)	7.1 (2)	20.6 (7)	0	8.3 (2)
> 0.81	0	7.1 (2)	2.9 (1)	0	0
Number of Items	16	28	34	16	24
Mean	0.29	0.47	0.57	0.26	0.48
SD	0.16	0.19	0.16	0.09	0.15
Min	0.11	0.19	0.24	0.12	0.20
Max	0.76	0.85	0.83	0.47	0.73

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

	Percentage of items (N)				
P-Values	Algebra	Biology	English I	Geometry	Government
< 0.30	75.0 (3)	100.0 (3)	0	100.0 (3)	83.3 (5)
0.30 to 0.40	25.0 (1)	0	50.0 (1)	0	16.7 (1)
0.41 to 0.50	0	0	50.0 (1)	0	0
0.51 to 0.60	0	0	0	0	0
0.61 to 0.70	0	0	0	0	0
0.71 to 0.80	0	0	0	0	0
> 0.81	0	0	0	0	0
Number of Items	4	3	2	3	6
Mean	0.23	0.14	0.41	0.24	0.22
SD	0.05	0.04	0.03	0.04	0.07
Min	0.18	0.11	0.39	0.19	0.14
Max	0.30	0.18	0.43	0.26	0.32

* Table information does not include items with collapsed levels

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

	Percentage of items (N)				
Correlation	Algebra	Biology	English I	Geometry	Government
< 0.15	37.5 (6)	7.1 (2)	2.9 (1)	31.3 (5)	8.3 (2)
0.15 to 0.24	25.0 (4)	10.7 (3)	11.8 (4)	31.3 (5)	29.2 (7)
0.25 to 0.34	25.0 (4)	42.9 (12)	14.7 (5)	6.3 (1)	8.3 (2)
0.35 to 0.44	12.5 (2)	28.6 (8)	17.6 (6)	18.8 (3)	12.5 (3)
0.45 to 0.54	0	10.7 (3)	52.9 (18)	6.3 (1)	33.3 (8)
> 0.55	0	0	0	6.3 (1)	8.3 (2)
Number of SR Items	16	28	34	16	24
Mean	0.20	0.33	0.40	0.24	0.35
SD	0.13	0.11	0.12	0.18	0.16
Min	-0.02	0.003	0.06	-0.10	-0.08
Max	0.45	0.52	0.54	0.56	0.57

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

	Percentage of items (N)				
Correlation	Algebra	Biology	English I	Geometry	Government
< 0.15	0	0	0	0	0
0.15 to 0.24	0	0	0	0	0
0.25 to 0.34	0	0	0	0	0
0.35 to 0.44	0	0	0	0	0
0.45 to 0.54	0	0	0	0	0
> 0.55	100.0 (4)	100.0 (3)	100.0 (2)	100.0 (3)	100.0 (6)
Number of Items	4	3	2	3	6
Mean	0.69	0.75	0.72	0.71	0.72
SD	0.05	0.02	0.01	0.12	0.03
Min	0.63	0.73	0.72	0.57	0.69
Max	0.74	0.76	0.73	0.81	0.77

* Table information does not include items with collapsed levels

Table 5.5. Distributions of P-Values for May Field Test SR Items

	Percentage of items (N)				
P-Values	Algebra	Biology	English I	Geometry	Government
< 0.30	15.6 (10)	8.0 (9)	0.6 (1)	28.1 (18)	5.3 (4)
0.30 to 0.40	15.6 (10)	17.9 (20)	10.6 (17)	17.2 (11)	16.0 (12)
0.41 to 0.50	17.2 (11)	22.3 (25)	16.3 (26)	14.1 (9)	21.3 (16)
0.51 to 0.60	25.0 (16)	22.3 (25)	26.3 (42)	18.8 (12)	22.7 (17)
0.61 to 0.70	14.1 (9)	12.5 (14)	20.6 (33)	9.4 (6)	21.3 (16)
0.71 to 0.80	9.4 (6)	16.1 (18)	22.5 (36)	10.9 (7)	13.3 (10)
> 0.81	3.1 (2)	0.9 (1)	3.1 (5)	1.6 (1)	0
Number of Items	64	112	160	64	75
Mean	0.50	0.51	0.58	0.43	0.53
SD	0.18	0.15	0.14	0.20	0.15
Min	0.11	0.16	0.20	0.05	0.16
Max	0.88	0.84	0.86	0.87	0.79

