



MeCAS

Maine's Comprehensive
Assessment System

2008–2009 Technical Report Part Two

Maine High School Assessment



TABLE OF CONTENTS

BACKGROUND AND OVERVIEW.....	1
CHAPTER 1. MAINE HIGH SCHOOL ASSESSMENT BACKGROUND AND OVERVIEW.....	1
SECTION I—MHSA DEVELOPMENT.....	3
CHAPTER 2. DEVELOPMENT AND DESIGN OF THE MHSA: SAT AND MATH–A.....	3
2.1 The MHSA: The SAT and Math–A Overview.....	3
2.2 Universal Design Specifications.....	4
2.3 SAT Critical Reading Test.....	4
2.4 SAT Writing Test.....	6
2.5 MHSA Mathematics Test: SAT and Math–A Components.....	12
2.6 Development.....	17
2.7 Item Writing and Review.....	18
2.8 Pre-testing the Items.....	19
2.9 Analysis of Pretest Information for the MHSA: SAT and Math–A.....	19
2.10 Item Difficulty.....	20
2.11 Item Discrimination and Item/Test Relationship.....	21
2.12 Differential Item Functioning.....	22
2.13 Evaluating Essay Pretests.....	24
2.14 Assembling the SAT Portion of the MHSA.....	24
2.15 Reviewing the MHSA: SAT Component.....	25
2.16 Test Production for the SAT Component.....	25
2.17 After the SAT Administration.....	26
2.18 Public Access to the SAT.....	26
CHAPTER 3. ALIGNMENT OF THE SAT TO THE REVISED <i>LEARNING RESULTS</i>	27
3.1 Design of SAT Critical Reading.....	27
3.2 Design of SAT Writing.....	28
3.3 Design of SAT Mathematics.....	28
CHAPTER 4. OVERVIEW OF THE SCIENCE TEST DESIGN.....	31
4.1 <i>Learning Results</i>	31
4.2 Test Design.....	32
4.3 Item Types.....	32
4.4 Test Session Times.....	33
CHAPTER 5. MHSA SCIENCE TEST DEVELOPMENT PROCESS.....	35
5.1 Item Development.....	35
5.1.1 External Review of Item Content.....	35
5.1.2 Bias and Sensitivity Review.....	35
5.1.3 Item Editing.....	35
5.1.4 Reviewing and Refining.....	36
5.2 Operational Test Assembly.....	36
5.2.1 Editing Drafts of Operational Tests.....	37
5.2.2 Braille and Large-Print Tests.....	37
SECTION II—MHSA TEST ADMINISTRATION.....	39
CHAPTER 6. ADMINISTRATION OF THE SAT.....	39
6.1 Preparation.....	39
6.2 Supervision.....	39
6.3 Physical Setting.....	40
6.4 Security.....	40
6.5 Calculator Policy for the SAT.....	41
6.6 Item Types.....	42
6.7 Instructions and Timing.....	42
6.8 Complaints and Irregularities.....	43
6.9 Subgroup Performance.....	43
6.10 Accommodations for Students on the MHSA.....	45
6.10.1 Process and Standards for College Board Approved Accommodations.....	45

6.10.2	Process and Standards for MPO Accommodations.....	47
6.10.3	Eligibility Process Additions to Incorporate MPO Accommodations.....	47
6.10.4	Accommodation Eligibility Form Submission Time Lines.....	48
6.10.5	Training and Technical Assistance.....	48
6.10.6	MHSA Accommodation Request and Approval Statistics.....	48
6.11	Participation.....	50
CHAPTER 7.	ADMINISTRATION OF MATH–A AND SCIENCE.....	51
7.1	Supervision and Security.....	51
7.2	Participation Requirements and Accommodations.....	52
SECTION III—SCORING.....		55
CHAPTER 8.	SCORING THE SAT.....	55
8.1	Receiving and Opening.....	55
8.2	Scanning and Editing.....	56
8.3	Matching.....	56
8.4	Machine-Scored Portions.....	57
8.5	Scoring the Essay.....	58
8.6	End to End Quality Control.....	63
8.7	Quality Assessments.....	64
8.8	Summary.....	64
CHAPTER 9.	SCORING MATH–A AND SCIENCE.....	65
9.1	Scoring MHSA Test Items.....	65
9.1.1	Machine Scored Items.....	65
9.1.2	Hand Scored Items.....	65
9.2	Scoring Locations and Staff.....	66
9.2.1	Scoring Locations.....	66
9.2.2	Staff Positions.....	66
9.2.3	Benchmarking Meetings with the MDOE.....	73
9.3	Methodology for Scoring Constructed-response Items.....	73
9.4	Scoring Reports.....	74
SECTION IV—PSYCHOMETRICS AND REPORTING.....		77
CHAPTER 10.	PSYCHOMETRIC TOPICS OF THE SAT.....	77
10.1	The Equating and Braiding Plan for SAT Mathematics, Critical Reading, and Writing.....	77
10.2	SAT Statistical Characteristics.....	78
10.3	Reliability and Standard Errors of Measurement.....	78
10.3.1	Reliability.....	78
10.3.2	Standard Errors of Measurement.....	79
10.4	Classification Accuracy and Consistency of Maine SAT Cut Scores.....	85
10.5	Completion Rates.....	87
10.6	Item Statistics.....	89
10.6.1	Item Difficulty: Equated Delta.....	89
10.6.2	Item Discriminating Power: Biserial Correlation.....	91
10.7	Differential Item Functioning (DIF).....	92
10.8	Summary.....	94
CHAPTER 11.	PSYCHOMETRIC TOPICS OF MATH–A AND SCIENCE.....	95
11.1	Formula Scoring.....	95
11.2	Standard Setting.....	96
11.2.1	Panel Membership.....	96
11.2.2	Calculation of Starting Cut Points.....	96
11.2.3	Orientation.....	96
11.2.4	Review of Assessment Materials.....	96
11.2.5	Completion of Item List Form.....	97
11.2.6	Review of Achievement Level Definitions and Definition of Borderline Students.....	97
11.2.7	Round 1 Judgments.....	97
11.2.8	Tabulation of Round 1 Results and Impact Data.....	98
11.2.9	Round 2 Judgments.....	98
11.2.10	Evaluation.....	98
11.2.11	Standard Setting Report.....	98

11.3	Deriving MHSA Scaled Scores	99
11.3.1	Item Response Theory Calibration.....	99
11.3.2	Scaling MHSA Mathematics and Science, and Equating the Two Mathematics Components	101
11.3.3	Scaling Additional Forms.....	102
11.4	Item Analyses	105
11.5	Subgroup Differences in Item Performance.....	107
11.6	Dimensionality Analyses.....	108
11.7	Item Response Theory Analyses.....	111
11.8	Reliability	114
11.9	Reliability and Standard Errors of Measurement.....	115
11.10	Classification Accuracy and Consistency of MHSA Cut Scores in Mathematics and Science	117
CHAPTER 12.	VALIDITY RESEARCH ON THE MHSA	119
12.1	Construct Validity.....	119
12.2	Verbal Reasoning.....	120
12.3	Quantitative Reasoning.....	121
12.4	Writing.....	123
12.5	Multiple-choice Questions	123
12.6	Essay Question.....	124
12.7	Does the Length of the SAT Result in a Fatigue Effect?	124
12.8	How Do SAT Scores Relate to College Performance?	125
12.9	Performance Over Multiple Time Periods	129
12.10	Longer Term Performance.....	133
12.11	Differential Validity for Subgroups	135
12.11.1	Gender.....	136
12.11.2	Race and Ethnicity	138
12.11.3	Students With Disabilities.....	139
12.11.4	Fatigue Effects	142
12.12	Summary of the MHSA SAT Component.....	142
12.13	MHSA Mathematics and Science Component Validities	143
12.14	State Level Results	144
12.15	MHSA Validity Studies Agenda.....	144
CHAPTER 13.	MHSA SCORE REPORTING.....	147
13.1	Primary Reports	147
13.2	Student Report for Parents/Guardians.....	147
13.3	Student Labels	147
13.4	Class Analysis Report.....	148
13.5	School and SAU Reports.....	148
13.6	Decision Rules.....	152
13.7	Quality Assurance.....	152
REFERENCES	155
APPENDICES	163
APPENDIX A	2008–09 TECHNICAL ADVISORY COMMITTEE MEMBERS	
APPENDIX B	ALIGNMENT ANALYSIS OF HIGH SCHOOL MATHEMATICS	
APPENDIX C	ALIGNMENT ANALYSIS OF HIGH SCHOOL READING	
APPENDIX D	POLICIES AND PROCEDURES FOR ACCOMMODATIONS AND ALTERNATE ASSESSMENT	
APPENDIX E	TEST ADMINISTRATOR, AND PRINCIPAL AND TEST COORDINATOR MANUALS	
APPENDIX F	SCORING SPECIFICATIONS AND DECISION RULES	
APPENDIX G	FORMULAS FOR RELIABILITY AND SEM	
APPENDIX H	INTERPRETING SCORES ON THE SAT	
APPENDIX I	MAINE SCIENCE STANDARD SETTING REPORT	
APPENDIX J	SAMPLE REPORTS	
APPENDIX K	ANALYSIS AND REPORTING DECISION RULES	
APPENDIX L	NATIONAL TABLES	

BACKGROUND AND OVERVIEW

Chapter 1. MAINE HIGH SCHOOL ASSESSMENT BACKGROUND AND OVERVIEW

The purpose of this report is to document the technical aspects of the 2008–09 Maine High School Assessment (MHSA), one of the three components of Maine’s Comprehensive Assessment System (MeCAS). The other two components are the Personalized Alternate Assessment Portfolio and the Maine Educational Assessment (MEA), each of which is documented in a separate technical report.

This report provides information about the technical quality of the MHSA, including a description of the processes used to develop, administer, and score the test and to analyze the test results. It is intended to serve as a guide for replicating and/or improving the procedural and analytical processes to be followed in subsequent years for the MHSA component of Maine’s testing program. It was written by staff at both the College Board, the SAT contractor, and Measured Progress, the MHSA testing contractor; reviewed by members of the Maine Technical Advisory Committee for Assessment (see Appendix A); and edited by Maine Department of Education (MDOE) staff.

While some sections of this technical report may be used by educated laypeople, it is intended for experts in psychometrics and educational research. The report assumes a working knowledge of measurement concepts such as *reliability* and *validity*, and statistical concepts such as *correlation* and *central tendency*. In some chapters, the reader is presumed also to have basic familiarity with advanced topics in measurement and statistics.

The MHSA is designed to measure student progress toward the achievement of the state standards, Maine’s *Learning Results: Parameters for Essential Instruction*. The *Learning Results* content standards are designed to identify the skills and knowledge that all Maine students will need to succeed in the 21st century and are intended to provide them the opportunity to be college, career, and citizenship ready upon graduation.

From 1985 through 2005, grade 11 students took the state developed MEA. The decision to use the SAT Reasoning Test® (SAT) as Maine’s high school assessment was made in 2005 by the commissioner of education, who determined that all third year high school students, not just the 75% typically taking the SAT for college admissions purposes, would benefit from participating in this testing program. Using the SAT to meet federal testing requirements detailed in the No Child Left Behind Act of 2001 (NCLB), while creating a culture that supported college readiness for all Maine students, fit the MDOE’s vision and provided high school students with a meaningful assessment.

Consequently, beginning in the spring of 2006, all Maine third year high school students were required to participate in the SAT program. The following year, 2006–2007, students were required to participate in the MHSA, which was composed of both the SAT (measuring mathematics, critical reading, and writing) and a mathematics augmentation (Math–A) designed to ensure alignment of the content area

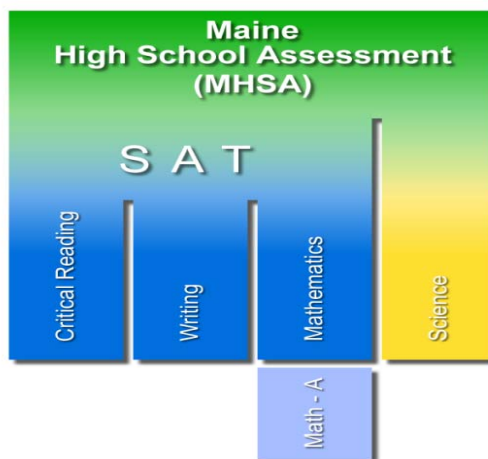
assessment to Maine’s mathematics standards. The Math–A portion was administered in each Maine high school on a school day in April 2007, and the SAT was administered on Saturday, May 5, 2007, with a makeup day in June.

In 2007–2008, a fourth discipline, science, was added to the MHSA compilation as required under NCLB and was administered along with the Math–A in each Maine high school on a school day(s) during a two week administration window in early April. The SAT was administered to Maine students on Saturday, May 3, 2008, with a makeup day on June 7, 2008. The same administration protocols and time lines were followed in 2008–2009: the Math–A and science components were administered during a two week window that ran from March 30 to April 10, 2009. As in previous years, all Maine public high schools were designated as SAT test centers. In two cases, schools opted to send students to nearby high schools/test centers.

Students who were approved for accommodations received the same accommodations on all components of the MHSA, as explained in Chapter 7. Details about the administration of the Math–A and science components and the SAT were communicated to schools on an ongoing basis through informational letters, the MDOE Web site, and ATM broadcast transmissions. Additionally, workshops were held throughout the state on all aspects of accommodations, the registration process, SAT test center supervisor training, and Math–A and science administration training.

After the May and June SAT administrations, students testing under standard conditions or with College Board approved accommodations received official SAT score reports from the College Board. Additionally, **all** students participating in the MHSA received individual score reports based on Maine’s achievement levels. The MHSA scores were then used for accountability purposes.

The three components of the 2009 MHSA (SAT, Math–A, and science) comprised a cohesive system with comparable item development, administration, and scoring protocols; similar test material formats; the same accommodations; and a seamless reporting system. Collaboration between the MDOE, the College Board, and Measured Progress assured that the entire process worked smoothly. This illustration has been used in public presentations to communicate the relationship between the SAT and the complete MHSA program.



SECTION I—MHSA DEVELOPMENT

Chapter 2. DEVELOPMENT AND DESIGN OF THE MHSA: SAT AND MATH–A

Beginning in the 2006–2007 academic year, the MHSA featured two components, the SAT and the mathematics augmentation (Math–A). Details on the content specifications and development of both components are featured in this chapter.

External alignment studies have indicated that the English language arts components of the SAT are reasonably aligned and adequately measure Maine’s *Learning Results*, while the mathematics component needs augmentation. The MHSA is intended to support good educational practice and is perceived as having an impact on instruction and curriculum. The SAT Committee, composed of teachers, academic administrators, measurement experts, admissions officers, college counselors, and students, provides the College Board with advice on any of the policies, practices, products, and services involving the SAT. In addition, the development of each of the three content areas on the SAT (mathematics, critical reading, and writing) is guided by the work of a test development committee composed of both secondary school and college teachers in that content area. The involvement of these development committees will be identified in the discussion of the test development process below. The current members of these committees can be found at www.collegeboard.com. The Math–A Committee is comprised of representatives from the MDOE, the College Board, and teachers and faculty from schools in the state of Maine. A list of committee members is included later in this chapter.

2.1 The MHSA: The SAT and Math–A Overview

Detailed content and statistical specifications for each of the three content areas define the parameters that ensure that each new form is comparable to all other forms of the SAT. That is, the detailed test specifications and statistical procedures ensure that different forms of the same test developed both within each academic year and across years are parallel in content and difficulty. These design features, plus SAT equating procedures, enable comparability of scores from different test administrations. For example, Maine scores from the May 2008 administration of the SAT can be directly compared with scores from the May 2009 administration. The MHSA designates the May and June (make-up only) SAT administration dates for state assessment purposes. Scores from these administrations can also be directly compared. The specifications for both the content and the psychometric characteristics of each test are provided later in this chapter. Examples of each type of question used on the test may be found at www.collegeboard.com.

2.2 Universal Design Specifications

The SAT and Math–A components of the MHSA are developed according to the following six principles of universal design defined by Thompson, Johnstone, and Thurlow (2002)

1. Inclusive assessment population—The MHSA provides assessment opportunities for all students, regardless of their cognitive abilities, cultural backgrounds, or linguistic backgrounds.
2. Precisely defined constructs—The MHSA measures the constructs it is intended to measure and does not measure irrelevant material.
3. Accessible, non-biased items—The MHSA uses appropriate accommodations to “level the playing field” for students with disabilities. These accommodations do not affect the validity of the assessments or the comparability of scores obtained on them.
4. Simple, clear, and intuitive instructions and procedures—The MHSA instructions are easy to understand regardless of a student’s experience, knowledge, language skills, or current concentration level. In addition, test development committees review SAT instructions to ensure that they are appropriate for the test-taking population.
5. Maximum readability and comprehensibility—SAT mathematics items including Math–A items, are developed with the minimal number of required words and the least amount of grammatical complexity for the task. For the critical reading and writing items, the level of readability and syntax is appropriate for the construct that is being measured by those items. Readability is part of the thorough review by content experts before and after the pretesting of items.
6. Maximum legibility—The text, tables, and figures that appear on the MHSA are designed to ensure maximum legibility. In the mathematics sections, figures that accompany problems are intended to provide information useful in solving the problems. All figures are drawn to scale unless otherwise indicated.

2.3 SAT Critical Reading Test

The May and June 2009 forms required by the MHSA, like all forms of the SAT critical reading test, met the specifications presented in Table 2-1.

**Table 2-1. 2008–09 MHSА: SAT
Critical Reading Content Specifications**

	<i>Number</i>	<i>Percentage of Test</i>
Time allotted	70 minutes	
Sentence completion	19 items	28
Passage based reading	48 items	72
Total	67 items	100
800 word passages*	2 passages	
650 word passages*	1 passage	
500 word passages*	1 passage	
Paragraph reading	2 passages	
Paired paragraph	1 pair	
Extended reasoning	36–40 items	54–60
Literal comprehension	4–6 items	6–9
Vocabulary in context	4–6 items	6–9

*Note: One of the long passages will actually be a pair of related passages (e.g., instead of an 800-word passage, there will be two related 400-word passages, etc.)

Each new form of the SAT critical reading test will continue to meet the listed specifications.

The passage based reading content is balanced across four categories: humanities, social studies, natural sciences, and literary fiction. Male and female references are balanced across the test. Representative minority relevant content is included. Approximately 80% of the passage based reading content (60% of the total test) measures extended reasoning skills through questions about primary purpose, rhetorical strategies, implication and evaluation, tone and attitude, application and analogy; the balance of the questions are concerned with literal comprehension or vocabulary in context. The three separately timed sections of a typical SAT critical reading test are configured as shown in Table 2-2.

An important constraint in the development of multiple parallel forms of a test is that the distribution of item difficulties be the same across forms. Using the equated delta¹ index, each SAT critical reading test must have questions with the distribution of difficulty indicated in Table 2-3.

Table 2-2. 2008–09 MHSА: SAT Critical Reading Section Configuration

<i>Reading 1 (25 minutes)</i>	<i>Reading 2 (25 minutes)</i>	<i>Reading 3 (20 minutes)</i>
Items 1–8: Sentence completion items (8)	Items 1–5: Sentence completion items (5)	Items 1–6: Sentence completion items (6)
Items 9–12: Either two paragraph reading passages with two items each OR one paired paragraph with four items (4)	Items 6–9: Either two paragraph reading passages with two items each OR one paired paragraph with four items (4)	Items 7–19: One 800-word passage with 13 items
Items 13–24: One 800-word passage with 12 items	Items 10–24: One 500-word passage and one 650-word passage with a total of 15 items	

Note: The actual number of passage-based reading questions in each section may vary by one or two, but the total number in each critical reading test will always be 48.

¹ Described more fully in Chapter 12, equated delta is a transformation of 1–p, with a mean of 13 and standard deviation of 4.

**Table 2-3. 2008–09 MHSA: SAT
Critical Reading Psychometric Specifications**

	<i>Item Type Difficulty</i>	
	<i>Sentence Completion</i>	<i>Passage-Based Reading</i>
<i>Mean equated delta by item type</i>	10.4–12.4	10.4–12.4
<i>Equated Delta Distribution for the Overall Test</i>		
Mean equated delta (SD)	11.4 (2.4)	
<i>Number and Percentage of Items by Delta Value</i>		
	DV	N (%)
	16	1 (1.5)
	15	4 (6.0)
	14	6 (9.0)
	13	7 (10.4)
	12	9 (13.4)
	11	12 (17.9)
	10	9 (13.4)
	9	7 (10.4)
	8	6 (9.0)
	7	4 (6.0)
	6	2 (3.0)
	Total	67 (100)

Note: The equated delta distribution, mean and standard deviation is provided for the overall reading test, while the equated delta mean is provided for the two item types. It is not necessary to specify the standard deviation of the mean equated delta by item type because the reading test is assembled to meet the overall point by point delta distribution.

2.4 SAT Writing Test

Although Maine does not use writing as an adequate yearly progress (AYP) measure for accountability under NCLB, Maine includes writing in its assessment system. The May and June 2009 forms required by the MHSA, like all forms of the SAT writing test, met the specifications presented in Table 2-4:

**Table 2-4. 2008–09 MHSA: SAT
Writing Content Specifications**

<i>Time Allotted-60 minutes</i>	<i>Number</i>	<i>Percent of MC Portion</i>
Improving sentences (sentence correction)	25 items	51
Identifying sentence errors (usage)	18 items	37
Improving paragraphs (revision in context)	6 items based on a passage*	12
Total	49 items	100
Essay (25 minutes)	1 essay	

* Passages can range from 150 to 250 words.
MC = multiple-choice

Each new form of the SAT writing test will continue to meet the listed specifications.

The essay portion of the test requires students to write an original first draft of an essay in which they develop a point of view on an issue that has been presented through a prompt. The prompt is written to be easily accessible to the general test taking population, including students for whom English is a second

language, and is free of figurative or technical language or specific literary references. The prompt presents an issue that engages students of high school age and allows them to draw on their knowledge and interests to respond. The prompt outlines a range of possible viewpoints within a single issue, and stimulates critical reflection on the issue. Following the prompt is an assignment that focuses the student on the issues addressed in the prompt. The essay is scored by trained readers using the essay scoring guide, displayed as Figure 2-1.

ESSAY SCORING GUIDE

The Scoring Guide expresses the criteria readers use to evaluate and score the student essays. The Guide is structured on a six-point scale. The language of the Scoring Guide provides a consistent and coherent framework for differentiating between score points, without defining specific traits or types of essays that define each score point.

Score of 6

An essay in this category demonstrates **clear and consistent mastery**, although it may have a few minor errors. A typical essay

- effectively and insightfully develops a point of view on the issue and demonstrates outstanding critical thinking, using clearly appropriate examples, reasons, and other evidence to support its position
- is well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas
- exhibits skillful use of language, using a varied, accurate, and apt vocabulary
- demonstrates meaningful variety in sentence structure
- is free of most errors in grammar, usage, and mechanics

Score of 5

An essay in this category demonstrates **reasonably consistent mastery**, although it will have occasional errors or lapses in quality. A typical essay

- effectively develops a point of view on the issue and demonstrates strong critical thinking, generally using appropriate examples, reasons, and other evidence to support its position
- is well organized and focused, demonstrating coherence and progression of ideas
- exhibits facility in the use of language, using appropriate vocabulary
- demonstrates variety in sentence structure
- is generally free of most errors in grammar, usage, and mechanics

Score of 4

An essay in this category demonstrates **adequate mastery**, although it will have lapses in quality.

A typical essay

- develops a point of view on the issue and demonstrates competent critical thinking, using adequate examples, reasons, and other evidence to support its position
- is generally organized and focused, demonstrating some coherence and progression of ideas
- exhibits adequate but inconsistent facility in the use of language, using generally appropriate vocabulary
- demonstrates some variety in sentence structure
- has some errors in grammar, usage, and mechanics

Score of 3

An essay in this category demonstrates **developing mastery**, and is marked by **one or more** of the following weaknesses:

- develops a point of view on the issue, demonstrating some critical thinking, but may do so inconsistently or use inadequate examples, reasons, or other evidence to support its position
- is limited in its organization or focus, or may demonstrate some lapses in coherence or progression of ideas
- displays developing facility in the use of language, but sometimes uses weak vocabulary or inappropriate word choice
- lacks variety or demonstrates problems in sentence structure
- contains an accumulation of errors in grammar, usage, and mechanics

Score of 2

An essay in this category demonstrates **little mastery**, and is flawed by **one or more** of the following weaknesses:

- develops a point of view on the issue that is vague or seriously limited, and demonstrates weak critical thinking, providing inappropriate or insufficient examples, reasons, or other evidence to support its position
- is poorly organized and/or focused, or demonstrates serious problems with coherence or progression of ideas
- displays very little facility in the use of language, using very limited vocabulary or incorrect word choice

- demonstrates frequent problems in sentence structure
- contains errors in grammar, usage, and mechanics so serious that meaning is somewhat obscured

Score of 1

An essay in this category demonstrates **very little** or **no mastery**, and is severely flawed by **one or more** of the following weaknesses:

- develops no viable point of view on the issue, or provides little or no evidence to support its position
- is disorganized or unfocused, resulting in a disjointed or incoherent essay
- displays fundamental errors in vocabulary
- demonstrates severe flaws in sentence structure
- contains pervasive errors in grammar, usage, or mechanics that persistently interfere with meaning

Score of 0

Essays not written on the essay assignment will receive a score of zero.

Figure 2-1. 2008–09 MHSA: Essay Scoring Guide

As illustrated in Table 2-5, writing process skills are assessed through both the improving paragraphs item type and through the essay that each student writes.

Table 2-5. 2008–09 MHSA: Alignment Between Writing Process Skills and SAT Writing Questions

<i>Writing Process Skill</i>	<i>Essay Prompt</i>	<i>Improving Paragraphs</i>
Writing personal narratives	X	
Using literal and figurative language appropriately	X	X
Using sentence variety	X	X
Demonstrating insight and/or creativity in the writing task	X	
Using topic sentences	X	X
Using appropriate voice, tone, and style	X	X
Focusing on a purpose for writing	X	
Writing persuasive and/or argumentative essays	X	
Organizing paragraphs and using appropriate transitions	X	X
Writing effective introductions and conclusions	X	X
Using writing and reading as tools for critical thinking	X	
Developing a logical argument	X	
Writing a unified essay	X	X
Using supporting details and examples	X	

Writing a clear and coherent essay	X	X
------------------------------------	---	---

The multiple-choice writing questions test a wide range of grammatical, usage, and sentence structure skills as shown in Table 2-6.

Table 2-6. 2008–09 MHSA: Alignment Between Grammar, Usage, and Sentence Structure Skills and the Problems Tested by SAT Writing Questions

<i>Grammar, Usage, and Sentence Structure Skills</i>	<i>Improving Sentences</i>	<i>Identifying Sentence Errors</i>	<i>Improving Paragraphs</i>
Avoiding faulty predication in sentences	X	X	X
Avoiding dangling modifiers	X		
Using comparative modifiers appropriately	X	X	
Using appropriate idiomatic words, phrases, or structures	X	X	X
Avoiding weak, passive constructions	X		
Using connectives appropriately	X	X	X
Avoiding illogical comparisons	X	X	
Subordinating and coordinating ideas in sentences	X	X	X
Avoiding pronoun shift	X	X	X
Combining sentences appropriately			X
Maintaining parallel structure in sentences	X	X	X
Using appropriate verb forms	X	X	X
Avoiding wordiness	X	X	X
			continued
Controlling errors in subject-verb agreement	X	X	
Avoiding errors in pronoun agreement, case, and reference	X	X	
Maintaining tense sequences	X	X	X
Making acceptable word choices	X	X	X
Avoiding run-on sentences	X		
Avoiding sentence fragments	X		X
Avoiding comma splices	X		X

The SAT writing test is administered in three separately timed sections as configured in Table 2-7.

Table 2-7. 2008–09 MHSA: SAT Writing Section Configuration

<i>Writing 1 (25 minutes)</i>	<i>Writing 2 (25 minutes)</i>	<i>Writing 3 (10 minutes)</i>
	Items 1–11: Improving sentence items (11)	
Essay	Items 12–29: Identifying sentence errors items (18)	Items 1–14: Improving sentences items (14)
	Items 30–35: Improving paragraphs (6)	

Multiple-choice items are spread across a variety of content areas, including science, practical affairs, human relations, geography, literature, art, legal, education, business, and history. Female and male references are balanced, and representative minority-relevant content is included.

In order to develop multiple parallel forms of a test, the distribution of item difficulties must be the same across forms. Using the equated delta index, each portion of the SAT writing multiple-choice portion of the test must have questions with the distribution of difficulty indicated in Table 2-8.

Table 2-8. 2008–09 MHSA: SAT Writing Psychometric Specifications

<i>Equated Delta Distribution for the Multiple-choice Portion of the Test</i>	
Mean equated delta (SD)	10.1 (2.5)
<i>Number and Percentage of Items by Delta Value</i>	
DV	N (%)
16	1 (2.0)
15	0 (0.0)
14	2 (4.1)
13	3 (6.1)
12	6 (12.2)
11	7 (14.3)
10	7 (14.3)
9	7 (14.3)
8	6 (12.2)
7	5 (10.2)
6	3 (6.1)
5	2 (4.1)
Total	49 (100)

2.5 MHSA Mathematics Test: SAT and Math–A Components

The MHSA mathematics test consists of two components: the traditional SAT mathematics test and Math–A. The following includes information related to the development of both portions for the 2008–09 administration. The content specifications for the SAT component stay relatively stable from year to year, with only slight differences due to a range of acceptable numbers of items measuring particular content specifications and routine variability as to whether the test form fell on the upper or lower end of the acceptable range. For a small number of content specifications, an item measuring that content may or may not be included on every form.

The Math–A portion of the MHSA is built each year after careful committee review of the operational SAT form to be administered in Maine for NCLB during May of that year. The Math–A test form includes content from the Maine *Learning Results* not routinely found in the SAT mathematics section, and any content that is determined to be underrepresented on that form due to the allowed variability in the acceptable number of items for any one SAT form, as described above.

The May and June 2009 SAT forms required by the MHSA, like all forms of the SAT mathematics test, met the specifications presented in Table 2-9.

Table 2-9. 2008–09 MHSA: SAT Mathematics Content Specifications

Time Allotted: 70 minutes	Number	Percent of Test
Multiple-choice	44 items	81
Student-produced-response	10 items	19
Total	54 items	
Number and Operations	11–13 items	20–24
Algebra and Functions	19–21 items	35–39
Geometry and Measurement	14–16 items	26–30
Data Analysis, Statistics, and Probability	6–7 items	11–13

Each new form of the SAT mathematics test will continue to meet the listed specifications listed. The four content areas specified in Table 2-9 are further defined in Table 2-10.

Table 2-10. 2008–09 MHSA: SAT Mathematics Content Description

<i>Number and Operations</i>
<ul style="list-style-type: none"> • Arithmetic word problems (including percent, ratio, and proportion) • Properties of integers (odd/even, prime numbers, divisibility, and so forth) • Rational numbers • Logical reasoning • Sets (union, intersection, elements) • Counting techniques • Sequences and series (including exponential growth) • Elementary number theory
<i>Algebra and Functions</i>
<ul style="list-style-type: none"> • Substitution and simplifying algebraic expressions • Properties of exponents • Algebraic word problems • Solutions of linear equations and inequalities • Systems of equations and inequalities • Quadratic equations • Rational and radical equations • Equations of lines • Absolute values • Direct and inverse variation • Concepts of algebraic functions • Newly defined symbols based on commonly used operations
<i>Geometry and Measurement</i>
<ul style="list-style-type: none"> • Area and perimeter of a polygon • Area and circumference of a circle • Volume of a box, cube, and cylinder • Pythagorean Theorem and special properties of isosceles, equilateral, and right triangles • Properties of parallel and perpendicular lines • Coordinate geometry • Geometric visualization • Slope • Similarity • Transformations

continued

- Data interpretation
- Descriptive statistics (mean, median, mode)
- Probability

The three separately timed SAT mathematics sections are configured as follows:

**Table 2-11. 2008–09 MHSА: SAT
Mathematics Section Configuration**

<i>Mathematics 1 (25 minutes)</i>	<i>Mathematics 2 (25 minutes)</i>	<i>Mathematics 3 (20 minutes)</i>
Items 1–20: Multiple-choice (20)	Items 1–8: Multiple-choice (8) Items 9–18: Student-produced- response (10)	Items 1–16: Multiple-choice (16)

Calculators are permitted on the SAT mathematics test, and basic geometric reference information is provided at the top of each separately timed section. Additional information on the calculator policy for the SAT is provided in Chapter 7.

In order to develop multiple parallel forms of a test, the distribution of item difficulties must be the same across forms. Using the equated delta index, each SAT mathematics section must have questions with the distribution of difficulty indicated in Table 2-12.

**Table 2-12. 2008–09 MHSА: SAT
Mathematics Psychometric Specifications**

		<u>Item Type Difficulty</u>	
		MC	SPR
<i>Mean equated delta (SD)</i>		12.2 (3.2)	13.6–14.2 (3)
<u>Number of Items by Delta Value and Item Type</u>			
MC		SPR	
18–20	1	18–20	1
17	2	16–17	2
16	2		
15	4	14–15	2
14	5		
13	5	12–13	2
12	5		
11	5	10–11	2
10	4		
9	3	8–9	1
8	3		
7	2		
6	2	<6–7	0
<6	1		
Total	44	Total	10

MC = multiple-choice; SPR = student-produced-response, SD = standard deviation

Table 2-13 summarizes the content of the 2009 Math–A form. The 2009 Math–A form consisted of 18 multiple-choice items—12 operational and 6 field test. In a given administration the number of forms of

the Math–A test will vary according to the need for field testing. However, all forms contain the same 12 operational items and differ only in the 6 field test items that are presented. The development procedure described in this chapter, except the separate pretest, occurs each year to determine if any gaps in mathematics coverage exist in the SAT forms that must be addressed to align with the Maine *Learning Results*. The subcontent codes referenced in Table 2-13 are further explained in Table 2-14.

**Table 2-13. 2008–09 MHSA:
Item Content and Difficulty of the Math–A**

<i>Item*</i>	<i>ACC#</i>	<i>Key</i>	<i>Classification</i>	<i>EQΔ</i>
1	ROC026	B	Data, Measurement	0.55
2	ROC031	E	Data, Measurement	0.81
3	ROC027	B	Data, Measurement	0.79
4	ROC006	B	Geometry, geometric figures	0.68
5	ROC012	C	Data, Measurement	0.56
6	ROC028	D	Data, Measurement	0.44
7	ADS030	A	Geometry, geometric figures	0.49
8	ADS040	C	Algebra	0.2
9	ADS045	E	Data, Measurement	0.49
10	ADS031	D	Geometry, geometric figures	0.47
11	ADS047	A	Data, Measurement	0.51
12**	ADS004	C	Algebra	0.22
			Mean	0.52
			SD	0.19

ACC = accession number, the College Board code for item id.; EQ Δ = equated delta; SD = standard deviation

*Reflects order of administration but not exact test placement due to the omission of the 6 field test items on each form.

**Item 12 was found to be flawed and was not included in any student results. Thus, the total raw score points possible changed from 66 to 65.

**Table 2-14. 2008–09 MHSAs:
Maine Learning Results for Mathematics**

Level	Content Description
A	<p>NUMBER: Students use numbers in everyday and mathematical contexts to quantify or describe phenomena, develop concepts of operations with different types of numbers, use the structure and properties of numbers with operations to <i>solve</i> problems, and perform mathematical computations. Students develop number sense related to magnitude, estimation, and the effects of mathematical operations on different types of numbers. It is expected that students use numbers flexibly, using forms of numbers that best match a situation. Students compute efficiently and accurately. <i>Estimation</i> should always be used when computing with numbers or solving problems.</p> <ul style="list-style-type: none"> ▪ Whole Number—No Performance Indicator ▪ Rational Number—No Performance Indicator ▪ Real Number—Students know how to represent and use real numbers.
B	<p>DATA: Students make measurements and collect, display, evaluate, analyze, and compute with data to describe or <i>model</i> phenomena and to make decisions based on data. Students compute statistics to summarize data sets and use concepts of probability to make predictions and describe the uncertainty inherent in data collection and measurement. It is expected that when working with measurements students <i>understand</i></p> <ul style="list-style-type: none"> ▪ that most measurements are approximations and that taking repeated measurements reveals this variability; ▪ that a number without a <i>unit</i> is not a measurement, and that an appropriate <i>unit</i> must always be attached to a number to provide a measurement; ▪ that the <i>precision</i> and <i>accuracy</i> of a measurement depends on selecting the appropriate tools and <i>units</i>; and use <i>estimation</i> comparing measures to <i>benchmarks</i> appropriate to the type of measure and <i>units</i>. ▪ Measurement and Approximation <ul style="list-style-type: none"> B1. Students <i>understand</i> the relationship between <i>precision</i> and <i>accuracy</i>. ▪ Data Analysis <ul style="list-style-type: none"> B2. Students <i>understand</i> correlation and cause and effect. B3. Students <i>understand</i> and know how to describe distributions and find and use descriptive statistics for a set of data. B4. Students <i>understand</i> that the purpose of random sampling is to reduce bias when creating a representative sample for a set of data. ▪ Probability <ul style="list-style-type: none"> B5. Students <i>understand</i> the relationship of probability to relative frequency and know how to find the probability of compound events
C	<p>GEOMETRY: Students use measurement and observation to describe objects based on their sizes and shapes; <i>model</i> or construct two-dimensional and three-dimensional objects; <i>solve</i> problems involving geometric properties; compute areas and volumes based on object properties and dimensions; and perform transformations on geometric figures. When making or calculating measures students use <i>estimation</i> to check the reasonableness of results.</p> <ul style="list-style-type: none"> ▪ Geometric Figures <ul style="list-style-type: none"> C1. Students <i>justify</i> statements about polygons and <i>solve</i> problems. C2. Students <i>justify</i> statements about circles and <i>solve</i> problems. C3. Students <i>understand</i> and use basic ideas of trigonometry. ▪ Geometric Measurement <ul style="list-style-type: none"> C4. Students find the surface area and volume of three-dimensional objects. ▪ Transformations—No Performance Indicator

continued

Level	Content Description
D	<p>ALGEBRA: Students use symbols to represent or <i>model</i> quantities, patterns, and relationships and use symbolic manipulation to <i>evaluate</i> expressions and <i>solve</i> equations. Students <i>solve</i> problems using symbols, tables, graphs, and verbal rules choosing the most effective representation and converting among representations.</p> <ul style="list-style-type: none"> ▪ Symbols and Expressions D1. Students <i>understand</i> and use polynomials and expressions with rational exponents. ▪ Equations and Inequalities D2. Students <i>solve</i> families of equations and inequalities. D3. Students <i>understand</i> and apply ideas of logarithms. ▪ Functions and Relations D4. Students <i>understand</i> and <i>interpret</i> the characteristics of functions using graphs, tables, and algebraic techniques. D5. Students express relationships <i>recursively</i> and use <i>iterative</i> methods to <i>solve</i> problems.

2.6 Development

Each new form of the MHSAs is developed through a multistage process that spans many months. The basic steps are similar for each of the three content areas (mathematics, critical reading, and writing), although the details of the process may vary somewhat among these three. Significant variations will be noted here as appropriate. The development process draws on the skills of content experts, psychometricians, and experienced educators in order to repeatedly develop new forms that are parallel, fair to students, and test the reasoning skills important to academic success in college. Current members of the SAT Test Development Committees and their respective affiliations are provided in Table 2-15.

Table 2-15. 2008–09 MHSAs: SAT Test Development Committee Members

Name	Affiliation
SAT Mathematics Committee	
Daniel Lotesto*	Riverside High School
Deborah Hughes Hallett	Harvard University
Brendan Murphy	John Bapst Memorial High School
Monica Stephens Cooley	Spelman College
Roxy Peck	California State Polytechnic University
Hung-Hsi Wu	University of California, Berkeley
J.T. Sutcliffe	St. Mark's School of Texas
Maria Randle	Bishop Kenny High School
MHSAs Mathematics Augmentation Committee	
Robin Callaghan	The College Board
Andrew Schwartz	The College Board
Daniel Hupp	MDOE

continued

<i>Name</i>	<i>Affiliation</i>
SAT Reading Committee	
Lance Balla*	Bellevue School District No. 405
Jacqueline Brice-Finch	Coppin State University
Eva Rodriguez Arce	James Bowie High School
Renee Shea	Bowie State University
Simone Waite	Cypress Bay High School
Guiyou Huang	St. Thomas University
Kevin Dettmar	Pomona College
SAT Writing Committee	
Noreen Duncan*	Mercer County Community College
Bernard Phelan	Homewood-Flossmoor High School
Veleeder Flythe	Heritage High School
Ed Coleman	North Central High School
Greg Hamilton	Marin Teaching Network
Sue Ham	University of Texas
Rosemarie Mundy-Shephard	Albany State University
Luis Torres	Metropolitan State College of Denver

*Chairperson

2.7 Item Writing and Review

Test development specialists at Educational Testing Service (ETS) write the test items for the SAT component and content specialists at the College Board write the items for the Math–A component. Some of the items are based on ideas from high school and college faculty and other qualified consultants. Faculty and consultants are selected for their knowledge of curriculum and for their expertise in a field. In general, the staff who work on a particular test are content specialists who have either high school or college teaching experience. In writing items, these people are guided by the content and statistical specifications for the particular portion of the MHSA (mathematics, critical reading, or writing) on which they are working.

Because such a high proportion of the questions on the critical reading test are tied to a reading passage, potential reading passages are first chosen and reviewed for suitability before any passage-based items are written.

Each newly written item (or set of items) is classified according to the appropriate category of the specifications. It is reviewed to maximize clarity and to eliminate ambiguity. It is further reviewed for sensitivity to members of gender and racial or ethnic subgroups. Each item is also examined to make sure that it has only a single correct answer. The student-produced-response items in mathematics may have more than one possible answer or more than one way to express the answer (see Chapter 6 for more information on student-produced-response items). During the review process, items may be discarded, accepted, or revised to eliminate ambiguity, improve wording, strengthen the correct answers, and so forth.

2.8 Pre-testing the Items

Every item used in an operational form of the MHSA SAT and Math–A components has been pretested; that is, the item has been tried out with an appropriate group of students to make sure that it is not ambiguous or confusing and to determine the difficulty level and the degree to which it differentiates more or less able students. The pretest responses are also analyzed to determine whether students of different racial/ethnic or gender groups respond to the question differently. MHSA SAT and Math–A item writing and review are ongoing activities throughout the year.

The multiple-choice items of the SAT (mathematics, critical reading, and writing), as well as the student-produced-response mathematics items, are pretested on a sample of actual SAT test takers. There are 10 separately timed sections in each SAT: three for the writing test, three for the critical reading test, and three for the mathematics test; the remaining section does not count toward the student’s score and is used either for pretesting, for providing calibration information for the equating of test scores, or for research. Pretests, each configured like one of the operational sections, are assembled from questions that have received a number of content, fairness, and editorial reviews prior to pretesting. Each pretest is administered as the unscored section of some fraction of all SATs administered on a particular date; that is, every n th test book will have a particular pretest or equating test in that unscored section. This pattern of administration provides item information on a large random sample of SAT test takers. Consequently, this item information provides an extremely accurate estimate of how the item will function when administered as part of a future SAT. The Math–A pretest items are embedded in the Math–A form administered operationally to students in Maine each year.

Each SAT writing essay prompt is reviewed by SAT staff at both the College Board and ETS. After all concerns raised during the review process are resolved, the essay prompt is pretested in a special administration in high school English classrooms. For each group of pretests, a diverse sample of schools is invited to participate by having students respond to a particular prompt during their English class. A sample of at least 300 responses to each essay prompt is obtained in order to determine whether the question is accessible to students and to provide exemplars of various levels of writing competence for use in the scoring process, described in Chapter 8.

2.9 Analysis of Pretest Information for the MHSA: SAT and Math–A

Data collected from multiple-choice and student-produced-response pretests are analyzed to provide important information about the appropriateness of items for use in operational forms of the SAT. Three statistical indices are computed: **equated delta** as an index of item difficulty within the SAT population, **r -biserial** as an index of whether the item discriminates between more and less able students, and Mantel-Haenszel **DIF** (differential item functioning) as an index of the relationship between group membership and the likelihood of answering the question correctly. These item statistics are used to judge whether a given

question is suitable for inclusion in the pool of items from which operational forms are assembled. The item statistics may also reveal problems with the conceptualization or wording of a question. Some of these items are revised and re-pretested. Others are discarded. SAT items are analyzed by ETS using data from the national administration of the test form. Math–A items are analyzed separately using the same procedures as followed for the SAT items. Since year two, the Math–A sample has included only students from Maine.

The statistical indices employed in analyzing and screening MHSA SAT and Math–A components follow. This section covers the procedures used and refers heavily to the SAT specifically; however, the same procedures are applied for the Math–A items where appropriate. Actual results of these analyses are presented in Chapter 11.

2.10 Item Difficulty

The difficulty of an item is a function of the percentage of test takers who answer it correctly (i.e., p -value). An item's difficulty should be appropriate for the population taking the test. When an item is too easy, virtually all test takers answer it correctly; thus, extremely easy items contribute very little information to the total test score. Similarly, inappropriately difficult items are not very useful in a test. Because items within a test are highly intercorrelated, it is best to select items with a moderate spread of difficulty around a mean p -value of 0.5 (or 50% correct). The required distribution of item difficulty for each part of the SAT is defined in the psychometric specifications found in Tables 2-3, 2-8, and 2-12.

Typically, p -values are converted to a standard scale that avoids negative values and decimals (Anastasi, 1976). The measure of difficulty used with the SAT is the delta index (Δ). This index is based on the percentage of test takers answering a given item correctly, where 1 minus the p -values are converted to z -scores and transformed to a scale with a mean of 13 and a standard deviation of 4. The delta scale is inversely related to the p -scale; thus, the more difficult the item, the greater the delta value and the smaller the p -value.

Because the samples to which specific pretest items are administered in a nonscored section may, to some degree, differ in ability level from the 1990 standard reference group used for the SAT, it is necessary to convert the raw delta values to equated delta values. To make this conversion, data from items in the scored sections is used since the equated delta values for all of those items are known. The raw delta values for the common items based on the current sample are then plotted against the known equated deltas from the previous equating. The resulting linear relationship between the pairs of raw and equated deltas is used to compute an equated delta for the new pretest item. An equated delta value is computed for each pretest item and is based on the standard reference population, permitting comparisons of items among samples (Thurstone, 1947).

Each form of the SAT is built to a well-defined distribution of item difficulty. While formula scored items include a correction for guessing, the delta scale (based on percentage correct) does not adjust for incorrect responses. As a result, the proportion-correct delta scale provides an estimate of difficulty that is

slightly lower than it would be if the formula scoring were taken into account. This is not a problem for the reading test and the multiple-choice portion of the writing test. All items in these sections are formula scored with the same amount subtracted ($\frac{1}{4}$ of a point) for an incorrect response (i.e., the k -factor is 0.25), and the statistical specifications have been designed to reflect this known difference. The mathematics section, however, contains both formula scored multiple-choice items (with a k -factor of 0.25) and student-produced-response items that do not penalize incorrect responses. For this reason, as shown in Table 2-12, psychometric specifications for the SAT mathematics test provide separate delta distributions for multiple-choice and student-produced-response items. For more detail on how statistical specifications were set for the SAT, see Lawrence and Schmitt (1994).

2.11 Item Discrimination and Item/Test Relationship

Although difficulty level is one important criterion in selecting items, item discrimination is essential to be able to distinguish among test takers at different levels of ability. The r -biserial correlation coefficient between the item and total test score is most often used to assess the item's utility in discriminating among test takers of differing ability levels and the homogeneity of test items (or extent that a student's performance on an item relates to his/her total test score). The biserial correlation ranges from 1 to -1. The more positive the correlation, the more the item distinguishes test takers with high total scores from those with low scores. A negative biserial correlation indicates that the item is measuring something different from the rest of the test; test takers with high scores are more likely to answer that item incorrectly than those with low scores. Correlations that are near 0 indicate that high scorers and low scorers have the same chance of correctly answering the item. Because of these results, the MHSA does not include items with low or negative biserial correlations.

Biserial correlations also provide an indication of the homogeneity of test items. If the correlation is very close to 1, all of the information provided by the item is redundant with that provided by the other test items. Items with moderate biserial correlations distinguish among ability levels, yet also supply unique information. Therefore, most items included on SAT operational forms fall within a biserial range of 0.30 to 0.80.

In determining whether to select, omit, or edit and refine an item based on results from pretests, test developers also consider the number and percentage of test takers who respond to the correct option and to each incorrect option (with all items on the SAT except student-produced responses and the essay). At each score level, the percentage of test takers selecting each option is plotted. For a correct option, it is expected that the percentage of students selecting the option will increase as the test score increases. Figure 2-2 displays an item with this increasing pattern. If the correct option does not display this pattern, the item is carefully reviewed. Similarly, if an incorrect option has this typical increasing pattern, then that option is closely evaluated. As a result of the evaluation, the item may be revised and then re-pretested, or it may be discarded entirely.

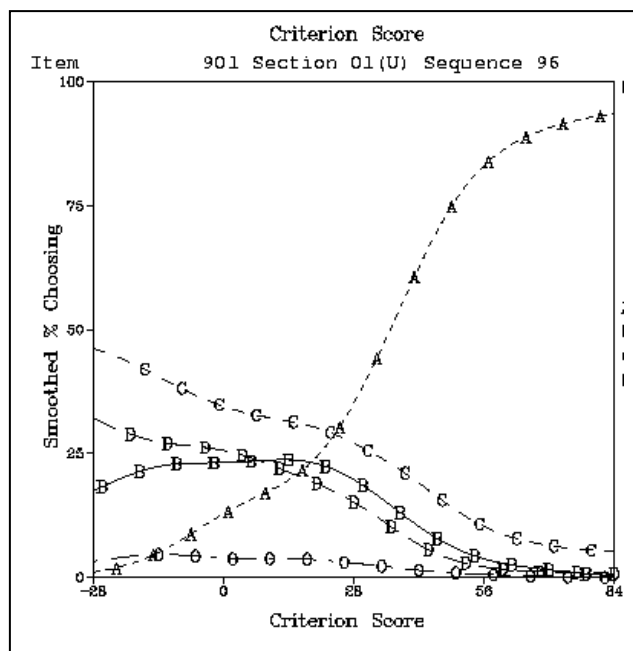


Figure 2-2. 2008–09 MHSAs Typical Discrimination Pattern Among Multiple-choice Response Options, Where Option A is the Key

2.12 Differential Item Functioning

Analyses of differential item functioning (DIF) are conducted to identify items that may function differently for members of different groups. DIF analyses compare the performance of two groups of test takers (e.g., males versus females, Asian American test takers versus White test takers) who have been matched on their reading, writing, or mathematical proficiency (SAT mathematics, critical reading, or writing total score²) on each item. The underlying assumption in conducting such analyses is that all test takers demonstrating the same level of proficiency in the content area should have similar chances of answering each item correctly regardless of gender, race, or ethnicity. DIF occurs when individuals with similar scores on the SAT critical reading, SAT writing (multiple-choice), or SAT mathematics tests differ notably in their performance on a specific test item (Crone and Schmitt, 1991). The presence of DIF indicates that an item functions differently for one subgroup than for another subgroup of the same proficiency. While the theoretical framework for explaining DIF is not yet well established, the assumption is that items exhibiting high levels of DIF may be measuring factors irrelevant to a test (such as culture) or more than one dimension for which the two groups have different strengths. For example, DIF may result from a mathematical word problem because the question measures language proficiency in addition to mathematical reasoning. One

² Groups of test takers are matched on some criterion that reflects the underlying dimension or construct of interest (e.g., critical reading, mathematical reasoning). Typically this “matching criterion” is total score on the relevant part of the SAT. However, the criterion may vary with the intent of the study. For example, in examining DIF associated with student-produced-response items on the SAT mathematics test, Lawrence, Lyu, and Feigenbaum (1995) used the raw score on 25 quantitative comparison items (an external matching criterion because it did not include the student-produced-response items under study) and the total raw score for SAT mathematics (an internal matching criterion because it included the student-produced-response items under study).

group of test takers may well be stronger in language proficiency. An item like this would be reviewed by one or more experts who have not been involved with the item and who are trained with respect to the construct being tested and item sensitivity. The experts would determine whether the amount of language proficiency required by the item is irrelevant to the dimension of interest, that is, mathematical reasoning.

DIF analyses begin by examining any differences in the performance on each individual item of two comparable groups, referred to as the reference group and the focal group. Typically, DIF analyses for the SAT compare groups based on gender (where males are the reference group and females are the focal group) or ethnicity/race (where White test takers are the reference group and African American, Hispanic, Asian American, or Native American test takers are the focal group). Occasionally DIF analyses are conducted with other groups (e.g., students with disabilities and those without disabilities; students for whom English is a second language (ESL) and non-ESL students). Items having extreme values of DIF—those items favoring one group over another for examinees of the same level of proficiency—undergo further review to determine whether some aspect of what the item is measuring is particularly related to subgroup membership and irrelevant to the dimension being measured. When an item is identified as exhibiting such characteristics, it is either revised and re-ptestted or eliminated. The final form of a test rarely includes an item that exhibits sizable DIF. All items with DIF, however, have been reviewed by experts and have been determined to be appropriate for administration.

The Mantel-Haenszel (1959) procedure (MH), adapted by Holland and Thayer (1988), is used for DIF analyses with the SAT.³ This procedure computes a ratio for the conditional probability of successful reference group performance on an item over the conditional probability of successful focal group performance on the item for each score level on the test. Thus, comparisons are made of test takers with equivalent scores (e.g., equivalent proficiency in mathematical reasoning) at each point on the test. Statistically optimal weights are then assigned to each ratio, and they are averaged across all score points. The MH statistic is transformed to the delta (Δ) scale described previously, and the resulting statistic is referred to as the Mantel-Haenszel delta DIF (MH D-DIF).

The MH D-DIF statistic ranges from negative infinity to infinity, with a value of 0 indicating no DIF. Both the magnitude of the MH D-DIF and a significance test are used to evaluate the presence or absence of DIF. For the SAT, MH D-DIF values are considered

- negligible if they are between 1.0 and -1.0 or are not statistically different from 0 at the 0.05 significance level;
- moderate if they fall between 1.0 and 1.5 or -1.0 and -1.5, or if they are greater than 1.5 or -1.5 and not statistically different from the absolute value of 1.0 at the 0.05 significance level; and
- sizable if they exceed 1.5 or -1.5 and are statistically different from the absolute value of 1.0 at the 0.05 significance level.

³ For a complete description of the DIF procedures used by the SAT, see Dorans and Holland (1993).

Items exhibiting sizable DIF are not included when a test is assembled. Items exhibiting moderate DIF are usually not selected for a final form unless items with negligible DIF are insufficient to meet particular specifications. The average DIF for each group comparison is constrained to be approximately 0 across all test items in a form when an internal matching criterion (e.g., total test score) is used.

2.13 Evaluating Essay Pretests

As indicated previously, essay pretests are administered in classrooms scattered throughout the country. The responses collected from students are read by a group of experienced teachers, including members of the SAT Writing Committee, to determine whether a particular prompt is readily understood by high school students and elicits responses that reflect differing degrees of writing skill. In other words, does the prompt lead to responses that can be scored reliably and that provide differentiation among better and poorer writers? The members of the pretest reading group individually read and score a substantial number of the responses using the essay scoring guide, displayed as Figure 2-1. As a group, they discuss each prompt and decide whether it should be used, revised, or discarded. From the student responses collected during the pretesting, exemplars are chosen for each point on the holistic scoring scale. These serve as anchor papers for training and monitoring the experienced high school and college teachers who serve as readers when the essay prompt is administered operationally. The scoring process is described in Chapter 8.

2.14 Assembling the SAT Portion of the MHSA

The ongoing process of writing, reviewing, and pretesting items results in a large pool of acceptable test questions that are ready to be used in future operational forms of the SAT. Each item is classified according to the content and skill specification(s) of the particular test (mathematics, critical reading, and writing) and by the statistical indices generated by the pretest administration. For each of the three parts of the SAT, each item is stored electronically with its associated classifications and statistics. This electronic system can be used to inventory the item pool to identify, for example, particular areas of the specifications where there are insufficient items. Such information can, in turn, guide item-writing assignments.

The electronic system also assists ETS test developers by assembling a draft test that meets the content and psychometric specifications. The test developer then refines the draft test, making sure, for example, that there is a balance of references to women and men, that a particular concept which can occur in a variety of contexts (e.g., absolute value) is included, or that one question does not inadvertently provide the answer or a clue to the answer of another question. The test developer then reviews the entire draft test for unintended patterns, e.g., a key run of five Bs. Because each question needs to provide a combination of content and psychometric characteristics, substituting one question for another may lead to the need for a number of other changes in the draft test in order to meet the overall test specifications. After the test developer has completed the draft test, other SAT staff review it. These reviewers consider the same elements as the test assembler, but specifically focus on whether the draft test fully meets both the content and the

psychometric specifications for the test, and whether there is an appropriate balance of gender references or subject contexts for reading passages or mathematics problems. There is, again, a review of the test with regard to whether it portrays members of gender or racial/ethnic groups in a sensitive manner and avoids stereotypes. Individual items are reviewed to ensure clarity and lack of ambiguity, and the test as a whole is reviewed to make sure that it is comparable overall to other forms of the SAT. After the resolution of these reviews, the draft test is ready to be reviewed by the SAT test development committees.

2.15 Reviewing the MHSA: SAT Component

Each draft test is reviewed independently by a substantial number of specialists. Members of the test development committees for each area of the test (mathematics, critical reading, and writing) review and discuss each new form of the test. These reviews are performed both by mail and at the site of the committee meeting. The reviews by mail provide time for consideration and reflection on each question and the test as a whole, plus an opportunity for a reviewer to check a reference or to make sure that no wrong answer on a multiple-choice question can be successfully defended as correct. The onsite reviews provide the opportunity for a reviewer to experience the test in much the same fashion as a student, i.e., with time constraints and a sense of pressure. The concerns identified during the review are discussed with the committee and with the staff of the SAT Program, College Board Test Development, and the MDOE. Each concern must be resolved before the test moves into production and printing for its scheduled administration.

2.16 Test Production for the SAT Component

The production of test booklets for any particular administration of the SAT is very complex. Within an administration, multiple forms of the SAT are produced for use in different settings, e.g., the Sunday test centers, the international test centers, Saturday Eastern U.S. centers, Saturday Western U.S. centers. For any given form, multiple variations are created for security reasons and to accommodate the pretest/equating sections. Preparing print ready copy for each of these distinct test booklets takes several months. Each distinct booklet must be carefully proofread to ensure that it has the correct sections in the correct sequence, and that no typographical errors have been introduced in the composition process.

The actual printing of SAT test books and answer sheets is performed at one of the few printers equipped to protect the security of the tests, to handle the collation of test form variants, and to package and ship the test books and answer sheets to the test centers. The actual administration of the SAT is described in Chapter 6. The Math–A component is provided in camera ready copy to Measured Progress each February. Measured Progress oversees the printing, distribution, and administration of the Math–A portion of the MHSA. Those processes are described in Chapter 7.

2.17 After the SAT Administration

A number of further checks are made after the administration of the SAT and also after the reporting of student scores. A preliminary item analysis of the multiple-choice and student-produced-response questions is done on a sample of the students taking the SAT. The results are used to make sure that each question behaved as expected in terms of the level of difficulty and its ability to differentiate between more and less able students. Items are again analyzed for DIF among subgroups of the population. All reports from test centers of student complaints of ambiguity or incorrectness are reviewed. If the complaint is valid, appropriate action (e.g., dropping the item from scoring) is taken.

After the preliminary analyses and the work of equating the current form(s) to baseline forms have been completed and the essays have been graded, individual tests are scored and reports are issued to the students, their schools, and the designated colleges.

2.18 Public Access to the SAT

A number of forms of the SAT are made public each year. This enables teachers, counselors, admissions officers, students, and parents to be aware of what is tested by the SAT. Such widely available information may be used by teachers planning curriculum, by college faculty in judging how the SAT corresponds to their expectations of students, or by students in preparing to do their best on the SAT.

Annually, the forms used in four SAT administrations are available through the SAT Question and Answer Service (QAS). This service gives a student a chance to review a copy of the SAT she or he took, a record of the student's answers, the correct answers, and scoring instructions. QAS also includes information about the types of questions and level of difficulty of each question. It does not include a copy of the student's essay, although that can be viewed as part of the online score report or requested via paper score report. The May SAT form used as part of the 2008–09 MHSA is one of the four released forms and will be available for Maine educators at no cost to inform teaching and learning of Maine's *Learning Results*. A link to the released form administered in Maine is also embedded in the MHSA online reporting tool for use by school administrators and classroom teachers.

Some published SAT forms are used as practice tests, either in a print publication or online at www.collegeboard.com. The Web site version of the practice test provides explanations or annotations for each question. Other published SAT forms contribute to the practice questions and explanations that are provided on the Web site. Yet other forms appear in *The Official SAT Online Course* and *The Official SAT Study Guide*, both of which include extensive explanations of questions. Copies of *The Official SAT Study Guide* have been provided to all Maine high schools, and *The Official SAT Online Course* is provided at no cost on a year round basis to all students (grades 9–12), as well as all high school teachers and administrators. In addition to giving explanations for all of the questions on the publicly available forms, SAT Program staff also prepare explanations for each SAT Question of the Day that appears on the Web site.

Chapter 3. ALIGNMENT OF THE SAT TO THE REVISED *LEARNING RESULTS*

In the summer of 2007, Maine’s newly revised *Learning Results* content standards were adopted into law with the proviso that they would be assessed by the state assessment program in the 2008–2009 school year in order to allow for adjustments in curriculum and instruction at the local level.

Alignment studies were conducted to compare the content of the SAT with these new standards. As in prior years, the studies revealed that the alignment between the state’s reading expectations and those of the SAT critical reading test fully satisfied the criteria of the Webb alignment model. They also revealed that the state’s mathematics expectations included topics not measured (or not measured fully) on the SAT mathematics test, and therefore some additional mathematics test items were required to satisfy the Webb alignment criteria. This set of additional mathematics questions, collectively known as the Math–A, consisted of 12 operational items for the 2008–2009 MHSA.

Complete documentation of the alignment studies conducted by Amy Burkam of Lothlorien Consulting is attached in the appendices. Appendix B details the mathematics alignment study, while Appendix C documents the reading analysis.

For historical reference, the alignment protocols used in each year of Maine’s SAT Initiative are extensively documented in the *MeCAS Technical Manuals* from 2005–2006 through the present.

3.1 Design of SAT Critical Reading

The 2007 Maine *Learning Results* accountability reading standards, covered by the SAT critical reading section, include the following:

- A1 Interconnected Elements: Comprehension, Vocabulary, Alphabetics, Fluency
- A2 Literary Texts
- A3/A4 Informational Texts/ Persuasive Texts

The number of items covering each performance indicator section of the reading standard is indicated in Table 3-1.

Table 3-1. 2008–09 MHSA: Number of Items on the SAT Coded to Maine’s *Learning Results* for Reading*

<i>Reading Content Standard</i>	<i>SAT Critical Reading (Grade 11)</i>
A1 Interconnected Elements:	23
A2 Literary Texts	28
A3/4 Informational Texts/ Persuasive Texts	16

3.2 Design of SAT Writing

The 2007 Maine *Learning Results* writing standard, covered by the SAT writing section, include the following:

- B1 Interconnected Elements
- B3 Argument/Analysis
- D1 Grammar and Usage

The number of items covering each standard is indicated in Table 3-2. As noted previously, writing is not used in Maine’s accountability system.

Table 3-2. 2008–09 MHSAs: Number of Points on the SAT Coded to Maine’s *Learning Results* for Writing

<i>Writing Content Standard</i>	<i>SAT Writing (Grade 11)</i>
▪ B1 Interconnected Elements	6
▪ B3 Argument/Analysis	12
▪ D1 Grammar and Usage	43

3.3 Design of SAT Mathematics

This section addresses only the SAT component of the MHSAs Mathematics assessment. Information on the design and alignment of the complete MHSAs Mathematics assessment, which includes the SAT and the Math–A, can be found in Chapter 2.

The 2007 Maine *Learning Results* accountability mathematics standards, covered by the SAT mathematics section, include the following:

- A Number
- B Data
- C Geometry
- D Algebra

Refer to the chart on page 16 for more complete standard-related information.

Table 3-3 displays the number of SAT items measuring each standard. Refer to Appendix B for the Webb alignment study on the MHSAs mathematics assessment.

**Table 3-3. Number of Items on the SAT
Coded to Maine's Learning Results for Mathematics**

<i>Mathematics Content Standard</i>	<i>SAT Mathematics (Grade 11)</i>
A Number	8
B Data	7
C Geometry	15
D Algebra	24

Chapter 4. OVERVIEW OF THE SCIENCE TEST DESIGN

The 2008–09 MHSAs Science test was administered along with the Math–A on March 30–April 10, 2009.

4.1 Learning Results

The 2008–09 MHSAs science test items are aligned to the content standards D: The Physical Setting and E: The Living Environment described in the Science and Technology section of Maine’s *Learning Results: Parameters for Essential Instruction*. No other science content standards are subject to statewide assessment. Content specialists use the content standards and performance indicators to help guide the development of test questions which may address one or more of the performance indicators listed below.

D. The Physical Setting

D1: Universe and Solar System—Students explain the physical formation and changing nature of our universe and solar system, and how our past and present knowledge of the universe and solar system developed.

D2: Earth—Students describe and analyze the biological, physical, energy, and human influences that shape and alter Earth systems.

D3: Matter and Energy—Students describe the structure, behavior, and interaction of matter at the atomic level and the relationship between matter and energy.

D4: Force and Motion—Students understand that the laws of force and motion are the same across the universe.

E. The Living Environment

E1: Biodiversity—Students describe and analyze the evidence for relatedness among and within diverse populations of organisms and the importance of biodiversity.

E2: Ecosystem—Students describe and analyze the interactions, cycles, and factors that affect short term and long term ecosystem stability and change.

E3: Cells—Students describe structure and function of cells at the intracellular and molecular level including differentiation to form systems, interactions between cells and their environment, and the impact of cellular processes and changes on individuals.

E4: Heredity and Reproduction—Students examine the role of DNA in transferring traits from generation to generation, in differentiating cells, and in evolving new species.

E5: Evolution—Students describe the interactions between and among species, populations, and environments that lead to natural selections and evolution.

The three reporting categories and distribution of score points on the 2008–09 MHSAs for science are shown in Table 4-1.

**Table 4-1. 2008–09 MHSAs:
Distribution of Possible Score Points**

<i>Reporting Categories</i>	<i>Score Points</i>
D1 and D2	14
D3 and D4	20
E1–E5	22
Total possible score points	56

4.2 Test Design

The current science test design is comprised of a common set of items and an embedded field test. The common items are part of the assessment for all students and the results are reported. The embedded field test items were distributed among five forms, to produce reliable data with which to inform the process of selecting common items for future tests. Embedding the field test items creates a pool of replacement items needed due to natural attrition caused by the release of approximately half of the common items each year. Embedding also ensures that students take the items under operational conditions.

4.3 Item Types

The item types used and the functions of each are described below.

- **Multiple-choice items** were used to provide breadth of coverage of a content area. Because they require no more than a minute for most students to answer, these items make efficient use of limited testing time and allow coverage of a wide range of knowledge and skills.
- **Constructed-response items** typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—in constructing a satisfactory response. Constructed-response items should take most students approximately seven to ten minutes to complete.

Fifty percent of the items were released from the 2008–09 MHSAs Science test. A practice test composed of released science items is available on the Maine Department of Education website www.maine.gov/education/mhsa/mathaugmentandsci/index.html. Schools are encouraged to incorporate the use of the released items in their instructional activities so that students will be familiar with them.

4.4 Test Session Times

Schools were able to schedule testing sessions at any time during the two-week window, provided they followed the sequence in the scheduling guidelines detailed in test administration manuals and that all testing classes within a school were on the same schedule.

The timing and scheduling guidelines for the MHSA Science test were based on estimates of the time it would take an average student to respond to each type of item that makes up the test:

- multiple-choice—1 minute
- constructed-response—7–10 minutes

While the guidelines for scheduling are based on the assumption that most students will complete the test within the time estimated, each of the two science test sections was scheduled to allow 60 minutes. Once testing began, test administrators were asked to mark the time remaining to guide the students, and they ended the sections after 60 minutes.

Table 4-2 summarizes the numbers and types of items that were used in the MHSA Science assessment for 2008–09. Each multiple-choice item is worth one point and each constructed-response item is worth four points.

**Table 4-2. 2008–09 MHSA:
Science Item Numbers and Types**

<i>Section</i>	<i>Common</i>		<i>Embedded Field Test</i>		<i>Testing Time</i>
	<i>MC</i>	<i>CR</i>	<i>MC</i>	<i>CR</i>	
Section 2	16	2	4	1	60 min
Section 3	24	2	4	0	60 min
Total Testing Time					120 min

MC = multiple-choice; CR = constructed-response

Chapter 5. MHPA SCIENCE TEST DEVELOPMENT PROCESS

5.1 Item Development

Curriculum and assessment content specialists at Measured Progress developed the 2008–09 MHPA science items and scoring guides for the constructed-response items. During the process of item development, the specialists consider many elements of an effective item, including science content and structure, alignment to the content standards, grade-level appropriateness, and universal design. They make sure multiple-choice items have only one correct answer, non-content related information in the options does not cue the answer, the graphics are appropriate and contain correct content, the scoring guides include expected student responses that are grade level appropriate and contain correct content. After the items are initially developed, they are subjected to a number of reviews, both external and internal.

5.1.1 External Review of Item Content

The external review involves a committee composed of classroom teachers, an MDOE content specialist, and curriculum and assessment specialists from Measured Progress. Teacher participants are selected based on their content area expertise and grade level familiarity.

The purpose of the external review is to evaluate new items for the embedded field test and determine their suitability for the assessment by answering the following four questions:

- Does the item align with the assigned content standard and performance indicator?
- Is the science content accurate?
- Is the science content grade level appropriate?
- Does the item provide maximum accessibility for all students?

5.1.2 Bias and Sensitivity Review

Bias review is another essential step in the external review process. MHPA science items were reviewed for bias by a committee of state educators and members of major constituencies representing the interests of legally protected and/or educationally disadvantaged groups. Assessment materials were also examined for sensitivity issues that might offend or dismay students, teachers, or parents. By including such groups in the review process many unduly controversial issues were avoided and unfounded concerns were allayed before the test forms were produced.

5.1.3 Item Editing

Following the external reviews described previously, editors at Measured Progress reviewed and edited the items for the embedded field test to ensure uniform style (based on *The Chicago Manual of Style*,

15th edition) and adherence to sound testing principles. These principles included the stipulation that the items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- contained unambiguous explanations to students as to what is required to attain a maximum score;
- were written at a reading level that would allow students to demonstrate their knowledge of the tested content matter, regardless of reading ability;
- had appropriate answer options or score point descriptors; and
- were free of potentially sensitive content.

5.1.4 Reviewing and Refining

In preparation for the face to face meeting with the MDOE content specialist Measured Progress curriculum and assessment specialists and psychometricians considered the following when selecting a proposed set of items for the common and embedded field test:

- **Content coverage/match to test design.** The test design stipulates a specific number of multiple-choice and constructed-response items from each content area. Item selection for the embedded field test was based on the number of items in the existing pool of items eligible for the common.
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously tested items were used to ensure similar levels of difficulty and complexity from year to year as well as quality psychometric characteristics.

The final step in the science item development process was the face to face meeting between the MDOE content specialist and Measured Progress curriculum and assessment specialists. At this meeting, the MDOE content specialist approved items for the common and for release. Approval was also given for the final wording of embedded field test items.

5.2 Operational Test Assembly

At Measured Progress, test assembly is the sorting and laying out of item sets into test forms. Criteria considered during this process included the following:

- **Visual balance.** Item sets were reviewed to ensure that each reflected a similar length and “density” of selected items (e.g., number of graphics).
- **Option balance.** Each item set was checked to verify that it contained a roughly equivalent number of key options (As, Bs, Cs, and Ds).

- **Bias.** Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socioeconomic status, and other factors.
- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Facing page issues.** For multiple items associated with a single graphic stimulus, consideration was given to whether those items needed to begin on a left or right hand page, as well as to the nature and amount of material that needed to be placed on facing pages. These considerations served to minimize the amount of page flipping required.
- **Relationship between forms.** Although embedded field test items differ across forms, they must take up the same number of pages so that sessions and content areas begin on the same page in every form. Therefore, the number of pages needed for the longest form often determines the layout of each form.
- **Visual appeal.** The visual accessibility of each page of the form was always taken into consideration, including such aspects as the amount of white space, the density of the text, and the number of graphics.

5.2.1 Editing Drafts of Operational Tests

Any changes made by the test construction specialist were reviewed and approved by the curriculum and assessment specialist. Once a form had been laid out in what was considered its final form, it was reread to identify any final considerations, including the following:

- **Editorial changes.** All text was scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Measured Progress’ publishing standards are based on *The Chicago Manual of Style*, 15th Edition.
- **“Keying” items.** Items were reviewed for any information that might “key” or provide information that would help answer another item. Decisions about moving keying items are based on the severity of the “key-in” and the placement of the items within the form.
- **Key patterns.** The final sequence of keys was reviewed to ensure that their order appeared random (e.g., no recognizable pattern, and no more than three of the same key in a row).

5.2.2 Braille and Large-Print Tests

Form 1 of the MHSA science test was transcribed into braille by a subcontractor that specializes in producing test materials for blind and visually impaired students. In addition, Form 1 was adapted into a large-print version.

SECTION II—MHSA TEST ADMINISTRATION

The MHSA is given in two administrations. The Math–A and science tests are given during the first two weeks of April each year. The third component, the SAT, is given the first Saturday of May, with the first Saturday of June serving as a makeup administration date. Chapter 7 documents the SAT administration, while Chapter 8 documents the Math–A and science test administrations.

Chapter 6. ADMINISTRATION OF THE SAT

The SAT component of the MHSA was offered to all Maine juniors or third year students on May 2 and June 6, 2009. There were 15,182 students who registered for the May date and 454 for the June makeup date. Of those 15,636 students, 13,670 (87.4%) registered under standard conditions, 1,205 (7.7%) registered with College Board approved accommodations, 302 (1.9%) preregistered with Maine Purposes Only (MPO) accommodations, and 761 (4.9%) tested with MPO accommodations.

Great care is taken to ensure that the SAT is administered to all Maine students in a fair, equitable, and standardized manner. The goal of this detailed process is to ensure that all students take the test under a uniform set of conditions so that the results are trustworthy and can be used with confidence in accountability reporting, counseling students, and making admissions and placement decisions. No one is to suffer a disadvantage or gain an advantage of any kind because of race, ethnicity, religion, gender, or disability.

6.1 Preparation

To promote its goal, the MDOE, in conjunction with the College Board, provides all students planning to take the SAT with extensive preparatory material in both online and print formats. These range from detailed descriptions of the test, to full length sample tests, to discussions of approaches to testing, to last minute tips (e.g., bring a snack) to help each student on the actual test day. The materials may be viewed at www.maine.gov/education/mhsa/studentrp.html and www.collegeboard.com.

6.2 Supervision

Each Maine public high school where the SAT was administered was supervised by an experienced educator trained by the College Board and provided with detailed instructions and scripts for administering the SAT. The supervisor was responsible for all aspects of the test administration, including hiring staff who met College Board qualification, planning the use of the facility, and ensuring the security of test materials from their arrival until their return. The test center staff reflected the diversity of the students being tested and were expected to act in a fair, courteous, nondiscriminatory, and professional manner.

The primary task of all test center staff was to provide an equitable, valid, and standardized test administration. The supervisor was assisted by associate (or room) supervisors and proctors. The associate supervisor checked student identification, read the test administration script verbatim, and managed all other

aspects of the administration in his or her assigned room. In large rooms, the associate supervisor was joined by one or more proctors; the ratio of proctors to students was 1 to every 35–50 students. During the course of the administration, the staff in each room distributed and collected test materials, told students when to begin and end each test section, walked around the room to guard against misconduct, ensured that each student was working on the appropriate section of the test and using appropriate pencils for marking the answer sheet, and made sure that no test material left the room.

In addition to standard testing rooms, most test centers had a separate room for students receiving College Board approved accommodations and/or for students receiving MPO accommodations, which are described later in this chapter. Finally, students whose disabilities could not be accommodated at the test center (e.g., 100% extended time) were tested (or completed testing) in school the following week.

6.3 Physical Setting

In order that testing take place in a familiar environment conducive to each student doing her or his best on the SAT, test centers were established in nearly every public high school in Maine. The test center supervisors were responsible for planning the use of the facility and selecting rooms with adequate seating, lighting, and ventilation; access to restrooms; and seclusion from noisy areas or distracting activities (e.g., band practice). To discourage copying, all seats in a testing room faced the same direction with at least four feet between each student. No material (e.g., charts, posters) that could be of assistance to a test taker was displayed in the room.

6.4 Security

Three important facets to the security of a test administration are ensuring that no test taker has had prior access to the content of the test, that the test taker is indeed the person registered for the test, and that the test taker receives no assistance in responding to the test.

The physical security of all testing materials is fundamental to a fair and equitable administration. The SAT test center supervisor was responsible for receiving the test materials, checking them to ensure that they corresponded with what was shipped, and storing the materials in a locked storage area that was not accessible to students or other staff. Test materials were accounted for several times during the day of testing—when the test books and answer sheets were distributed to students, when they were collected from the students, and as they were packed for return to the SAT Program. Supervisors were encouraged to return test materials to the SAT Program immediately after the test, although many had to be picked up for return shipping to the SAT Program on the Monday following the test (or even later for students whose accommodations required that they be tested in school during the week).

Even though nearly all students tested in their own high school, admission to the test center was carefully monitored. Students were instructed to bring their SAT admission ticket and an acceptable photo ID,

which was checked against both the admission ticket and the attendance roster previously provided to the supervisor.

Students were not permitted to choose their own seats; rather, they were assigned seating by the supervisory staff to minimize the opportunity for preplanned collaboration among friends. No unauthorized person was permitted to enter the testing room after the administration had begun.

The materials that students could have on their desk during testing were very limited: the test book, answer sheet, No. 2 pencils (pens were not permitted), erasers, and, for the SAT mathematics sections, a calculator. Although all mathematics questions on the SAT can be solved without a calculator, students were encouraged to bring a graphing or scientific calculator. The only exceptions to this rule were materials approved as an accommodation for students with disabilities.

Test takers were strictly prohibited from using alarm watches or watches containing cameras; protractors; compasses; rulers; dictionaries or other books; pamphlets; papers of any kind; highlighters; colored pens or pencils; recording, copying, or photographic devices; pagers; handheld computers; electronic devices of any type; or cell phones. Handheld computers had to be turned off and stored out of sight. When approved to address a specific disability, students could use a computer to write their essays. Pagers and cell phones were not allowed at the test center. Violation of these prohibitions could lead to dismissal from the testing session and/or cancellation of test scores.

As a further step to prevent students from helping each other (deliberately or inadvertently), a number of test book variants were used during any one administration. At any given time some students could be working on a mathematics section, some on a critical reading section, and some on a writing section.

6.5 Calculator Policy for the SAT

Calculators are permitted for the entire mathematics section of the SAT. It is recommended that students use a graphing calculator or a scientific calculator. Four-function calculators are not recommended. Every question on the test can be solved without a calculator; however, using a calculator on some questions may be helpful. Students are encouraged to bring a calculator with which they are familiar and should know how and when to use their calculator.

Most calculators, even those with computer algebra systems (CAS) are permitted on the SAT. Unacceptable calculators are those that

- use QWERTY (typewriter-like) keypads;
- require an electronic outlet;
- “talk” or make unusual noises;
- use paper tape; or
- are electronic writing pads, pen input/stylus-driven devices, pocket organizers, cell phones, power books, or handheld laptop computers.

6.6 Item Types

The mathematics section of the SAT contains two types of questions:

- Standard multiple-choice (44 questions)
- Student-produced-response questions that provide no answer choices (10 questions)

For student-produced-response questions, no answer choices are provided. Students must solve the problem and fill in the answer on a special grid. The directions are fairly simple, and the gridding technique is similar to the way other machine readable information is entered on forms.

A primary advantage of this format is that it allows students to enter the form of the answer that they obtain, whether whole number, decimal, or fraction. For example, a student who obtains an answer of $\frac{2}{5}$ can grid $\frac{2}{5}$. If a student obtains an answer of 0.4 to the problem, the answer can be gridded in that form as well.

It is virtually impossible to guess an answer to a student-produced-response question, so they are highly reliable. There are no points deducted for incorrect answers to these questions. Figure 7-1 shows the actual test directions for student-produced-response items.

Each of the remaining questions requires you to solve the problem and enter your answer by marking the circles in the special grid, as shown in the examples below. You may use any available space for scratchwork.

Write answer in boxes.

Grid in result.

Fraction line

Decimal point

Note: You may start your answers in any column, space permitting. Columns not needed should be left blank.

- Mark no more than one circle in any column.
- Because the answer sheet will be machine-scored, you will receive credit only if the circles are filled in correctly.
- Although not required, it is suggested that you write your answer in the boxes at the top of the columns to help you fill in the circles accurately.
- Some problems may have more than one correct answer. In such cases, grid only one answer.
- No question has a negative answer.
- Mixed numbers such as $3\frac{1}{2}$ must be gridded as 3.5 or $\frac{7}{2}$. (If $3\frac{1}{2}$ is gridded, it will be interpreted as $\frac{31}{2}$, not $3\frac{1}{2}$.)
- **Decimal Answers:** If you obtain a decimal answer with more digits than the grid can accommodate, it may be either rounded or truncated, but it must fill the entire grid. For example, if you obtain an answer such as 0.6666..., you should record your result as .666 or .667. A less accurate value such as .66 or .67 will be scored as incorrect.

Acceptable ways to grid $\frac{2}{3}$ are:

Figure 6-1. 2008–09 MHSAs: Instructions for Student-produced Responses

6.7 Instructions and Timing

Central to the concept of standardized testing is the notion that all students should receive exactly the same instructions and be given precisely the same amount of time to work on the several parts of a test. To

achieve standardization, the SAT Program provides a script for associate supervisors to read and instructions about the amount of time allowed for each of the 10 sections of the test. This rule also applies to students receiving extended time as an approved accommodation; they are permitted 50% or 100% additional time for each section of the test, while the room supervisor strictly controls when they start and stop each section.

6.8 Complaints and Irregularities

Because hundreds of people were involved in administering the SAT in Maine, certain situations did not conform to the standardized model. Each irregularity was documented, including any action taken at the test center to remedy the situation. Supervisors were provided with instructions for dealing onsite with many common irregularities. All reports of irregularities are reviewed by Test Administration Services and SAT Program staff to determine whether the occurrence was severe enough to invalidate the test scores of the students involved. None of the irregularities at Maine test centers required the cancellation of scores or the scheduling of makeup tests.

6.9 Subgroup Performance

In accordance with NCLB legislation that subgroup performance be analyzed and reported, Table 6-1 presents the number of examinees from Maine in each subgroup along with the mean and standard deviation for each subgroup in mathematics, critical reading, and writing. To protect student confidentiality of test scores, the MDOE does not report mean scores and standard deviations for subgroups containing fewer than 5 examinees.