Table 5.6. Distributions of P-Values for May Field Test CR Items

	Percentage of items (N)				
P-Values	Algebra	Biology	English I**	Geometry	Government
< 0.30	53.3 (8)	56.3 (9)		46.7 (7)	76.9 (10)
0.30 to 0.40	6.7 (1)	12.5 (2)		13.3 (2)	23.1 (3)
0.41 to 0.50	20.0 (3)	0		33.3 (5)	0
0.51 to 0.60	0	0		6.7 (1)	0
0.61 to 0.70	0	0		0	0
0.71 to 0.80	0	0		0	0
> 0.81	0	0		0	0
Number of Items	12	11		15	13
Mean	0.28	0.24		0.35	0.27
SD	0.10	0.07		0.09	0.06
Min	0.11	0.14		0.24	0.15
Max	0.46	0.38		0.53	0.37

* Table information does not include items with collapsed levels

** No CR items were scored

Table 5.7. Distributions of Item-Total Correlations for May Field Test SR Items

	Percentage of items (N)				
Correlation	Algebra	Biology	English I	Geometry	Government
< 0.15	6.0 (4)	5.4 (6)	2.5 (4)	12.5 (8)	5.3 (4)
0.15 to 0.24	14.1 (9)	13.4 (15)	4.4 (7)	9.4 (6)	14.7 (11)
0.25 to 0.34	26.6 (17)	28.6 (32)	20.6 (33)	9.4 (6)	13.3 (10)
0.35 to 0.44	31.3 (20)	40.2 (45)	37.5 (60)	35.9 (23)	30.7 (23)
0.45 to 0.54	12.5 (8)	12.5 (14)	33.1 (53)	14.1 (9)	30.7 (23)
> 0.55	9.4 (6)	0.0 (0)	1.9 (3)	18.8 (12)	5.3 (4)
Number of SR Items	64	112	160	64	75
Mean	0.36	0.34	0.40	0.39	0.38
SD	0.13	0.10	0.10	0.16	0.13
Min	0.09	0.06	-0.04	0.01	0.02
Max	0.65	0.53	0.57	0.72	0.57

Table 5.8. Distributions of Item-Total Correlations for May Field Test CR Items

	Percentage of items (N)				
Correlation	Algebra	Biology	English I**	Geometry	Government
< 0.15	0	0		0	0
0.15 to 0.24	0	0		0	0
0.25 to 0.34	0	0		0	0
0.35 to 0.44	0	0		0	0
0.45 to 0.54	8.3 (1)	0		0	0
> 0.55	91.7 (11)	100.0 (11)		100.0 (15)	100.0 (13)
Number of CR Items	12	11		15	13
Mean	0.65	0.68		0.71	0.71
SD	0.06	0.04		0.07	0.05
Min	0.53	0.63		0.55	0.62
Max	0.74	0.75		0.79	0.78

* Table information does not include items with collapsed levels

** No CR items were scored

Table 5.9 Field Test Items Excluded from Analyses: January

	Algebra		Biology		English I		Geometry		Government	
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
Not Scored				1		1		1		
Low/Neg. P-biserial	1		1		1		1			
Collapsed levels		1		3						2

Table 5.10 Field Test Items Excluded from Analyses: May

	Algebra		Biology		English I		Geometry		Government	
	SR	CR	SR	CR	SR	CR	SR	CR	SR	CR
Not Scored		3				9*		1		
Low/Neg. P-biserial			1		1				1	
Collapsed levels		1		6		1				5

* English 10 test will be replacing English I test after May 2005 administration so no English I field test CR items were scored for May 2004 administration.

Appendix 5.A Maryland High School Assessment Special Study: Directional Statements accompanying the Government Constructed Responses

Maryland High School Assessment Special Study:

Directional Statements Accompanying the Government Constructed Responses

December 2004

Educational Testing Service

Maryland High School Assessment Special Study: Directional Statements accompanying the Government Constructed Responses

Background

The HSA is based on a pre-equated design– that is, items were not recalibrated following administrations, and instead bank parameters were used for scoring. As a result, the items must appear exactly as they did in the administration associated with the bank parameters. Any change to the item can result in change how students interact with the item and the resulting item parameters. Therefore, items cannot be modified: text cannot be edited or revised or graphics altered.

In the evolution of the item writing process, the directional statements associated with the Government brief and extended constructed response items were modified to be more specific, beginning with the May, 2004 administration (see Figure 5.A.1). In reviewing the item bank, there were several items that could be used on future forms, however, these items included the previous directional statements and formatting (see Figure 5.A.2). As a result, available items have two different formats and future test forms could include items with both types of formatting. While changing all of the items to the "new" format would be desirable, MSDE was concerned that this change could impact item performance. To obtain new item parameters, the items would need to be re-field tested, which would decrease the numbers of items available for form construction in the short term, would delay the field testing of newly written items, and would increase the development costs associated with these existing items. In reviewing the change it was hypothesized that item performance would not differ based on the modification. Therefore, a study was completed during the May 2004 administration that involved printing two items in both the old and new formats to help evaluate whether or not the items would need to be field tested if they were reformatted to the new style guidelines.

Figure 5.A.1. Government Brief Constructed Response Item: With Instruction

69. Read the sentences below and use them to complete the BRIEF CONSTRUCTED RESPONSE that follows.

Read the scenario below.
Recently a city ordinance [law] was passed that banned skateboard riding on most city streets and sidewalks. You and your friends believe this is an unjust law.

- **Describe two legal ways you and your friends could try to get this law changed.**
- **Explain why each of your choices would be effective.**
- **Include details and examples to support your answer.**

Write your answer on the lines in your Answer Book.

Figure 5.A.2. Government Brief Constructed Response Item: Without Instruction

69.

Read the scenario below.
Recently a city ordinance [law] was passed that banned skateboard riding on most city streets and sidewalks. You and your friends believe this is an unjust law.

- **Describe two legal ways you and your friends could try to get this law changed.**
- **Explain why each of your choices would be effective.**
- **Include details and examples to support your answer.**

Write your answer on the lines in your Answer Book.

Method and Results

In May 2004, two BCR items were selected and included in the field test sections in both the old and new formats. The classical item statistics in Table 5.A.1 show that the two versions of the items were very similar in terms of p-values and poly-serial correlations. We also compared the IRT parameter estimates of the items in each format, and noted that these values were very similar as well (see Table 5.A.2). Figures 5.A.3 and 5.A.4 show the item characteristic curves for the two different versions of items 1 and 2. Figure 5.A.5 and 5.A.6 show the item characteristic curves for each response option for the two different versions of items 1 and 2. Figure 5.A.7 and 5.A.8 show the item information function for the two versions of item 1 and 2.

Table 5.A.1: Classical Item Statistics

	P value		Poly-serial correlation	
	New	Old	New	Old
Item 1 Reappointment/Political Power	MD52236 N=6813	MD68796 N=6378	MD52236 N=6813	MD68796 N=6378
	0.19	0.20	0.31	0.31
Item 2 Due Process/ Public Safety v Rights	MD52234 N=6378	MD68795 N= 6307	MD52234 N=6378	MD68795 N=6307
	0.78	0.76	0.65	0.65

Appendix 5.A

Table 5.A.2. Frequency Distribution of Score Points

	Percent Score 0		Percent Score 1		Percent Score 2		Percent Score 3		Percent Score 4	
	New	Old	New	Old	New	Old	New	Old	New	Old
Item 1 Reappointment/Political Power	0.50	0.50	0.26	0.24	0.21	0.22	0.03	0.03	0.00	0.00
Item 2 Due Process/ Public Safety v Rights	0.19	0.20	0.42	0.40	0.34	0.35	0.04	0.04	0.00	0.00

Table 5.A.3. IRT Parameter Estimates

	A-Value		B1-Value		B2-Value		B3-Value		B4-Value	
	New	Old	New	Old	New	Old	New	Old	New	Old
Item 1 Reappointment/Political Power	0.03125	0.02926	414.1	420.0	435.3	433.8	497.0	498.2	561.7	553.4
Item 2 Due Process/ Public Safety v Rights	0.02321	0.02237	371.1	373.1	424.3	422.5	500.6	507.1	578.7	570.3

Conclusion and Recommendation

Any time item parameter estimates are obtained with different samples, some differences occur due to sampling error. This study found that there were minimal differences between the two sets of item parameters. Thus, this change in the directions does not appear to have had an important or systematic effect on item performance.