**Table 6-1. 2008–09 MHSA: SAT and Math–A Components
Subgroup Performance by Content Area**

Category of Participation		Mathematics*			Critical Reading			Writing			Science		
		N	Mean [†]	SD	N	Mean [†]	SD	N	Mean [†]	SD	N	Mean [†]	SD
Total group of students		15,008	1141	11	14,660	1141	14.6	14,663	1140	14.1	14,867	1140	11.2
Ethnicity	African American/Black	315	1134	9.6	303	1133	14	302	1133	13.1	311	1133	9.7
	American Indian/ Native Alaskan	106	1134	12.2	100	1135	14.3	100	1134	12.4	102	1135	9.2
	Asian/Pacific Islander	227	1144	13.1	219	1141	16.5	219	1141	15.1	225	1141	12.8
	Hispanic	157	1136	10.8	151	1137	14.3	151	1135	14.2	152	1136	10.3
	Caucasian/White	14,203	1141	10.9	13,887	1141	14.5	13,891	1140	14.1	14,077	1141	11.2
	Not reported	0			0			0			0		
Identified disability	Yes	1,959	1130	9.6	1865	1127	12.1	1,861	1125	10.8	1,928	1131	8.4
	No	13,049	1142	10.3	12,795	1143	13.7	12,802	1142	13.2	12,939	1142	10.9
Current LEP	Yes	239	1132	10.8	225	1126	11.1	224	1127	10.8	234	1129	7.8
	No	14,769	1141	11	14,435	1141	14.5	14,439	1140	14.1	14,633	1140	11.2
Economically disadvantaged	Yes	4306	1136	9.8	4,120	1136	13.5	4,121	1134	13	4,264	1136	9.9
	No	10,702	1142	11	10,540	1143	14.4	10,542	1142	13.9	10,603	1142	11.3
Migrant	Yes	4	1140	14.8	3	1147	22.7	3	1143	19	4	1143	15.9
	No	15,004	1141	11	14,657	1141	14.6	14,660	1140	14.1	14,863	1140	11.2
Gender	Female	7,248	1140	10.1	7,098	1142	14	7,103	1143	13.4	7,179	1139	10.1
	Male	7,760	1141	11.8	7,562	1140	15	7,560	1138	14.4	7,688	1142	12
	Not reported	0			0			0			0		
Title 1A Program	Yes	293	1137	8.3	291	1135	13.3	291	1135	12.9	287	1136	9.8
	No	14,715	1141	11	14,369	1141	14.6	14,372	1140	14.1	14,580	1140	11.2
Gifted/talented program	Yes	521	1157	10.7	520	1161	10.8	520	1159	10.4	517	1156	10.4
	No	14,487	1140	10.6	14,140	1140	14.2	14,143	1139	13.7	14,350	1140	10.8

*Including the Math–A items

†The MHSA reporting scale was changed from the traditional 200–800 College Board scale to 1100–1180 to be consistent with Maine’s reporting scales in other grades

6.10 Accommodations for Students on the MHSA

Accommodations for students who cannot access state assessments through standard administration are available on the MHSA, as they are for the state assessment in grades 3 through 8. They are designed to allow all students with unique learning needs a fair opportunity to demonstrate what they know and can do at the high school level. The decision to allow the use of accommodations by an individual on any state assessment must be made by the student's IEP team.

There are two categories of accommodations for the MHSA: (1) those approved by the College Board through the Eligibility Form process, and (2) those approved only by the State of Maine, designated as MPO. The accommodations listed for either category are equivalent. In order to assure the opportunity for all Maine students to participate in the SAT component of the MHSA, the College Board agreed to allow some Maine third year high school students to use accommodations selected from a state approved MPO list, with the understanding that the scores would be used strictly for Maine adequate yearly progress (AYP) purposes and not result in scores reportable to colleges for admissions. The same accommodations are included in both categories.

Students with an identified disability are instructed to apply first for College Board approval by submitting a Student Eligibility Form to the College Board. Students may include any MPO accommodations under the category "Other" on the Student Eligibility Form. College Board approval of the accommodations allows students to take the SAT portions of the MHSA and receive college reportable scores. Students whose accommodations requests have not met College Board criteria, who are categorized as limited English proficient, or who did not apply for accommodations through the College Board are still eligible for MPO accommodations if approved by a local district team. For state assessment reporting purposes there is no difference based on the type of accommodation used. However, only those students using College Board approved accommodations receive official SAT scores that can be reported to colleges. School personnel are instructed to provide the same accommodations on all components of the MHSA (both SAT and Math–A), as appropriate.

Historically, about 10% of those taking the state administered 11th grade MHSA tests have qualified for testing accommodations. Nationally, approximately 2% of SAT test takers qualify for College Board approved Services for Students With Disabilities (SSD) accommodations. In the 2008–09 administration, 9.6% of those taking the MHSA qualified for testing accommodations: 6.4% in reading, 6.3% in mathematics, and 6.4% in writing.

6.10.1 Process and Standards for College Board Approved Accommodations

Generally, to be eligible for College Board approved accommodations, the student must

- have a disability that necessitates testing accommodations,

- have documentation on file at school that supports the need for the requested accommodations and meets the Guidelines for Documentation, and
- receive and use the requested accommodations, due to the disability, for school-based tests, for at least four school months.

The College Board Guidelines for Documentation require that documentation

- state the specific disability, as diagnosed;
- be current (in most cases, the evaluation and testing should be completed within five years of the request for accommodations). For psychiatric disabilities, an annual evaluation update must be within 12 months of the request for accommodations;
- provide relevant educational, developmental, and medical history;
- describe the comprehensive testing and techniques used to arrive at the diagnosis, including evaluation date[s] and test results with subtest scores.
- describe the functional limitations (how the disability impacts learning).
- describe the specific accommodations requested, including the amount of extended time required if applicable. State why the disability qualifies the student for such accommodations on standardized tests; and
- establish the professional credentials of the evaluator, including information about license or certification and area of specialization.

The guidelines are included in the instructions for the Student Eligibility Form and are also available on the College Board Web site at www.collegeboard.com. The College Board offers two ways for a student to be determined eligible for accommodations on its tests.

1. School verification: When a student's school generated individualized education program (IEP), 504 plan, or other formal written educational plan/program and its supporting documentation align with the College Board's eligibility criteria and guidelines, and officials at the student's school verify this to be accurate, the College Board generally does not need further documentation. The College Board processes the form and notifies the student and school of the approved accommodations.
2. Documentation review: If all of the above requirements are not met, a student may still be eligible for accommodations on College Board tests. The student's disability documentation is submitted with the Student Eligibility Form, and a panel of experts in educating and assessing students with disabilities reviews the documentation and advises whether the guidelines are met. The College Board reviews the panel's recommendation, makes a determination, and notifies the student and school whether any of the requested accommodations are approved. Documentation review is also available for students who want the College Board to make a determination without their school's involvement.

6.10.2 Process and Standards for MPO Accommodations

Maine has historically allowed testing accommodations to be provided to students, regardless of disability identification, if approved by a local team of educators. As these accommodations are not necessitated by limitations on the ability to participate in College Board tests due to disability, they would not be available on any ordinary, college reportable administration of a College Board test. These accommodations include

- services for students who are limited English proficient (e.g., bilingual dictionaries, word lists); and
- services for “at risk” students who perform poorly under standardized testing conditions but have no identified or suspected disabilities (e.g., extra time).

Maine’s state assessment policies and practices allow accommodations for students other than those with disabilities. Such students include those who are ill or incapacitated in some way, those with Limited English Proficiency, those with a 504 plan, or those for whom classroom accommodations are necessary on a daily basis to measure academic achievement. The “Policies and Procedures for Accommodations and Alternate Assessment” is presented in Appendix D. The MPO accommodations have been designed to be comparable to those available to students approved by the College Board through the Eligibility Form process.

6.10.3 Eligibility Process Additions to Incorporate MPO Accommodations

Maine students with disabilities were encouraged to apply for testing accommodations through the College Board’s SSD eligibility process. Maine students who were approved for testing accommodations through the SSD eligibility process were allowed to be tested through existing College Board processes for SSD center based SAT testing and SSD school based SAT testing. Tests administered through these processes with approved accommodations were considered valid by the College Board and became part of the student’s SAT record maintained by the College Board.

Maine students who desired testing accommodations not approved by the SSD eligibility process were, as noted above, allowed to take the test if the additional or alternate accommodations were approved by a local team of Maine educators. Refer to Appendix D for a list of specific MPO accommodations. Under this process, the test was scored by the College Board but was not considered a valid SAT administration and did not become part of the student’s SAT record.

MPO accommodations were granted both in cases for which the College Board SSD approved no accommodations and in cases for which the College Board SSD approved fewer accommodations than did an IEP team. In both cases, the student’s family and school IEP team were afforded the final decision whether to take the test with the level of accommodations approved by the College Board and have the test applied to the student’s SAT record, or to take the test with the MPO accommodations and forfeit the SAT record.

Each Maine high school coordinator was assigned ultimate responsibility by the MDOE for ensuring all students with disabilities were processed through the College Board SSD and Maine specific eligibility processes (working directly with the designated College Board SSD coordinator and/or Maine eligibility coordinator as necessary).

6.10.4 Accommodation Eligibility Form Submission Time Lines

In order to assist Maine in organizing its students' requests for accommodation and providing for sufficient time for students to choose between College Board approved accommodations and MPO accommodations, an earlier submission deadline was established for accommodation eligibility forms to be submitted to the College Board SSD.

Specifically, a January 14, 2009, deadline was established for Maine high school junior eligibility form submissions. March 14, 2009, was the standard deadline for eligibility form submissions for the May 2, 2009, SAT.

6.10.5 Training and Technical Assistance

Workshops were conducted by College Board program staff in collaboration with MDOE personnel in order to fully inform individual school representatives about the MHSA and associated deadlines. Rather than conducting separate workshops for issues involving students with disabilities, this information was incorporated into the regularly scheduled training workshops. Workshops were conducted via the Web on February 23 and 24, 2009, and March 18, 2009.

6.10.6 MHSA Accommodation Request and Approval Statistics

Table 6-2 presents the numbers of accommodations requested and approved and the types of accommodations approved for Maine public school juniors or third year students for the 2009 MHSA administration. It includes any approvals for students who chose to take the test under MPO conditions.

**Table 6-2. 2008–09 MHSA: Summary
of Accommodations for 2009 MHSA Administration**

Total number of accommodations requested for College Board approval	1,350
Total number of accommodations approved by College Board	1,130
Total number of students using College Board accommodations	1,102
Total number of students using MPO accommodations	761
<hr/>	
Total number of students using accommodations	
Some students moved to MPO accommodations even though they had been approved for accommodations by the College Board because either they were not approved by the College Board for all of the accommodations they requested or they were absent from the May 3 testing and chose to test during the MPO window the following week. The resulting scores were not reportable for college admissions purposes.	1,863
<hr/>	
<i>MPO ACCOMMODATIONS</i>	
<hr/>	
MPO accommodations for May 2009	240
MT1–Extended time same day	100
MT2–Extended time over several days	47
MT3–Multiple or frequent breaks	35
MT4–Flexible test day or start time	2
MT5–Flexible ordering of test sections	5
MS1–School location other than classroom	7
MS2–Offsite location with school personnel	7
<hr/>	
MP1–Individual testing	10
MP2–Small group testing	176
MP3–Human reader	43
MP4–Sign language (not for reading test)	1
MP5–Stand, move, pace during testing	1
MP7–Proctored by special education or ESL personnel	24
MP10–Bilingual dictionary	15
MP11–Translation into native language	8
MP12–"Sheltered English" content	12
MR1–Scribe/recording device, nonessay	20
MR3–Other assistive devices	5
MR4–Word processor	3
MR7–Bilingual dictionary	2
MR8–Verification directions understood	87
MO1–Accommodations based on test content	2
<hr/>	
<i>COLLEGE BOARD ACCOMMODATIONS</i>	
<hr/>	
Large print–photo enlarged to 14 point	7
Large print–20 point	6
Large block answer sheet	9
Magnifying machine	2
Braille test	1
Braille graphs and figures	1
Braille device for written responses	1
Reader	188
Cassette test version	18
Writer to record responses	79
Computer to record written responses	54
Reading–50% extended time	761
Writing–50% extended time	768
Mathematical calculations–50% extended time	760
Listening–50% extended time	415
<hr/>	

continued

COLLEGE BOARD ACCOMMODATION (continued)

Reading–100% extended time	88
Writing–100% extended time	95
Mathematical calculations–100% extended time	81
Extra breaks	291
Written directions, bring sign language interpreter	4
Extended breaks	53
Snacks and/or fluids permitted	13
Preferential seating	27
Write answers in the test book	6
Separate location	14
School based testing	24
Auditory amplification, including FM system	2
Test blood sugar level	3
Small group setting	533
Two consecutive test days	1

*Students may be granted more than one accommodation and therefore may appear in multiple counts within the table. The listing of accommodations is not comprehensive. Accommodations with counts of 0 were omitted.

6.11 Participation

The intent of the MHSA is for all students in their third year of high school to participate in all components of the test. However, on those occasions where it was necessary to grant a waiver to students from taking the SAT due to special considerations, such as hospitalization or a death in the family, schools were asked to seek the approval of the MDOE MHSA coordinator. Approved students' nonparticipation was reported in the MHSA results.

Chapter 7. ADMINISTRATION OF MATH–A AND SCIENCE

After the College Board had completed the Math–A test development cycle, print ready files were produced and transferred to Measured Progress, where they were combined with the two science sections to complete the production of the test booklets. As the contractor responsible for the administration of the test, Measured Progress completed tasks such as printing and shipping the test materials, arranging for the return and login of test materials, scanning the answer documents, and transferring the scan files back to the College Board so that item analysis could be completed.

The combined Math–A and science test was administered at all Maine high schools during the testing window of March 30 to April 10, 2009. As indicated in the *Principal and Test Coordinator Manual*, principals and/or their designated MHSA coordinators were responsible for the proper administration of the Math–A and science portions of the MHSA. Manuals containing explicit directions and scripts for test administrators to read aloud to test takers were used to ensure the uniformity of administration procedures from school to school.

7.1 Supervision and Security

To ensure the administration of the Math–A and science tests in a fair, equitable, and standardized manner, principals and/or schools' designated MHSA coordinators were instructed to read the *Principal and Test Coordinator Manual* prior to testing and to be familiar with the instructions given in the *Test Administration Manual*. The *Principal and Test Coordinator Manual* provided checklists to help schools prepare for testing before, during, and after test administration. Along with these checklists, the *Principal and Test Coordinator Manual* outlined the nature of the testing material being sent to each school, how to inventory the material, how to track it during administration, and how to return the material once testing was complete. The *Test Administration Manual* also included checklists for administrators to ready themselves, their classrooms, and the students for the administration of the test. *The Test Administration Manual* contained sections detailing the procedures to be followed during testing as well as instructions on preparing the material for its return to Measured Progress. The *Principal and Test Coordinator* and *Test Administration Manuals* are included as Appendix E.

In addition to distributing the *Principal and Test Coordinator* and *Test Administration Manuals*, the MDOE conducted a series of live and broadcast test administration workshops across the state to train and inform school personnel about the Math–A and science tests. The test coordinator was responsible for the security of the tests while within the schools. Information concerning test security and ethical administration is clearly spelled out in both manuals and stressed during test administration workshops. Principals were required to complete an online Principal Certification of Proper Administration form at the conclusion of testing, certifying that all testing was administered according to MHSA protocols.

7.2 Participation Requirements and Accommodations

The intent of the MHSA is for all students in their third year of high school to participate in testing through standard administration or administration with accommodations. Any student who is absent during the test session is expected to take a makeup test within the testing window.

Eligibility for taking the Math–A and science tests with accommodations was determined during the registration process for the SAT conducted by the College Board. (Please see Chapter 6 for a complete description of this process and a chart showing the numbers of students who tested using accommodations.) School personnel were advised in the *Principal and Test Coordinator Manual*, in test administration workshops run by the College Board and the MDOE, and by information posted on the MDOE Web site that students were to take the Math–A and science tests using the same approved accommodations documented during the SAT registration process.

On those occasions where it was necessary to grant a student a waiver from taking the Math–A and science tests due to special considerations, such as hospitalization or a death in the family, schools were asked to seek the approval of the MDOE MHSA coordinator. The names of the excluded students were forwarded to Measured Progress, so they would not be included in any reports. Table 7–1 shows MHSA participation rates for each content area tested.

Table 7-1. 2008–09 MHSA: State Participation Rates May and June 2009

Student Category and Mode of Participation	Enrollment	Number Tested				Percentage Enrolled ⁴	Percentage Tested ^{5,6}			
		Mathematics	Critical Reading	Writing	Science		Mathematics	Critical Reading	Writing	Science
Category of participation										
Total number of students	15,632	15,274	14,928	14,926	15,079	100	98	96	96	97
Ethnicity										
African American/Black	341	322	310	309	317	2	95	91	91	93
American Indian /Native Alaskan	111	107	101	101	103	1	96	91	91	93
Asian or Pacific Islander	241	229	221	221	227	2	95	92	92	94
Hispanic	166	162	156	156	155	1	98	94	94	93
Caucasian/White	14,773	14,454	14,140	14,139	14,277	95	98	96	96	97
Not reported	0	0	0	0	0	0	0	0	0	0
Identified disability	2,327	2,200	2,108	2,099	2,140	15	95	91	91	92
Current LEP	262	246	232	231	240	2	94	89	88	92
Economically disadvantaged	4,634	4,451	4,263	4,262	4,383	30	96	92	92	95
Migrant	5	5	4	4	5	0	100	80	80	100
Mode of participation										
Participation without accommodations		13,417	13,079	13,084	13,288		86	84	84	85
Identified disability (PET/IEP)		814	727	725	802		6	6	6	6
LEP		181	170	170	177		1	1	1	1
504 Plan		245	238	238	241		2	2	2	2
Participation with accommodations		1,636	1,626	1,624	1,579		10	10	10	10
Identified disability (PET/IEP)		1,165	1,158	1,156	1,126		71	71	71	71
LEP		59	56	55	57		4	3	3	4
504 Plan		79	79	80	77		5	5	5	5
Other		360	360	360	345		22	22	22	22
Participation through alternate assessment (PAAP)		221	223	218	212		1	1	1	1
Identified disability (PET/IEP)		221	223	218	212		100	100	100	100
LEP		6	6	6	6		3	3	3	3
504 Plan		0	0	0	0		0	0	0	0
Approved non-participation in reading—1st year LEP			0					0		
Approved nonparticipation—special consideration		34	24	24	26		0	0	0	0
Non-participation—other		324	680	682	527		2	4	4	3

⁴ The percentage of students enrolled in each participation category.

⁵ The percentage of students, including those who took the PAAP, who participated in the content area.

⁶ The percentage of students in each content area by mode.

SECTION III—SCORING

Chapter 8. SCORING THE SAT

Most students, parents, teachers, guidance counselors, and college admissions officers are familiar with the SAT score scale of 200 to 800. How do the responses made by a student on an answer sheet become a score between 200 and 800? This chapter will describe that process.⁷ The first portion of the chapter focuses on the process of receiving the completed answer sheets and materials and the associated quality control process; the second portion focuses on the majority of the test—those questions and responses that can be scored by machine; the third portion describes scoring the essay section of the SAT writing test—a process that involves experienced teachers facilitated by electronic technology.

8.1 Receiving and Opening

Upon completion of the SAT, test center supervisors begin to pack the answer sheets and ancillary materials into shipping cartons with pre-affixed tracking labels. Each test center shipment is routed to the answer sheet processing center in Austin, Texas. The tracking labels are tied to each unique testing center. The tracking labels are scanned, matching them to test centers, which enables the identification of missing or incomplete shipments from the center.

Shipments are then moved into opening, where materials are removed from the shipping cartons. Representatives perform a quality review of the Supervisor Report Form and visually inspect answer sheets for obvious *n*-count discrepancies. Discrepancies are isolated to the individual test taker and held for resolution. Answer sheets are batched and placed on carts in preparation for scanning.

Ancillary materials are reviewed and forwarded to the applicable departments. Ancillary materials include but are not limited to the following:

- Standby registrations
- Cancellation forms
- Supervisor Irregularity Report (SIR)
- Supervisor Report Form (SRF)
- Student Information Correction Forms
- Seating charts
- Test Question Ambiguity/Error Form

⁷ Chapter 11 describes how scores are transformed to the MHSA scale of 1100 to 1180.

8.2 Scanning and Editing

Scanning is a single pass operation that captures demographic data, form data, item response data, and essay images from each side of the answer sheet. Answers sheets are held in a climate controlled environment and scanned twice. Discrepant items are reviewed by an editor to determine which scan value should be captured. The following quality controls regulate the scanning process:

- Prior to starting a batch of answer sheet documents on a scanner, the operator must successfully run 10 diagnostic sheets to ensure scanner calibration. The scanner must accurately read 59,220 ovals without an error; the scan program does not proceed unless the diagnostic sheets have been read successfully.
- Prior to the scanning of each batch, the scanner operator performs a multisheet test to ensure the scanner halts if two or more sheets pass through at the same time.
- Each answer sheet has anchor points and timing tracks, which ensure it is properly aligned.
- Periodically, answer sheets receive a hand scan accuracy review, ensuring the scan values match the item responses on the answer sheet.
- Quality control check sheets are placed in every stack to ensure the scanner continues to operate correctly.

Additional quality checks at edit include the following:

- Resolve conditions where the information was written but not gridded. Fields include name, social security number, date of birth, gender, and registration number.
- Validate that the test form and form code on the answer sheet matches the valid values for the administration date.
- Ensure that only those students with authorized accommodations receive the Student Services With Disabilities form.

8.3 Matching

Matching is the term applied to the process used to associate a candidate's complete and scanned answer sheet with his or her complete and valid registration. There are three types of matches.

1. Auto matching occurs when a specific set of demographic information from the answer sheet matches exactly to the corresponding information from the candidate's registration with a high confidence interval as specified by quality control. There are 10 such data combinations that can result in a high confidence match. Data elements to be matched include, but are not limited to, registration number, last name, first name, date of birth, and gender.
2. Manual matching occurs when combinations of various data elements exactly match the information from the registration, but one or more major data elements (such as

registration number) do not match exactly to the registration data. These cases are reviewed to ensure that the correct match is being made even though some data elements are incongruous.

3. Force matching occurs when a registration is neither high confidence nor low confidence matched and is considered to be in an unmatched status. The College Board investigates all unmatched answer documents. The document stays in an unmatched status until it can be high confidence or low confidence matched to a created registration or the College Board declares the need for a force match. Force matching is necessary because it is possible that incomplete demographic information, or major discrepancies between registration and answer sheet data, will prevent an answer sheet from ever being high or low confidence matched. During the course of a College Board investigation, it can be determined that a candidate registration and answer sheet should be matched, but the matching cannot take place within established matching rules. At this point, the College Board performs a force match, or override, to associate the answer sheet with the identified registration. This process is subjected to rigorous quality control oversight.

8.4 Machine-Scored Portions

All of the SAT mathematics (including the student-produced responses), critical reading, and writing questions, except the essay, are scored by machines. Each student answer sheet is optically scanned and converted to a digital file. These digital files are processed by computer, comparing the student response to each item with the official scoring key to determine the number of questions answered correctly, the number answered incorrectly, and the number omitted.

For all multiple-choice questions (each with five options), each wrong answer results in a deduction of $\frac{1}{4}$ of a point from the total number of right answers to give the corrected raw score, also known as formula scoring. Formula scores are calculated based on the rights, wrongs, or omits, taking into account the penalty for incorrect responses. For SAT mathematics, the total number right among the student-produced-response questions is added to the corrected raw score for the multiple-choice questions to produce the total raw score. For SAT writing, the corrected raw score for the multiple-choice questions is combined with the essay score to produce the total raw score.

Prior to each administration, a test set of answer sheets consisting of all right and all wrong answers is run through the formula score process. This quality control check is designed to determine if the correct score keys within the system are valid. Upon successful completion of this check, the administration is approved for answer sheet processing.

The raw score for each of the three sections is converted to the 200–800 score scale through a statistical process called equating. Equating ensures that the varying difficulty levels of different forms of the test do not affect the scaled score that is reported. Equating allows comparisons among test takers who take

different editions of the test across different administrations. This process is described in more detail in Chapter 11.

Conversion is a system activity that applies the conversion tables produced during equating to raw formula and essay scores to generate the scaled scores. Conversion quality assurance for each administration uses a randomly selected statistically valid sample to manually convert each answer sheet through independently generated tables, which are compared to the systematic results produced.

8.5 Scoring the Essay

The SAT writing essays are scored by experienced high school teachers and college faculty members who teach either English or another subject that requires a substantial amount of writing. To be considered for the position of essay reader, a person must

- hold a bachelor’s degree or higher;
- teach or have taught a high school or college level course that requires writing;
- have taught for at least a three year period;
- reside in the continental United States, Alaska, or Hawaii; and
- be a U.S. citizen, a resident alien, or authorized to work in the U.S

In addition, readers must complete a rigorous online training course on the principles of holistic scoring that teaches them to evaluate essays according to the agreed-upon standards.

The qualification process, which takes 10 to 15 hours, requires readers to score 30 papers that have previously been scored by leadership and approved by the College Board. To qualify to serve as a reader, a person must score these qualifying papers consistently with leadership, either assigning the same exact score to at least 70% of the papers OR scoring at least 50% exactly, with at least 90% within one point (exact or adjacent).

The pool of readers available for essay scoring is very large, and every effort is made to ensure diversity in terms of gender, ethnicity, education level, and teaching experience. The exact breakdown of rater characteristics for any one administration varies due to demand for and availability of readers. Confidentiality requirements permit readers to omit or choose not to answer some background questions, and therefore the exact percentages in the pool may vary from those reported. The reader pool for a recent large administration was approximately 23% male and 77% female. The ethnic breakdown was approximately 59% White, 1.5% Native American, 2% Asian, 2% Black, 2% Hispanic, 1.5% Pacific Islander, and 32% unspecified. Approximately 76% of the readers held advanced degrees, with 14% of those at the doctoral level. In terms of teaching experience, 27% of readers reported 3 to 5 years at the high school or college level, 28% reported 6 to 10 years, and 45% reported 11 or more years.

Essays are scored in a fair and consistent manner using a holistic approach. A piece of writing is considered as a total work, the whole of which is greater than the sum of its parts. Readers take into account such aspects as complexity of thought, the substantiality of the development, and facility with language. Holistic scoring recognizes that the real merit of a piece of writing cannot be determined by merely adding together the values assigned to such separate factors as word choice, organization, use of evidence, and adherence to the conventions of written English. A reader does not judge a work based on such separate traits but rather on the total impression it creates, with an emphasis on how these separate factors blend together to become the whole piece of writing.

Readers are trained to be mindful of the conditions under which students wrote the essays and to keep a number of guidelines in mind when scoring essays, including the following:

- Use the scoring guide (displayed in Chapter 2) in conjunction with the sample essays selected for training.
- Read quickly to gain an impression of the whole essay.
- Read the entire essay before scoring, and then score immediately.
- Read supportively, looking for and rewarding what is done well rather than what is done badly or omitted.
- Ignore the quality of handwriting.
- Judge an essay by its quality, not by its length.
- Understand that no one aspect of writing (coherence, diction, grammar) is more important than another, and that no aspect of writing is to be ignored.

Each essay is scored independently by two qualified readers on a scale of 1 to 6, with the combined score for both readers ranging from 2 to 12. (An essay not written on the assignment receives a score of 0.) If the two readers' scores differ by more than one point, a third reader scores the essay. During scoring, readers are also asked to be cognizant of special circumstances that may require flagging due to the following alerted condition codes:

- Off topic, unrelated, or suspected cheating
- Cheating—wrong prompt; valid for a different administration
- On topic but similar to essays read before
- Cry for help—response suggests a situation that warrants investigation, such as the possibility of abuse, depression, or contemplation of suicide
- Confidential data—response contains confidential information such as social security numbers, malicious information about another student, etc.

The accuracy and fairness of the readers are evaluated regularly and frequently through a number of processes. Some of these checks are apparent to a reader, while others are embedded in the flow of student papers. For each administration of the SAT essay, readers are trained by scoring a set of prescored calibration essays on the topic(s) used for that administration. The calibration papers are used to clarify issues and provide feedback to the readers.

An additional aid to maintaining scoring accuracy is the use of prompt specific anchor papers. Anchor papers are 16 prescored essays selected to represent the full range of performance, across all 6 score points, that a reader is likely to see on a given prompt. By comparing operational essays to prescored anchor papers, readers are able to assign scores on a given prompt with maximum accuracy. To ensure accuracy across prompts as well, anchor papers are selected by consensus agreement of a test development committee during a process known as range finding. Essays are only selected as anchor papers if members of the range finding committee, a diverse group of secondary and university teachers, unanimously agree that the level of performance of an essay at a score point matches that expected for essays at the same score point for other prompts. (The range finding committee works to ensure that an anchor paper at the 3 score point for prompt A demonstrates the same level of performance as a corresponding anchor paper at the 3 score point for prompt B.)

As a further step in maintaining reader accuracy throughout the scoring process, validity papers—clear examples of score points—are interspersed randomly with other student responses. Scoring leaders review readers' scoring of selected essays and provide feedback via phone and the Web when appropriate. If a reader is unable to accurately score the papers consistently, he or she will not continue as a reader. Web based scoring enables leaders to monitor readers in real time, informed by extensive real-time and summary reports on interrater reliability, validity, and calibration statistics. This robust training and monitoring program ensures the highest quality of performance from the readers. Confirming this rigorous training, qualification process, and continuous monitoring of readers, only about 2% of the 2009 SAT essays required a third reading (Figure 8-1). For the Maine specific population of students who received official score reports, the percentage of essays requiring a third reading was the same (Figure 8-2).

Essays are scanned and distributed to readers via the Web. By working with readers via the Web, the College Board is able to attract and involve a larger reader pool from across the country than would be possible at a common site.

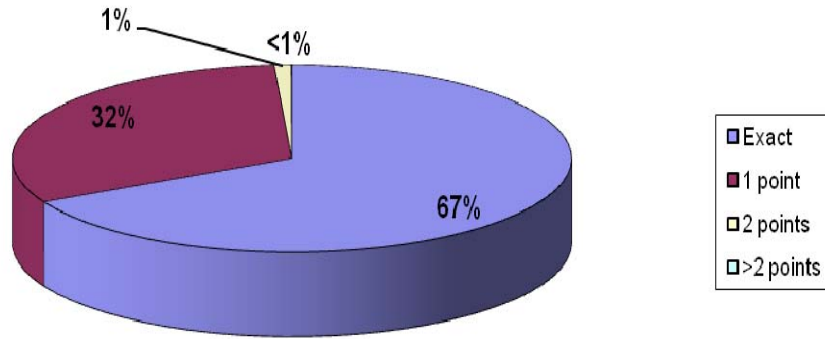


Figure 8-1. Differences in Reader Scores for National Sample in May and June 2009

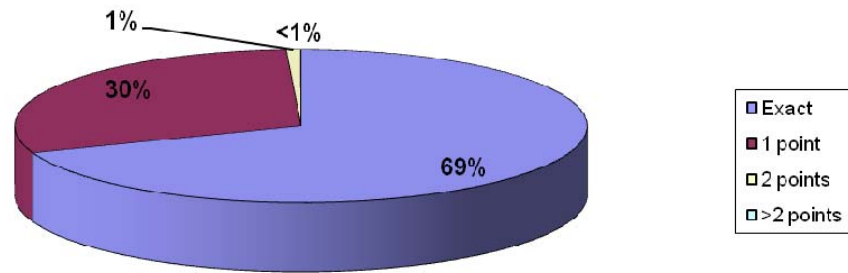


Figure 8-2. 2008–09 MHSAs: Differences in Reader Scores for Maine Specific Sample in May and June 2009

*Includes data for students receiving official college reportable scores only. Scores for students receiving Maine Purposes Only accommodations cannot be used for college admission or placement purposes.

The scores assigned by the two readers are combined into an essay subscore ranging from 2 to 12. The distribution of scores assigned in the May and June 2009 national administrations for all test takers are shown in Figure 9-3. The Maine specific distributions for May and June 2009 are displayed in Figure 8-4. It should be noted that Figure 8-4 is based only upon students in Maine who received official College Board score reports for the May and June 2009 administrations.

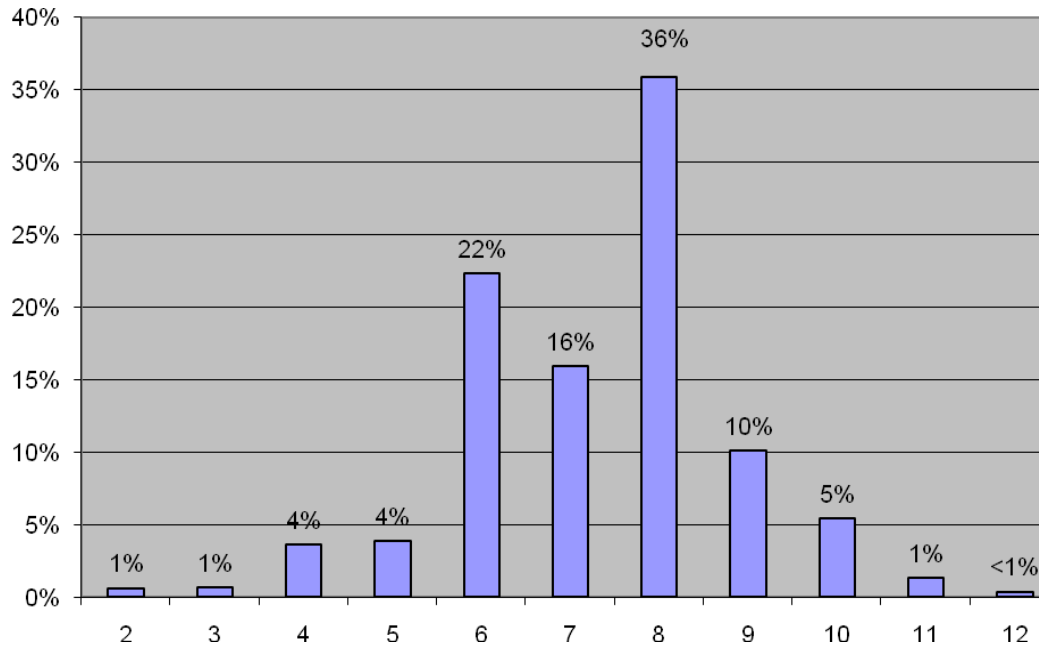


Figure 8-3. National Distribution of SAT Essay Scores for May and June 2009

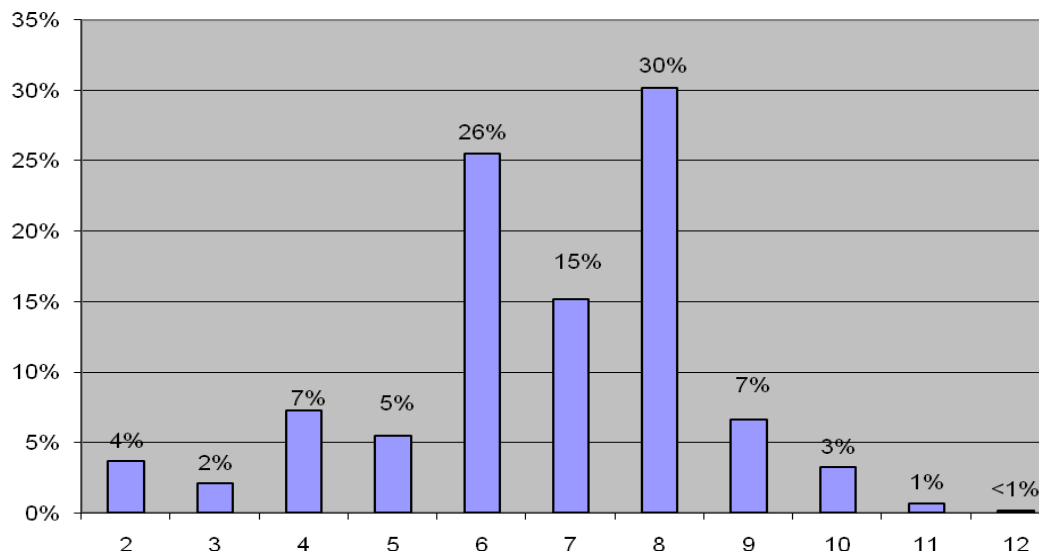


Figure 8-4. 2008–09 MHSAs: Maine Specific Distribution* of SAT Essay Scores for May and June 2009

*Includes data for students receiving official college reportable scores only.

The essay score is combined with the raw score earned on the multiple-choice portion of SAT writing and converted to the 200–800 reporting scale. This conversion is described in Chapter 11. The essay score constitutes approximately 30% of the total raw score, and the multiple-choice section makes up the remaining 70%. The distribution of SAT writing scores for the national 2008 College Board college bound seniors cohort and the associated percentile ranks are shown in Table 8-1.

Table 8-1. 2008–09 National SAT Writing Percentile Ranks

<i>Score</i>	<i>Writing Percentile Rank</i>	<i>Score</i>	<i>Writing Percentile Rank</i>	<i>Score</i>	<i>Writing Percentile Rank</i>
800	99+	590	79	380	13
790	99	580	77	370	11
780	99	570	74	360	9
770	99	560	71	350	8
760	99	550	69	340	6
750	99	540	65	330	5
740	98	530	62	320	4
730	98	520	59	310	3
720	97	510	55	300	3
710	97	500	52	290	2
700	96	490	49	280	2
690	95	480	45	270	1
680	94	470	42	260	1
670	93	460	38	250	1
660	92	450	34	240	1
650	90	440	31	230	1
640	89	430	28	220	-1
630	87	420	24	210	-1
620	85	410	22	200	–
610	84	400	18	Mean	494
600	82	390	16	SD	110

SD = standard deviation

For the SAT writing scores from the 2007 college bound seniors cohort, the mean was 494 and the standard deviation was 109.

8.6 End to End Quality Control

In addition to specific quality checks at each functional step, the College Board has an end to end quality assurance program that follows selected cases from receipt through reporting. The program selects answer sheets from all variations of forms and spirals to ensure that what is gridded on the answer sheet is accurately represented in the final delivered score report.

8.7 Quality Assessments

Starting with registration and continuing through score reporting, the College Board's quality engineering department performs onsite process reviews to ensure that all documented procedures have been followed. These assessments include reviewing the results of quality control checks, ensuring that the processes are performing as specified.

8.8 Summary

The SAT component of the MHSA is scored through a combination of electronic technology and human readers. The resulting raw scores are then converted to the familiar 200–800 scale using statistical procedures that ensure the comparability of scores across administrations. These steps allow students, parents, teachers, counselors, and admissions officers to utilize the scores while providing a common yardstick to augment other student information. These SAT component scores are then translated into Maine's 80 point achievement scale used for accountability purposes at all grade levels from three through eight and high school. Details of this process are found in Chapter 11.

Chapter 9. SCORING MATH–A AND SCIENCE

Scoring is a critical process in any large assessment. This chapter defines the scoring process used for the MHSA. Complete scoring specifications are provided in Appendix F.

9.1 Scoring MHSA Test Items

Upon receipt of completed MHSA Math–A and science answer booklets following testing, Measured Progress scanned all student responses, along with student identification and demographic information. Imaged data for multiple-choice responses were machine scored. Images of constructed-response items were processed and organized by iScore, a secure server-to-server electronic scoring software designed by Measured Progress, for hand scoring.

Student responses that could not be physically scanned (e.g., answer documents damaged during shipping) and typed responses submitted according to applicable test accommodations were physically reviewed and scored on an individual basis by trained, qualified readers. These scores were linked to the student’s demographic data and merged with the student’s scoring file by Measured Progress’s data processing department.

9.1.1 Machine Scored Items

Multiple-choice responses were compared to scoring keys using item analysis software. (Chapter 11 describes how correct, incorrect, and blank responses were assigned scores.)

The hardware elements of the scanners monitored themselves continuously for correct read, and the software driving these scanners monitored the correct data reads. Standard checks included recognition of a sheet that did not belong, was upside down, or was backward; identification of critical data that was missing, including a student ID number or test form that was out of range or missing; and identification of page/document sequence errors. When a problem was detected, the scanner stopped and displayed an error message directing the operator to investigate and correct the situation.

9.1.2 Hand Scored Items

The images of student responses to constructed-response items were hand scored through the iScore system. Using iScore minimized the need for readers to physically handle actual answer booklets and related scoring materials. Student confidentiality was easily maintained, as the MHSA scoring was “blind” (i.e., district, school, and student names were not visible to readers). The iScore system maintained the linkage between the student response images and their associated test booklet numbers.

Through iScore, qualified readers accessed electronically scanned images of student responses at computer terminals. The readers evaluated each response and recorded each student’s score via keypad or

mouse entry through the iScore system. When a reader finished one response, the next response appeared immediately on the computer screen.

Imaged responses from all answer booklets were sorted into item specific groups for scoring purposes. Readers reviewed responses from only one item at a time; however, when necessary, imaged responses from a student’s entire booklet were available for viewing, and the actual physical booklet was also available to the onsite chief reader.

For scoring of 2008–09 MHSA equating items, at least 200 responses from previous MHSA administrations were “seeded” among all equating item responses (constructed-response items) for scaling and equating purposes.

The use of iScore also helped ensure that access to student response images was limited to only those who were scoring or who were working for Measured Progress in a scoring management capacity.

9.2 Scoring Locations and Staff

9.2.1 Scoring Locations

The iScore database, its operation, and its administrative controls are all based in Dover, New Hampshire; however, 2008–09 MHSA test item responses were scored in Troy, New York.

The iScore system monitored accuracy, reliability, and consistency across all scoring sites. Constant daily communication and coordination were accomplished through e-mail, telephone, faxes, and secure Web sites, to ensure that critical information and scoring modifications were shared and implemented across all scoring sites.

9.2.2 Staff Positions

The following staff members were involved with scoring the 2008–09 MHSA responses:

- The **MHSA scoring project manager**, an employee of Measured Progress based in Dover, New Hampshire, oversaw communication and coordination of scoring across all scoring sites.
- The **iScore operational manager and iScore administrators**, employees of Measured Progress based in Dover, New Hampshire, coordinated technical communication across all scoring sites.
- A **scoring center manager**, an employee of Measured Progress located in Troy, New York, provided logistical coordination for his or her scoring site.
- A **chief reader** in science ensured consistency of scoring for this grade and content area. The chief reader, an employee of Measured Progress, also provided read behind activities for quality assurance coordinators.
- Numerous **quality assurance coordinators (QACs)**, selected from a pool of experienced senior readers for their ability to score accurately and their ability to instruct and train readers, participated in benchmarking activities for this specific grade and content area combination.

QACs provided read behind activities for senior readers at the site. The ratio of QACs and senior readers to readers was approximately 1 to 11.

- Numerous **senior readers (SRs)**, selected from a pool of skilled and experienced readers, provided read behind activities for the readers at their scoring tables (2–12 readers at each table). The ratio of QACs and SRs to readers was approximately 1 to 11.
- **Readers** scored the operational 2008–09 MHSA science test responses. The recruitment of readers is described in section 10.2.2.1.

9.2.2.1 Reader Recruitment and Qualifications

For scoring of the 2008–09 MHSA tests, Measured Progress actively sought a diverse scoring pool that was representative of the population of Maine. The broad range of readers included scientists, editors, business professionals, writers, teachers, graduate school students, and retired educators. Demographic information for readers (e.g., gender, race, educational background) was electronically captured and reported.

Although a four year college degree or higher was preferred, readers of the grade 11 responses were required to have successfully completed at least two years of college and to demonstrate knowledge of the content area they were to score. This permitted the recruitment of readers who were enrolled in a college program, a sector of the population that had relatively recent exposure to current classroom practices and current trends in their field of study. In all cases, potential readers submitted documentation (e.g., resume and/or transcripts) of their qualifications.

Table 9-1 summarizes the qualifications of the 2008–09 MHSA scoring leadership and readers.

Table 9-1. 2008–09 MHSA: Qualifications of Scoring Leadership and Readers—March 2009 Administration

Scoring Responsibility	Educational Credentials				Total
	Doctorate	Masters	Bachelors	Other	
Scoring leadership	0%	38%	63%	0%	100%
Readers	9%	31%	52%	7%	100%

Readers were temporary employees secured through the services of one or more temporary employment agencies. All readers signed a nondisclosure or confidentiality agreement.

9.2.2.2 Reader Training

Reader training began with an introduction of onsite scoring staff and an overview of the MHSA program’s purpose and goals, including a discussion about the security, confidentiality, and proprietary nature of testing and scoring materials and procedures.

Next, readers thoroughly reviewed and discussed the scoring guide for the item to be scored. Each item specific scoring guide included the item itself and score point descriptions.

Following review of the item specific scoring guide for any constructed-response item, readers began reviewing or scoring the following response sets, organized for specific training purposes:

- Anchor set
- Training set
- Qualifying set

During training, readers were able to highlight or mark hard copies of the anchor, training, and first qualifying sets, even if all or part of the set was also presented online via computer.

Anchor Set

Readers first reviewed an anchor set of exemplary responses, approved by the specialists representing the MDOE, for the item to be scored. Responses in anchor sets were typical, rather than unusual or uncommon; solid, rather than controversial or borderline; and true, meaning that they had scores that could not be changed by anyone other than MDOE test development staff.

Responses were read aloud to the room of readers and presented in descending score order. Trainers then announced the true score of each anchor response and facilitated a group discussion of the response in relation to the score point descriptions to allow readers to internalize typical characteristics of each score point.

This anchor set served as a reference for readers as they continued with calibration, scoring, and recalibration activities for that item.

Training Set

Next, readers practiced applying the scoring guide to responses in the training set. The training set typically included 10 to 15 student responses designed to help establish the score point range and the range of responses within each score point. The training set often included unusual responses that were less clear or solid (e.g., were shorter than normal, employed atypical approaches, contained both very low and very high attributes, or included difficult handwriting). Responses in the training set were presented in randomized score point order.

After readers had independently read and scored a training set response, trainers polled readers or used online training system reports to record the initial range of scores. Then they led a group discussion of the responses, directing reader attention to scoring issues that were particularly relevant to the specific scoring group, such as the line between two score points. Trainers modeled for readers how to discuss scores by referring to the anchor set and to scoring guides.

Qualifying Set

After the training set had been completed, for all items, readers were required to measurably demonstrate their ability to accurately and reliably score the item according to its scoring rubric by scoring responses in the qualifying set. The 10 responses in the qualifying set, selected from an array of responses that clearly illustrated the range of score points for that item, were chosen in accordance with the responses reviewed and approved by the MDOE. Hard copies of the responses were also available to readers so that they could make notes and refer back to specific responses during the postqualifying discussion.

To be eligible to score operational 2008–09 MHSA responses, readers of all items were required to demonstrate scoring accuracy rates of at least 80% exact agreement and at least 90% exact or adjacent agreement. In other words, exact scores were required on at least eight of the qualifying set responses and either exact or adjacent scores were required on at least nine; readers were allowed one discrepant score, as long as they had at least eight exact scores.

Readers who met the percentage requirements were allowed to score operational student responses.

Retraining

Readers who did not pass the first qualifying set were retrained as a group by reviewing their performance with scoring leadership and then scoring a second qualifying set of responses. If they achieved a minimum scoring accuracy rate of 80% exact and 90% exact or adjacent agreement on this second set, they were allowed to score operational responses.

If readers did not achieve the required scoring accuracy rates on the second qualifying set, they were not allowed to score responses for that item; instead, they either began training on a different item or were dismissed.

9.2.2.3 Senior QAC and SR Training

QACs and select SRs were trained in a separate training session that occurred immediately prior to reader training. In addition to discussing the items and their responses, QAC and SR training included emphasis on the MDOE’s rationale behind the score points. This rationale was discussed in greater detail with QACs and SRs than with regular readers to better equip leadership to handle questions from the regular readers.

9.2.2.4 Monitoring of Scoring Quality Control and Consistency

Readers were monitored for continued accuracy rates and scoring consistency throughout the scoring process, using the following methods and tools:

- Embedded committee reviewed responses (CRRs)
- Read behind procedures

- Double scoring
- Recalibration sets
- Scoring reports

If readers met or exceeded the expected accuracy rate, they continued scoring operational responses. Any reader whose accuracy fell below the expected rate for the particular item and monitoring method was retrained on that item and, upon approval by the QAC or chief reader, as appropriate (see below), was allowed to resume scoring.

There is a difference between the accuracy rate each reader must have achieved to qualify to begin scoring live responses and the accuracy rate each reader must have maintained to continue scoring live responses. Specifically, the accuracy rate to qualify was stricter than the accuracy rate a reader was expected to maintain while scoring live responses. The reason for this difference was that an “exact score” in double blind statistics required that two readers both identify the same score for a response during live scoring, whereas an exact score during qualification did not require agreement between two readers; instead, it required only that each individual reader match the score that was predefined by scoring leadership.

The accuracy rates of readers were monitored using an array of techniques, thereby providing a more complete picture of a reader’s performance than would be the case by relying on just one technique.

Embedded CRRs

Previously scored “embedded” CRRs were selected and loaded into iScore for blind distribution to readers as a way to monitor reader accuracy. CRRs, either chosen before scoring had begun or selected by scoring leadership during scoring, were inserted into the scoring queue such that they appeared the same as all the other live student responses.

Between 5 and 30 CRRs were distributed at random points throughout the first full day of operational scoring to ensure that readers were sufficiently calibrated at the beginning of the scoring period. Individual readers often received up to 20 CRRs within the first 100 responses scored, and up to 10 additional responses within the next 100 responses scored on that first day of scoring.

If any reader fell below the required scoring accuracy rate, he or she was retrained before being allowed by the QAC to continue scoring. Once the reader was allowed to resume scoring, leadership carefully monitored him or her by increasing the number of read behinds.

Read Behind Procedures

Read behind scoring refers to the practice of scoring leadership, usually an SR, scoring a response after a reader has already scored that same response.

Responses to be placed in the read behind queue were randomly selected by scoring leadership; readers were not made aware as to which of their responses was to be reviewed by their SR. The iScore system allowed one, two, or three responses per reader to be placed in the read behind queue at a time.

The SR entered his or her score into iScore before being allowed to see the score assigned by the reader for whom the read behind was being performed. The SR then compared scores, and the reported score was determined as follows:

- If there was exact agreement between the scores, no action was taken; the regular reader’s score remained.
- If the scores were adjacent (i.e., the difference was not greater than 1), the SR’s score became the score of record; if there were a significant number of adjacent scores for this reader, an individual scoring consultation was held with the reader, and the QAC determined whether or when the reader could resume scoring.
- If there was a discrepant difference between the scores (greater than 1 point), the SR’s score became the score of record (see Table 9-2). An individual consultation was held with the reader, with the QAC determining whether or when the reader could resume scoring.

Table 9-2. 2008–09 MHSA: Examples of MHSA Read-Behind Scoring Resolutions

<i>Reader</i>	<i>QAC/SR Resolution</i>	<i>Final*</i>
4	4	4
4	3	3
4	2	2

* QAC score is score of record.

An average of 5 read behinds per reader was conducted throughout each half scoring day, with an average of 10 read behinds per reader conducted throughout each full scoring day. If a reader’s scoring rate fell below the required accuracy percentage, additional read behinds were performed.

In addition to the daily read behinds, scoring leadership could choose to read behind any reader at any point during the scoring process, thereby providing an immediate, real-time “snapshot” of a reader’s accuracy.

Double Scoring

Double scoring refers to the practice of two readers independently scoring a response, each without knowing the identity of the other reader or the score assigned to the response by the other reader. Readers were not told which of their responses was to be reviewed by a second reader.

If there was a discrepancy (a difference greater than 1) between scores, the response was placed in an arbitration queue. Arbitration responses were reviewed by scoring leadership (SR or QAC) without any background knowledge of scores assigned by the two previous readers. Table 9-3 provides the 2008–09 MHSA percentages of agreement between readers for the science test.

Table 9-3. 2008–09 MHSA Grade 11 Science Open-response Double Scoring Interrater Agreement by Item

<i>Item Number</i>	<i>Percent Exact + Adjacent</i>	<i>Percent Exact</i>	<i>Percent Adjacent</i>	<i>Percent >1</i>
21	98.8	79.8	19.0	1.2
23	98.7	88.1	10.7	1.3
52	97.8	79.5	18.3	2.2
53	99.6	93.1	6.4	0.4
TOTAL	98.8	86.4	12.5	1.2

Scoring leadership consulted individually with any reader whose scoring rate fell below the required accuracy percentage, and the QAC determined whether or when the reader could resume scoring. Once the reader was allowed to resume scoring, leadership carefully monitored him or her by increasing the number of read behinds.

Recalibration Sets

Recalibration (recal) sets were used when readers were scoring an item on which they had been trained (and qualified) on a previous day. To determine whether they were still calibrated to the scoring standard, readers took an online recal set at the start and midpoint of each scoring shift. .

Each recal set consisted of five responses representing the entire range of possible scores, including some with a score point of 0.

- Readers who were discrepant on two responses or were exact on only two or less of the five responses of the first recal set did not score on that item for that day and were either reassigned or dismissed from scoring for that day.
- Readers who were discrepant on only one response and/or exact on just three of the five responses of the first recal set were retrained by their SR by discussing the recal set responses in terms of the score point descriptions and the original anchor set. After this retraining, such readers began scoring operational responses with the caveat that their scores for that day and that item would be kept only if they were exact on all five of the responses on the second recal set, to be administered at the midpoint of the shift. The QAC determined whether or when readers had received enough retraining to begin scoring operational responses. Scoring leadership also carefully monitored the accuracy of such readers by significantly increasing the number of read behinds..
- Readers who were not discrepant on any response and were exact on at least four of the five responses of the first recal set began scoring operational responses but were still required to meet the accuracy standard on the second recal set, administered at the midpoint of the shift, to determine whether their scores for that day and that item were to be kept.

Readers were required to achieve an 80% exact and 90% adjacent standard overall for the responses presented in the two recal sets for that item for that day.

The scoring project manager voided scores posted (on that item for that day) by readers who did not meet the accuracy requirement for an item. The responses associated with these voided scores were reset and redistributed for scoring by readers who had met the scoring accuracy standard for that item for that day.

9.2.3 Benchmarking Meetings with the MDOE

In preparation for implementing MDOE guidelines for the scoring of operational responses, Measured Progress scoring staff prepared and facilitated benchmarking meetings with MDOE representatives. The purpose of the meetings was to establish item specific guidelines for scoring each MHSA item for current and future scoring sessions.

Prior to the benchmarking meetings, the scoring staff collected a set of several dozen student responses that chief readers identified as being illustrative midrange examples of their respective score points. As a matter of practice, each of these authoritative sets is included as part of the scoring training materials and used to train readers each time that item is scored—both as a field test item and as part of a future MHSA administration.

This repeated use of MDOE approved sets of midrange score point exemplars helps ensure that each time a particular MHSA item is scored readers follow the established guidelines.

9.3 Methodology for Scoring Constructed-response Items

Constructed-response items were scored based on possible score points and scoring procedures, as shown in Table 9-4

Table 9-4. 2008–09 MHSA: Possible Score Points for MHSA Science Constructed-response Items

<i>Item Type</i>	<i>Possible Score Points</i>	<i>Possible Highest Score</i>
Constructed-response	0–4	4
Non-scorable items	0	0

Nonscorable Items

Readers could designate a response as non-scorable for any of the following reasons:

- Response was blank (no attempt to respond to the question)
- Response was unreadable (illegible, too faint to see, or only partially legible/visible)—see note below
- Response was written in the wrong location (seemed to be a legitimate answer to a different question)—see note below
- Response was written in a language other than English
- Response was completely off task or off topic
- Response included an insufficient amount of material to make scoring possible
- Response was an exact copy of the assignment

- Response was incomprehensible
- Student made a statement refusing to write a response to the question

Note: Unreadable and wrong location responses were eventually resolved, whenever possible, by researching the actual answer document (electronic copy or hard copy, as needed) to identify the correct location or to more closely examine the response and then assign a score.

Scoring Procedures

Scoring procedures for constructed-response items included both single scoring and double scoring. Single scored items were scored by one reader. Double scored items were scored independently by two readers, whose scores were tracked for agreement.

Table 9-5 shows by which method(s) common and equating constructed-response science items responses were scored.

**Table 9-5. 2008–09 MHSA: Methods of Scoring
Common and Equating Science Constructed-response Items**

<i>Grade</i>	<i>Test/Pilot Test Name</i>	<i>Responses Single Scored (per grade and test/pilot test)</i>	<i>Responses Double Scored (per grade and test/pilot test)</i>
HS	Science	100%	10% randomly
All	Unreadable responses	100%	100%
All	Blank responses	100%	100%

For each embedded field test item, 1,200 responses were scored.

9.4 Scoring Reports

Measured Progress’s electronic scoring software, iScore, generated multiple reports that were used by scoring leadership to measure and monitor readers for scoring accuracy, consistency, and productivity..

Reports Generated During Scoring

For the 2008–09 MHSA science test administration, computer generated reports were run to ensure all of the following:

- Overall group level accuracy, consistency, and reliability of scoring were maintained and acceptable
- Immediate, real-time individual reader data were available to allow early reader intervention when necessary
- Scoring schedules were maintained

The following reports were produced by iScore:

- The **Read Behind Summary** showed the total number of read behind responses for each reader, and noted the numbers and percentages of scores that were exact, adjacent, and discrepant between that reader and the SR or QAC. Scoring leadership could choose to generate this report by choosing options such as “Today,” “Past Week,” or “Cumulative” from a pull down menu. The report could also be filtered to select data for a particular item or across all items. This report was used in conjunction with other reports to determine whether a reader’s scores would be voided (i.e., responses scored by that reader would be sent back out to the floor to be rescored by other readers). The benefit of this report is that it measures the degree to which individual readers agree with their QAC or SR on how to best score live responses.
- The **Double Blind Summary** showed the total number of double score responses scored by each reader, and noted the numbers and percentages of scores that were exact, adjacent, and discrepant between that reader and the second reader. This report was used in conjunction with other reports to determine whether a reader’s scores would be voided. The benefit of this report is that it reveals the degree to which readers are in agreement with each other about how to best score live responses.
- The **Accuracy Summary** combined read behind and double score data, showing the total number of double score and read behind responses scored for each reader, and noting his or her accuracy percentages and score point distributions.
- The **Embedded CRR Summary** showed, for each reader and for either a particular item or across all items, the total number of responses scored, the number of CRRs scored, and the numbers and percentages of scores that were exact, adjacent, and discrepant between the reader and the SR or QAC. This report was used in conjunction with other reports to determine whether a reader’s scores would be voided. The benefit of this report is that it measures the degree to which individual readers agree with their chief reader on how to best score live responses—and since embedded responses are administered during the first hours of scoring, this report provides an early illustration of agreement between readers and their chief reader.
- The **Qualification Statistics Report** listed all readers by name and ID number, identifying which qualifying set(s) they did and did not take and, for the ones they did take, whether they passed or failed. The total number of qualifications passed and failed was noted for each reader, as was the total number of individuals passing or failing a particular qualifying set.

The QAC could use this report to determine how the readers within his or her specific scoring group performed on a specific qualifying set.

- The **Summary Report** showed the total number of student responses for an item and identified, for the time at which the report was generated, (1) the number of single and double scorings that had been performed, and (2) the number of single and double scorings yet to be performed.

SECTION IV—PSYCHOMETRICS AND REPORTING

This section discusses the psychometric properties of the MHSA, which contains the restructured SAT and the Math–A component.

Chapter 10. PSYCHOMETRIC TOPICS OF THE SAT

The use of the SAT supports Maine’s vision of graduating all high school students as college, career, and citizenship ready by assessing how students apply what they have learned in high school to analyze and solve problems they will likely encounter in college. The 2005 changes to the test were initiated to “strengthen the alignment of the SAT to the instructional practices in today’s classroom and to address the importance of writing skills” (College Board, 2005c). The critical reading section represents increasingly heavier reliance on a reading construct, with approximately 72% reading comprehension items. Examinees are allotted 70 minutes to answer the 67 multiple-choice items in the critical reading section. The SAT mathematics section contains 54 items in total—44 multiple-choice and 10 student-produced responses—with an allotted time of 70 minutes to answer the items. The mathematics section covers mathematical concepts through third year college preparatory mathematics. The SAT mathematics is augmented by Math–A in order to fully measure Maine’s standards. The writing section contains 49 multiple-choice questions with an allotted time of 60 minutes, and a 25 minute section in which the student produces a response to an essay prompt. The writing section is intended to measure how well students use standard written English.

10.1 The Equating and Braiding Plan for SAT Mathematics, Critical Reading, and Writing

This section outlines the equating and braiding plan for the SAT forms. *Equating* refers to the statistical process used to ensure that the reported scores on each version of the SAT have the same meaning as on every other version. SAT equating employs two types of data collection: the nonequivalent groups anchor test (NEAT) design and equivalent groups (EG) design. At each SAT administration of one new form, the new form is linked to multiple old SAT forms through a NEAT design. One of the old forms was administered to a similar sample from a similar population—that is, to a sample of students who were administered the SAT during the same month in a previous year. Each of the other old forms was administered at one of the core administrations of the SAT that contribute large numbers of scores to the SAT cohort. The final conversion line is the weighted average line of the four individual lines, with more weight (usually 50%) given to the link to the old form that was administered to a sample from the similar population, defined as the group of students testing in the same administration one year previously. This data collection design has been shown to produce stable equating results because it directly acknowledges the important role that the old form linking plays in placing a new form on scale (Dorans, Liu, & Hammond, in press).

An EG design is usually employed in an SAT administration with two or more new forms, where the first new form is equated using the NEAT design and the second new form is equated to the first one through an EG design. The spiraling procedure used in the SAT administration and the large numbers of test takers who take each form usually ensure equivalent groups in the same administration.

The Math–A

The Math–A was added to the MHSA in 2007 to supplement the SAT in meeting the Maine *Learning Results*. Details of the development of the Math–A and information on the development and contents of the Math–A form for 2008–09 are presented in Chapter 2.

10.2 SAT Statistical Characteristics

The statistical characteristics of the SAT, based on the two forms administered in May and June 2009, are examined in this section. The test level statistics include reliability, standard errors of measurement (SEM), and test speededness. The item level statistics include item difficulty, item discriminating power, and differential item functioning (DIF). Analyses for the SAT conducted on the national SAT population and not specific to Maine are presented in Appendix G. Tables G-1 through G-3 provide summaries of the scores for examinees participating in SAT testing in May and June 2009 by section for each form. Tables G-4 through G-6 present the rounded scaled score conversion tables by section for each SAT form.

10.3 Reliability and Standard Errors of Measurement

10.3.1 Reliability

Reliability is an indicator of the consistency or stability of test scores. Test scores that are used for making important decisions should be very reliable. The estimates of reliability detailed in this report are internal consistency measures, which are 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 are they responsive to day to day variation due, for example, to state of health or testing environment.

The reliability and SEM on the national equating sample for the mathematics, critical reading, and writing sections are within normally acceptable ranges (see Table G-7 of Appendix G). Due to makeup testing administrations and special forms for students with disabilities, students in Maine took one of up to four test forms. Using recommendations in the literature as to the size of the sample needed to obtain stable estimates, reliability estimates were calculated only for test forms and subgroups with at least 200 examinees (Kline, 1986; Charter, 1999). The reliability estimates for Maine students only are reported in Table 10-1. These values range from 0.69 to 0.94 for mathematics, 0.81 to 0.94 for critical reading, and 0.75 to 0.91 for writing. This supports the use of SAT scores for students in Maine and is evidence that the reliability of scores for Maine students is

comparable to that of the national sample. Reliability estimates were also computed for subgroups that met the minimum sample size requirements: males, females, students with disabilities, economically disadvantaged students, and students with limited English proficiency (beyond the first year). Subgroup reliabilities range from 0.87 to 0.94 in mathematics, 0.88 to 0.94 in critical reading, and 0.82 to 0.91 in writing, with students with disabilities showing the lowest reliability coefficients and males generally showing the highest. Maine subgroup reliabilities are reported in Table 10-2. Average SAT scores and standard deviations on the raw score scale for Maine students are reported in Table 10-3.

10.3.2 Standard Errors of Measurement

The SEM is an estimate of the amount of variation that can be expected in obtained scores for the same individual if the person were to retake the test with no change in knowledge between administrations or for individuals with the same true score. The interpretation of the SEM is usually made in terms of a statement of probability that the score obtained by an individual is within a certain distance of his or her true score (that is, the score he or she would obtain on a perfectly reliable test). The probability is 0.68 that an individual's score will be within one SEM of his or her true score and 0.95 that it will be within two SEMs (assuming a normal distribution). The SEMs for Maine students only are reported in Tables 10-1 and 10-2 for the total Maine group and Maine subgroups, respectively. All raw score SEMs for the total Maine group and for the Maine subgroups ranged from 2.2 to 4.2 for mathematics, 1.9 to 4.1 for critical reading, and 2.0 to 3.6 for writing. Form 2 reliabilities and SEMs were not provided for the Maine specific sample due to small sample size.

Table 10-1. 2008–09 MHSА: Reliability Coefficients and SEMs* for Sections of the MHSА**

			Form 1—May 2009	
			N= 13,230	
Test Section			Reliability	SEM
Math 1	Dressel-KR20	Raw	0.79	2.2
Math 2	Dressel-KR20	Raw	0.81	1.8
Math 3	Dressel-KR20	Raw	0.76	2.0
Math–A	Dressel-KR20	Raw	0.61	1.8
Total MHSА mathematics (includes augment)	Alpha	Raw	0.93	3.3
	Var. components	Raw	0.93	3.9
Critical Reading 1	Dressel-KR20	Raw	0.81	2.5
Critical Reading 2	Dressel-KR20	Raw	0.83	2.5
Critical Reading 3	Dressel-KR20	Raw	0.80	2.3
Total critical reading	Alpha	Raw	0.93	3.4
	Var. components	Raw	0.93	4.2
Writing 1	Dressel-KR20	Raw	0.88	3.1
Writing 2	Dressel-KR20	Raw	0.74	1.9
Total writing MC	Alpha	Raw	0.91	2.9
	Var. components	Raw	0.91	3.6

* See Appendix H for formulas used to compute reliability coefficients and SEMs.

** Estimates are computed based on Maine students only for the two forms that were taken by the majority of Maine students and had sufficient sample size.

MC = multiple-choice

Table 10-2. 2008–09 MHSА: Reliability Coefficients and SEMs for Sections of the MHSА*

		Form 1—May 2009					
Test Section	Subgroup	N	KR-20		Variance Components		
			Reliability	SEM	Reliability	SEM	
Total MHSА mathematics (includes augment)	Male	6,714	0.93	3.3	0.93	3.9	
	Female	6,516	0.92	3.3	0.92	3.9	
	Students with disabilities	1,192	0.85	3.2	0.85	3.9	
	Economically disadvantaged	3,489	0.90	3.3	0.90	3.9	
	Limited English proficiency (beyond first year)	243	0.91	3.3	0.91	4.0	
Total critical reading	Male	6,714	0.93	3.4	0.93	4.2	
	Female	6,516	0.93	3.4	0.93	4.2	
	Students with disabilities	1,192	0.88	3.4	0.88	4.2	
	Economically disadvantaged	3,489	0.91	3.5	0.91	4.2	
	Limited English proficiency (beyond first year)	243	0.90	3.4	0.90	4.2	
Total writing MC	Male	6,714	0.91	3.0	0.91	3.6	
	Female	6,516	0.90	2.9	0.90	3.6	
	Students with disabilities	1,192	0.84	3.0	0.84	3.7	
	Economically disadvantaged	3,489	0.89	3.0	0.89	3.7	
	Limited English proficiency (beyond first year)	243	0.87	3.0	0.87	3.7	

* Estimates are calculated based on Maine students only for subgroups where sufficient sample sizes were present.
MC = multiple-choice

**Table 10-3. 2008–09 MHSA: Raw Score Summary
Statistics for Total Group and Subgroups**

Form 1 May 2009		Mathematics			Critical Reading			Writing		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
Gender	Male	6714	21.3	12.7	6714	23.8	16.1	6714	19.6	12.0
	Female	6516	19.9	11.5	6516	25.8	15.4	6516	22.3	11.6
	All	13230	20.6	12.1	13230	24.8	15.8	13230	20.9	11.9
Students with disabilities	Yes	1192	8.7	8.7	1192	9.9	11.8	1192	8.9	9.2
	No	12038	21.8	11.8	12038	26.3	15.3	12038	22.1	11.5
	All	13230	20.6	12.1	13230	24.8	15.8	13230	20.9	11.9
Economically disadvantaged	Yes	3489	16.0	10.7	3489	19.0	14.2	3489	16.2	11.0
	No	9741	22.2	12.2	9741	26.9	15.8	9741	22.6	11.7
	All	13230	20.6	12.1	13230	24.8	15.8	13230	20.9	11.9

SD = standard deviation

The SEMs reported in these tables represent the average of the *conditional* standard errors of measurement; that is, the SEM is not the same at all score levels. The term *conditional standard error of measurement* (CSEM) indicates the SEM that is associated with a particular score level. Scaled scores are more or less accurate at different points on the scale, typically more accurate in the middle of the scale and less accurate at the ends of the scale. CSEM bands are reported in Tables 10-4 through 10-6 for the total Maine group. (Note that MHSA scaled scores take on only even values within the 1100–1180 range.) Due to the small sample size, CSEMs for the total Maine group in June 2009 are not reported.

**Table 10-4. 2008–09 MHSA: Scaled Score
Conditional Standard Error Bands for MHSA Mathematics**

Scaled Score	Lower Bound*	Upper Bound*	Scaled Score	Lower Bound	Upper Bound
1100	1100.0	1104.3	1142	1140.0	1144.0
1102	1105.5	1125.0	1144	1141.8	1146.2
1104	1105.5	1125.0	1146	1144.8	1147.2
1106	1105.5	1125.0	1148	1146.2	1149.8
1108	1105.5	1125.0	1150	1147.9	1152.1
1110	1105.5	1125.0	1152	1149.9	1154.1
1112	1105.5	1125.0	1154	1151.8	1156.2
1114	1105.5	1125.0	1156	1154.4	1157.6
1116	1105.5	1125.0	1158	1156.3	1159.7
1118	1105.5	1125.0	1160	1158.1	1161.9
1120	1105.5	1125.0	1162	1159.4	1164.6
1122	1105.5	1125.0	1164	1160.9	1167.1
1124	1111.0	1137.0	1166	1162.4	1169.6
1126	1119.4	1132.6	1168	1164.0	1172.0
1128	1126.4	1129.6	1170	1166.5	1173.5
1130	1128.5	1131.5	1172	1168.5	1175.5
1132	1130.7	1133.3	1174	1170.9	1177.1
1134	1132.0	1136.0	1176	1173.5	1178.0
1136	1134.5	1137.5	1178	1175.0	1180.0
1138	1136.2	1139.8	1180	1178.7	1180.0
1140	1138.8	1141.2			

* Because there are a variety of ways to achieve any particular mathematics scaled score (due to the combination of SAT and augment items), the upper and lower bound values in this table reflect the averages of their respective distributions, each calculated using the Lord (1980) binomial method.

**Table 10-5. 2008–09 MHTA: Scaled Score
Conditional Standard Error Bands for MHTA Critical Reading**

<i>Scaled Score</i>	<i>Lower Bound*</i>	<i>Upper Bound*</i>	<i>Scaled Score</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
1100	1109.0	1111.0	1142	1137.0	1147.0
1102	1109.0	1111.0	1144	1139.0	1149.0
1104	1109.0	1111.0	1146	1141.0	1151.0
1106	1109.0	1111.0	1148	1143.0	1153.0
1108	1109.0	1111.0	1150	1145.0	1155.0
1110	1109.0	1111.0	1152	1147.0	1157.0
1112	1111.0	1113.0	1154	1149.0	1159.0
1114	1112.0	1117.0	1156	1151.0	1161.0
1116	1114.0	1119.0	1158	1153.0	1163.0
1118	1115.5	1121.5	1160	1155.0	1165.0
1120	1116.5	1124.0	1162	1157.0	1167.0
1122	1118.0	1126.0	1164	1160.0	1168.5
1124	1121.0	1128.0	1166	1162.0	1170.0
1126	1122.0	1130.0	1168	1164.5	1172.0
1128	1124.7	1132.0	1170	1166.5	1174.0
1130	1125.0	1135.0	1172	1168.0	1176.0
1132	1127.0	1137.0	1174	1171.0	1178.0
1134	1129.0	1139.0	1176	1172.0	1180.0
1136	1131.0	1141.0	1178	1176.0	1180.0
1138	1133.0	1143.0	1180	1179.0	1180.0
1140	1135.0	1145.0			

* Because there are a variety of ways to achieve any particular reading scaled score (the same raw score can be arrived at in multiple ways due to formula scoring), the upper and lower bound values in this table reflect the averages of their respective distributions, each calculated using the Lord (1980) binomial method.

**Table 10-6. 2008–09 MHTA: Scaled Score
Conditional Standard Error Bands for MHTA Writing**

<i>Scaled Score</i>	<i>Lower Bound*</i>	<i>Upper Bound*</i>	<i>Scaled Score</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
1100	1112.0	1113.0	1142	1137.0	1147.0
1102	1112.0	1113.0	1144	1139.5	1149.0
1104	1112.0	1113.0	1146	1142.0	1151.0
1106	1112.0	1113.0	1148	1143.0	1153.0
1108	1112.0	1113.0	1150	1145.5	1155.0
1110	1112.0	1113.0	1152	1148.0	1157.0
1112	1112.0	1113.0	1154	1149.5	1159.0
1114	1113.0	1115.0	1156	1151.0	1161.0
1116	1114.0	1118.0	1158	1154.0	1163.0
1118	1116.0	1121.0	1160	1155.5	1165.0
1120	1117.5	1123.5	1162	1157.0	1167.0
1122	1118.0	1126.0	1164	1160.0	1168.5
1124	1121.0	1128.0	1166	1163.0	1170.0
1126	1123.0	1130.0	1168	1165.0	1172.0
1128	1125.0	1132.0	1170	1167.0	1174.0
1130	1127.0	1134.0	1172	1169.0	1176.0
1132	1128.5	1136.5	1174	1171.0	1178.0
1134	1129.5	1139.0	1176	1174.0	1178.0
1136	1131.0	1141.0	1178	1176.0	1180.0
1138	1134.0	1143.0	1180	1179.0	1180.0
1140	1135.5	1145.0			

* Because there are a variety of ways to achieve any particular writing scaled score (the same raw score can be arrived at in multiple ways due to formula scoring), the upper and lower bound values in this table reflect the averages of their respective distributions, each calculated using the Lord (1980) binomial method.

Tables 10-7 through 10-12 present additional descriptive information on scale score results.

Table 10-7. 2008–09 MHSА: SAT and Math–A Summary Statistics

<i>Content Area</i>	<i>N*</i>	<i>Mean</i>	<i>SD</i>	<i>Minimum</i>	<i>Maximum</i>
Mathematics**	15,032	1140.7	11.0	1100	1180
Critical reading	14,698	1141.2	14.6	1110	1180
Writing	14,705	1140.1	14.1	1112	1180

* From the May 2, 2009 SAT

** Includes data for the MHSА mathematics test (SAT and Math–A).

SD = standard deviation

**Table 10-8. 2008–09 MHSА: Frequency Distribution of MHSА Scores—
Mathematics (SAT/Math–A)**

<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>
1100	0	0.00
1102	0	0.00
1104	0	0.00
1106	0	0.00
1108	0	0.00
1110	150	1.00
1112	0	0.00
1114	1	0.01
1116	0	0.00
1118	4	0.03
1120	0	0.00
1122	0	0.00
1124	212	1.41
1126	553	3.68
1128	701	4.66
1130	772	5.14
1132	1,730	11.51
1134	276	1.84
1136	1,128	7.50
1138	1,315	8.75
1140	1,909	12.70
1142	655	4.36
1144	1,067	7.10
1146	798	5.31
1148	820	5.46
1150	642	4.27
1152	419	2.79
1154	451	3.00
1156	342	2.28
1158	262	1.74
1160	228	1.52
1162	96	0.64
1164	93	0.62
1166	91	0.61
1168	86	0.57
1170	44	0.29
1172	38	0.25
1174	27	0.18
1176	15	0.10
1178	12	0.08
1180	95	0.63

**Table 10-9. 2008–09 MHSА: Frequency Distribution of MHSА Scores—
Critical Reading**

<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>
1100	0	0.00
1102	0	0.00
1104	0	0.00
1106	0	0.00
1108	0	0.00
1110	389	2.65
1112	64	0.44
1114	111	0.76
1116	161	1.10
1118	261	1.78
1120	300	2.04
1122	273	1.86
1124	485	3.30
1126	365	2.48
1128	865	5.89
1130	338	2.30
1132	654	4.45
1134	866	5.89
1136	357	2.43
1138	931	6.33
1140	1,030	7.01
1142	573	3.90
1144	1,028	6.99
1146	498	3.39
1148	871	5.93
1150	718	4.89
1152	451	3.07
1154	565	3.84
1156	410	2.79
1158	471	3.20
1160	322	2.19
1162	217	1.48
1164	277	1.88
1166	225	1.53
1168	143	0.97
1170	176	1.20
1172	61	0.42
1174	81	0.55
1176	0	0.00
1178	92	0.63
1180	69	0.47

Table 10-10. 2008–09 MHTA: Frequency Distribution of MHTA Scores—Writing

<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>	<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>
1100	0	0.00	1142	454	3.09
1102	0	0.00	1144	989	6.73
1104	0	0.00	1146	923	6.28
1106	0	0.00	1148	302	2.05
1108	0	0.00	1150	846	5.75
1110	0	0.00	1152	587	3.99
1112	516	3.51	1154	621	4.22
1114	225	1.53	1156	274	1.86
1116	119	0.81	1158	353	2.40
1118	320	2.18	1160	363	2.47
1120	283	1.92	1162	203	1.38
1122	242	1.65	1164	252	1.71
1124	455	3.09	1166	188	1.28
1126	479	3.26	1168	51	0.35
1128	795	5.41	1170	142	0.97
1130	352	2.39	1172	82	0.56
1132	832	5.66	1174	30	0.20
1134	794	5.40	1176	45	0.31
1136	512	3.48	1178	22	0.15
1138	1,030	7.00	1180	48	0.33
1140	976	6.64			

Table 10-11. 2008–09 MHTA: Range of Scores for Each Achievement Level

<i>Content Area</i>	<i>Does Not Meet</i>	<i>Partially Meets</i>	<i>Meets</i>	<i>Exceeds</i>
Mathematics	1100–1132	1134–1140	1142–1160	1162–1180
Critical reading	1100–1128	1130–1140	1142–1160	1162–1180
Writing	1100–1128	1130–1140	1142–1160	1162–1180

Table 10-12. 2008–09 MHTA: Number and Percentage* of Students by Achievement Level

<i>Content Area</i>	<i>Does Not Meet</i>		<i>Partially Meets</i>		<i>Meets</i>		<i>Exceeds</i>	
	<i>N</i>	<i>percent</i>	<i>N</i>	<i>percent</i>	<i>N</i>	<i>percent</i>	<i>N</i>	<i>percent</i>
Mathematics	4,123	27.43	4,628	30.79	5,684	37.81	597	3.97
Critical Reading	3,274	22.28	4,176	28.41	5,907	40.19	1,341	9.12
Writing	3,434	23.35	4,496	30.57	5,712	38.84	1,063	7.23

*Percentages in this table may not always sum to 100 due to rounding.

10.4 Classification Accuracy and Consistency of Maine SAT Cut Scores

The Livingston and Lewis (1995) procedure was used in calculating classification accuracy and consistency of the cut scores. In their article Livingston and Lewis define *accuracy* as “the extent to which the actual classifications of test takers (on the basis of their single-form scores) agree with those that would be made on the basis of their true scores, if their true scores could somehow be known” and *consistency* as “the agreement between the classifications based on two non-overlapping, equally difficult forms of the test” (p. 180).

Besides student raw scores, the Livingston and Lewis (1995) method requires an estimate of reliability as part of the input for performing calculations. The reliability measures used in the analyses were 0.94 for both mathematics and critical reading and 0.91 for writing, numbers based on reliability estimates from Maine test takers.

Table 10-13 presents the accuracy contingency table for each section of the MHSAs (Livingston & Lewis, 1995). The rows represent categorizations based on true scores, estimated using the beta-binomial model, while the columns represent categorizations based on the actual observed scores in the data. The cells in the table show detailed information for adjacent and nonadjacent classification errors, comparing the observed classifications and true achievement categories.

Table 10-13. 2008–09 MHSAs: Accuracy Contingency for MHSAs Mathematics, Critical Reading, and Writing

Content Area	True	Observed				Total
		Does Not Meet	Partially Meets	Meets	Exceeds	
Mathematics	Does Not Meet	0.21	0.04	0.00	0.00	0.25
	Partially Meets	0.04	0.17	0.06	0.00	0.27
	Meets	0.00	0.05	0.39	0.01	0.45
	Exceeds	0.00	0.00	0.00	0.02	0.02
	Total	0.26	0.26	0.46	0.03	1.00
Critical reading	Does Not Meet	0.19	0.03	0.00	0.00	0.22
	Partially Meets	0.03	0.21	0.05	0.00	0.28
	Meets	0.00	0.04	0.36	0.03	0.42
	Exceeds	0.00	0.00	0.01	0.07	0.08
	Total	0.22	0.27	0.41	0.09	1.00
Writing	Does Not Meet	0.19	0.04	0.00	0.00	0.23
	Partially Meets	0.04	0.21	0.06	0.00	0.31
	Meets	0.00	0.04	0.34	0.03	0.41
	Exceeds	0.00	0.00	0.01	0.05	0.06
	Total	0.23	0.29	0.40	0.07	1.00

Table 10-14 presents the consistency contingency table for each section of the MHSAs. The numbers in the cells were estimated by comparing the actual classifications and classifications based on expected observed scores to a hypothetical parallel form estimated from the model. This represents an estimate of what the parallel form reliability of achievement classifications would be.

**Table 10-14. 2008–09 MHSА: Consistency
Contingency Table for MHSА Mathematics, Critical Reading, and Writing**

Content Area	True	Observed				Total
		Does Not Meet	Partially Meets	Meets	Exceeds	
Mathematics	Does Not Meet	0.20	0.05	0.01	0.00	0.26
	Partially Meets	0.05	0.14	0.07	0.00	0.26
	Meets	0.01	0.07	0.37	0.01	0.46
	Exceeds	0.00	0.00	0.01	0.02	0.03
	Total	0.26	0.26	0.46	0.03	1.00
Critical reading	Does Not Meet	0.18	0.04	0.00	0.00	0.22
	Partially Meets	0.04	0.17	0.06	0.00	0.27
	Meets	0.00	0.06	0.33	0.03	0.41
	Exceeds	0.00	0.00	0.03	0.07	0.09
	Total	0.22	0.27	0.41	0.09	1.00
Writing	Does Not Meet	0.18	0.05	0.00	0.00	0.23
	Partially Meets	0.05	0.17	0.07	0.00	0.29
	Meets	0.00	0.07	0.31	0.03	0.40
	Exceeds	0.00	0.00	0.03	0.05	0.07
	Total	0.23	0.29	0.40	0.07	1.00

Table 10-15 presents the overall accuracy and consistency across all achievement levels for each section of the MHSА. For mathematics, 83% of students in the sample were classified correctly when comparing their observed achievement levels with their true levels. If the students took another form of the test, 77% would be consistently classified. The kappa for mathematics indicates that a minimum of 65% of students were classified correctly after factoring out chance.