Appendix 5.A

Figure 5.A.3: Item Characteristic Curve for CR item 1.

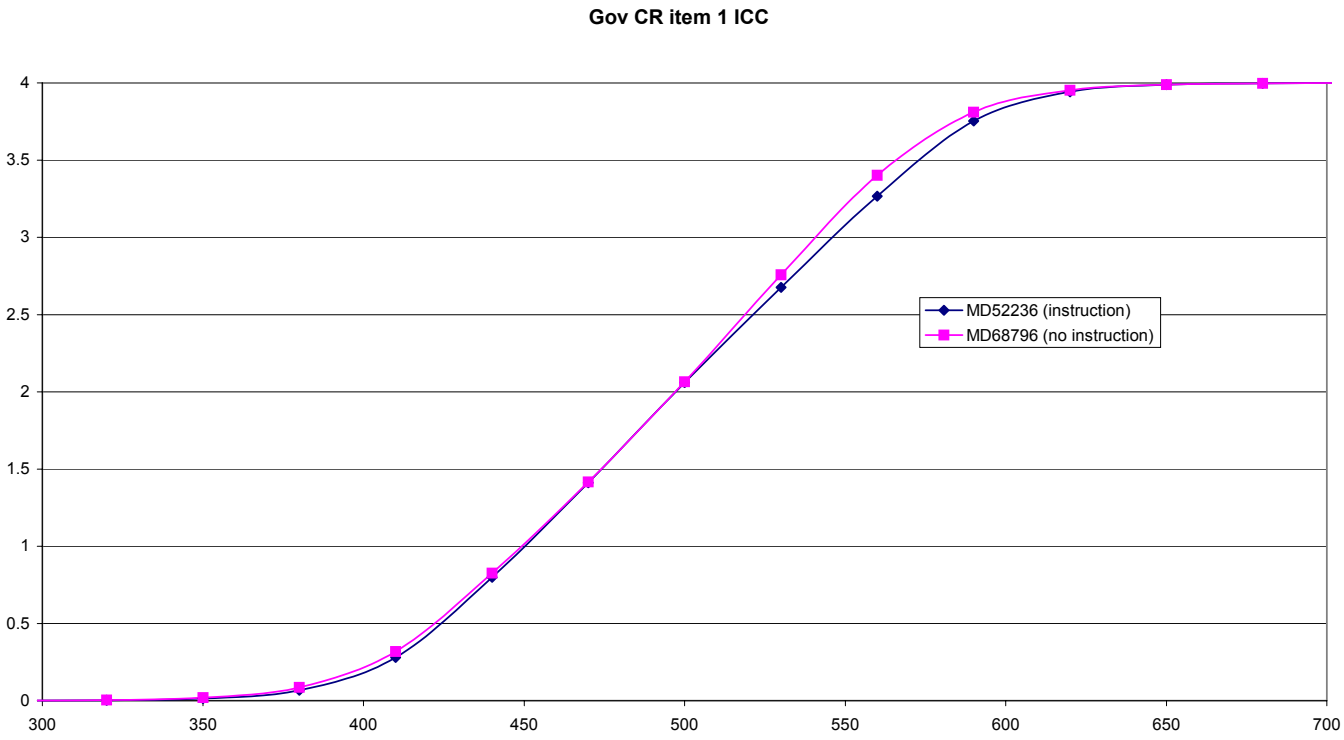
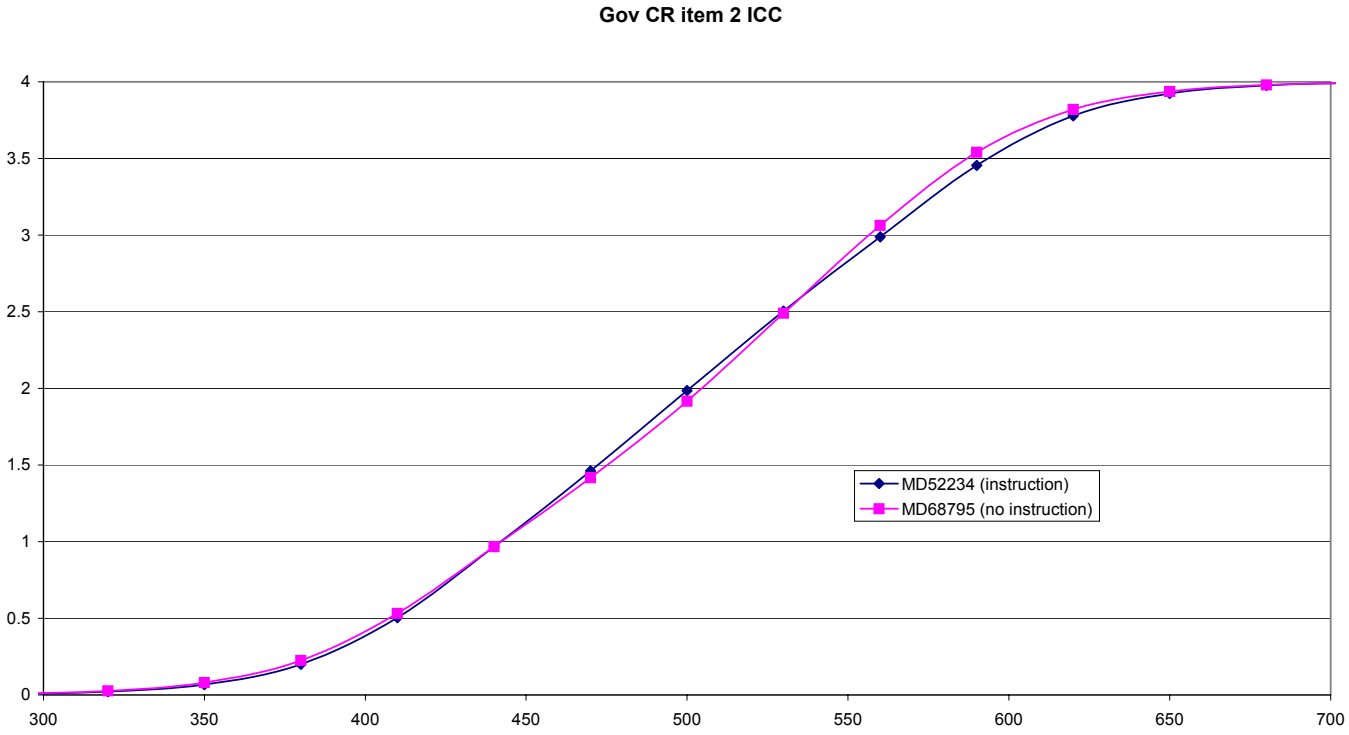