**Table 10-15. 2008–09 MHSА: Summary of
Overall MHSА Classification Accuracy and Consistency**

Content Area	Accuracy	Consistency	Kappa
Mathematics	0.79	0.72	0.57
Critical reading	0.82	0.75	0.64
Writing	0.79	0.70	0.57

Table 11-16 presents the classification accuracy and consistency for various dichotomous classifications. The Does Not Meet/Partially Meets dichotomy represents two groups: those who do not meet the standard and all other students. The Partially Meets/Meets dichotomy compares all students who are below fully meeting the standard and all students above it. The Meets/Exceeds dichotomy compares the students who exceed the standard to all students below. For accuracy, the Meets/Exceeds cut ranges from 0.97 to 0.99 across four tests, the Partially Meets/Meets cut ranges from 0.89 to 0.92, and the Does Not Meet/Partially Meets cut ranges from 0.90 to 0.94. Compared to Meets/Exceeds, the Does Not Meet/Partially Meets and Partially Meets/Meets cuts have relatively lower accuracy because there are significantly larger numbers of students with achievement levels near those two cuts. This can be seen by looking around those cuts (found in Table 11-11) within the score distributions in Tables 11-7 through 11-10. Compared to the accuracy results, the consistency measures have a similar pattern, but are generally lower than the accuracy

measures. This is expected because the consistency measures include fallibility in the parallel form as well as the existing form, whereas accuracy assumes true scores. Also shown in Table 11-16 are the false positive and false negative rates. For the mathematics Partially Meets/Meets cut, 5% of students were classified incorrectly into the Meets category and 4% were classified incorrectly into the Partially Meets category.

Table 10-16. 2008–09 MHTA: Accuracy and Consistency of Dichotomous Categorizations

<i>Content Area</i>	<i>Performance Level</i>	<i>Accuracy</i>	<i>False Positive</i>	<i>False Negative</i>	<i>Consistency</i>
Mathematics	D/P	0.91	0.04	0.04	0.88
	P/M	0.89	0.06	0.05	0.85
	M/E	0.98	0.01	0.00	0.97
Critical reading	D/P	0.94	0.03	0.03	0.91
	P/M	0.92	0.05	0.04	0.88
	M/E	0.96	0.03	0.01	0.95
Writing	D/P	0.92	0.04	0.04	0.89
	P/M	0.90	0.06	0.04	0.86
	M/E	0.96	0.03	0.01	0.95

D/P = Does Not Meet/Partially Meets; P/M = Partially Meets/Meets; M/E = Meets/Exceeds

10.5 Completion Rates

Completion rate refers to the extent to which the test takers are able to complete each section of the test in the time allotted. Because there is no generally accepted index of acceptable or adequate completion rates, several criteria are reported. Each is arbitrary and by itself should not be too strictly applied. However, taken together, the criteria can be useful. When considering these criteria, the relative ability of the group, as defined by the analysis sample scaled score mean and median, needs to be taken into account.

One statistic reported is the percentage of the analysis sample reaching the items at the end of each test section. These results may be confounded with item difficulty because one or two very difficult items at the end of the test section may make it appear more speeded than it really is. This case would be shown by a sharp decrease in the number of test takers completing the last few items, rather than a gradual tapering off.

Additional completion rate data are based on the items that are not reached. Information presented in Table 11-17 includes the percentage of the group who completed each section (answered the last item in the section), the percentage of the group who completed 75% of the section (answered one or more items that were at least three-quarters of the way through the section), and the number of items that were reached by 80% of the group. The ratio of the variance of the number of items not reached to the variance of the formula scores (given as “NR variance/score variance”) is presented in the table as another index of completion rate. The total number of items in each section and the mean and standard deviation of the number of items not reached are also given in the table.

As a rule of thumb, a test is usually regarded as essentially unspeeded if at least 80% of the test takers reach the last question and if virtually everyone reaches at least three-quarters of the items. Swineford (1974)

determined that a variance index less than 0.15 may be taken to indicate an unspeeded test, while an index greater than 0.25 usually means that the test is clearly speeded. Values between 0.16 and 0.25 generally indicate a moderately speeded test. However, these are only arbitrary indices, and judgments of appropriateness of timing should be made in the context of additional data. For example, lack of motivation among the test takers may make sections appear more speeded.

Table 11-17 provides the speededness data for the state of Maine. The critical reading portion is unspeeded for Section 2 in May and moderately speeded (variance index of 0.23) during the June 2009 administration, though this should be interpreted with caution given the small sample size of Maine students testing in June. The low percentage of students completing each section in the SAT mathematics portion of the test indicates that the mathematics test is speeded, though the variance index for May 2009 on Section 1 and for both May and June 2009 on Section 3 indicates a lack of speededness. The variance index for the remaining mathematics sections in May and June 2009 suggest it is only moderately speeded. The writing portion is unspeeded for Section 2 in May 2009 and only slightly speeded, as judged by the percentage completing the section, for Section 1 in both May and June 2009 and for Section 2 in June 2009. Completion rate data for the national SAT population are provided in Appendix G, Table G-8.

**Table 10-17. 2008–09 MHSA: Maine Completion
Rate Statistics for Sections of the College Board SAT**

Form	1	2	1	2	1	2
Administration	5/09	6/09	5/09	6/09	5/09	6/09
Sample size*	12,104	205	12,104	205	12,104	205
	<i>Mathematics 1</i>		<i>Critical Reading 1</i>		<i>Writing 1</i>	
% completing section	62.1	71.7	79.7	83.9	81.4	77.1
% completing 75%	98.5	97.6	99.4	98.5	100.0	100.0
Number of items reached by 80%	18	18	22	24	35	34
Mean not reached	0.9	0.8	0.4	0.4	0.5	0.6
SD not reached	1.5	1.8	1.1	1.6	1.2	1.2
NR variance/score variance	0.10	0.16	0.04	0.07	0.02	0.02
Number of items	20	20	23	24	35	35
	<i>Mathematics 2</i>		<i>Critical Reading 2</i>		<i>Writing 2</i>	
% completing section	55.7	60.5	83.2	78.0	92.0	93.2
% completing 75%	93.0	93.2	97.8	97.1	98.8	98.5
Number of items reached by 80%	15	17	24	23	14	14
Mean not reached	1.5	1.0	0.6	0.7	0.2	0.1
SD not reached	1.9	1.7	1.8	1.8	0.7	0.6
NR variance/score variance	0.21	0.21	0.09	0.10	0.03	0.03
Number of items	18	18	24	24	14	14
	<i>Mathematics 3</i>		<i>Critical Reading 3</i>			
% completing section	44.3	89.3	82.7	81.5		
% completing 75%	96.5	99.0	97.5	99.5		
Number of items reached by 80%	13	16	20	19		
Mean not reached	1.2	0.2	0.6	0.4		
SD not reached	1.5	0.8	1.6	1.0		
NR variance/score variance	0.15	0.04	0.09	0.05		
Number of items	16	16	20	19		

SD = standard deviation; NR = number of items not reached

*The sample size is the final sample of Maine NCLB students taking the test and answering at least one question in each respective section of the test.

10.6 Item Statistics

10.6.1 Item Difficulty: Equated Delta

The simplest measure of item difficulty for a given group of test takers is the p -value—the proportion of test takers who attempted to answer the item correctly to those who attempted to answer the item. For the SAT, p -values are converted onto a standard scale called the delta index.

$$\text{Delta} = 13 + 4 \times z$$

where
 z is computed based on item difficulty, p .

First, $(1-p)$ is converted to a normalized z -score and then linearly transformed to a scale with a mean of 13 and a standard deviation of 4. Deltas are inversely related to p -values; that is, the lower the p -value, the higher the delta, and the more difficult the item.

The conversion of p -values provides raw delta values that reflect the difficulty of the items for the particular test takers from a particular administration. This measure of item difficulty then must be adjusted to correct for differences in the abilities of different test taking populations. Delta equating is a statistical procedure used to convert raw delta values to equated delta values. This procedure involves administering some old items with known equated delta values along with new items. Each old item now has two difficulty measures: the observed delta that reflects the difficulty of the item for the current group of test takers and the equated delta that is an estimate of how difficult the items would have been for the initial reference group. The linear relationship between the pairs of observed and equated deltas on the old items is used to determine the scaled values for each of the new items. Delta equating is essential because the groups taking a particular test may differ substantially in ability from one administration to another. Through delta equating, item difficulties can be compared directly.

As described in Chapter 2, new forms of the SAT are built to detailed content and statistical specifications. Each item in the new form has already been administered and has an associated difficulty estimate (equated delta). SAT statistical specifications set target means and standard deviations of the equated deltas for mathematics, critical reading, and writing. In addition, each measure has a specific requirement for the particular number of items at each delta level across the range of the delta scale. For each measure, the delta distribution is a unimodal distribution with more middle difficulty items and fewer very easy or very difficult items. The target mean delta is 11.4 (standard deviation of 2.4) for critical reading. The means and standard deviations of the deltas for critical reading in May and June 2009 range from 11.3 (2.5) to 11.5 (2.4), which is very close to the specification. For mathematics and writing, the mean deltas for the two forms administered in May and June 2009 are also very close to the specifications. Table 11-18 summarizes the mean equated delta and standard deviation for each content area by form for students testing on the MHSA in Maine only. Data in Table 11-18 cover only SAT items on the MHSA, but the sample was composed strictly of students with scores for both the SAT and Math–A components of the MHSA.

Table 10-18. 2008–09 MHSA: Maine Summary Statistics of Equated Deltas (Δ) for Mathematics, Critical Reading, and Writing Sections of the College Board SAT*

Content Area		Specified Equated Delta	<u>Form 1</u>	<u>Form 2</u>
			May 2009 N=13,230 Equated Delta	June 2009 N=227 Equated Delta
Mathematics MC	N	44	44	44
	Mean	12.2	12.1	12.2
	SD	3.2	3.2	3
Mathematics SPR	N	10	10	10
	Mean	13.6–14.2	14.3	14.8
	SD	3	3.5	3.1
Total critical reading	N	67	67	67
	Mean	11.4	11.4	12.8
	SD	2.4	2.3	2.3
Total writing	N	49	49	49
	Mean	10.1	10.1	10.3
	SD	2.5	2.5	2.5

MC = multiple-choice; SPR = student-produced-response; SD = standard deviation

*Estimates are based on students who took the MHSA SAT component and answered at least one item in each section.

10.6.2 Item Discriminating Power: Biserial Correlation

Another important characteristic of an item is item discrimination. Each item in a test should be able to distinguish higher ability test takers from lower ability test takers with respect to the construct being tested. An item is considered discriminating if proportionately more test takers who are high in the ability being measured answer the item correctly than do test takers low in the ability measured. The total score is generally used as the criterion for judging levels of ability on the construct being tested. Item difficulty can constrain item discrimination power, in that if most or very few examinees are responding correctly to an item, the discrimination is restricted.

A number of indices are used in assessing the discriminating power of an item. The index currently used on the SAT is the biserial correlation coefficient (r_{bis}), which measures the strength of the relationship (correlation) between examinees' performance on a single item and the formula score, excluding the item being analyzed. A very low or negative correlation indicates that the item does not add any precision to the measurement of the test as a whole.

During assembly of new forms, there are specifications concerning discrimination. The specified mean r_{bis} for both critical reading and writing is 0.49 to 0.53. For mathematics, the specified mean of r_{bis} is 0.53 to 0.57 on the multiple-choice items and 0.60 to 0.70 on the student-produced-response items. Table 10-19 presents the biserial coefficients for the May and June 2009 forms of the SAT for students taking the MHSA in Maine only. Data in Table 10-19 cover only SAT items on the MHSA, but the sample was strictly composed of students with scores for both the SAT and Math–A components of the MHSA.

**Table 10-19. 2008–09 MHSAs: Maine Summary
Statistics for Biserial Coefficients* for Mathematics,
Critical Reading, and Writing Sections of the College Board SAT**

Content Area			Form 1	Form 2
			May 2009 N = 13,230	June 2009 N = 227
Mathematics MC	N	0.53–0.57	44	44
	Mean		0.5	0.45
	SD		0.13	0.19
Mathematics SPR	N	0.60–0.70	10	10
	Mean		0.67	0.55
	SD		0.09	0.15
Total critical reading	N	0.49–0.53	67	66
	Mean		0.51	0.46
	SD		0.12	0.16
Total writing	N	0.49–0.53	49	48
	Mean		0.51	0.47
	SD	0.53–0.57	0.11	0.15

MC = multiple-choice; SPR = student-produced-response; SD = standard deviation

An *r*-biserial is not calculated when the percentage correct is greater than 95 or less than 5, or when dropout exceeds 50%.

*Estimates are based on students who took the MHSAs and answered at least one item in each section.

10.7 Differential Item Functioning (DIF)

Measures of differential item functioning (DIF) are used to help ensure test and item fairness. DIF indicates “a difference in item performance between two comparable groups of examinees; that is, the groups that are matched with respect to the construct being measured by the test” (Dorans & Holland, 1993, p. 35). Theoretically, if test takers from two different groups have the same ability level, they should have the same probability of getting an item correct. The two groups are referred to as the focal group and the reference group, where the focal group is the focus of analysis and the reference group is the basis for comparison.

Currently, the SAT uses the Mantel-Haenszel (MH) approach (Holland & Thayer, 1988) for DIF detection (D-DIF). On the basis of the MH D-DIF statistic, which can be interpreted as a difference in deltas, items are classified into the following categories based on specific criteria:

- Category A—Negligible DIF. Items are classified in this category for a particular combination of reference and focal groups if either MH D-DIF is not statistically different from 0 or if the magnitude of the MH D-DIF values is less than 1.0 delta unit in absolute value.
- Category B—Intermediate DIF. This category is composed of items that are not classified as A or C
- Category C—Large DIF. Items are classified as C if MH D-DIF both exceeds 1.5 in absolute value and is statistically significantly larger than 1.0 in absolute value.

A minus sign (e.g., B- or C-) indicates that the item tended to favor the reference group (male or White), while a plus sign (e.g., B+ or C+) indicates the item tended to favor the focal group (female or non-White).

The current practice for the SAT is to run DIF for selected ethnicities, with Whites as the reference group. Separate DIF analyses are performed with African Americans, Hispanics, Asian Americans, and Native Americans as the focal groups. In Maine, the population is not as diverse as that found nationally, and so subgroup sample sizes permitted only analyses for the African American versus White ethnicity comparison. DIF analyses are also performed with males as the reference group and females as the focal group. The DIF analyses completed using all students who took the May and June 2009 SAT test forms for the national population are listed in Tables G-11 and G-12 of Appendix G. Table 10-20 represents DIF analyses for Form 1 of the SAT using only students from Maine. DIF analyses for the June administration, Form 2, were not conducted due to insufficient sample size.

For the analysis using only Maine students, fewer students were available for the analysis. The low number of students had two immediate impacts upon the analysis. First, comparisons across all groups were not possible. A standard minimum applied when completing DIF analysis is that 200 or more students must exist in each group being analyzed. Using a sample of students fewer than 200 would yield unreliable results. While the sample for the Black students exceeds the criteria of 200 students, some caution should be used in the interpretation of these results as well. The second impact of the small sample sizes is that more items were classified with C-DIF.

For the analysis using only Maine students, two items from the critical reading test were classified as C-DIF. Both items were classified as C-DIF in the male versus female student comparison. For Math–A and writing, no items were classified as C-DIF.

Table 10-20. 2008–09 MHSAs: Maine Differential Item Functioning (DIF) Summary Form: 1 Administration: 5/08

<i>Category of Maximum Absolute DIF Value for All Comparisons</i>				<i>Female</i>	<i>Black</i>
				N = 6,516	N = 255
<i>Content Area</i>				<i>Male</i>	<i>White</i>
				N = 6,714	N = 12,568
<i>Category</i>	<i>Number</i>	<i>% of Items</i>	<i>Number of Items by DIF Category</i>		
Total mathematics*	+C	0	0.0	0	0
	+B	2	3.0	0	2
	A	59	89.4	63	61
	-B	4	6.1	2	3
	-C	1	1.5	1	0
	Total	66	100.0	66	66
Total critical reading	+C	0	0.0	0	0
	+B	1	1.5	1	0
	A	61	91.0	62	66
	-B	5	7.5	4	1
	-C	0	0.0	0	0
	Total	67	100.0	67	67
Total writing	+C	0	0.0	0	0
	+B	2	4.1	0	2
	A	45	91.8	49	45
	-B	2	4.1	0	2
	-C	0	0.0	0	0
	Total	49	100.0	49	49

*Mathematics = 54 SAT items and 11 Math–A items

10.8 Summary

The scores reported for SAT test takers must be accurate and must be comparable regardless of which form is administered or at which administration the student takes the examination. The intention of this chapter was to describe the intense scrutiny that each item, form, and reported score must undergo. The care and the thought required in establishing a new scale, such as the new writing section, and in maintaining the meaning of established scales, such as the mathematics and critical reading sections, were also described. The information in this chapter should help the reader to understand the psychometric rigor required to ensure that the interpretations of the score results are valid and fair. In addition, the statistical results that were reported concerning items and forms provide assurance that the test scores are reliable. For information on interpreting SAT scores, see Appendix H or visit www.collegeboard.com/prod_downloads/sat/sat-program-handbook.pdf.

Chapter 11. PSYCHOMETRIC TOPICS OF MATH–A AND SCIENCE

The purpose of this chapter is to provide a detailed technical description of the psychometric procedures used for the MHSA mathematics, with particular focus on the Math–A portion, and MHSA science tests. As described earlier in this report, the full MHSA mathematics test included both multiple-choice items (44 provided by the SAT and 11 provided by Math–A) and student-produced-response items (10 provided by the SAT). The science test included 40 multiple-choice items and 4 constructed-response items. The chapter is intended for those with a working knowledge of item response theory and psychometric methods.

11.1 Formula Scoring

Of critical importance in any testing program is the method for scoring the assessment instrument. The following describes the rationale and procedure of formula scoring.

For mathematics multiple-choice items, students were asked to choose the one correct answer from a list of five possible choices. In this case, a student who has no knowledge of the correct answer and randomly guesses will have a probability of $1/5$ for choosing the correct answer and a probability of $4/5$ for choosing an incorrect answer. If students were awarded 1 point for a correct response and 0 points for an incorrect response, the expected item score for the student who randomly guesses would be 0.2. On the student-produced-response items, the student who randomly guesses has essentially no chance of a correct answer. To correct for this difference between multiple-choice and student-produced-response items, the MHSA applies a scoring procedure known as formula scoring to the multiple-choice items. In formula scoring for multiple-choice items, a student receives a score of 1 for a correct response; $-1/(k-1)$ for an incorrect response, where k is the number of choices; or 0 for not recording an answer, also known as an omitted response, or “omit.” The use of formula scoring results in an expected score of 0 across multiple-choice items for a student who randomly guesses on the items—the same expected score as for student-produced-response items. The total raw score is a student’s formula score on the multiple-choice items (both SAT and Math–A items) plus the number of correct responses on the student-produced-response items.

Formula scoring was used on the MHSA science test to be consistent with the scoring of MHSA mathematics.⁸ In this case, however, students chose one correct answer from *four* possible choices. Given that k for science is 4, the previous discussion for item scores applies for science as well, and the sum of science item formula scores is the total raw science score.

⁸ For a discussion on the scoring of constructed-response items on the MHSA science test, see Chapter 9.

11.2 Standard Setting

Standard setting to establish cut scores for the MHSA in science was conducted on Thursday and Friday, May 21 and 22, 2009 at the Department of Education offices in Augusta, Maine.

The standard setting method implemented was a modified version of the bookmark method. A validation approach was used, in which panelists were presented with starting cuts and were asked to either validate those cuts or recommend adjustments. An overview of the method is described below.

11.2.1 Panel Membership

Panelists were selected prior to standard setting by the Maine Department of Education. The goal was to recruit approximately 15 participants, representing a range of geographic areas, demographic groups, etc. The majority of the panelists were science teachers, but some higher education and special education teachers, English Language Learner (ELL) teachers, as well as school administrators, also participated. A total of 12 panelists participated in the standard setting.

11.2.2 Calculation of Starting Cut Points

Starting cut points were calculated using an equipercentile approach. Specifically, the percentage of students who were classified into each achievement level category according to the 2007-2008 test results were calculated, and starting cuts were chosen based on the 2008-2009 administration such that the previous year's percentages were matched as closely as possible.

Panelists were informed that the purpose of the starting cuts was 1) to make use of the information gathered at the previous standard setting meeting; and 2) to streamline the standard setting process by giving panelists information about the probable approximate location of the cuts. The facilitators made sure the panelists understood that, because new content standards had been implemented since the previous cuts had been determined, they were free to recommend changes to those cuts as appropriate.

11.2.3 Orientation

The standard setting session began with a general orientation for the panelists. The purpose of the orientation was to provide background information, an introduction to the issues of standard setting, and a brief overview of the bookmark procedure and the activities that would occur during standard setting.

11.2.4 Review of Assessment Materials

The first step after the opening session was for the panelists to take the test. The purpose of this step was to make sure the panelists were thoroughly familiar with what the assessment asks of students. Once panelists completed the test an answer key was distributed. At this point, panelists were encouraged to discuss any issues that came to mind regarding items or scoring.

11.2.5 Completion of Item List Form

The purpose of the next step was to ensure that panelists became very familiar with the ordered item booklet and understood the relationships among the ordered items. The ordered item booklet contained one item per page, ordered from the easiest to the most difficult. The ordered item booklet was created by sorting items by their IRT-based difficulty values.

The item list form listed the items in the same order they were presented in the ordered item booklet and had spaces for the panelists to write in the knowledge, skills, and abilities required to answer each item correctly. There was also a space for the panelists to write in why they felt the current ordered item was more difficult than the previous one.

11.2.6 Review of Achievement Level Definitions and Definition of Borderline Students

Next, panelists reviewed the Achievement Level Definitions (ALDs). This important step of the process was designed to ensure that panelists thoroughly understood the needed knowledge, skills, and abilities to be classified as Partially Meets the Standard, Meets the Standard, and Exceeds the Standard. Panelists began by individually reviewing the ALDs; they then discussed the descriptors as a group, clarifying each level. Afterwards, panelists developed consensus definitions of borderline students, i.e. students who are “just able enough” to be categorized into an achievement level. Bulleted lists of characteristics for each level were generated based on the whole group discussion and posted in the room for reference throughout the bookmark process.

11.2.7 Round 1 Judgments

In the first round, panelists worked as a group with the ALDs, the item list form they completed earlier, and the ordered item booklet. Beginning approximately five ordered items before the Does Not Meet/Partially Meets starting cut, and considering the skills and abilities needed to complete that first item, they asked the question, “Would at least 2 out of 3 students performing at the borderline of Partially Meets the Standard answer this question correctly?” Panelists considered each ordered item in turn, asking the same question until their answer changed from “yes” (or predominantly “yes”) to “no” (or predominantly “no”). A bookmark was placed there.

Although the panelists worked as a group, the facilitator made sure it was understood that they should set the bookmark according to their individual best judgments, and that they need not come to consensus. They were encouraged to listen to the points made by their colleagues but not feel compelled to change their bookmark placements.

Panelists then repeated the process for the other two cuts and used the provided rating form to record their ratings for each cut.

11.2.8 Tabulation of Round 1 Results and Impact Data

After the Round 1 ratings were complete, Measured Progress staff calculated the average cut-points for the room as well as impact data. The impact data consisted of the percentage of students who would be classified into each achievement level category according to the Round 1 average cut points. This information was shared with the group to assist them in Round 2.

11.2.9 Round 2 Judgments

The purpose of Round 2 was for panelists to discuss the Round 1 placements and impact data and revise their ratings, if necessary. Panelists shared their individual rationales for their bookmark placements in terms of the necessary knowledge and skills for each classification. Panelists were asked to pay particular attention to how their individual ratings compared to those of the others and get a sense for whether they were unusually stringent or lenient within the group. Room average cut-points and impact data were to be considered as well.

The facilitator once again emphasized that the panelists should set the bookmark according to their *individual* best judgments, and that they need *not* come to consensus. They were encouraged to listen to the points made by their colleagues but not feel compelled to change their bookmark placements. Once discussions were completed, panelists recorded their Round 2 ratings on the rating form. Results of the Round 2 ratings are presented in Table 11-1.

Table 11-1. 2008–09 MHSA: Round 2 Results

Grade	Achievement Level	Theta Cut	Standard Error	Raw Score		Percent of Students
				Min	Max	
HS	Does Not Meet			-13.33	16.67	33.3
	Partially Meets	0.3318	0.0552	17	25.33	26.0
	Meets	0.3616	0.0624	25.67	46	38.3
	Exceeds	2.7352	0.1996	46.33	55	2.4

11.2.10 Evaluation

As the last step in the standard setting process, panelists anonymously completed an evaluation asking for their perceptions of the standard setting process.

11.2.11 Standard Setting Report

Upon completion of the standard setting sessions, Measured Progress submitted a report to the DOE that documented all aspects of the standard setting process. Documentation included all procedures completed prior to, during, and after the standard setting meeting, the recommended cut points, impact data that resulted

from the standard setting, and the results of the panelist evaluation of the standard setting process. The standard setting report can be found in Appendix I.

Maine DOE personnel, including the Commissioner of Education, met to review the standard setting report from Measured Progress. Based on their review, one minor adjustment was made to the Round 2 cut scores: The Meets/Exceeds cut was decreased slightly such that the raw score required for a student to be classified into the Exceeds the Standards category was 44 instead of 46.33.

11.3 Deriving MHS A Scaled Scores

For the 2008–09 administration of the MHS A, five combined Math–A and science test forms were constructed. Student scores are based on a common set of items across these forms. The items that differentiate one form from another are referred to as matrix items and do not contribute to student scores. Consequently, the same set of Math–A and science items counts toward a student’s score in any given year, irrespective of the test form taken. However, the common items differ across years. Because of this, the tests can vary slightly in difficulty year to year. It is important, therefore, that student raw scores be converted to a common scale that takes these crossyear differences into account. To accomplish this goal, the raw scores were converted to a common scale after the first operational administration. This conversion from raw scores to scaled scores is called scaling. Adjusting the raw scores to account for differences in difficulty across test forms is called equating. Psychometric procedures based on IRT were used to accomplish the scaling and equating.

11.3.1 Item Response Theory Calibration

IRT models student item response behavior by means of an item characteristic curve (ICC). An ICC is an equation for the probability that a student gives a particular scored response on an item, conditional on the student’s level of achievement in the construct being measured by the test (in this case, mathematics or science achievement). There were three possible scored responses (correct, incorrect, or omit) for multiple-choice items, two for student-produced-response items (correct or incorrect, the latter including omits), and five for constructed-response items. For all three item types (multiple-choice, student-produced-response, and constructed-response), item response behavior can be modeled by using Samejima (1997)’s graded response model (GRM) for the ICC.

In the GRM, an item is scored in $k+1$ categories, which can be viewed as sets of k dichotomies. In the case of MHS A mathematics and science, $k = 2$ for the multiple-choice items, $k = 1$ for the student-produced-response items, and $k = 4$ for the constructed-response items. At each point of dichotomization (i.e., at each threshold), a two parameter model can be used, with one parameter modeling difficulty and the other modeling discrimination power. This implies that an item with $k+1$ score categories can be characterized by k item category threshold curves (ICTCs) of the following two parameter logistic form:

$$P_{ik}^* (1 | \theta_j, a_i, b_i, d_{ik}) = \frac{\exp Da_i (\theta_j - b_i + d_{ik})}{1 + \exp Da_i (\theta_j - b_i + d_{ik})}$$

where
i indexes the items,
j indexes students,
k indexes threshold,
a represents item discrimination,
b represents item difficulty,
d represents threshold step difficulty parameters,
D is a normalizing constant equal to 1.701, and
 θ_j is a continuous real-valued variable representing student achievement level.

After computing *k* ICTCs in the GRM, *k*+1 ICCs are derived by subtracting adjacent ICTCs:

$$P_{ik} (1 | \theta_j) = P_{i(k-1)}^* (1 | \theta_j) - P_{ik}^* (1 | \theta_j)$$

where
 P_{ik} represents the probability that the score on item *i* falls in category *k*, and
 P_{ik}^* represents the probability that the score on item *i* falls above the threshold *k*
($P_{i0}^* = 1$ and $P_{i(m+1)}^* = 0$).

The GRM is also commonly expressed as

$$P_{ik} (k | \theta_j, \xi_i) = \frac{\exp [Da_i (\theta_j - b_i + d_k)]}{1 + \exp [Da_i (\theta_j - b_i + d_k)]} - \frac{\exp [Da_i (\theta_j - b_i + d_{k+1})]}{1 + \exp [Da_i (\theta_j - b_i + d_{k+1})]}$$

where
 ξ_i represents the set of item parameters for item *i*.

Finally, the ICC is computed as a weighted sum of ICCs, where each ICC is weighted by a score assigned to a corresponding category:

$$P_i (1 | \theta_j) = \sum_k^{m+1} w_{ik} P_{ik} (1 | \theta_j)$$

This model was applied to the MHSA mathematics and science tests under the assumption that all the items measured a common unitary construct, namely mathematics or science achievement. This assumption of measurement of a single construct is commonly referred to as unidimensionality. Dimensionality analyses were run on the full set of MHSA mathematics items (SAT mathematics plus Math–A) and the science items, the results of which are presented later in this chapter.

Application of a unidimensional model results in the items from the SAT and Math–A being placed onto a common scale through what is typically referred to as concurrent calibration. This widely used calibration method ultimately allows for within-year form equating. The PARSCALE program was used for

calibration. In order to accommodate the several types of scoring employed on the MHSA (dichotomously scored student-produced-response items for mathematics, formula scored multiple-choice items for mathematics and science, and hand scored constructed-response items for science), the GRM was used with the default settings in PARSCALE. Upon completion of IRT calibration, the resulting estimated item parameters were used to construct a test characteristic curve (TCC) by summing the ICCs. To account for the specific scoring scheme used in the SAT, the item parameter outputs by PARSCALE were multiplied by -1/4 (for mathematics) or -1/3 (for science), 0, and 1 for the 1st, 2nd, and 3rd step difficulty parameters, respectively. For more information about item calibration and determination, refer to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

The calibration method just described was used not only to place all items onto a common scale, but to rank order items for evaluation during the science standards validation using the bookmark method. The bookmark standards validation resulted in cut scores on the theta (θ) science achievement metric, the metric of the item parameters. The reported MHSA score scale, on the other hand, ranges from 1100 to 1180, where the cut score between Partially Meets and Meets is associated with 1142 (i.e., a student must obtain a score of 1142 in order to be in the Meets category), and the cut score between Meets and Exceeds is associated with 1162. The following section describes the relationship between the θ achievement variable used in the ICCs (and TCC) and the MHSA score scale.

11.3.2 Scaling MHSA Mathematics and Science, and Equating the Two Mathematics Components

In the case of the MHSA mathematics test, students are administered a base form—the form of the SAT taken by the majority of third year high school students (i.e., all students except those who take a makeup form)—and the Math–A items that are exclusively administered to Maine students. Scaling the SAT and Math–A onto the single, reported MHSA scale requires three main steps.

1. Construct the TCC for the SAT and Math–A (described above)
2. Map through the TCC to determine the θ scores for each obtainable expected total raw score (i.e., SAT raw plus Math–A raw)
3. Use the slope and intercept terms (derived from the θ cuts and the 1142 and 1162 scaled score cuts) to place each obtainable expected total score onto the reporting metric via simple linear transformation

From here, the procedures for scaling the MHSA science and mathematics tests are identical. The equation for converting θ values to scaled scores for the tests can be written as follows:

$$\text{Scaled Score} = \text{Slope} * \theta_x + \text{Intercept}$$

Here, θ_X is the θ value for which the TCC = X at some particular total raw score. In the defining of the slope and intercept, let $\theta_{PM/M}$ stand for the θ value at the cut between Partially Meets and Meets and $\theta_{M/E}$ stand for the θ value at the cut between Meets and Exceeds. (These θ values were determined during the respective test's standard setting.) The equations for the slope and intercept are then given as

$$\text{Slope} = (1162 - 1142) / (\theta_{M/E} - \theta_{PM/M})$$

and

$$\text{Intercept} = (1142 * \theta_{M/E} - 1162 * \theta_{PM/M}) / (\theta_{M/E} - \theta_{PM/M})$$

This linear transformation results in a one to one relationship between achievement as measured on the θ metric and achievement as measured on the MHSA scaled score metric.

11.3.3 Scaling Additional Forms

Each year, it is necessary to administer additional—makeup—SAT forms to a small population of students. Scaling for these makeup forms is accomplished using the SAT scale itself as the equating link. That is, the SAT raw score to scaled score tables provided by the College Board are used to determine a raw score equivalent on the base form for each raw score on any given additional form, which in turn is made possible because scores on the base form and all additional forms have been placed on the same SAT scale by the College Board. In the case of MHSA mathematics, once the base form equivalent score is found, it is combined with the Math–A score, following which the base form/Math–A TCC is used to estimate a θ score. Finally, the linear transformation previously described is applied to obtain an MHSA scaled score.

Because the College Board provides a common scale for all SAT forms, the SAT scale is used for converting MHSA raw scores to the MHSA scale by following the series of steps below.

- Perform a concurrent calibration on the 2009 SAT base form and Math–A raw scores to obtain a 2009-specific scale. In other words, a given 2009 raw composite score, call it X, corresponds to a 2009 θ_X . (In this step, the Math–A items are combined with the SAT base form items to maximize the reliability of the θ_X .)
- Derive an expected 2009 SAT base form raw score using the 2009 TCC, the 2009 θ_X value.
- Translate the expected 2009 SAT base form raw score to an expected 2009 SAT base form scaled score using 2009 SAT conversion tables.
- Translate the expected 2009 SAT base form scaled score (one and the same as a 2008 SAT base form scaled score) to a 2008 SAT base form raw score using the 2008 SAT conversion tables.
- Translate the 2008 SAT base form raw score to a 2008 θ_X using the 2008 TCC.
- Finally, transform the 2008 θ_X to a 2008 comparable scaled score on the MHSA scale, based on a linear transformation using the 2008 slope and intercept terms described earlier.

For makeup mathematics tests, a student’s raw makeup score is simply converted to the equivalent SAT base form score for 2009, which is added to the student’s Math–A raw score, yielding the equivalent total score the student would have received if he or she had taken the base form of 2009. From there, the steps above should be followed to obtain the MHSA scaled score.

Table 11-2 displays descriptive statistics for MHSA mathematics and science, and Tables 11-3 and 11-4 give the scaled score frequency distributions (mathematics is repeated from Chapter 10). Again, note that MHSA scaled scores take on only even values within the 1100–1180 range.

Table 11-2. 2008–09 MHSA: Mathematics and Science Scaled Score Summary Statistics

<i>Content Area</i>	<i>N</i>	<i>SD</i>	<i>Std Dev</i>	<i>Minimum</i>	<i>Maximum</i>
Mathematics*	15032	1140.7	11.0	1100	1180
Science	14867	1140.3	11.2	1100	1180

* Includes data for the MHSA mathematics test (SAT and Math–A)
SD = standard deviation

Table 11-3. 2008–09 MHSA: Frequency Distribution of MHSA Scores—Mathematics (SAT/Math–A)

<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>	<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>
1100	0	0.00	1142	655	4.36
1102	0	0.00	1144	1067	7.10
1104	0	0.00	1146	798	5.31
1106	0	0.00	1148	820	5.46
1108	0	0.00	1150	642	4.27
1110	150	1.00	1152	419	2.79
1112	0	0.00	1154	451	3.00
1114	1	0.01	1156	342	2.28
1116	0	0.00	1158	262	1.74
1118	4	0.03	1160	228	1.52
1120	0	0.00	1162	96	0.64
1122	0	0.00	1164	93	0.62
1124	212	1.41	1166	91	0.61
1126	553	3.68	1168	86	0.57
1128	701	4.66	1170	44	0.29
1130	772	5.14	1172	38	0.25
1132	1730	11.51	1174	27	0.18
1134	276	1.84	1176	15	0.10
1136	1128	7.50	1178	12	0.08
1138	1315	8.75	1180	95	0.63
1140	1909	12.70			

Table 11-4. 2008–09 MHA: Frequency Distribution of MHA Scores—Science

<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>	<i>Scaled Score</i>	<i>Frequency</i>	<i>Percent</i>
1100	3	0.02	1142	560	3.77
1102	0	0.00	1144	817	5.50
1104	1	0.01	1146	868	5.84
1106	1	0.01	1148	678	4.56
1108	1	0.01	1150	605	4.07
1110	5	0.03	1152	573	3.85
1112	2	0.01	1154	395	2.66
1114	19	0.13	1156	362	2.43
1116	24	0.16	1158	260	1.75
1118	106	0.71	1160	313	2.11
1120	115	0.77	1162	90	0.61
1122	244	1.64	1164	101	0.68
1124	409	2.75	1166	119	0.80
1126	494	3.32	1168	51	0.34
1128	603	4.06	1170	81	0.54
1130	919	6.18	1172	34	0.23
1132	2012	13.53	1174	47	0.32
1134	0	0.00	1176	10	0.07
1136	1125	7.57	1178	29	0.20
1138	1074	7.22	1180	40	0.27
1140	1677	11.28			

Table 11-5 presents the MHA mathematics and science scaled cut scores (mathematics is repeated from Chapter 10). The values in the table do not change from year to year because the cut scores along the θ scale do not change. In any given year, it may not be possible to attain a particular scaled score, but the scaled score cuts will remain the same.

Table 11-5. 2008–09 MHA: Mathematics and Science Range of Scaled Scores for Each Achievement Level

<i>Content Area</i>	<i>Does Not Meet</i>	<i>Partially Meets</i>	<i>Meets</i>	<i>Exceeds</i>
Mathematics	1100–1132	1134–1140	1142–1160	1162–1180
Science	1100–1132	1134–1140	1142–1160	1162–1180

Based on the cut score ranges, Table 11-6 shows the 2009 achievement level frequency distributions for mathematics and science.

Table 11-6. 2008–09 MHA: Mathematics and Science Number and Percentage of Students in Each Achievement Level

<i>Content Area</i>	<i>Does Not Meet</i>		<i>Partially Meets</i>		<i>Meets</i>		<i>Exceeds</i>	
	<i>N</i>	<i>Percent</i>	<i>N</i>	<i>Percent</i>	<i>N</i>	<i>Percent</i>	<i>N</i>	<i>Percent</i>
Mathematics	4123	27.4	4628	30.8	5684	37.8	597	4.0
Science	4958	33.4	3876	26.1	5431	36.5	602	4.1

*Percentages in this table may not always sum to 100 due to rounding.

11.4 Item Analyses

As noted in Brown (1983), “A test is only as good as the items it contains.” A complete evaluation of a test’s quality must include an evaluation of each question. Both *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) and *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) include standards for identifying quality questions. Questions should assess only knowledge or skills that are identified as part of the domain being measured and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Previous sections on development described various qualitative procedures that were conducted to ensure that MHSA questions met these standards. The following discussion focuses on several categories of quantitative evaluation: difficulty indices, item-test correlations, subgroup differences in item performance (differential item functioning [DIF]), dimensionality analyses, and IRT analyses.

The expected item difficulty, also known as the p -value, is the main index of item difficulty under the classical test theory (CTT) framework. This index measures an item’s difficulty by averaging the proportion of points received across all students who took the item. The difficulty index for both formula scored multiple-choice items and polytomously scored constructed-response items is on a 0–1 scale and computed as the average score on the item divided by its maximum possible score. The p -value is traditionally called a measure of difficulty but is properly interpreted as an easiness index, because larger values indicate easier items. An index of 0.0 indicates that no student received credit for the item. At the opposite extreme, an index of 1.0 indicates that every student received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student ability, but they do indicate knowledge or skills that have been mastered by most students. The converse is true of items that are incorrectly answered by most students. In general, to provide the most precise measurement, difficulty indices should range from near chance performance (0.25 for four-option multiple-choice items, 0.00 for constructed-response items) to 0.90. Experience has indicated that items conforming to this guideline tend to provide satisfactory statistical information for the bulk of the student population. However, on a criterion referenced test such as the MHSA, it may be appropriate to include some items with difficulty values outside this region in order to measure well, throughout the range, the skill present at a given grade. Having a range of item difficulties also helps to ensure that the test does not exhibit an excess of scores at the floor or ceiling of the distribution.

Another desirable feature of an item is that higher ability students should perform better than lower ability students. A commonly used measure of this characteristic is the correlation between total test score and student performance on the item. Within CTT, this item-test correlation is referred to as the item’s

discrimination, because it indicates the extent to which successful performance on an item discriminates between high and low scores on the test. For formula scored and polytomous items on the MHSA, the Pearson product-moment correlation was used as the item discrimination index. The theoretical range of discrimination is -1.0 to 1.0, with a typical range from 0.2 to 0.6.

One can think of a discrimination index as a measure of how closely an item assesses the same knowledge and skills as other items that contribute to the criterion total score; in other words, the discrimination index can be interpreted as a measure of construct consistency. In light of this, it is quite important that an appropriate total score criterion be selected. For the MHSA, raw score was selected.

Table 11-7 presents CTT statistics for MHSA mathematics. As shown, the items on the SAT were similar to those of the Math–A in terms of overall difficulty. The SAT items had a mean difficulty of 0.38 and the Math–A items a mean of 0.31. The standard deviations of the difficulty indices were also similar (0.24 for the SAT items, 0.17 for the Math–A items). An important finding in the table is that the Math–A items had a slightly lower mean discrimination (0.37) than the SAT items (0.43). Although both values are within an acceptable range, it does suggest that the Math–A items were slightly less related to overall test score than the SAT items. This could indicate that the augment items tapped into the measured construct in a different yet meaningful way.

Table 11-7. 2008–09 MHSA: Summary and CTT Statistics for the SAT Mathematics, Math–A, and MHSA Mathematics Test Items

<i>Part of Test</i>	<i>SAT Items</i>	<i>Math-A Items</i>	<i>All Items</i>
Number of items	54	11	65
Multiple-choice items	44	11	55
Student-produced-response items	10	0	10
Possible formula-score range	-11~54	2.75 ~ 11	-13.75 ~ 65
Mean	20.5	3.5	23.9
SD	12.14	2.93	14.38
Median	19.5	3.5	23.25
Skewness	0.34	0.35	0.39
Correlation coefficient	0.71		
<i>Difficulty Index*</i>			
<i>Part of Test</i>	<i>SAT Items</i>	<i>Math-A Items</i>	<i>All Items</i>
Mean	0.38	0.31	0.37
SD	0.24	0.17	0.23
Minimum*	0.01	0.02	0.01
Maximum	0.91	0.62	0.91
<i>Discriminating Power</i>			
<i>Part of Test</i>	<i>SAT Items</i>	<i>Math-A Items</i>	<i>All Items</i>
Mean	0.43	0.37	0.42
SD	0.11	0.07	0.11
Minimum	0.11	0.29	0.11
Maximum	0.61	0.49	0.61

SD = standard deviation

* Note: Due to formula scoring, item scores (and therefore *p*-values) can be less than 0.

CTT statistics for MHSA sScience are presented in Table 11-8:

Table 11-8. 2008–09 MHSA: Summary and CTT Statistics for the MHSA Science Test Items

Number of Items	44
Multiple-choice items	40
Constructed-response items	4
Possible formula score range	-13.3 to 56
Mean	22.8
SD	11.8
<i>Difficulty Index*</i>	
Mean	0.54
SD	0.18
Minimum*	0.21
Maximum	0.88
<i>Discriminating Power</i>	
Mean	0.36
SD	0.09
Minimum	0.18
Maximum	0.62

SD = standard deviation

* Note: Due to formula scoring, item scores (and therefore p -values) can be less than 0.

11.5 Subgroup Differences in Item Performance

Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 1988) explicitly states that subgroup differences in performance should be examined when sample sizes permit, and action should be taken to ensure that differences in performance are due to construct relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines. As part of the effort to identify such problems, MHSA items were evaluated in terms of DIF statistics.

As explained in Chapter 11, DIF procedures are designed to identify items for which subgroups of interest perform differently beyond the impact of differences in overall achievement. DIF analyses on the Math–A were reported in that chapter.

For MHSA science, it was possible to use the standardization DIF procedure (Dorans & Kulick, 1986) to evaluate differences between male and female and White and Black students, because the procedure requires a minimum of 200 students in each subgroup. The procedure calculates the difference in item performance for groups of students matched for achievement on the total test. That is, the average item performance is calculated for students at every total score, and then an overall average is calculated weighting the total score distribution so it is the same for the two groups. The index ranges from -1.0 to 1.0 for multiple-choice and short-answer items and is adjusted to the same scale for constructed-response items. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance of MHSA items fell within this range. Dorans and Holland further stated that items with values between -0.10 and -0.05 and between 0.05 and 0.10 (i.e., low DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values outside the -0.10 to 0.10 range (i.e., high DIF) are

more unusual and should be examined very carefully. If a field test item displays high DIF, it is flagged for review by a Measured Progress content specialist. The content specialist consults with the MDOE to determine whether to include the flagged item in a future test administration.

Differential performances revealed by DIF analyses may or may not be indicative of bias in the test. Course taking patterns, group differences in interests, or differences in school curricula can lead to DIF. If subgroup differences in performance are related to construct relevant factors, the items should be considered for inclusion on a test.

For the 2008–09 MHSA science test, each item was categorized according to the guidelines adapted from Dorans and Holland (1993). Table 11-9 more specifically provides the number of items in each of the three DIF categories for each item type. Some MHSA items were categorized as low or high DIF. These indices must not be interpreted as indisputable evidence of bias. Both *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) and *Standards for Educational and Psychological Testing* (AERA et al., 1999) assert that test items must be free from construct irrelevant sources of differential difficulty. If subgroup differences in performance can be plausibly attributed to construct relevant factors, the items may be included on a test. Thus, it is important to determine whether the cause of this differential performance is construct relevant.

Table 11-9. 2008–09 MHSA Science: Number of Items Classified into DIF Categories—Within Pairwise Comparisons

	All A	All B	All C	MC A	MC B	MC C	Y	CR A	CR B	CR C	Y
Male versus female	31	8	5	29	6	5	0	2	2	0	0
White versus Black	36	8	0	33	7	0	0	3	1	0	0

MC = multiple-choice items; CR = constructed-response items; All = MC and CR items
A = negligible DIF; B = low DIF; C = high DIF

11.6 Dimensionality Analyses

The MHSA mathematics and science tests were each designed to measure and report a single score on mathematics and science achievement, respectively, with each test using a unidimensional scale from 1100 to 1180. Thus, these tests are said to each measure a single dimension, and the term *unidimensional* is used to describe each test.

Because these tests were constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for a large number of secondary dimensions being invoked beyond the primary dimension (mathematics or science) that all the items have in common. Of particular interest for the MHSA mathematics test is the use of the SAT mathematics section in conjunction with the Math–A. Generally, the scores on such subtests are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the

unidimensional IRT models that are used for calibrating, linking, scaling, and equating the 2008–09 MHSA mathematics and science test forms.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (1) the degree to which unidimensionality is violated and (2) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2008–09 MHSA mathematics and science tests are reported below.

Dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Nonparametric techniques were preferred for this analysis because such techniques avoid strong parametric modeling assumptions while still adhering to the fundamental principles of IRT. Parametric techniques, including nonlinear factor analysis, make strong assumptions that are often inappropriate for real data, such as assuming a normal distribution for ability and lower asymptotes of zero for the ICCs.

Both DIMTEST and DETECT use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on expected total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Nonzero conditional covariances are essentially violations of the principle of local independence, and such local *dependence* implies multidimensionality. Thus, nonrandom patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis testing procedure for detecting violations of local independence. For exploratory analyses, the data are first randomly divided into a training sample and a crossvalidation sample. An analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The crossvalidation sample is then used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the nonclustered items. For confirmatory analyses, the practitioner selects a group of items suspected to represent a secondary dimension, and the whole sample is used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the nonclustered items. The DIMTEST statistic follows a standard distribution under the null hypothesis of unidimensionality.

DETECT is an effect size measure of multidimensionality. For exploratory analyses, as with DIMTEST, the data are first randomly divided into a training sample and a crossvalidation sample (if a DIMTEST exploratory analysis has been conducted, one could use the same training and crossvalidation samples, but using new samples is also permissible). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional

covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the crossvalidation sample data to average the conditional covariances. The within cluster conditional covariances are summed, and from this sum the between cluster conditional covariances are subtracted. The resulting difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. For confirmatory analyses, the practitioner selects the clusters, and then the DETECT statistic is calculated in the same way as for exploratory analyses, but using all the data, not just the crossvalidation sample. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality); values of 0.2 to 0.4, weak to moderate multidimensionality; values of 0.4 to 1.0, moderate to strong multidimensionality; and values greater than 1.0, very strong multidimensionality.

DIMTEST was applied to the 2008–09 MHSa mathematics test in a confirmatory analysis using data for a random sample of 12,000 students (the upper limit of the DIMTEST software), with the Math–A section selected as a cluster suspected of measuring a secondary dimension. This resulted in a DIMTEST statistic of 5.8074, indicating rejection of the DIMTEST null hypothesis of unidimensionality with a p -value less than 0.00005. These results were not surprising because strict unidimensionality is an idealization that almost never holds exactly for a given data set. Thus, it was important to use DETECT to estimate the effect size of the violation of local independence found by DIMTEST.

A two cluster confirmatory DETECT analysis was conducted on the MHSa mathematics test using all the students (more than 13,000), with the SAT items selected as one cluster and the Math–A items selected as the second cluster. This resulted in a DETECT statistic of 0.06, a value indicative of a nearly unidimensional test. Furthermore, the ratio of the DETECT statistic to the maximum possible value of the DETECT statistic was only 0.23, and the percentage of conditional covariance pairs having positive signs for item pairs in the same cluster and negative signs for items coming from different clusters was only 54.1%.

Taken together, the DIMTEST and DETECT results indicate that the SAT mathematics section and the Math–A section are measuring very similar dimensions, but there is also a very small amount of difference between them, detectable by DIMTEST because of the very large sample size employed in the analyses. Finally, an exploratory DETECT analysis was run using equal numbers of students in randomly sampled training and crossvalidation samples. The resulting DETECT statistic was 0.14, the ratio of this value to the maximum possible DETECT value was 0.47, and the percentage of signs corresponding to positive within-cluster and negative between-cluster conditional covariances was 65.1%. These results indicate some multidimensionality, though very weak in magnitude. The clusters reported in the analysis showed no correspondence to a separation of the SAT items and the Math–A items as different dimensions. The clusters that were reported appeared to be uninterpretable. The estimated signs of the conditional covariances indicated little systematic violation of local independence, and the size of the violation was clearly quite small. Because the violations of local independence, as evidenced by the DETECT effect sizes, were very small, they do not warrant any changes in test design or scoring. In particular, the dimensionality analysis

results support the application of unidimensional IRT to the combined set of SAT and Math–A items for purposes of calibrating, linking, scaling, and equating. Indeed, the results support using unidimensional IRT to place the SAT and Math–A items onto a single score scale for reporting purposes.

The MHSA science test was analyzed similarly. DIMTEST was first applied in an exploratory analysis using training and crossvalidation samples of more than 7,000 students each. This resulted in a DIMTEST statistic of 5.7257, indicating rejection of the DIMTEST null hypothesis of unidimensionality with a p -value less than 0.00005. DETECT was used to estimate the effect size of the violation of local independence found by DIMTEST. The results of the DETECT analysis gave a DETECT value of 0.17, which is indicative of very weak multidimensionality. Moreover, the ratio of this DETECT value to the maximum possible value was 0.50, and approximately 69% of the signs of the conditional covariances corresponded to the ideal of all the within cluster covariances being positive and all the between cluster values being negative. These statistics are indicative of very weak but noticeable multidimensionality. Furthermore, while the 2007–08 analysis indicated that the multiple-choice and constructed-response items primarily formed two separate clusters, the separation of multiple-choice and constructed-response items was much less strong in this year’s analysis. One cluster consisted of mostly constructed-response items, but it also included a substantial number of multiple-choice items, and few other constructed-response items fell in another cluster dominated by multiple-choice items.

Taken together, the DIMTEST and DETECT results for science indicate that the test has very small, though detectable, violations of unidimensional local independence, and that these violations are only somewhat related to the two item types used on the test. The weak multidimensionality detected does not warrant changes in test design or scoring..

11.7 Item Response Theory Analyses

Subsection 11.3.1 introduced IRT and gave a thorough description of the topic. It was noted there that all MHSA mathematics and science items were calibrated using IRT and that the calibrated item parameters were ultimately used to scale both the items and students onto a common framework. In this section, the statistical characteristics of the MHSA mathematics and science tests are presented from an IRT perspective.

Before the IRT results are presented, it is important to reiterate that because the mathematics standards were established on the θ scale and the SAT items and the Math–A items were calibrated together onto a common θ scale, the two item sets had to be in concert with one another from a statistical perspective. Psychometricians at Measured Progress spent significant time studying the adequacy of the model fit. This process compared observed data to ICCs using the estimated item parameters. The extent to which these two curves are similar is a reflection of model fit. Given that a polytomous IRT model was used, model fit curves were drawn for each score point for each item. Figure 11-1 shows the q-q plot for the two parts of the test. A q-q plot is helpful for comparing two sets of data. In the figure, formula scores for Math–A are shown on the vertical axis and formula scores for the SAT on the horizontal axis. Each point on the graph represents the

same quantile for each section, where a quantile is the percentage of students who scored below the given value. If both data sets had exactly the same distribution, the resultant plot would be a straight line. In any real data set, some variability around the line is expected. The points in Figure 11-1 tend to fall below the line, indicating that the mean score on the Math–A items is somewhat lower than the mean score on the SAT items. Despite the difference between the SAT and Math–A items revealed by the q-q plot, the analysis does not lead to the conclusion that one set of items is better than the other. Bear in mind that in this context, the Math–A items were designed to serve the critically important role of ensuring that adequate content area coverage was obtained.

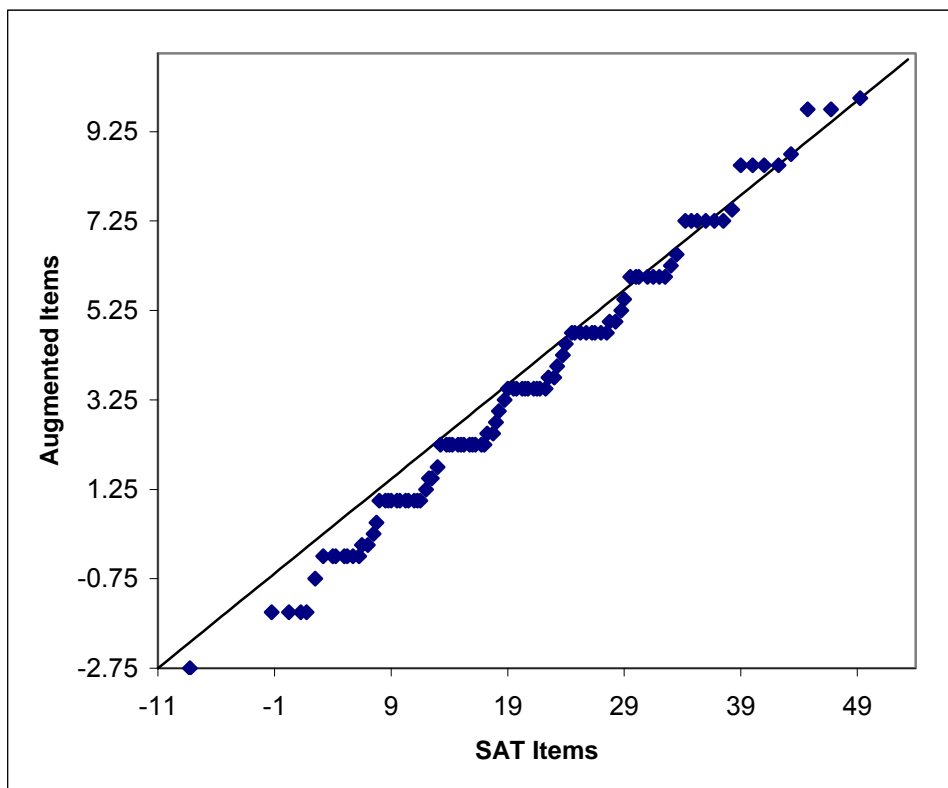


Figure 11-1. 2008–09 MHSA: q-q Plot of Math–A and SAT Items

Table 11-10 summarizes the IRT item parameters from the two parts of the test. The b parameter is an index of item difficulty and represents the point on the ability scale at which an examinee has a 50% probability of answering that item correctly. The item discrimination parameter, a , is proportional to the slope of the ICC at the point of inflection. High values of a are characteristic of steep ICCs and increase sharply as a function of ability, while low values of a are characteristic of ICCs that increase gradually as a function of ability.

Table 11-10. 2008–09 MHSA: Summary of Item Parameters for the MHSA Mathematics Test

<i>Parameter</i>	<i>Statistic</i>	<i>SAT Items</i>	<i>Math–A Items</i>
<i>a</i> parameter	Average	0.72	0.47
	SD	0.28	0.13
	Minimum	0.23	0.34
	Maximum	1.51	0.71
<i>b</i> parameter	Average	0.01	0.16
	SD	1.07	0.76
	Minimum	-3.17	-1.08
	Maximum	2.60	1.26

SD = standard deviation

As seen in Table 11-10, the Math–A items have lower discrimination and are harder than the SAT items (as indicated by their mean *a* and *b* parameters). An examination of overall model fit indicated adequate fit of the IRT model to both item sets, lending support to the IRT analysis results.

Finally, Figure 11-2 shows the proportional TCCs for the two sets of items. (Note: The expected scores were transformed onto the unit scale since the maximum raw scores on the two parts of the test were different. This common practice essentially standardizes TCCs based on different numbers of score points so that more-direct comparisons can be made.)

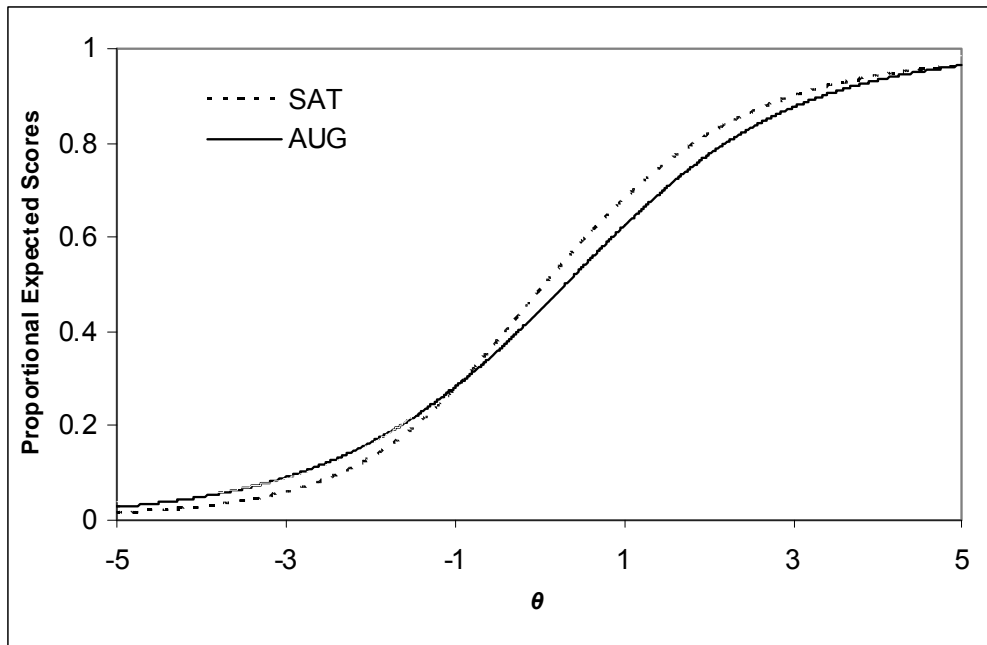


Figure 11-2. 2008–09 MHSA: TCC Plots of Math–A and SAT Items

Figure 11-2 confirms the IRT parameter summary in Table 11-10. The Math–A items have lower discrimination power than the SAT items, as indicated by the more gradual slope of the Math–A item TCC compared to that of the SAT TCC. The Math–A items are also slightly harder than the SAT items, as indicated by the slightly higher inflection point of the Math–A item TCC than that of the SAT item TCC.

Though the Math–A item set had less discrimination power, psychometricians at Measured Progress were satisfied with the resulting parameters, the relationship between the SAT and Math–A items, and as noted above, overall item-level model fit.

Table 11-11 summarizes the IRT item parameters from the MHSA science test.

Table 11-11. 2008–09 MHSA: Summary of Item Parameters for the MHSA Science Test

<i>Parameter</i>	<i>Statistic</i>	<i>Science Items</i>
<i>a</i> parameter	Average	0.54
	SD	0.17
	Minimum	0.23
	Maximum	0.97
<i>b</i> parameter	Average	-0.04
	SD	1.03
	Minimum	-1.55
	Maximum	3.06

SD = standard deviation

11.8 Reliability

Although an individual item’s performance is an important focus for evaluation, a complete evaluation of a test must also address the way in which items function together and complement one another. Any measurement includes some amount of measurement error. No academic test can measure student performance with perfect accuracy; some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. Items that function well together produce tests that have less measurement error (i.e., the error is small on average). Such tests are described as *reliable*.

There are a number of ways to estimate a test’s reliability. One approach is to split all test items into two groups and then correlate students’ scores on the two half-tests. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, items on the two half-tests are likely measuring very similar knowledge or skills. Such a correlation is evidence that the items complement one another and suggest that measurement error will be minimal.

The split-half method requires psychometricians to select items that contribute to each half-test score. This decision may have an impact on the resulting correlation, since each different possible split of the test halves will result in a different correlation. Another problem with the split-half method is that it underestimates reliability, because test length is cut in half. All else being equal, a shorter test is less reliable than a longer test. Cronbach (1951) provided a statistic, alpha (α), which avoids this concern of the split-half method. By comparing individual item variances to total test variance, Cronbach’s α coefficient estimates the average of all possible split-half reliability coefficients. Alpha was used to assess the reliability of the 2008–09 MHSA tests. The formula for computing alpha is as follows:

$$\alpha \equiv \frac{n}{n-1} \left[1 - \frac{\sum_{i=1}^n \sigma^2_{(Y_i)}}{\sigma_x^2} \right]$$

where
i indexes the item
n is the total number of items,
 $\sigma^2_{(Y_i)}$ represents individual item variance, and
 σ_x^2 represents the total test variance.

11.9 Reliability and Standard Errors of Measurement

Table 11-12 presents descriptive statistics, Cronbach’s α coefficient, and raw score standard errors of measurement (SEMs) for the MHSAs mathematics and science tests.

Table 11-12. 2008–09 MHSAs Mathematics and Science Raw Score Descriptive Statistics, Reliabilities, and SEMs

<i>Content Area</i>	<i>N</i>	<i>Possible Score</i>	<i>Min Score</i>	<i>Max Score</i>	<i>Mean Score</i>	<i>Score SD</i>	<i>Reliability (α)</i>	<i>SEM</i>
Mathematics	15032	55	0	55	33.3	16.4	0.98	2.1
Science	14867	56	0	55	22.8	11.8	0.88	4.1

SD = standard deviation

Conditional standard error of measurement (CSEM) bands for mathematics are reported in Tables 11-13 and 11-14 (mathematics is repeated from Chapter 10).

**Table 11-13. 2008–09 MHSA: Scaled Score
Conditional Standard Error Bands for MHSA Mathematics**

<i>Scaled Score</i>	<i>Lower Bound*</i>	<i>Upper Bound*</i>	<i>Scaled Score</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
1100	1100.0	1104.3	1142	1140.0	1144.0
1102	1105.5	1125.0	1144	1141.8	1146.2
1104	1105.5	1125.0	1146	1144.8	1147.2
1106	1105.5	1125.0	1148	1146.2	1149.8
1108	1105.5	1125.0	1150	1147.9	1152.1
1110	1105.5	1125.0	1152	1149.9	1154.1
1112	1105.5	1125.0	1154	1151.8	1156.2
1114	1105.5	1125.0	1156	1154.4	1157.6
1116	1105.5	1125.0	1158	1156.3	1159.7
1118	1105.5	1125.0	1160	1158.1	1161.9
1120	1105.5	1125.0	1162	1159.4	1164.6
1122	1105.5	1125.0	1164	1160.9	1167.1
1124	1111.0	1137.0	1166	1162.4	1169.6
1126	1119.4	1132.6	1168	1164.0	1172.0
1128	1126.4	1129.6	1170	1166.5	1173.5
1130	1128.5	1131.5	1172	1168.5	1175.5
1132	1130.7	1133.3	1174	1170.9	1177.1
1134	1132.0	1136.0	1176	1173.5	1178.0
1136	1134.5	1137.5	1178	1175.0	1180.0
1138	1136.2	1139.8	1180	1178.7	1180.0
1140	1138.8	1141.2			

* Because there are a variety of ways to achieve any particular mathematics scaled score (due to the combination of SAT and Math–A items), the upper and lower bound values in this table reflect the averages of their respective distributions, each calculated using the test information method.

**Table 11-14. 2008–09 MHSA: Scaled Score
Conditional Standard Error Bands for MHSA Science**

<i>Scaled Score</i>	<i>Lower Bound*</i>	<i>Upper Bound*</i>	<i>Scaled Score</i>	<i>Lower Bound</i>	<i>Upper Bound</i>
1100	1100.0	1101.1	1142	1139.8	1144.3
1102	1100.0	1104.5	1144	1141.5	1146.5
1104	1101.0	1107.0	1146	1143.7	1148.3
1106	1102.0	1110.0	1148	1145.8	1150.2
1108	1104.5	1111.5	1150	1147.7	1152.3
1110	1107.0	1113.0	1152	1149.7	1154.3
1112	1109.3	1114.7	1154	1151.8	1156.2
1114	1111.3	1116.8	1156	1153.4	1158.6
1116	1113.8	1118.3	1158	1155.8	1160.3
1118	1115.8	1120.2	1160	1156.8	1163.2
1120	1117.8	1122.2	1162	1159.0	1165.0
1122	1119.8	1124.2	1164	1160.7	1167.3
1124	1122.3	1125.7	1166	1162.8	1169.3
1126	1124.0	1128.0	1168	1164.5	1171.5
1128	1125.9	1130.1	1170	1166.0	1174.0
1130	1128.3	1131.8	1172	1169.0	1175.0
1132	1130.5	1133.5	1174	1170.0	1178.0
1134	1132.0	1136.0	1176	1172.0	1180.0
1136	1133.4	1138.6	1178	1175.0	1180.0
1138	1136.0	1140.0	1180	1178.9	1180.0
1140	1138.5	1141.5			

* Because there are a variety of ways to achieve any particular science scaled score (the same raw score can be arrived at in multiple ways due to formula scoring), the upper and lower bound values in this table reflect the averages of their respective distributions, each calculated using the Lord binomial method.

11.10 Classification Accuracy and Consistency of MHSAs Cut Scores in Mathematics and Science

In Section 10.4, the Livingston and Lewis (1995) procedure for calculating classification accuracy and consistency of cut scores was described in detail. Tables 11-15 through 11-18 present the results for MHSAs mathematics and science (mathematics is repeated from Chapter 10).

**Table 11-15. 2008–09 MHSAs: Accuracy
Contingency Table for Mathematics and Science**

<i>Content Area</i>	<i>True</i>	<i>Observed</i>				<i>Total</i>
		<i>Does Not Meet</i>	<i>Partially Meets</i>	<i>Meets</i>	<i>Exceeds</i>	
Mathematics	Does Not Meet	0.21	0.04	0.00	0.00	0.25
	Partially Meets	0.04	0.17	0.06	0.00	0.27
	Meets	0.00	0.05	0.39	0.01	0.45
	Exceeds	0.00	0.00	0.00	0.02	0.02
	Total	0.26	0.26	0.46	0.03	1.00
Science	Does Not Meet	0.27	0.05	0.01	0.00	0.33
	Partially Meets	0.05	0.10	0.06	0.00	0.21
	Meets	0.00	0.05	0.38	0.02	0.44
	Exceeds	0.00	0.00	0.00	0.01	0.02
	Total	0.33	0.20	0.45	0.03	1.00

**Table 11-16. 2008–09 MHSAs: Consistency
Contingency Table for Mathematics and Science**

<i>Content Area</i>	<i>True</i>	<i>Observed</i>				<i>Total</i>
		<i>Does Not Meet</i>	<i>Partially Meets</i>	<i>Meets</i>	<i>Exceeds</i>	
Mathematics	Does Not Meet	0.20	0.05	0.01	0.00	0.26
	Partially Meets	0.05	0.14	0.07	0.00	0.26
	Meets	0.01	0.07	0.37	0.01	0.46
	Exceeds	0.00	0.00	0.01	0.02	0.03
	Total	0.26	0.26	0.46	0.03	1.00
Science	Does Not Meet	0.25	0.06	0.02	0.00	0.33
	Partially Meets	0.06	0.08	0.06	0.00	0.20
	Meets	0.02	0.06	0.35	0.01	0.45
	Exceeds	0.00	0.00	0.01	0.02	0.03
	Total	0.33	0.20	0.45	0.03	1.00

**Table 11-17. 2008–09 MHSAs: Summary of Overall
Mathematics and Science Classification Accuracy and Consistency**

<i>Content Area</i>	<i>Accuracy</i>	<i>Consistency</i>	<i>Kappa</i>
Mathematics	0.79	0.72	0.57
Science	0.77	0.70	0.54

Table 11-18. 2008–09 MHSAs: Accuracy and Consistency of Mathematics and Science Dichotomous Categorizations

<i>Content Area</i>	<i>Performance Level</i>	<i>Accuracy</i>	<i>False Positive</i>	<i>False Negative</i>	<i>Consistency</i>
Mathematics	D/P	0.91	0.04	0.04	0.88
	P/M	0.89	0.06	0.05	0.85
	M/E	0.98	0.01	0.00	0.97
Science	D/P	0.90	0.05	0.05	0.85
	P/M	0.88	0.07	0.05	0.84
	M/E	0.98	0.02	0.00	0.97

D/P = Does Not Meet/Partially Meets; P/M = Partially Meets/Meets; M/E = Meets/Exceeds

Chapter 12. VALIDITY RESEARCH ON THE MHSA

This chapter seeks to bring together a wide range of validity evidence regarding the MHSA, (consisting of the SAT, Math–A, and science tests) in a logical and systematic manner. It is guided by the concept of validity articulated in *Standards for Educational and Psychological Testing* (AERA et al., 1999), which provides the following definition: “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests.” Further, “The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations” (AERA et al., 1999, p. 9). The initial sections (12.1–12.12) of this chapter will provide some of the more recent evidence supporting the interpretation of SAT scores and will refer to other evidence that has accumulated over past decades. Section 12.13 will also provide information regarding Math–A (the augmentation to the SAT mathematics section), and the science section. Some evidence relates to test content, some to the processes used in responding to the test, some to the internal structure of the test, and still more to the relationship of test scores to other variables, especially criteria such as performance in particular content areas or college grades. Section 12.14 contains a validity studies agenda for future consideration.

12.1 Construct Validity

The SAT is described variously as “a measure of the critical thinking skills you will need for academic success in college,”⁹ or as an assessment of “student reasoning based on knowledge and skills developed by the student in school coursework.”¹⁰ What is the nature of the reasoning or critical thinking that is measured by the SAT? Powers and Dwyer (2003) seek to delineate a construct of reasoning, broadly conceived. They point out that “a construct provides a target for a particular assessment; it is not synonymous with the test itself” (p. 1). They identify several definitions of reasoning that have been used by educators, philosophers, and psychologists and note that more recent conceptions of reasoning have emphasized the importance of domain specific reasoning, i.e., reasoning that is knowledge based. Similarly, a considerable range of definitions for thinking or critical thinking exists. They conclude that there is no single construct of reasoning but that any of the several formulations may be useful and informative depending on the context and purpose.

Powers and Dwyer (2003) argue for the importance of reasoning in academic contexts, such as performance in college. “But of the many things that matter, two of the most important, we believe, are: (a) academic knowledge and skill in the domain of study, and (b) the ability to reason well in the symbol systems used to communicate new knowledge. Reasoning tests correlate with academic success because reasoning abilities are very often required in school learning, whether for understanding a story, inferring the meaning of an unfamiliar word, detecting patterns and regularities in information, going beyond the information given

⁹ From *SAT preparation booklet 2004–2005* by the College Board, 2004, p. 3.

¹⁰ From *About the new SAT*, retrieved from www.collegeboard.com on January 21, 2005.

to form more general rules or principles, or applying mathematical concepts to solve a problem. In these ways and in hundreds of others, successful learning requires reasoning strategies” (p. 12).

This argument seems particularly apropos to the stated purpose of the SAT as a tool in counseling and admissions decisions regarding future learning opportunities. Out of the many possible facets of reasoning, the College Board has chosen to assess three dimensions that are closely related to academic performance: verbal reasoning in the form of critical reading, quantitative reasoning using a defined domain of academic knowledge, and writing—the productive use of a symbol system to communicate one’s ability to present and support a point of view.

12.2 Verbal Reasoning

The critical reading section is based on written discourse. Male and female references are balanced, and representative minority relevant content is included in each test. Approximately 72% (48) of the items are based on passages, while 28% (19) of the items are in the sentence completion format.

Sentence completion items are useful for measuring an understanding of the relationships among words and concepts, an understanding of the structure of the text, and knowledge of vocabulary. Within a given form of the critical reading section, a balance exists between those items that primarily measure vocabulary and those that measure reasoning about the logic of a sentence.

The passage based reading content is balanced across four categories: humanities, social studies, natural sciences, and literary fiction. The preponderance of the items (approximately 80%) measure higher level reading skills of the following types:

- **Primary purpose:** These questions ask about the main idea of a passage or about the author’s primary purpose in writing the passage. They address the passage as whole, or an entire paragraph, rather than focusing on a smaller part of the passage. These questions tap both the process of understanding discourse and of interpreting discourse.
- **Rhetorical strategies:** These questions usually focus on a specific part of a passage, often on a particular word, image, phrase, example, or quotation and ask why this particular element is present or what purpose it serves, rather than simply on what it means. Such questions involve the processes of interpreting discourse and evaluating discourse.
- **Implication and evaluation:** These questions go beyond the passage by asking what the information presented in the passage suggests, or what can be inferred about the author’s view. They might also ask the test taker to evaluate ideas or assumptions in a passage, or to evaluate the relationship between a pair of passages. These questions involve the process of evaluating the discourse and may involve aspects of creating new understandings.
- **Tone and attitude:** These questions ask about the author’s tone or attitude in the whole or a specific part of the passage. Such questions tap into the test taker’s ability to interpret discourse.
- **Application and analogy:** These questions may address a specific idea or relationship in a passage and ask the test taker to recognize a parallel idea or relationship in a different context. Such questions may also ask the test taker to recognize an additional example that would support an idea presented in the passage or may ask about an analogy that is used. Alternatively, these

questions may ask how ideas presented in one passage apply to another passage, or how the author of one passage would be likely to react to an idea expressed in a related passage. Such questions draw on the test taker’s ability to evaluate discourse and to create new understandings.

A few questions in each critical reading section test the literal comprehension of what is being said in a particular part of the passage. A few others, vocabulary in context questions, probe what a specific word means as it is used in a passage. Both of these question types draw on the process of understanding discourse.

The critical reading section taps several of the underlying dimensions posited by Burton, Welsh, Kostin, and Van Essen (2004), especially the breadth and depth of understanding in a receptive mode. The critical reading section samples the construct of verbal reasoning in a variety of ways. The detailed specifications (see Tables 2-1 through 2-3) ensure that each succeeding form or version of the test samples similar aspects of that construct. In addition, key aspects of the process of communicating are addressed in the writing portion of the SAT (see Section 12.3).

12.3 Quantitative Reasoning

Dwyer, Gallagher, Levin, and Morley (2003) have reviewed the research on quantitative reasoning in an effort to better define the construct for assessment purposes. They observe, “Although the assessment of quantitative reasoning has been a measurement goal from early in the 20th century, systematic treatment of quantitative reasoning as a cognitive process distinct from mathematics as content or curriculum did not begin to take shape until much later” (p. 7). Further, they point out “that it is critical to the interpretation of reasoning tests to differentiate between elements of the reasoning construct itself that is the target of the assessment and the common core of content knowledge that all test takers are assumed to bring to the test” (p. 12). They recognize that “it is not possible, however, to assess quantitative reasoning without the content since it is the manipulation and application of the content that allows test takers to demonstrate their reasoning” (p.13). Dwyer et al. define quantitative reasoning “as the ability to analyze quantitative information” and note that it includes six capabilities.

1. Reading and understanding information given in various formats, such as in graphs, tables, geometric figures, mathematical formulas or in text
2. Interpreting quantitative information and drawing appropriate inferences from it
3. Solving problems, using arithmetical, algebraic, geometric, or statistical methods
4. Estimating answers and checking answers for reasonableness
5. Communicating quantitative information verbally, numerically, algebraically, or graphically
6. Recognizing the limitations of mathematical or statistical methods (p.13)

Dwyer et al. (2003) stress that the validity and fairness of an assessment of quantitative reasoning depends on limiting the content of the assessment to a level of mathematical knowledge that is explicitly

assumed to be common throughout the testing population (p.15). Independent of any particular mathematical content or level of mathematical achievement, Dwyer et al. posit a problem solving process of three multifaceted steps.

1. Understanding and defining the problem
2. Solving the problem
3. Understanding results

This problem solving process becomes the target for any assessment of quantitative reasoning even though the authors acknowledge that “in practice, most tests are designed to assess only a portion of the quantitative reasoning process” (Dwyer et al., 2003, p.15). In responding to the SAT mathematics questions, students need to apply this process in the context of two different item types: multiple-choice questions, and student-produced responses—in which a student must solve the problem and fill in the numeric response (no options are provided). There are 44 items in multiple-choice format and 10 in student-produced-response format.

Students must apply this problem solving process to questions drawn from a particular content domain within mathematics. In broad terms, they must have knowledge of numbers and operations, algebra and functions, geometry, measurement, statistics, probability, and data analysis. The boundaries of this domain were somewhat expanded in creating the new SAT, first administered in March 2005, to reflect the fact that 98% of the college bound seniors cohort have taken three or more years of mathematics in secondary school, including 96% who have studied algebra and 95% who have studied geometry.¹¹ Consequently, the educators who helped to define the new test thought it appropriate to slightly increase the level of mathematical content assumed on the test to include such content from a third-year high school mathematics course as exponential growth, absolute value, and functional notation. The new test also places greater emphasis on such other topics as linear functions, manipulations with exponents, and properties of tangent lines.

Two aspects of the SAT underscore that this is a test of quantitative reasoning rather than solely mathematical knowledge: (1) students are permitted to use a four function, scientific, or graphing calculator on the test—although it is possible to solve every question without a calculator; and (2) students are provided with commonly used formulas in the test book itself, so that they do not have to memorize them. The purpose of these two “helps” is to send a clear signal to the test taker about the reasoning nature of the test.

The specifications for the mathematics section of the new SAT were presented in Chapter 2, Tables 2-9 through 2-11. Each form of the test is defined in terms of the item types to be used, the mathematical content that provides the opportunity for demonstrating quantitative reasoning, as well as the distribution of questions of different levels of difficulty.

¹¹ From “2005 college bound seniors: Total group profile report,” by the College Board, 2005, table 3-1.

12.4 Writing

Although the College Board has previously offered tests of writing¹² for use in making admissions and placement decisions, the 2005 revision of the SAT was the first to incorporate a writing test that includes a direct measure of writing proficiency. Writing is an extremely complex activity: it can include different modes of discourse (e.g., narration, argumentation, description), while calling on a range of cognitive skills (e.g., interpreting, analyzing, synthesizing, organizing) and requiring various kinds of knowledge (e.g., understanding linguistic structures). Thus, it is not useful to think of writing as a unitary construct. Breland, Bridgeman, and Fowles (1999) observe, “Even if a unitary construct of writing could be defined, no single test could possibly assess the full domain” (p. 1).

The several groups of educators who helped to define the new SAT writing test chose particular aspects of writing to be tested. The student is asked to write a first draft essay and respond to multiple-choice questions that assess the ability to identify errors in sentences and to improve sentences and paragraphs. These skills relate closely to the cognitive operation of communication described by Burton et al. (2004). The specifications for the writing test may be found in Tables 2-4 through 2-8.

12.5 Multiple-choice Questions

The multiple-choice questions assess how well students use standard written English and test students’ ability to identify sentence errors, improve sentences, and improve paragraphs. The multiple-choice writing questions are used to evaluate a student’s ability to

- use language that is consistent in tenses and pronouns;
- understand parallelism, noun agreement, and subject-verb agreement;
- understand how to express ideas logically;
- avoid ambiguous and vague pronouns, wordiness, improper modification, and sentence fragments; and
- understand proper coordination and subordination, logical comparisons, diction, idiom, and modification and word order.

The multiple-choice writing questions do not ask the students to define or use grammatical terms and do not test spelling and capitalization. Using the multiple-choice format, the test assesses a student’s control of different levels of writing. Focused on improving sentences, some (25) questions ask the student to recognize and correct faults in usage and sentence structure, as well as recognize effective sentences that follow the conventions of standard written English. Others (18) ask the student to recognize and correct errors

¹² The Test of Standard Written English was administered with the SAT from 1974 to 1995. The English Composition Test (sometimes with and sometimes without an essay) was part of the Achievement Test series from the 1940s to 1995. The SAT II: Writing Test was offered from 1995 to 2005.

of grammar and usage in sentences. The third type of multiple-choice question asks the student to improve paragraphs. This type of question assesses a student's ability to edit and revise sentences in the context of a paragraph or entire essay, organize, and develop paragraphs in a coherent and logical manner, while applying the conventions of standard written English (College Board, 2004, pp. 27–30).

12.6 Essay Question

The SAT writing test provides 25 minutes for a student to write a first draft essay in response to an assignment question. The student is presented with a short paragraph adapted from a published text that offers a perspective on an issue and with a question that asks for his or her point of view. The student is asked to think critically about the issue and develop a point of view, using reasoning and examples taken from reading, studies, experience, or observation to support that point of view. The essay measures a student's ability, under timed conditions, to do the kind of writing required in most college courses—writing that emphasizes precise use of language, logical presentation of ideas, development of a point of view, and clarity of expression. SAT essay prompts are developed according to the following guidelines:

- They should be accessible to the general test taking population, including students for whom English is not a first or best language.
- They should be relevant to a wide range of fields and interests, and neither require specialized knowledge nor advantage students who have completed a specific course of study.
- They should engage high school age students while stimulating critical reflection about important topics.
- They should be free of figurative or technical language or specific literary references.
- They should give the students the opportunity to use a broad spectrum of experiences, learning, and ideas to support their points of view.

The elements of writing that can be assessed through this direct measure are reflected in the scoring guide that readers use to evaluate and score the student essays holistically. The scoring guide used by the readers is displayed in Chapter 2.

12.7 Does the Length of the SAT Result in a Fatigue Effect?

With the introduction of the new SAT total testing time was increased for all examinees. Wang (2006) examined the effect of increased testing time by comparing four performance indices calculated using randomly equivalent examinee subpopulations on sections of similar content and difficulty administered at different times on three SAT administrations. This study was conducted to address concerns that the increased test length of the new SAT was resulting in increased fatigue and poorer performance. A variety of analyses were conducted in this study, and the researcher found no evidence that the current SAT test length had affected examinee performance at the population level or differentially across gender, racial/ethnic, and

language subgroups. On the contrary, this study produced consistent findings, indicating that examinees performed the same on sections of similar content and difficulty, both in terms of direct group comparisons and comparisons conditional on total score, throughout the entire SAT. Furthermore, the findings from the March and October 2005 SAT data were replicated using the May 2002 SAT I data, indicating no significant changes in performance trends between the two tests.

12.8 How Do SAT Scores Relate to College Performance?

Much of the empirical evidence for the validity of the SAT is based on analyses of the relationship of test scores to performance in college (Angoff, 1971; Wilson, 1983; Donlan, 1984; Willingham, Lewis, Morgan, & Ramist, 1990; Hezlett, Kuncel, Vey, Ahart, Ones, Campbell, & Camara, 2001; Young & Kobrin, 2001). Drawing heavily on the Young and Kobrin review, evidence gathered since 1994 is presented below.