Figure 5.A.4: Item Characteristic Curve for CR item 2



Appendix 5.A

Figure 5.A.5: Item Characteristic Curve for each Response Option of Item 1

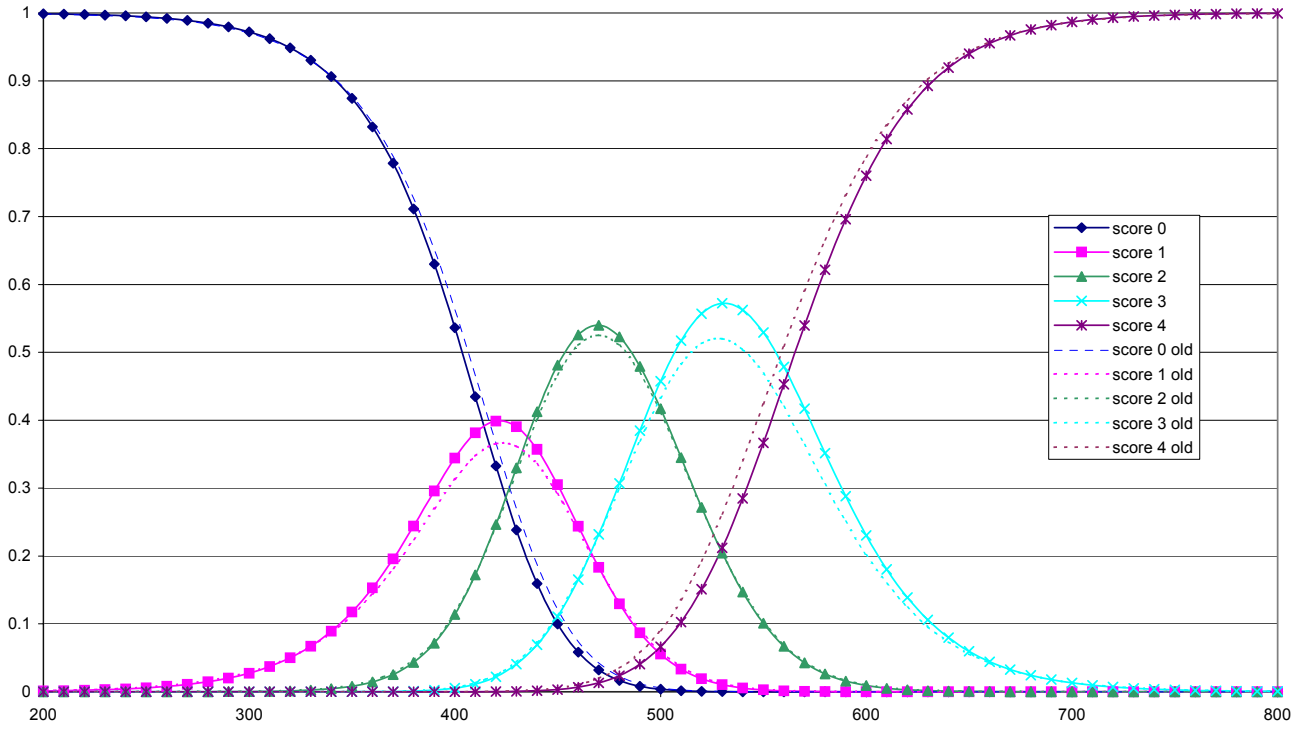
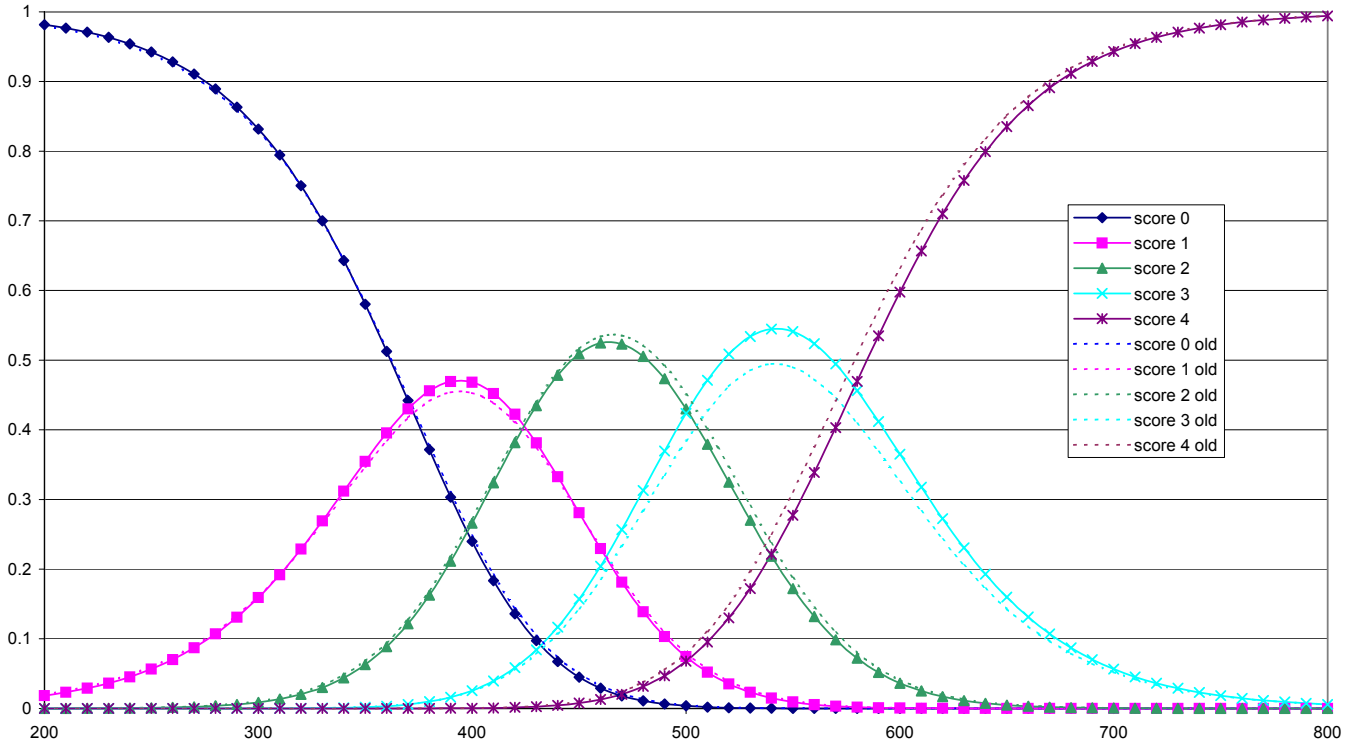