Kobrin and Michel (2006) explored the question of whether the SAT or high school grade point average (HSGPA) is a better predictor of freshman grade point average (FGPA) for students with high FGPA compared to students with lower FGPA. Employing logistic regression, they predicted the probability of a student successfully achieving a FGPA at various levels, based on that student's SAT scores and HSGPA. They found that in the total sample, at all success criterion levels except the 2.5 level, the SAT was equal to or slightly more accurate than HSGPA in predicting successful students, but generally less accurate than HSGPA in predicting unsuccessful students. However, at the highest FGPA level, 3.75 or higher, neither the SAT nor the HSGPA was able to predict successful students. Across each of the racial/ethnic groups, the SAT was typically a better predictor of successful students, and HSGPA was typically a better predictor of unsuccessful students. For students attending the most selective colleges, the SAT was more effective than or equally effective as HSGPA in predicting success at nearly all FGPA criterion levels. However, for students attending the least selective colleges, HSGPA tended to be a better predictor of success.

Norris, Oppler, Kuang, Day, and Adams (2006) studied the predictive and incremental validity of a prototype version of the recently introduced SAT writing section. Data were collected in 2003–2004 from 13 institutions, both public and private, from different sections of the country. The study included institutions of different levels of selectivity and of different size freshman classes. Data were available for a total of 1,572 students who took the SAT writing prototype and who also took the operational SAT. Note that the SAT verbal (SAT-V) and SAT mathematics (SAT-M) scores were earned in a standard administration with high motivation, whereas the writing score was earned in an experimental administration with only an unspecified monetary incentive. The incremental validity could be different if all three scores had been earned under the same motivational condition. Such data should become available in the near future.

Norris et al. (2006) obtained two criteria—FGPA and English composition grade point average (ECGPA). Because of the variability across participating institutions, all analyses were conducted within each institution, and then weighted averages were calculated and pooled across institutions to derive the overall

estimate. Statistical procedures to correct for multivariate range restriction (Lord & Novick, 1968) and shrinkage (Rozeboom, 1978) were applied.

The relationship of each of the predictors with FGPA and ECGPA is shown in Table 12-1. The values in the table represent the weighted-average validity coefficients across all of the participating institutions..

Table 12-1. 2008–09 MHSA SAT Component—Weighted Average Correlations for Predictors With FGPA and ECGPA

Predictor	FGPA			ECGPA		
	N	Corrected	Uncorrected	N	Corrected	Uncorrected
SAT verbal	1,248	0.49	0.32	891	0.30	0.20
SAT mathematics	1,248	0.47	0.29	891	0.23	0.10
SAT total	1,248	0.51	0.35	891	0.28	0.17
SAT essay	1,248	0.20	0.16	891	0.18	0.14
SAT multiple-choice	1,248	0.45	0.30	891	0.31	0.22
SAT writing total	1,248	0.46	0.32	891	0.32	0.24
HSGPA	1,248	0.43	0.38	891	0.35	0.32

Note: Corrected for multivariate range restriction (Lord & Novick, 1968). Source: Norris et al. (2006), Table 9.

These data show very similar corrected correlations with FGPA for each of the section scores and HSGPA. In other words, SAT writing (SAT-W) is about as strongly related to freshman performance as are SAT-V, SAT-M, and HSGPA. The SAT-W, the writing multiple-choice section, as well as the SAT-V, are fairly predictive of English composition grades with corrected validity coefficients of 0.32, 0.31, and 0.30, respectively.

To assess the incremental validity of the SAT-W for predicting FGPA, a series of hierarchical regression analyses were conducted. Model A examined the incremental validity of adding SAT-W to a traditional SAT-V + SAT-M + HSGPA regression analysis. The results are shown in Table 12-2.

Table 12-2. Weighted Average Incremental Validity Results Across Institutions for Predicting First Year GPA (Model A)

Step	Adjusted		Unadjusted		
	R	ΔR	R	ΔR	
Corrected	1. SAT-V + SAT-M + HGSPA		0.59	0.63	
	2. SAT-V + SAT-M + HGSPA + SAT-W		0.60	0.01	0.64
Uncorrected	1. SAT-V + SAT-M + HGSPA		0.46	0.51	
	2. SAT-V + SAT-M + HGSPA + SAT-W		0.47	0.01	0.53

Note: N = 1,248

Corrected correlations were corrected for multivariate range restriction (Lord & Novick, 1968). Adjusted correlations were adjusted for shrinkage using Rozeboom (1978) Formula 8. Source: Norris et al. (2006), Table 11.

As shown in Table 12-2, the incremental validity of the SAT-W scores when added to SAT-V, SAT-M, and HSGPA was 0.01 when corrections for range restriction and shrinkage were made. As a further exploration of the relative contribution of each of the predictors to predicting FGPA, Norris et al. (2006)

performed a regression analysis in which SAT-W was the first variable introduced. The results are shown in Table 12-3.

Table 12-3. Weighted Average Incremental Validity Results Across Institutions for Predicting FGPA (Model D)

	Step	Adjusted		Unadjusted	
		R	ΔR	R	ΔR
Corrected	1. SAT-W	0.43		0.46	
	2. SAT-W + HSGPA	0.54	0.11	0.58	0.12
	3. SAT-W + HSGPA + SAT-V + SAT-M	0.60	0.06	0.64	0.07
Uncorrected	1. SAT-W	0.28		0.32	
	2. SAT-W + HSGPA	0.43	0.16	0.47	0.16
	3. SAT-W + HSGPA + SAT-V + SAT-M	0.47	0.04	0.53	0.06

Note: N = 1,248

Corrected correlations were corrected for multivariate range restriction (Lord & Novick, 1968). Adjusted correlations were adjusted for shrinkage using Rozeboom (1978) Formula 8. Source: Norris et al. (2006), Table 14.

This study demonstrates that the SAT writing section is substantially related to both FGPA and to English composition grades. When used in a multiple regression analysis in combination with SAT-V, SAT-M, and HSGPA, the writing score makes only a small incremental improvement to the prediction of FGPA.

The immediate predecessor of the SAT-W was the SAT II: Writing Subject Test. Breland, Kubota, and Bonner (1999) examined the usefulness of this test as a predictor of writing performance in college English courses. Because of the great similarity of the SAT-W and the Writing Subject Test, it is likely that their results will be indicative of how SAT-W will perform in this regard.

The Breland, Kubota, and Bonner (1999) study emphasized criteria data from actual student performance in writing. Other data on course grades, student self-assessment of writing ability, and student accomplishments in writing were also collected. Data were obtained from eight colleges for students entering college in fall 1996. Each institution was asked to collect from each of approximately 40 students a total of four different writing samples from regular course work in first semester English composition courses. Although topics would be different in each institution, three general types of writing were requested: (1) response to text, (2) argument or persuasion, and (3) analysis. Two experienced readers read and scored each of eight samples for each student independently. Two criteria were developed—a total writing performance variable, based on all students who submitted all eight writing samples, and an average writing performance variable, based on all students who submitted at least four writing samples. A writing experience questionnaire, completed by the students, was used to generate scores for overall GPA, writing GPA, writing self-assessment, and self reported writing accomplishments. These self reported variables, along with student scores on SAT-V, writing total, writing essay, and writing multiple-choice, were correlated with the two performance criteria. The results are shown in Table 12-4.

Table 12-4. Correlations of Test Scores and Student Self-Reports With College Writing Performance

Predictor Variables		Total Writing Performance			Average Writing Performance		
		Total Sample	Females	Males	Total Sample	Females	Males
		N = 112	N = 69	N = 43	N = 154	N = 93	N = 61
Test scores	SAT I verbal score	0.58*	0.64*	0.49*	0.54*	0.58*	0.58*
	SAT II: Writing score	0.48*	0.48*	0.47*	0.48*	0.49*	0.46*
	SAT II: Writing multiple-choice score	0.44*	0.39*	0.52*	0.43*	0.40*	0.46*
	SAT II: Writing essay score	0.21*	0.31*	-0.02	0.30*	0.37*	0.20
Self reports	High school GPA	0.10	0.25*	-0.10	0.25*	0.28*	0.21
	High school writing GPA	0.35*	0.39*	0.26	0.45*	0.44*	0.46*
	Writing self-assessment	0.27*	0.34*	0.11	0.30*	0.35*	0.25*
	Writing accomplishments	0.09	0.16	0.03	0.19*	0.18	0.20
	Self report composite	0.31*	0.40*	0.11	0.42*	0.43*	0.42*

* $p < 0.05$; Source: Breland, Kubota, & Bonner (1999), Table 3.

The SAT-V score, SAT II: Writing score, and SAT II: Writing multiple-choice subscore all predicted total writing performance reasonably well, while the SAT II: Writing essay subscore correlation was significantly lower, and virtually nonexistent for the males in the sample. The SAT-V correlation of 0.58 was significantly different from the writing GPA correlation of 0.35, and from the correlations of the SAT II: Writing score (0.48) and the writing multiple-choice subscore (0.44). For males, the self report variables showed little relationship to the total writing performance criterion. For the average writing performance criterion, the test score variables showed a pattern of correlations similar to that with the total writing performance criterion.

The relative contribution of the predictor variables was also examined through multiple regression analyses. The results of combining SAT-V and the SAT II: Writing Test score in the prediction of total writing performance and average writing performance are shown in Table 12-5.

Table 12-5. Predictive Contributions of SAT-Verbal and SAT II: Writing Test

Criterion Variable	Predictors	N	r	beta	R
Total writing performance	SAT-VI	127	0.57	0.50**	0.59
	SAT II: Writing		0.50	0.39*	(0.58)
Average writing performance	SAT-V	173	0.52	0.43**	0.56
	SAT II: Writing		0.49	0.40**	(0.55)

* $p < 0.05$; ** $p < 0.01$

Note: Figures for R in parentheses corrected for shrinkage. Source: Breland, Kubota, & Bonner (1999), Table 9.

Both the SAT-V and SAT II: Writing scores contributed significantly to the prediction of total writing performance. The multiple correlation of 0.59 is significantly higher than the zero-order correlation of 0.57 at the 0.05 level. Similarly, both predictors contributed significantly to the prediction of average writing performance. The multiple correlation of 0.56 is significantly higher than the zero-order correlation of 0.52 at

the 0.01 level of confidence. In summary, both SAT-V and SAT II: Writing make statistically significant contributions to the prediction of college writing performance.

12.9 Performance Over Multiple Time Periods

Hezlett, Kuncel, Vey, Ahart, Ones, Campbell, and Camara (2001) conducted a comprehensive meta-analysis of approximately 3,000 studies of the predictive validity of the SAT, involving over one million students. The observed correlations were corrected for range restriction. The reliability of the criterion measures was also taken into consideration. The results demonstrated that the validity coefficients of the SAT composite, SAT verbal, and SAT mathematics for predicting FGPA ranged from 0.44 to 0.62. The meta-analysis also confirmed that the SAT is a valid predictor of performance throughout college, showing a positive relationship not only with FGPA but also with the noncumulative GPA for second, third, and fourth year; the cumulative GPA at second and fourth year; the GPA in major; persistence; degree attainment; and State Nursing Board exams. However, evidence also suggested that the correlation between the SAT and college GPA declines over time.

Bridgeman, Pollack, and Burton (2004) took a somewhat unusual approach to demonstrating the relationship of college performance to preadmissions information such as SAT scores and HSGPA. Rather than the traditional multiple regressions methods that seek to demonstrate the percentage of “explained variance” associated with each variable, they categorized each of the variables into a limited number of levels and examined the variability in college performance for each combination of categories. Their goal was “to determine how many students at different levels of SAT score reached different criteria of success in college, after controlling on the selectivity of the college, the academic intensity of the students’ high school curriculum, and the students’ high school grades” (p. 1). They used data from over 60,000 students who had begun college in 1995 at 41 colleges that submitted course grades for the cohort over a multiyear period. The sample was geographically diverse, included both public and private institutions, and covered a fairly broad range of ability levels, although all of the colleges in the sample were somewhat selective. From the submitted data, the researchers computed a college grade point average (CGPA) for the end of the freshman year and the end of the senior year. As the criteria of college success, they used the number and percentage of students who earned a CGPA greater than 2.5 and greater than 3.5.

Their analysis categories were as follows.

- The college selectivity level used combined SAT verbal and mathematics scores. Level 1 colleges had mean combined SAT scores between 965 and 1093; Level 2 ranged from 1110 to 1195; Level 3 ranged from 1201 to 1249, and Level 4 scores ranged from 1256 to 1406.
- The academic intensity of the high school curriculum was defined in three levels based on the number of advanced placement (AP) exams taken and the number of years of study in the several disciplines. To be classified in Level 3 (high), a student had to have at least two AP exams in one area (mathematics/science or humanities/social science) and one AP exam in the other area. Level

2 represented strong coursework, while Level 1 represented less than the commonly recommended course work in the several disciplines.

- HSGPA was classified in four categories: Level 4 represented HSGPAs above 3.70; Level 3 included HSGPAs of 3.30 to 3.70; Level 2 included those between 2.71 and 3.29; and Level 1 included HSGPAs of 2.70 and below.
- SAT scores were divided into the following five levels:

Level 5—1410–1600	Level 2—810–1000
Level 4—1210–1400	Level 1—400–800
Level 3—1010–1200	

The relationship of college selectivity to each of the other variables is shown in Table 12-6. Since the SAT scores were used to define the college selectivity levels, it is no surprise to see the strong relationship between the combined SAT scores and those levels. However, a similar relationship can be seen between the college selectivity levels and both the academic intensity of students' high school curriculum and the GPA earned in high school. For example, over a third of the students in the most selective college category had an intensive (three or more AP courses) academic preparation, in contrast to 2% of the students in the least selective college category. Similarly, half of the students in the most selective colleges had a HSGPA of 3.7 or higher, while in the least selective colleges, only 20% had performed as well.

Table 12-6. Percentage of Students by Academic Intensity, HSGPA, SAT Score and Level of College Selectivity

		College Selectivity Level				
		Total	1 (low)	2	3	4 (high)
Academic Intensity	3 (high)	14	2	9	26	35
	2	62	52	68	64	59
	1 (low)	23	46	22	10	5
HSGPA	4 above 3.70)	36	20	34	51	50
	3 (3.30–3.70)	38	35	41	40	31
	2 (2.71–3.29)	18	28	19	8	17
	1 (2.70 and below)	7	17	6	2	3
SAT (V+M) scores	5 (1410–1600)	7	1	4	8	28
	4 (1210–1400)	34	12	34	49	52
	3 (1010–1200)	42	49	49	36	19
	2 (810–1000)	16	35	13	6	2
	1 (400–800)	1	4	1	0	0

Source: Bridgeman, Pollack, & Burton (2004) Tables 2–4.

Bridgeman, Pollack, and Burton (2004) calculated the number and percentage of students achieving freshman and senior CGPAs greater than 3.5 and 2.5 within every combination of the four variables previously discussed. A complete display of this data is provided in their report. However, a good sense of the

relationship of each of the preadmissions variables to performance in college can be gained by examining subsets of the data. For example, the contribution of SAT to understanding college performance can be observed by holding constant the college selectivity level, the intensity of academic preparation, and the HSGPA and observing how the different levels of SAT scores relate to the CGPA criteria. Table 12-7 presents the relationship of SAT score level to CGPA for students who achieved well in high school (Category 4) while taking the standard curriculum (Intensity 2) and attending Level 1 colleges

**Table 12-7. Freshman Success Rates in Level 1 Colleges
by SAT Score for Students in HSGPA Category 4 and Academic Intensity Level 2**

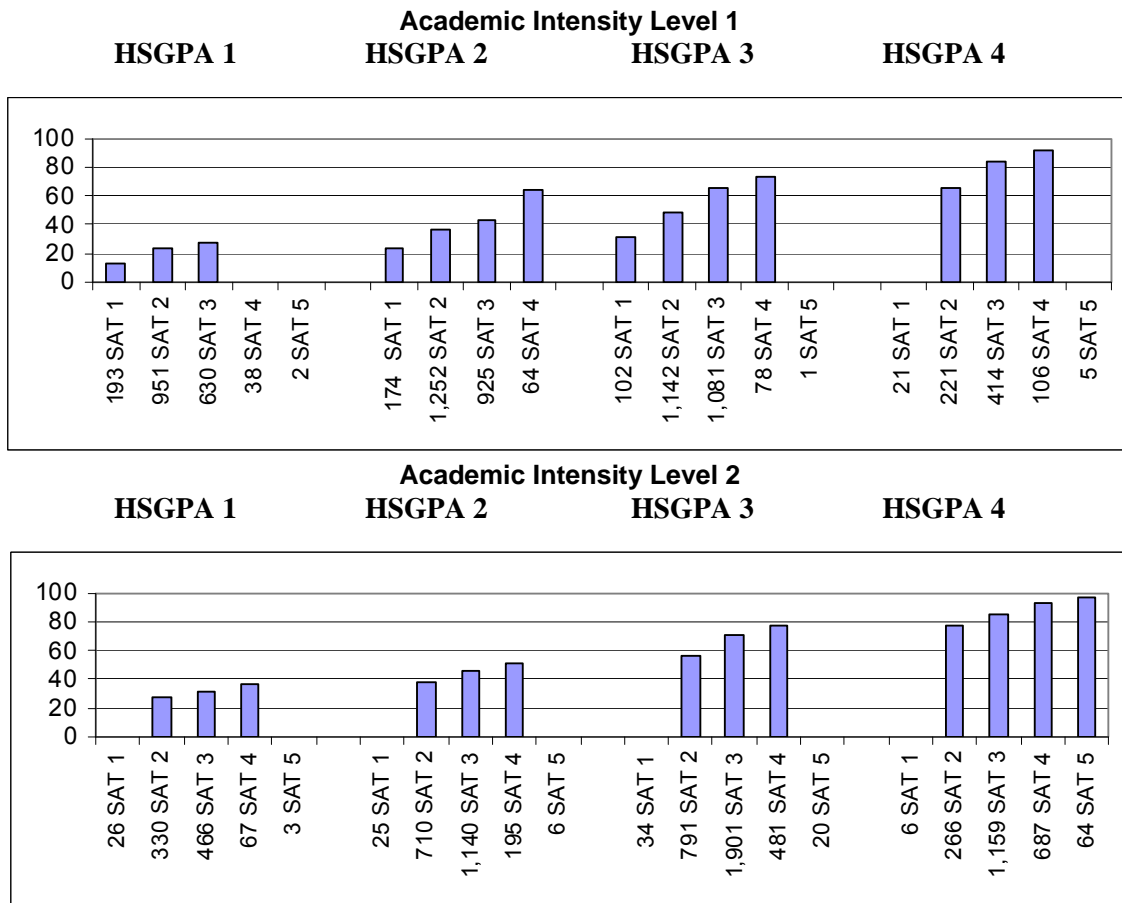
<i>Academic Intensity</i>	<i>HSGPA Category</i>	<i>Student SAT Level</i>	<i>Total N</i>	<i>N CGPA ≥ 3.5</i>	<i>N CGPA ≥ 2.5</i>	<i>Percent CGPA ≥ 3.5</i>	<i>Percent CGPA ≥ 2.5</i>
2	4	1(<800)	6	0	3	0.0	50.0
2	4	2(800–1000)	266	37	204	13.9	76.7
2	4	3(1010–1200)	1,159	345	1,001	29.8	86.4
2	4	4(1210–1400)	687	348	642	50.7	93.4
2	4	5(>1400)	84	49	62	76.6	96.9

Source: Bridgeman, Pollack, & Burton (2004) Table 5

Within this group of students who are relatively homogenous with respect to the level of college attended, the intensity of their academic preparation, and their HSGPA, it seems clear that SAT score is related to collegiate performance. Fewer than 14% of the students with SAT scores of 1000 or lower earned a freshman year CGPA of 3.5 or higher. More than half of the students with SAT scores over 1200 and 77% of the students with SAT scores over 1400 performed at this high level in college.

Figure 12-1 shows the percentage of students achieving a freshman CGPA of 2.5 or higher when the variables are fully crossed (SAT score level within HSGPA level within academic intensity level).¹³ This figure makes it clear that both high school grades and SAT scores are strong indicators of who will do well in college. For example, examining the students at academic intensity Level 1 who were in HSGPA Category 3 (3.3 to 3.7), one can observe that twice as many students at SAT Level 4 (1210–1400) earned a 2.5 CGPA than those in SAT Level 1. Almost 20% more of the students at SAT Level 3 achieved at this level than those at SAT Level 2. Similarly, if one examines the students at SAT Level 3 and academic intensity Level 1 across the HSGPA categories, there is a progression of 30% at HSGPA Category 1, 40% at Category 2, 60% at Category 3, and 80% at Category 4 completing the freshman year with a CGPA of 2.5 or higher. Similar relationships can also be observed for academic intensity Level 2.

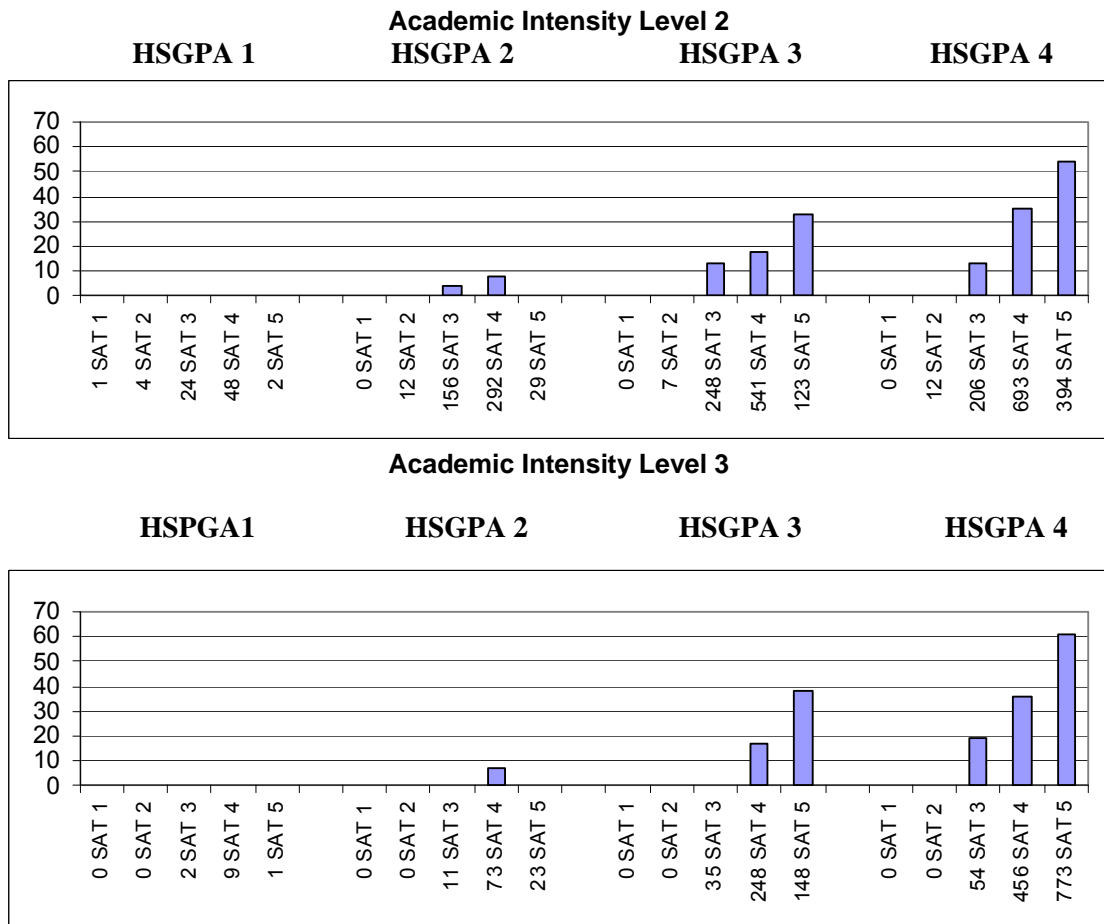
¹³ Academic intensity Level 3 is not included because of the small number of students in this category in Level 1 schools.



Note: The first number at the base of each bar indicates the sample size for that bar; the second number indicates the level of the predictor. Bars are omitted for sample sizes below 50. Source: Bridgeman, Pollack & Burton (2004) Figure 2 and Table A1.

Figure 12-1. Level 1 Colleges—Percentage of Freshmen With CGPAs of 2.5 or Higher by the First Two Levels of Academic Intensity, HSGPA, and SAT Score

SAT is not just related to performance as freshmen, but shows a strong relationship to achievement at the end of senior year. To illustrate this, we will examine students in colleges in selectivity Level 4 who achieved a four year CGPA of 3.5 or higher. Figure 12-2 omits academic intensity Level 1 because there are so few students with that kind of preparation in these colleges. There are also relatively few students in HSGPA Levels 1 and 2 among the students in academic intensity Levels 2 and 3. Even among these students at the most selective colleges who were in the highest two levels of academic intensity and who were in the top HSGPA category, the SAT differentiates college performance. Three times as many students in SAT Level 5 achieved the 3.5 criterion than did students in SAT Level 3.



Note: The first number at the base of each bar indicates the sample size for that bar; the second number indicates the level of the predictor. Bars are omitted for sample sizes below 50. Source: Bridgeman, Pollack & Burton (2004) Figure 6 and Table A2.

Figure 12-2. Level 4 Colleges—Percentage of Students With CGPAs of 3.5 or Higher by Highest Two Levels of Academic Intensity, HSGPA, and SAT Score

The Bridgeman, Pollack and Burton (2004) study demonstrates how the preadmission measures are related to academic performance in college at the end of the freshman year and at the end of the senior year. Although the authors acknowledge that other methods, such as ordinary least squares regression or logistic regression, may be more useful for making predictions about the likely success of individual students (p. 10), their data effectively demonstrate the practical importance of grades and SAT scores as indicators of which students are most likely to succeed in college.

12.10 Longer Term Performance

A major study of long term validity did not report correlation coefficients for individual preadmissions variables. Bowen and Bok (1998) analyzed data on the academic performance of 32,000 students who entered 28 selective undergraduate institutions in 1989. In addition to SAT scores and high school records, they included a set of control variables (gender, race/ethnicity, socioeconomic status, selectivity of the college attended, and major) in their study. They reported for the total set of variables a

correlation of 0.45 with cumulative college rank in class. When controlled for gender, race, socioeconomic status, college selectivity, major, and high school rank in class, a 100 point increase in combined SAT-V and SAT-M scores resulted in a 5.9 point increase in percentile rank in college. Bowen and Bok observed that “among both black and white students, those in the highest SAT interval had an appreciably higher average rank in class [based on cumulative four year GPA] than did those who entered with lower SAT scores” (p. 74).

Bowen and Bok (1998) also observed a mildly positive relationship between combined SAT-V and SAT-M scores and the rate of graduation. These data are shown in Table 12-8. When they adjusted the data for other variables, however, Bowen and Bok found that “above a threshold of 1100, SAT scores have a very limited role to play in explaining differences in graduation rate. The college or university that a student attends is a much better predictor of the odds of graduating than is the student’s own SAT score” (p. 65). For their sample of selective undergraduate institutions, they noted, “Most students who fail to graduate do not drop out because they were incapable of meeting academic requirements. They leave for many other reasons. Inability to do the academic work is often much less important than loss of motivation, dissatisfaction with campus life, changing career interests, family problems, financial difficulties, and poor health” (p. 55).

Table 12-8. Combined SAT Score and Actual Graduation Rate

<i>Combined SAT Score Range</i>	<i>Graduation Rate</i>
<1000	76%
1000–1099	82%
1100–1199	85%
1200–1299	86%
1300+	90%

Source: Bowen & Bok (1998), Figure 3.6

Burton and Ramist (2001) examined the long term validity of the SAT and other variables in predicting “success in college.” They examined the usefulness of preadmissions measures in predicting cumulative undergraduate grade averages as well as comparing the results for cumulative grades with the results for first year grades. They also examined studies that correlated admissions predictors with graduation from college. They aggregated data from 16 different studies involving a total of 30,000 students graduating since 1980 from 174 undergraduate institutions. The weighted average correlations (uncorrected) with cumulative undergraduate GPA were 0.40 for SAT-V, 0.41 for SAT-M, 0.42 for high school record, and 0.52 for the combination of SAT-V, SAT-M, and high school record. This pattern of the high school record having a slightly higher correlation with college performance than the test scores has been observed in many past studies. However, when highly selective institutions are studied, this pattern of higher correlations for high school record does not hold (Bridgeman, McCamley-Jenkins, & Ervin, 2000).

Burton and Ramist (2001) compared their aggregated data with earlier studies of the prediction of cumulative college performance and observed an increase in the predictive importance of SAT-M. Studies

reported by French (1957) show SAT-M correlations in the 0.2 range. Those reported by Wilson (1983) show SAT-M correlations in the 0.3 range, compared to the average of 0.4 in the more recent studies included in the Burton and Ramist analysis. The authors suggest that this trend may be explained by “the increased importance of quantitative areas in the college curriculum and the increased level of preparation in high school mathematics for virtually all SAT takers” (p. 9). Another possible explanation is that the SAT population is much more diverse now than in 1957; a more diverse group allows for higher correlations, especially if there is no correction for range restriction.

Burton and Ramist (2001) also analyzed eight studies that correlated admissions predictors with four-, five-, or six-year degree attainment in classes graduating between the 1980s and the mid 1990s. They found weighted average correlations for SAT-V, SAT-M, and high school record, singly and in combination, to range from 0.27 to 0.33. According to Burton and Ramist, the correlations with graduation are lower than the correlations with cumulative grade point averages due to the influence of nonacademic factors such as finances, motivation, social adjustment, family problems, or health.

12.11 Differential Validity for Subgroups

A considerable amount of research in the last decade and a half has examined the question of whether the SAT scores, as well as other predictors, have differential validity for various subgroups of the test taking population. In other words, is there a different relationship between the predictors and the criterion of college grades for men than for women, or among members of different racial or ethnic groups? Ramist, Lewis, and McCamley-Jenkins (1994) analyzed a database of course grades from 38 colleges and universities to determine if group differences occurred in the prediction of individual course grades as well as FGPA. This was the same database that was used in the earlier study by Ramist et al. (1990). A sample of over 46,000 students was used to investigate differences by gender and by five ethnic/racial groups (Native American, African American, Hispanic, Asian American, and White). The uncorrected and corrected correlations with FGPA and with a course grade criterion (adjusted for the grading difficulty of the courses) are shown in Table 12-9. Since the total sample for Native American students was only 184, results for this group should be considered tenuous at best.

The courses taken by these students in their first year of college were assigned to 37 categories based on subject, skills required, and level. For example, there were five categories for mathematics (based on level) and nine for English (based on level as well as whether the emphasis was on reading/literature, writing/composition, or both). Their results showed differences in course taking behavior for the different gender and ethnic/racial groups.

Table 12-9. Effectiveness by Student Group Correlation with FGPA

N	Gender			Ethnic Group				
	All Students	Male	Female	Native American	Asian American	African American	Hispanic	White
	46,379	22,412	23,967	184	3,848	2,475	1,599	36,743
Correlations* With FGPA								
SAT-V	0.50	0.48	0.55	0.42	0.47	0.44	0.39	0.50
SAT-M	0.53	0.53	0.58	0.36	0.56	0.44	0.38	0.52
SAT (V+M)	0.57	0.56	0.62	0.49	0.58	0.49	0.43	0.56
HSGPA	0.61	0.58	0.61	0.49	0.60	0.46	0.53	0.61
V+M+H	0.68	0.65	0.71	0.63	0.69	0.56	0.58	0.68
Correlations* With Course Grade Criterion								
SAT-V	0.50	0.48	0.53	0.39	0.49	0.47	0.44	0.49
SAT-M	0.54	0.53	0.57	0.32	0.59	0.48	0.48	0.53
SAT (V+M)	0.60	0.59	0.64	0.48	0.63	0.57	0.55	0.59
HSGPA	0.58	0.57	0.59	0.59	0.63	0.46	0.55	0.57
V+M+H	0.70	0.69	0.74	0.70	0.76	0.64	0.68	0.69

*Correlations corrected for restriction of range and criterion unreliability. Source: Ramist, Lewis, & McCamley-Jenkins (1994), Tables 1 and 4.

12.11.1 Gender

Drawn from the Ramist, Lewis, and McCamley-Jenkins (1994) study, Table 13-9 shows that the correlations between the predictor variables and both the FGPA and the course grade criteria were higher for females than for males, more so for the SAT than for HSGPA, and more so for the verbal score than for the mathematics score. For both criteria, the correlation of HSGPA exceeded the correlation of the combined SAT-V and SAT-M for males, but for females, the SAT showed a stronger correlation than did HSGPA. Using both HSGPA and SAT scores, the corrected correlation for predicting FGPA was higher for females (0.71) than for males (0.65), as was the corrected correlation for predicting course grade (0.74 versus 0.69).

In a 1994 report, Pennock-Román investigated gender differences in the prediction of college grades at four universities: two in California, one in Massachusetts, and one in Texas. As in the Ramist et al. (1994) study, Pennock-Román found that males were more likely to take courses in the physical sciences and engineering, while females were more likely to take courses in the humanities and social sciences.

Since it has been widely observed at many institutions that the average grade earned by students in courses varies considerably from department to department, one explanation for the underprediction of women’s grades is that this is due to differences in course selection. Because it is more common for women to enroll in courses where the average grade is higher than in the courses that men take, the underprediction of women’s grades may result from differences between men and women in the courses used to compute FGPA or CGPA. Pennock-Román (1994) sought to examine this hypothesis by developing and using a variable (MAJSCAL) that reflected the “degree of grading toughness” of the student’s category of college major. Separate prediction equations, by sex, of FGPA from SAT scores and HSGPA were used to calculate MAJSCAL. The average residual for the students who majored in a given department was used as an

indication of the “grading toughness” of that department. The magnitude of the residual for each department was then converted to the ordinal scale used for MAJSCAL.

The FGPA of women were underpredicted using all predictors (HSGPA, SAT verbal, SAT mathematics, or all three combined) at all four universities. This finding was also true for three subgroups of women (Asian American, White, and a combined group of African American and Latino students), with the exception of Asian American female students at the Texas university. For example, when all three predictors were used, the average underprediction of women’s grades ranged from 0.019 for Asian American females at one of the California schools to 0.185 for White females at the Texas institution. When MAJSCAL was used as an additional predictor, the underprediction of women’s FGPA was significantly reduced but not completely eliminated. This study provided further evidence that gender differences in the selection of college courses and majors may be the main reason behind the underprediction of women’s grades. The use of MAJSCAL, a measure that is a relatively easy to construct and understand, substantially reduced the degree of underprediction. In addition, by incorporating information on college majors through a measure such as MAJSCAL, a reasonable, practical procedure for controlling departmental grading differences may be available for use in future studies of differential prediction.

The recent study by Bridgeman, McCamley-Jenkins, and Ervin (2000) examined the impact of changes to the content and scale of the SAT on the predictive validity of the SAT overall as well as for subgroups of students. Results indicated that the correlations of SAT verbal, SAT mathematics, and SAT composite with FGPA, averaged across all the schools, were higher by 0.03 to 0.05 for women than for men. The average correlation of HSGPA with FGPA was slightly higher (by 0.02 to 0.03) for men than for women. When less selective institutions were analyzed separately, these correlations were found to be higher for females. Other studies of differential validity that have examined data from highly selective institutions have also found that gender differences in validity are often smaller than is the case at less selective institutions (Ramist et al., 1994).

The combination of SAT score and HSGPA was about equally effective in predicting FGPA for men (multiple correlation of 0.44) and for women (0.45). At the most selective institutions (with an average SAT composite score over 1250), the grades of men and women were predicted equally well. In contrast, at schools with lower average SAT scores, the grades of females were more predictable than the grades of males. As with other studies of differential prediction, Bridgeman et al. (2000) found that the grades of women were underpredicted from SAT scores alone (with an average underprediction of 0.11); from SAT scores and HSGPA (0.07); and from SAT scores, HSGPA, and an adjustment factor for course difficulty (0.05).

In Young and Kobrin’s review (2001) of the literature on differential validity and prediction with regard to gender differences, the correlations between predictors and criterion were generally higher for women than for men. In terms of prediction, the typical finding in these studies was that women’s college grades were underpredicted. However, in the most selective universities, the correlations for men and women appeared to be equal, and the degree of underprediction for women’s grades appeared to be noticeably less

than at other institutions. Compared with earlier studies on this topic, gender differences in validity and prediction appear to have persisted, although the magnitude of the differences seems to have recently decreased.

12.11.2 Race and Ethnicity

In the Ramist, Lewis, and McCamley-Jenkins (1994) study reported in Table 13-9, the highest correlation of SAT-V with FGPA was for White students (0.50) and the lowest was for Hispanic students (0.39). For SAT-M, the lowest correlation was for Native American students (0.36) and the highest was for Asian American students (0.56). This may reflect the fact that Asian American students took more quantitatively oriented courses than the other subgroups, a fact confirmed by Bridgeman, Pollack, and Burton's (in press) *Predicting Grades in Different Types of College Courses*. Asian American students had the highest multiple correlation for test scores combined with HSGPA (0.69), while African American students had the lowest (0.56). Results for predicting individual course grades were comparable to those for predicting FGPA, with the highest corrected correlations for the combination of SAT-V, SAT-M, and HSPGA for Asian American (0.76), Native American (0.70), and White (0.69) students. For four of the five ethnic groups, the combination of SAT-V and SAT-M scores was equal to or better than HSGPA in predicting course grades.

Both FGPA and course grades of Native American, African American, and Hispanic students were overpredicted; that is, they earned lower grades in college than was predicted, using any predictor, alone or in combination, while the grades of Asian American students were underpredicted. The magnitude of the overprediction was largest for Native American, followed by African American, and finally Hispanic students. Performance for Native American students was overpredicted in a variety of science, foreign language, English, and mathematics courses; African American student performance was overpredicted, especially in quantitative and science courses; Hispanic student performance was overpredicted in most courses. Course performance of Asian American students was underpredicted in mathematics and science but overpredicted in English, architecture, and physical education. The performance of White students was slightly underpredicted in English and overpredicted in mathematics and technical/vocational courses.

The Bridgeman, McCamley-Jenkins, and Ervin (2000) study found that correlations of SAT-V, SAT-M, and SAT composite with FGPA were uniformly higher for women than for men in the four subgroups studied (African American, Asian American, Hispanic, and White). However, the results for HSGPA were mixed, with some correlations higher for one gender or the other, depending on the ethnic/racial subgroup. The combination of SAT score and HSGPA appeared to be equally effective across all of the ethnic/racial subgroups and for men and women within each subgroup. The single exception to this finding was the somewhat lower multiple correlation for Hispanic men (0.38) as compared to Hispanic women (0.44).

The differential prediction findings indicated that, using SAT score and HSGPA, the grades of women from three of the subgroups were underpredicted. On average, the largest underprediction was for

White (0.09), then Asian American (0.07), and finally African American (0.05) women. The grades of Hispanic women were slightly overpredicted at 0.02. Adding the adjustment factor served to reduce the underprediction (or increase the overprediction) by 0.01 to 0.03. For men, the largest overprediction occurred in African American (0.16), followed by Hispanic (0.12), then White (0.09) students. The grades of Asian American men were accurately predicted. Adding the adjustment factor changed the overprediction only slightly for African American, Hispanic, and White men (by 0.02 or less) but caused the grades of Asian American men to be underpredicted by 0.05.

In 2001, Young and Kobrin produced a comprehensive review and analysis of all of the available differential validity and prediction studies published between 1974 and 2000. (See also Young [2004] for a further discussion of these differential validity and prediction studies.) In all, 29 studies of ethnic/racial differences and 37 studies of gender differences were reviewed. Young and Kobrin provided detailed information on each of the studies in the review, including type of study, name of institution(s), specific cohorts, sample sizes, predictors and criterion used, and values of validity coefficients and prediction results reported by each study's author(s). In addition, a short descriptive summary of each study was included. In another section of the report, Young summarized the findings from five earlier research reviews on differential validity and prediction (Breland, 1979; Duran, 1983; Linn, 1973; Linn, 1982; Wilson, 1983).

With regard to ethnic and racial differences, Young and Kobrin (2001) reported that the subgroups that have been studied include Asian American, African American, Hispanic, and Native American students. Some studies used a combined sample of minority students composed primarily of African American and Hispanic students. Overall, there was no common pattern to the results for validity and prediction for the different subgroups. Correlations between predictors and criterion were different for each subgroup, with generally lower values for African American and Hispanic students and similar values for Asian American students compared to White students. Too few studies of Native American or of combined samples of minority students were available to reliably determine typical validity coefficients for these groups. In terms of grade prediction, the common finding was one of overprediction of college grades for all minority groups with the exception of Asian American students, although the magnitude differed for each group. With Asian American students, studies that adjusted grades to account for differences in course difficulty found that grades were underpredicted.

12.11.3 Students With Disabilities

Increased attention to testing procedures for students with disabilities occurred in 1977 when the U.S. Department of Education issued regulations implementing Section 504 of the Rehabilitation Act of 1973. The regulations require individualized testing accommodations, validation of admissions tests for examinees with disabilities, and assurance that the tests are measuring aptitude and achievement without the impact of extraneous variables attributed to disability (Willingham, Ragosta, Bennett, Braun, Rock, & Powers, 1988). In response to Section 504, the College Board and the ETS sponsored a four year study that focused primarily on

students with different disabilities who had taken admissions testing program exams. Data on score reliability and validity did not show dependable differences in precision between students with disabilities and those without (Bennett, Ragosta, & Strickler, 1984). For most students with disabilities, the combination of high school grades and test scores remained the best predictor of college performance. Some exceptions noted were an underprediction of college freshman grades for deaf or hearing impaired students, an overprediction for students with specific learning disabilities, and a slight overprediction for students with physical disabilities.

Other studies investigated the validity of SAT scores for examinees with and without disabilities. Bennett, Rock, and Kaplan (1985) examined verbal and mathematics scores for groups of examinees (with and without disabilities) to discern whether SAT scores were comparable across individuals tested under standard administration procedures versus those tested under special administrations (including extended time). Findings suggested that SAT scores are generally equally reliable and valid for predicting the performance of students with and without disabilities. Similarly, Ragosta, Braun, and Kaplan (1991) tested the validity of SAT scores for predicting overall performance and persistence of college students with disabilities and found that scores were a good predictor of both variables.