Figure 5.A.6: Item Characteristic Curve for each Response Option of Item 2



Appendix 5.A

Figure 5.A.7: Information function for CR item 1

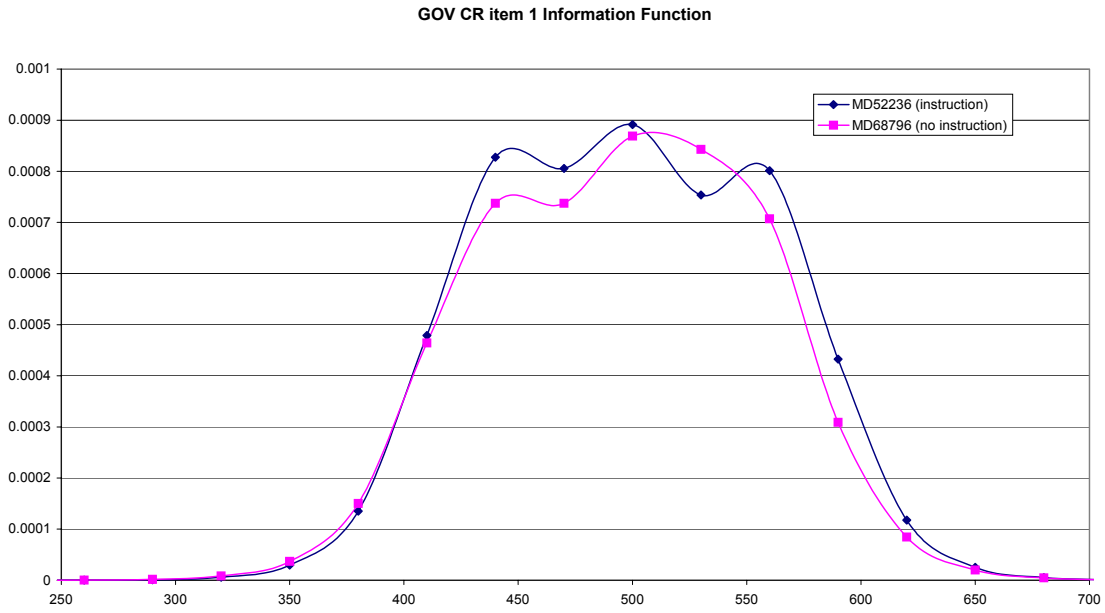


Figure 5.A.8: Information function for CR item 2