Extended Time Accommodations

Students with specific learning disabilities comprise approximately 90% of examinees who request accommodations on the SAT (Camara & Schneider, 2000) and account for the largest percentage of college freshmen with disabilities (Cahalan, Mandinach, & Camara, 2002). In addition, extended time is the most often requested and granted accommodation on college admissions tests. As such, more recent studies have focused on students with specific learning disabilities who take the SAT with extended time to determine the impact that providing extra time may have on performance.

Providing extended time accommodations for SAT I examinees with documented disabilities is based on the notion that test timing is a primary source of noncomparability between test scores (i.e., certain disabilities may lead to slower processing of test content). Data from test administration timing records were used to establish empirically derived testing times for special administrations of the SAT for examinees with disabilities and to establish eligibility guidelines for individuals requesting special administrations (Ragosta & Wendler, 1992). This research established that comparable testing time for students with disabilities was between 1.5 and 2 times that for students without disabilities. These time limits assured that approximately equal percentages of students from both groups would complete each section of the SAT. An exception was students with visual impairments or blindness using braille or cassette versions of the test, who required between double and triple the normal time limits.

Camara, Copeland, and Rothschild (1998) examined the impact of extended time on SAT performance. They compared the mathematics and verbal section score gains for students who received an extended time accommodation and completed each SAT section in standard time (75 minutes), up to time and

a half (an additional 1 to 38 minutes), time and a half to double time (an additional 39 to 75 minutes), and greater than double time (an additional 76 or more minutes). Findings revealed that time and a half to double time produced the highest score gains on the mathematics section, and greater than double time produced the highest score gains on the verbal section.

In a study on the effects of taking the SAT I with extended time for students with specific learning disabilities, Camara and Schneider (2000) cited important conclusions about extended time administrations. One conclusion is that allowing students to retest using extended time does lead to SAT I score improvement, but the amount of improvement is modest. Average score gains with extended time are 32 points on the verbal scale and 26 points on the mathematics scale. Overall, there is a positive correlation between the amount of extended time allowed and the amount of score gain. While extended time does enable students with learning disabilities to perform better on the SAT I, the standard allowance of time and a half or double time may overcompensate for some students and result in overprediction of college performance. Finally, the study found that students who scored higher on their initial SAT I examination used more time in a subsequent administration and experienced larger score gains than their peers who received lower scores on the initial examination.

Cahalan, Mandinach, and Camara (2002) examined the predictive validity of scores from the SAT I for students who received special testing accommodations. Particularly, they were interested in students with specific learning disabilities who had taken the SAT I between 1995 and 1998 with an extended time accommodation. The study provided evidence that scores from the SAT I are a valid tool for helping admissions officers select students with specific learning disabilities (who received extended time accommodations) for college admission. While SAT scores alone are a good predictor of FGPA, the prediction is increased by using both SAT scores and HSGPA.

Morgan and Huff (2002) compared the reliability and dimensionality of the SAT I verbal and mathematics sections for examinees tested under standard timing conditions and examinees tested with extended time accommodations. Four comparisons were conducted between the standard time and extended time groups for May 2001 verbal and mathematics and October 2001 verbal and mathematics. Reliability and standard error of measurement estimates across the two groups of examinees differed slightly for all four comparisons, with the extended time group showing slightly more measurement error than the standard time group. Results from item-level factor analyses and multidimensional scaling analyses produced no evidence to suggest that the scores on the SAT I have different interpretations when the examinees have an extended time administration compared to the standard.

Lindstrom (2006) used data from the initial administration of the new SAT (administered March 17, 2005) to analyze a sample of 4,952 examinees. First, confirmatory factor analysis was used to assess the fit of a single-factor structure model for the mathematics, critical reading, and writing sections to each of the two groups. Next, a study of factorial invariance examined whether a common factor model for the mathematics, critical reading, and writing sections holds across the two groups at increasingly restrictive levels of

constraint. Invariance across the two groups was supported for factor loadings, thresholds, and factor variances. Thus, there was no real evidence to suggest that the scores on the mathematics, critical reading, and writing sections of the SAT have different interpretations when examinees have an extended time administration as opposed to the standard time administration.

12.11.4 Fatigue Effects

Cahalan-Laitusis, Morgan, Bridgeman, Zanna, and Stone (2007) examined operational data from the SAT to determine if students who tested under extended time conditions were suffering from excessive fatigue relative to students who tested under standard time conditions. Excessive fatigue was defined by significant increases in differential item functioning (DIF) and decreases in item completion rates, for items at the end of testing compared to the beginning of testing. Both of these factors were examined by comparing the performance of students who tested under standard time to students testing with extended time on items administered early in the test (Sections 2 or 3) and different items administered late (Sections 8, 9, or 10) during the 10-section test administration. Results indicated few changes in the level of DIF. In addition, item completion rates for students who received extra time were comparable to test takers without disabilities who tested under standard time on both early and late sections.

12.12 Summary of the MHSA SAT Component

This section began with a discussion of what is measured by the SAT. The substance of the test represents a complex interaction between the particular reading, mathematical, and writing skills; the content through which students are asked to demonstrate their skills; and the types of questions used to elicit that demonstration of skills. The test does not include esoterica but rather focuses on the application of content and skills that are part of a typical high school experience.

The second portion of the section reviewed evidence of the relationship of the substance of the test to what teachers judge to be important in each domain, and the intensity with which it is treated in the classroom. The third portion of the section reported on evidence demonstrating that the scores on the revised (2005) SAT can be interpreted in the same way as earlier scores and argued that the predictive validity evidence collected over past decades can be used to support the interpretation of the revised test.

The final portion of the section examined the relationship of SAT scores to performance in college, as measured by different criteria such as freshman GPA, four year cumulative GPA, college graduation, or performance in an English composition course. Research on the differential validity of the test by gender and racial/ethnic group was also presented.

Overall, there is a substantial body of evidence that supports the use of the SAT in the college admissions process. Even within homogeneous groups with similar high school preparation and grades attending a particular stratum of colleges, the SAT differentiates between those who are academically more successful and those who are less so. The SAT does not account for all the variation in college performance,

but it does provide a good indicator of how a student is likely to perform in the particular context of a college or university.

12.13 MHSA Mathematics and Science Component Validities

The MHSA consists of two additional sections besides the SAT. Because the interpretations of test scores, and not the test itself, are evaluated for validity, the purpose of reporting on the mathematics and science sections of the 2008–09 MHSA is to describe several of their technical aspects in support of score interpretations (AERA et al., 1999). Important components in the investigation of score validation include test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

As described in Chapter 2, the MHSA mathematics test is composed of SAT mathematics items plus additional “augmentation” items (the Math–A) that were necessary to adequately cover Maine *Learning Results* mathematics content standards. The Science test, as described in Chapters 4 and 5, was written and aligned in its entirety to *Maine’s Learning Results* science accountability standards. MHSA mathematics and science results are intended to facilitate inferences about student achievement on the mathematics and science standards, which in turn serve the evaluation of school accountability and inform the improvement of programs and instruction.

Standards for Educational and Psychological Testing (AERA et al., 1999) provides a framework for describing sources of evidence that should be considered when evaluating validity. These evidentiary sources include five general areas: test content, response processes, internal structure, consequences of testing, and relationship to other variables. Although each of these sources may speak to a different aspect of validity, they are not distinct types of validity. Instead, each contributes to a comprehensive body of evidence about the validity of score interpretations.

Evidence on test content validity is meant to help determine how well the test items represent the curriculum and standards for each content area and grade level. Content validation is informed by the item development process, including how test blueprints and test items align with the curriculum and standards. MHSA mathematics validation through the content lens was extensively described in Chapter 2, and science validation was extensively described in Chapters 4 and 5. Item alignment with grade level expectations; item bias, sensitivity, and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training are all components of validity evidence based on test content. As discussed earlier, all MHSA science test questions were aligned by Maine educators to state standards and underwent several rounds of review for content fidelity and appropriateness. Items were presented to students in multiple formats (constructed-response and multiple-choice). Finally, tests were administered according to state mandated standardized procedures with allowable accommodations.

Chapter 7 provided additional content validation evidence in describing mandated standardized testing procedures, including the requirement that all test coordinators and test administrators familiarize themselves with and adhere to the procedures outlined in the *Principal and Test Coordinator Manual* and *Test Administrator Manual*. The quality control procedures related to scanning and machine scoring, as well as the training and monitoring of readers, presented with the scoring information in Chapter 9 added to the body of content validation evidence.

Evidence on internal structure was extensively detailed in Chapter 11, where classical and item response theory statistics as well as the results of dimensionality analyses were presented, and overall and subgroup reliability coefficients, as well as conditional standard errors of measurement, were given. In general, the MHSA's Math-A and science item difficulty and discrimination indices were in acceptable and expected ranges. Very few items were answered correctly at near chance or near perfect rates. Similarly, the positive discrimination indices indicate that students who performed well on individual items tended to perform well overall. Chapter 11 also described the planned procedure for maintaining the MHSA score scale through future administrations of the combined SAT/Math-A test.

Evidence on the consequences of testing was addressed, to some extent, in the discussions of scaled scores in Chapters 10 and 11. Each of these speaks to efforts undertaken for providing the public with accurate and clear test score information. Scaled scores simplify results reporting across content areas, grade levels, and successive years. Achievement levels give reference points for mastery at each grade level—another useful and simple way to interpret scores. Several different standard reports were provided to stakeholders.

12.14 State Level Results

The state level results by achievement level and reporting subgroup are presented in Table 6.1. Given that the high school science test is new (i.e., standards have just been established) the results in Table 6.1 can be used as evidence of validity. In particular the results for science can be compared to the other content areas. The extent to which we see similar trends among the content areas suggests that the science assessment is functioning similarly to the operational critical reading, writing, and mathematics assessments. For example, in science we see that students classified in the gifted/talented programs perform much better compared to those not classified as such across content areas. Although this particular finding may not be of particular interest or very surprising, the reader is encouraged to make other comparisons to further evaluate where the science assessment is similar to the other content areas.

12.15 MHSA Validity Studies Agenda

The remainder of this discussion addresses further studies that could enhance the body of validation evidence for the MHSA. The proposed areas fall into four categories: external validity, convergent and discriminant validity, structural validity, and procedural validity. These will be discussed in turn.

External validity could be investigated by identifying additional variables for correlating with MHSA results. For example, data could be collected on the course grades of students who took the MHSA tests. Cross-tabulations of MHSA achievement levels with course grades, or average MHSA scaled scores with assigned grades (A, B, C, etc.), could be constructed. MHSA scores could also be correlated with appropriate classroom tests. Further evidence of external validity might come from correlating MHSA scores with scores on another standardized test, such as the Iowa Tests of Educational Development (ITED).

The concepts of convergent and discriminant validity were defined by Campbell and Fiske (1959) as specific types of validity that fall under the umbrella of construct validity. Convergent validity states that measures or variables that are intended to align should actually be aligned in practice. Discriminant validity, on the other hand, is the idea that measures or variables that are intended to differ should not be too highly correlated. Evidence for validity comes from examining whether the correlations among variables are as expected in direction and magnitude.

Campbell and Fiske (1959) introduced the study of different traits and methods as the means of assessing convergent and discriminant validity. Traits refer to the constructs that are being measured (e.g., mathematical ability), and methods are the instruments of measuring them (e.g., a mathematics test or course grade). To utilize the framework of Campbell and Fiske, it is necessary that more than one trait and more than one method be examined. Analysis is performed through the multitrait/multimethod matrix, which gives all possible correlations of the different combinations of traits and methods. Campbell and Fiske defined four properties of the multitrait/multimethod matrix that serve as evidence of convergent and discriminant validity.

- The correlation among different methods of measuring the same trait should be sufficiently different from zero. For example, scores on a mathematics test and grades in a mathematics class should be positively correlated.
- The correlation among different methods of measuring the same trait should be higher than that of different methods of measuring different traits. For example, scores on a science test and grades in a science class should be more highly correlated than scores on a science test and grades in a reading class.
- The correlation among different methods of measuring the same trait should be higher than the same method of measuring different traits. For example, scores on a mathematics test and grades in a mathematics class should be more highly correlated than scores on a mathematics test and scores on an analogous reading test.
- The pattern of correlations should be similar across comparisons of different traits and methods. For example, if the correlation between test scores in reading and writing is higher than the correlation between test scores in mathematics and reading, it is expected that the correlation between grades in reading and writing would also be higher than the correlation between grades in mathematics and reading.

For the MHSA, convergent and discriminant validity could be examined by constructing a multitrait/multimethod matrix and analyzing these four pieces of evidence. The traits examined would be

mathematics, reading, and writing; different methods could include MHSAs score and such variables as grades, teacher judgments, and/or scores on another standardized test.

Though the aspects of validity described above examine the concurrence among different measures of the same content area, structural validity focuses on the relation between strands within a content area. Standardized tests are carefully designed to ensure that all appropriate strands of a content area are adequately covered, and structural validity is the degree to which related elements of a test are correlated in the intended manner. For instance, it is desired that performance on different strands of a content area be positively correlated; however, as these strands are designed to measure distinct components of the content area, it is reasonable to expect that each strand would contribute a unique component to the test. Additionally, it is desired that the correlation between different item types (multiple-choice, student-produced-response) of the same content area be positive.

As an example, an analysis of MHSAs structural validity would investigate the correlation between performance in Numbers and Operations and performance in Patterns. Additionally, the concordance between performance on multiple-choice items and student-produced-response items would be examined. The dimensionality analyses of Chapter 11 could be expanded to further study a variety of issues surrounding the structural validity of the MHSAs program.

As mentioned earlier, the *MHSAs Test Coordinator* and *Test Administrator Manuals* delineated the procedures to which all MHSAs test coordinators and administrators were required to adhere. A study of procedural validity would provide a comprehensive documentation of the procedures that were actually followed throughout the MHSAs administration. The results of the documentation would then be compared to the manuals, and procedural validity would be confirmed to the extent that the two were in alignment. Evidence of procedural validity is important because it verifies that the actual administration practices are in accord with the intentions of the design.

Possible instances where discrepancies can exist between design and implementation include the following: a teacher may spiral test forms incorrectly within a classroom; cheating may occur among students; answer documents may be scanned incorrectly. These are examples of administration error. A study of procedural validity involves capturing any administration errors and presenting them within a cohesive document for review.

Chapter 13. MHPA SCORE REPORTING

All students who participate in the MHPA receive score reports that contain Maine specific scores on all components of the MHPA: SAT, Math–A (combined with SAT mathematics), and science. Those students who took the SAT under college reportable conditions (i.e., without Maine purposes only [MPO] accommodations) also received SAT score reports directly from the College Board.

13.1 Primary Reports

There were five primary reports for the 2008–09 MHPA.

- Student Report for Parents/Guardians
- Student results labels
- Class Analysis Reports
- School Report
- School Administrative Unit (SAU) Report

All reports were distributed to schools and SAUs via a secure Web site hosted by Measured Progress. In addition, printed copies of the student reports were produced for distribution to parents and guardians by schools. Printed student labels were also produced for use by schools. Each of these reports is described in the following subsections, and sample reports are provided in Appendix J.

13.2 Student Report for Parents/Guardians

The front side of the single-page Student Report includes a letter from the commissioner of education and the MDOE, a description of the achievement levels, and a graph showing state summary results. The second side provides a complete picture of an individual student’s performance on the MHPA, divided into two sections. The first section gives the student’s overall performance for each content area. The student’s scaled scores and achievement levels are shown, both in a table and graphically. The graph shows the range of possible scaled scores, divided up into the four achievement levels. This section also displays the standard error of measurement (SEM) bar for each content area.

The second section of the student report displays the student’s achievement level by content area relative to the percentage of students at each achievement level for the school, SAU, and state. For science only, student level data is displayed by content standard cluster as the number of points attained.

13.3 Student Labels

To aid schools in keeping track of student scores, schools were supplied with student score information on individual labels that they could affix to school files, if desired.

13.4 Class Analysis Report

The Class Analysis Report includes a roster of all the students in a school and indicates their performance on the Math–A and science common items in the assessment. The student names and MEDMS identification numbers are listed down the left side of the report, and the released items, comprising 50% of the common items, are listed across the top in the order in which they appear on the released item CD (not the position they appeared on the test). For each item, the following information is provided: the cluster level, the grade level or span expectation measured by the item, the item type, the correct response for multiple-choice items, and the total possible points for the item. For each student, each multiple-choice item is marked either with a plus sign (+), indicating that the student chose the correct response, or a letter (A–E for Math–A and A–D for science), indicating which incorrect response the student chose. For constructed-response science items, the number of points the student attained is shown. Also displayed are each student’s total points earned, scaled score, and achievement level, as well as the school, SAU, and state percentage correct for each item.

For the science Class Analysis Report, additional item level statistics are displayed that are currently not available for the SAT content area tests for Maine. These data include the number and percentage of points possible for each content standard cluster. Also displayed are the school, SAU, and state average points attained for each cluster by number and percentage. (This data is also displayed at the student level for science only on the Student Report for Parents/Guardians.)

13.5 School and SAU Reports

Prior to the release of the school and SAU reports to the secure Web site, each SAU office and school received a notification containing a user name and password allowing access to these reports. The school and SAU reports consist of three parts: The first part gives an overall summary of scores, the second provides a summary of student participation, and the third includes a report for each content area with scores by reporting subgroups.

The summary of scores includes a table that is designed to show, for each content area, the average scaled score for the school, SAU, and state for each of the last three years, as well as a cumulative average across the three years. (Note: Math–A was added in 2006–07 and mathematics standards reset in June 2007, so only two years of data is displayed. For science, only one year of data is displayed, as 2008 was the first year of this assessment.) In addition, a bar graph for each content area shows the percentage of students in each achievement level at the school, SAU, and state levels. For the SAU version of this report, the school information is blank.

The summary of student participation gives the number and percentage of students who participated at the school, SAU, and state levels for each content area. These numbers are provided for the overall group of students and broken down by the following categories:

- Ethnic group
- Identified disability
- LEP status
- Socioeconomic status
- Migrant status

These numbers are also provided for the overall groups of students as well as by the following modes:

- Students who took the assessment without accommodations
- Students who took the assessment with accommodations
- Students who took an alternate assessment
- Approved nonparticipation in reading for first year limited English proficient (LEP) students
- Approved nonparticipation for special considerations
- Nonparticipation for other reasons

For all three participation modes, data were captured for whether the student had an identified disability, LEP, or a 504 plan. Again, for the SAU version of this report, the school information is blank.

For each content area, there is a two page report showing results in more detail. The first page gives a definition of each of the achievement levels along with a table showing the number and percentage of students in the school, SAU, and state who scored at each level. The second page of the content area report breaks the results down by a number of different reporting categories: gender, ethnicity, socioeconomic status, Title 1 program, migrant status, gifted/talented, disability, and LEP status. This information is provided for the school, SAU, and the state on the school level report and for the SAU and the state on the SAU level report. To protect student confidentiality, results are displayed on this page only for groups with five or more students.

For each reporting category, the following information is given at the school or SAU level and at the state level:

- The percentage of students in that category
- The average scaled score for the group
- The percentage in the response category who exceeded, met, partially met, or did not meet the standard

These state level data by reporting category are displayed in Tables 13-1 and 13-2.

Table 13-1. 2008–09 MHSAs: State Achievement Results for Mathematics and Critical Reading

Reporting Categories		Mathematics					Critical Reading						
		Tested N	% E	% M	% P	% D	Mean Scaled Score	Tested N	% E	% M	% P	% D	Mean Scaled Score
All students		15,008	4	38	31	27	1141	14,660	9	40	28	22	1141
Ethnicity	African American/Black	315	1	15	29	56	1134	303	3	23	27	47	1133
	American Indian or Native Alaskan	106	1	20	31	48	1134	100	5	27	30	38	1135
	Asian or Pacific Islander	227	11	41	28	21	1144	219	11	34	28	26	1141
	Hispanic	157	1	27	25	46	1136	151	3	34	33	30	1137
	Caucasian/White	14,203	4	39	31	27	1141	13,887	9	41	28	21	1141
	Not reported	0						0					
Identified disability	Yes	1,959	0	7	19	73	1130	1,865	1	11	24	64	1127
	No	13,049	5	42	33	21	1142	12,795	10	45	29	16	1143
Current LEP	Yes	239	0	14	24	62	1132	225	0	9	22	68	1126
	No	14,769	4	38	31	27	1141	14,435	9	41	29	21	1141
Economically disadvantaged	Yes	4,306	1	24	33	42	1136	4,120	3	30	32	35	1136
	No	10,702	5	43	30	21	1142	10,540	11	44	27	17	1143
Migrant	Yes	4	0	25	25	50	1140	3	33	33	0	33	1147
	No	15,004	4	38	31	27	1141	14,657	9	40	28	22	1141
Gender	Female	7,248	3	38	33	27	1140	7,098	10	43	29	18	1142
	Male	7,760	5	38	29	28	1141	7,562	9	37	28	26	1140
	Not reported	0						0					
Title 1A program	Yes	293	1	23	37	39	1137	291	3	28	28	41	1135
	No	14,715	4	38	31	27	1141	14,369	9	40	28	22	1141
Gifted/talented program	Yes	521	31	63	4	2	1157	520	52	45	3	1	1161
	No	14,487	3	37	32	28	1140	14,140	8	40	29	23	1140

E = Exceeds the Standard; M = Meets the Standard; P = Partially Meets the Standard; D = Does Not Meet the Standard

**Table 13-2. 2008–09 MHSA:
State Achievement Results for Writing and Science**

Reporting Categories		Writing					Science						
		Tested N	% E	% M	% P	% D	Mean Scaled Score	Tested N	% E	% M	% P	% D	Mean Scaled Score
All students		14,663	7	39	31	23	1140	14,867	4	37	26	33	1140
Ethnicity	African American/Black	302	2	22	32	44	1133	311	1	18	20	61	1133
	American Indian or Native Alaskan	100	2	23	35	40	1134	102	1	19	30	50	1135
	Asian or Pacific Islander	219	10	37	27	26	1141	225	5	40	20	36	1141
	Hispanic	151	4	29	32	35	1135	152	2	23	18	57	1136
	Caucasian/White	13,891	7	40	31	23	1140	14,077	4	37	26	32	1141
	Not reported	0						0					
Identified disability	Yes	1,861	0	8	21	71	1125	1,928	0	9	18	72	1131
	No	12,802	8	43	32	16	1142	12,939	5	41	27	28	1142
Current LEP	Yes	224	0	8	28	64	1127	234	0	10	11	79	1129
	No	14,439	7	39	31	23	1140	14,633	4	37	26	33	1140
Economically disadvantaged	Yes	4,121	2	27	33	38	1134	4,264	2	24	26	47	1136
	No	10,542	9	44	30	18	1142	10,603	5	41	26	28	1142
Migrant	Yes	3	33	33	0	33	1143	4	25	25	0	50	1143
	No	14,660	7	39	31	23	1140	14,863	4	37	26	33	1140
Gender	Female	7,103	9	43	31	17	1143	7,179	2	32	29	37	1139
	Male	7,560	6	35	30	30	1138	7,688	6	40	23	30	1142
	Not reported	0						0					
Title 1A program	Yes	291	3	25	36	35	1135	287	2	23	26	49	1136
	No	14,372	7	39	30	23	1140	14,580	4	37	26	33	1140
Gifted/talented program	Yes	520	43	52	3	1	1159	517	28	65	6	1	1156
	No	14,143	6	38	32	24	1139	14,350	3	35	27	35	1140

E = Exceeds the Standard; M = Meets the Standard; P = Partially Meets the Standard; D = Does Not Meet the Standard

13.6 Decision Rules

To ensure that reported results for the 2008–09 MHSAs are accurate relative to collected data and other pertinent information, a document that delineates analysis and reporting rules was created. These decision rules were observed in the analyses of MHSAs test data and in reporting the assessment results. Moreover, these rules are the main reference for quality assurance checks.

The decision rules document used for reporting results of the May 2009 administration of the MHSAs can be found in Appendix K.

The first set of rules pertains to general issues in reporting scores. Each issue is described and pertinent variables are identified. The actual rules applied are described by the way they impact analyses and aggregations and by their specific impact on each of the reports. The general rules are further grouped into issues pertaining to test items, school type, student exclusions, and number of students for aggregations.

The second set of rules pertains to reporting student participation. These rules describe which students were counted and reported for each subgroup in the student participation report.

13.7 Quality Assurance

Quality assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician working on the MHSAs implement quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within Psychometrics and Research, the sending function verifies that the data are accurate prior to handoff. Additionally, when a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality assurance measure is parallel processing. Students' scaled scores for each content area are assigned by the psychometrician through a process of equating and scaling. The scaled scores are also computed by the data analyst to verify that scaled scores and corresponding achievement levels are assigned accurately. Respective scaled scores and achievement levels are compared across all students for 100% agreement. Different exclusions assigned to students that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each content area, the exclusions assigned by each data analyst are compared across all students. Only when 100% agreement is achieved can the rest of data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality assurance group to check the veracity and accuracy of reported data. Using a sample of schools and SAUs, the quality assurance group verifies that reported information is correct. The step is conducted in two parts: (1) verify that the computed information was obtained correctly through the appropriate application of different decision rules, and (2) verify that the correct data points populate each cell in the MHSAs reports. The selection of

sample schools and SAUs for this purpose is very specific and can affect the success of the quality control efforts. There are three sets of samples selected that may not be mutually exclusive. The first set includes those that satisfy the following criteria:

- One-school SAU
- Two-school SAU
- Multi-school SAU

If reporting includes class level reports, then the set also includes the following:

- Multi-class school, multi-school SAU
- One-class school, one-school SAU
- Multi-class school, one-school SAU
- One-class school, multi-school SAU
- Private school
- Special school (e.g., the “Big 11”)
- Small school that receives no School Report
- Small SAU that receives no SAU Report
- SAU that receives a report, but all schools are too small to receive a School Report
- School with excluded (not tested) students
- School with home schooled students

The second set of samples includes SAUs or schools that have unique reporting situations as indicated by decision rules. This set is necessary in order to check that each rule is applied correctly. The third set includes SAUs and schools identified by the MDOE for its review and approval before reports are produced for distribution.

The quality assurance group uses a checklist to implement its procedures. Once the checklist is completed, sample reports are circulated for psychometric checks and program management review. The appropriate sample reports are then sent to the MDOE for review and signoff. Once the MDOE gives the approval to proceed, the reports are posted to Measured Progress’s Web site for school and SAU access. Prior to public release, schools and SAUs have a two week review period in which to examine their results and, if necessary, to report any data issues.

REFERENCES

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.
- Angoff, W. H. (Ed.). (1971). *The College Board admissions testing program: A technical report on research and development activities related to the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.
- Baker, F. B. & Kim, S-H. (Eds.). (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.
- Bennett, R. E., Ragosta, M., & Strickler, L. (1984). *The test performance of handicapped people* (Report No. 84-32). Princeton, NJ: Educational Testing Service.
- Bennett, R. E., Rock, D. A., & Kaplan, B. A. (1985). *The psychometric characteristics of the SAT for nine handicapped groups* (ETS Research Report RR-85-49). Princeton, NJ: Educational Testing Service.
- Bowen, W. G., & Bok, D. (1998). *The shape of the river: Long-term consequences of considering race in college and university admissions*. Princeton, NJ: Princeton University Press.
- Breland, H. M. (1979). *Population validity and college entrance measures* (Research Monograph No. 8). New York: The College Board.
- Breland, H., Bridgeman, B., & Fowles, M. (1999). *Writing assessment in admission to higher education: Review and framework* (CBR No. 99-3). New York: The College Board.
- Breland, H. M., Kubota, M. Y., & Bonner, M. W. (1999). *The performance assessment study in writing: Analysis of the SAT II: Writing Subject Test* (CBR 99-4). New York: The College Board.
- Bridgeman, B., McCamley-Jenkins, L., & Ervin, N. (2000). *Predictions of freshman grade-point average from the revised and recentered SAT I: Reasoning Test* (CBRR 2000-1). New York: The College Board.
- Bridgeman, B., Pollack, J., & Burton, N. (2004). *Understanding what SAT Reasoning Test scores add to high school grades: A straightforward approach* (CBRR 2004-4). New York: The College Board.
- Bridgeman, B., Pollack, J., & Burton, N. (in press). *Predicting grades in different types of college courses*. Princeton, NJ: Educational Testing Service.
- Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd ed.). Fort Worth: Holt, Rinehart and Winston.
- Burton, N. W., & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980* (CBRR 2001-2). New York: The College Board.
- Burton, N., Welsh, C., Kostin, I., & Van Essen, T. (2004). *Toward a definition of verbal reasoning in higher education*. Unpublished manuscript.
- Cahalan, C. (2000). Geographic clusters of learning disabled test takers in the United States. Paper presented at the meeting of the American Educational Research Association, New Orleans, LA. (ERIC Document Reproduction Service No. ED 443 841).

- Cahalan, C., Mandinach, E. B., & Camara, W. J. (2002). *Predictive validity of SAT I: Reasoning Test for examinees with learning disabilities and extended time accommodations* (CBRR No. 2002-5). New York: College Entrance Examination Board.
- Cahalan-Laitusis, C., Morgan, D. L., Bridgeman, B., Zanna, J., & Stone, E. (2007). *Examination of fatigue effects from extended time accommodations on the SAT Reasoning Test* (CBRR 2007-1). New York: The College Board.
- Camara, W. J., Copeland, T., & Rothschild, B. (1998). *Effects of extended time on the SAT I: Reasoning Test score growth for students with disabilities* (CBRR No. 98-7). New York: College Entrance Examination Board.
- Camara, W. J., & Schneider, D. (2000). *Testing with extended time on the SAT I: Effects for students with learning disabilities* (College Board Research Note No. RN-08). New York: College Entrance Examination Board.
- Campbell, D. T., and Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81–105.
- Charter, R. A. (1999). Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559–566(8).
- College Board. (2004). *SAT preparation booklet 2004–2005 for the new SAT*. New York: Author.
- College Board. (2005a). *2005 college bound seniors: Total group profile report*. New York: Author.
- College Board. (2005b). *The new SAT: Implemented for the class of 2006*. Retrieved January 21, 2005, from www.collegeboard.com.
- College Board. (2005c). *The new SAT: A guide for admission officers*. New York: Author.
- College Board. (2005d). *Report for the State of Maine on the alignment of the SAT and PSAT/NMSQT to the Maine Learning Results*. Internal report provided to the Maine Department of Education in September 2005.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.
- Crone, C. R., & Schmitt, A. P. (1991). *Alternative verbal aptitude item types: DIF issues*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.
- Donlan, T. F. (Ed.). (1984). *The College Board technical handbook for the Scholastic Aptitude Test and Achievement Tests*. New York: College Entrance Examination Board.
- Dorans, N. J. (2000). *Distinctions among classes of linkages* (College Board Research Note RN-11). New York: The College Board.
- Dorans, N. J. (2002). Recentering and realigning the SAT score distributions: How and why. *Journal of Educational Measurement*, 39(1), 55–84.
- Dorans, N. J. (2004a). Equating, concordance and expectation. *Applied Psychological Measurement*, 28(4), 227–246.
- Dorans, N. J. (2004b). Using subpopulation invariance to assess test score equity. *Journal of Educational Measurement*, 41(1), 43–68.

- Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P. W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Dorans, N. J., Liu, J., & Hammond, S. (in press). The role of the anchor test in achieving population invariance across subpopulations and test administrations. *Applied Psychological Measurement*.
- Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equitability of tests: Basic theory and the linear case. *Journal of Educational Measurement*, 37(4), 281–306.
- Dorans, N. J., and Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23, 355–368.
- Duran, R. P. (1983). *Hispanics' education and background: Predictors of college achievement*. New York: The College Board.
- Dwyer, C. A., Gallagher, A., Levin, J., & Morley, M. (2003). *What is quantitative reasoning? Defining the construct for assessment purposes* (RR-03-30). Princeton, NJ: Educational Testing Service.
- French, J. W. (1957). *Validation of the SAT and new item types against four-year academic criteria* (RB-57-4). Princeton, NJ: Educational Testing Service.
- Gulliksen, H. (1950). *Theory of mental tests*. New Jersey: Erlbaum.
- Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer Academic Publishers.
- Hezlett, S. A., Kuncel, N. R., Vey, M., Ahart, A. M., Ones, D. S., Campbell, J. P., & Camara, W. (2001). *The effectiveness of the SAT in predicting success early and late in college: A meta-analysis*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle, WA.
- Holland, P. W., and Thayer, D. T. (1988) Differential item performance and the Mantel-Haenszel procedure. In H. Wainer and H. Braun (Eds.) *Test validity*, (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Joint Committee on Testing Practices (1988). *Code of Fair Testing Practices in Education*. Washington, D.C.: National Council on Measurement in Education.
- Khaliq, S., & Reshetar, R. (2003). *Summary of testing years 1998/1999 through 2002/2003 DIF statistics for the SAT* (Research memorandum). Princeton, NJ: Educational Testing Service.
- Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design*. London: Methuen.
- Kobrin, J. L., Camara, W. J., & Milewski, G. B. (2002). *The utility of the SAT I and SAT II for admissions decisions in California and the nation* (CBRR 2002-6). New York: The College Board.
- Kobrin, J. L., & Michel, R. S. (2006). *The SAT as a predictor of different levels of college performance* (CBRR 2006-3). New York: The College Board.
- Kobrin, J. L. & Schmidt, A. E. (2005). *The research behind the new SAT* (Research Summary RS-11). New York: The College Board.
- Kolen, M. J. (1990). Does matching in equating work? A discussion. *Applied Measurement in Education*, 3, 97–104.

- Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer-Verlag.
- Lawrence, I. D., Lyu, C. F., & Feigenbaum, M. D. (1995). *DIF data on free response SAT I mathematical items*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.
- Lawrence, I., Rigol, G., Van Essen, T. & Jackson, C. (2002). *A historical perspective on the SAT 1926–2001* (CBRR 2002-7). New York: The College Board.
- Lawrence, I. D., & Schmitt, A. (1994). Setting statistical specifications for the new SAT and PSAT/NMSQT. In Lawrence et al. (Eds.) *Technical issues related to the introduction of the new SAT and PSAT/NMSQT* (RM 94-10). Princeton, NJ: Educational Testing Service, 1–25.
- Lindstrom, J. H. (2006). *The role of extended time on the SAT Reasoning Test for students with disabilities*. Unpublished research report completed as part of the College Board Student Grant Fellowships Program.
- Linn, R. L. (1973). Fair test use in selection. *Review of Educational Research*, 43, 139–161.
- Linn, R. L. (1982). Ability testing: Individual differences, prediction and differential prediction. In A.K. Wigdor & W.R. Garner (Eds.), *Ability testing: Uses, consequences, and controversies (Part 2, pp.335-388)*. Washington, DC: National Academy Press.
- Liu, J. (2004). *Examination of long leg and short leg equatings for SAT verbal and math by administration for the 2002–03 testing year* (Research memorandum). Princeton, NJ: Educational Testing Service.
- Liu, J., Cahn, M. F., & Dorans, N. J. (2006). An application of score equity assessment: Invariance of linkage of new SAT[®] to old SAT across gender groups. *Journal of Educational Measurement*, 43(2), 113–129.
- Liu, J., Feigenbaum, M., & Cook, L. (2004). *A simulation study to explore configuring the SAT[®] I: Verbal Test without analogy items* (College Board Research Report No. 2004-2, ETS Research Report RR-04-01). Princeton, NJ: Educational Testing Service.
- Liu, J., Feigenbaum, M. D., & Dorans, N. J. (2003). *Equitability analysis of the new SAT to the current SAT I* (Statistical Report 2003-73). Princeton, NJ: Educational Testing Service.
- Liu, J., Feigenbaum, M., & Walker, M. E. (2004). *New SAT and PSAT/NMSQT spring 2003 field trial design* (Statistical Report 2004-95). Princeton, NJ: Educational Testing Service.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179–198.
- Lord, F. M. (1980). *Application of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.
- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Mantel, N., and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.
- Mathematical Sciences Education Board. (1989). *Everybody counts: A report to the nation on the future of mathematics education*. Washington, DC: National Academy Press.

- Milewski, G., Johnsen, D., Glazer, N., & Kubota, M. (2005). *A survey to evaluate the alignment of the new SAT writing and critical reading sections to curricula and instructional practices* (RR 2005-1). New York: The College Board.
- Morgan, D. L., & Huff, K. (2002). *Reliability and dimensionality of the SAT for examinees tested under standard timing conditions and examinees tested with extended time*. Unpublished research conducted at the Educational Testing Service documented in a memorandum on July 15, 2002.
- Morgan, R. (1994). *Effect of scale choice on predictive validity*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.
- Norris, D., Oppler, S., Kuang, D., Day, R., & Adams, K. (2006). *The College Board SAT writing validation study: An assessment of predictive and incremental validity* (CBRR 2006-2). New York: The College Board.
- Oh, H., & Sathy, V. (2006). *Construct comparability and continuity in the SAT* (Statistical Report SR-2006-22). Princeton, NJ: Educational Testing Service.
- Pennock-Román, M. (1994). *College major and gender differences in the prediction of college grades* (CBR 94-2). New York: The College Board.
- Powers, D., & Dwyer, C. A. (2003). *Toward specifying a construct of reasoning* (RM-03-01). Princeton, NJ: Educational Testing Service.
- Ragosta, M., Braun, H., & Kaplan, B. (1991). *Performance and persistence: A validity study of the SAT for students with disabilities* (College Board Report No. 91-3, ETS Research Report No. 91-41). New York: College Entrance Examination Board.
- Ragosta, M., & Wendler, C. (1992). *Eligibility issues and comparable time limits for disabled and nondisabled SAT examinees* (College Board Research Report No. 92-5, ETS Research Report RR-92-35). New York: College Entrance Examination Board.
- Ramist, L., Lewis, C., & McCamley, L. (1990). Implications of using freshman GPA as the criterion for the predictive validity of the SAT. In W. W. Willingham et al. (Eds.) *Predicting college grades: An analysis of trends over two decades* (pp. 253–288). Princeton, NJ: Educational Testing Service.
- Ramist, L., Lewis, C., & McCamley-Jenkins, L. (1994). *Student group differences in predicting college grades: Sex, language, and ethnic groups* (CBR 93-1). New York: The College Board.
- Rozeboom, W. W. (1978). Estimation of cross-validated multiple correlation: A clarification. *Psychological Bulletin*, 85(6), 1348–1351.
- Samejima, F. (1997). Graded response model. In Van Linden, W. J. & Hambleton, R. K. (Eds.) *Handbook of modern item response theory* (pp. 85-100). New York: Springer-Verlag.
- Silver, E. A., Kilpatrick, J., & Schlesinger, B. (1990). *Thinking through mathematics: Fostering inquiry and communication in mathematics classrooms*. New York: The College Board.
- Steen, L. A. (Ed.). (1997). *Why numbers count: Quantitative literacy for tomorrow's America*. New York: The College Board.
- Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika*, 52, 589–617.

- Stout, W. F., Froelich, A. G., & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duign, & T. A. B. Snijders (Eds.) *Essays on item response theory* (pp. 357–375). New York: Springer-Verlag.
- Swineford, F. (1974). *The test analysis manual* (SR-74-06). Princeton, NJ: Educational Testing Service
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved March 25, 2003, from www.education.umn.edu/NCEO/OnlinePubs/Synthesis44.html.
- Thurstone, L. L. (1947). The calibration of test items. *American Psychologist*, 2, pp. 103–104.
- Walker, M. E. (2003). *Scaling issues associated with the SAT I: Writing Test* (Statistical Report SR-2003-12). Princeton, NJ: Educational Testing Service.
- Walker, M. E. (2005). *Evaluation of decision tree items for March 2005 writing section scaling*. Unpublished manuscript. Princeton, NJ: Educational Testing Service.
- Walker, M. E., & Allspach, J. R. (2005). Scaling the SAT writing section. Presentation at College Board offices for College Board staff members, New York, NY.
- Walker, M. E., Allspach, J., & Liu, J. (2004). *Scaling the new SAT[®] writing section: Finding the best solution* (Statistical Report 2004-61). Princeton, NJ: Educational Testing Service.
- Walker, M. E., & Liu, J. (2003). *Scaling the new SAT writing test: Evidence from the 2003 field trial* (Statistical Report SR-2003-94). Princeton, NJ: Educational Testing Service.
- Walker, M. E., & Liu, J. (2004). *Scaling issues associated with the new SAT writing test*. Paper presented at the annual meeting of the National Council on Measurement in Education, April 13–15, 2004, San Diego, CA.
- Walker, M. E., Liu, J., & Allspach, J. R. (2005). *Scaling tests via nonlinear post-stratification methods*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.
- Wang, X. B. (2006). *Investigating the effect of new SAT test lengths on the performance of regular SAT examinees* (CBRR 2006-9). New York: The College Board.
- Webb, N. L. (2006a). *Alignment analysis of secondary language arts standards and the SAT Reasoning Test: Maine*. External report provided to the Maine Department of Education on April 10, 2006.
- Webb, N. L. (2006b). *Alignment analysis of secondary mathematics standards and the SAT Reasoning Test: Maine*. External report provided to the Maine Department of Education on April 10, 2006.
- Willingham, W. W., Lewis, C., Morgan, R., & Ramist, L. (1990). *Predicting college grades: An analysis of institutional trends over two decades*. Princeton, NJ: Educational Testing Service.
- Willingham, W. W., Ragosta, M., Bennett, R. E., Braun, H., Rock, D. A., & Powers, D. E. (1988). *Testing handicapped people*. Boston, MA: Allyn and Bacon.
- Wilson, K. M. (1983). *A review of research on the prediction of academic performance after the freshman year* (CBRR 83-2). New York: The College Board.

- Young, J. W. (2004). Differential validity and prediction: Race and sex differences in college admissions testing. In R. Zwick (Ed.) *Rethinking the SAT: The future of standardized testing in university admissions* (pp. 289–301). New York: Routledge/Falmer.
- Young, J. W., with Kobrin, J. L. (2001). *Differential validity, differential prediction, and college admission testing: A comprehensive review and analysis* (CBRR 2001-6). New York: The College Board.
- Zhang, J., & Stout, W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213–249.

APPENDICES

