

MeCAS
Maine's Comprehensive
Assessment System

**Maine High School
Assessment
MeCAS Part II
2009–10 TECHNICAL REPORT**



TABLE OF CONTENTS

CHAPTER 1. OVERVIEW OF THE MAINE HIGH SCHOOL ASSESSMENT	1
1.1 Purpose of the Assessment System.....	1
1.2 Purpose of this Report	2
CHAPTER 2. TEST DESIGN AND DEVELOPMENT OF THE MHSA: SAT	5
2.1 The MHSA: The SAT Overview	5
2.2 Universal Design Specifications.....	5
2.3 SAT Critical Reading Test	6
2.4 SAT Writing Test	8
2.5 MHSA Mathematics Test: SAT.....	13
2.6 Development.....	16
2.7 Item Writing and Review	16
2.8 Pretesting the Items	17
2.9 Analysis of Pretest Information for the MHSA: SAT	17
2.10 Item Difficulty	18
2.11 Item Discrimination and Item/Test Relationship	19
2.12 Differential Item Functioning	20
2.13 Evaluating Essay Pretests	22
2.14 Assembling the SAT Portion of the MHSA.....	22
2.15 Reviewing the MHSA: SAT Component.....	23
2.16 Test Production for the SAT Component	23
2.17 After the SAT Administration	23
2.18 Public Access to the SAT	24
2.19 Alignment of the SAT to the NECAP Standards	24
2.19.1 Design of SAT Critical Reading	25
2.19.2 Design of SAT Mathematics	26
CHAPTER 3. TEST DESIGN AND DEVELOPMENT OF THE MHSA: SCIENCE	27
3.1 Test Specifications.....	27
3.1.1 Criterion Referenced Test	27
3.1.2 Item Types	27
3.1.3 Description of Test Design	27
3.2 Science Test Specifications	27
3.2.1 Standards.....	27
3.2.2 Items Types	28
3.2.3 Test Design	29
3.2.4 Blueprints.....	29
3.2.5 Depth of Knowledge	29
3.2.6 Use of Calculators and Reference Sheets.....	30
3.3 Test Development Process.....	30
3.3.1 Item Development.....	30
3.3.2 Item Reviews at Measured Progress	30
3.3.3 Item Reviews at State Level.....	31
3.3.4 Bias and Sensitivity Review.....	31
3.3.5 External Expert Review	31
3.3.6 Reviewing and Refining.....	32
3.3.7 Item Editing	32
3.3.8 Item Selection and Operational Test Assembly	32
3.3.9 Operational Test Draft Review	33
3.3.10 Alternative Presentations.....	33
CHAPTER 4. TEST ADMINISTRATION: SAT	35
4.1 Preparation.....	35
4.2 Supervision.....	35
4.3 Physical Setting	36
4.4 Security.....	36
4.5 Calculator Policy for the SAT	37
4.6 Item Types.....	37
4.7 Instructions and Timing.....	38

4.8	Complaints and Irregularities	39
4.9	Subgroup Performance	39
4.10	Accommodations for Students on the MHSA.....	39
4.10.1	Process and Standards for College Board–Approved Accommodations.....	40
4.10.2	Process and Standards for MPO Accommodations.....	41
4.10.3	Eligibility Process Additions to Incorporate MPO Accommodations.....	41
4.10.4	Accommodation Eligibility Form Submission Time Lines.....	42
4.10.5	Training and Technical Assistance.....	42
4.10.6	MHSA Accommodation Request and Approval Statistics	43
4.11	Participation.....	44
CHAPTER 5.	TEST ADMINISTRATION: SCIENCE.....	45
5.1	Responsibility for Administration	45
5.2	Participation Requirements and Documentation	45
5.3	Test Security.....	46
5.4	Test and Administration Irregularities.....	46
5.5	Test Administration Window	46
5.6	Service Center	47
CHAPTER 6.	SCORING: SAT.....	49
6.1	Receiving and Opening	49
6.2	Scanning and Editing.....	49
6.3	Matching.....	50
6.4	Machine-Scored Portions	51
6.5	Scoring the Essay	51
6.6	End-to-End Quality Control	57
6.7	Quality Assessments.....	57
6.8	Summary	58
CHAPTER 7.	SCORING: SCIENCE	59
7.1	Machine-Scored Items	59
7.2	Person-Scored Items	59
7.2.1	Scoring Location and Staff	59
7.2.2	Benchmarking Meetings	60
7.2.3	Reader Recruitment and Qualifications	61
7.2.4	Methodology for Scoring Polytomous Items	61
7.2.5	Reader Training	62
7.2.6	Leadership Training	64
7.2.7	Monitoring of Scoring Quality Control.....	64
7.2.8	Reports Generated During Scoring	67
CHAPTER 8.	PSYCHOMETRIC TOPICS: SAT.....	71
8.1	The Equating and Braiding Plan for SAT Mathematics, Critical Reading, and Writing.....	71
8.2	SAT Statistical Characteristics	71
8.3	Reliability and Standard Errors of Measurement.....	72
8.3.1	Reliability.....	72
8.3.2	Standard Errors of Measurement	72
8.4	Classification Accuracy and Consistency of MHSA: SAT Cut Scores	75
8.4.1	Accuracy and Consistency	76
8.5	Completion Rates	79
8.6	Item Statistics	80
8.6.1	Item Difficulty: Equated Delta.....	80
8.6.2	Item Discriminating Power: Biserial Correlation.....	82
8.7	Differential Item Functioning (DIF).....	83
8.8	Summary	85
CHAPTER 9.	PSYCHOMETRIC TOPICS: SCIENCE TEST.....	86
9.1	Classical Item Analysis	86
9.1.1	Classical Difficulty and Discrimination Indices	86
9.1.2	Differential Item Functioning	88
9.1.3	Dimensionality Analysis	89
9.2	IRT Scaling and Equating.....	92
9.2.1	Item Response Theory	92
9.2.2	Item Response Results	93

9.2.3	Achievement Standards.....	94
9.2.4	Scaled Scores	94
9.3	Reliability	96
9.3.4	Reliability and Standard Errors of Measurement	98
9.3.5	Subgroup Reliability	98
9.3.6	Subcategory Reliability	98
9.3.7	Interrater Consistency	99
9.3.8	Reliability of Achievement-Level Categorization	99
CHAPTER 10.	MHSA SCORE REPORTING.....	104
10.1	Primary Reports	104
10.2	Individual Student Report for Parents/Guardians	104
10.3	Student Labels	105
10.4	Interactive Reporting	105
10.4.1	Item Analysis Report.....	105
10.4.2	Achievement Level Summary	106
10.4.3	Item Analysis Data.....	106
10.4.4	Longitudinal Data Report.....	106
10.5	School and SAU Reports	106
10.6	Decision Rules	108
10.7	Quality Assurance.....	108
CHAPTER 11.	VALIDITY RESEARCH ON THE MHSA SAT COMPONENT	112
11.1	Construct Validity.....	112
11.2	Verbal Reasoning.....	113
11.3	Quantitative Reasoning.....	114
11.4	Writing.....	116
11.5	Multiple-Choice Questions	116
11.6	Essay Question.....	117
11.7	Does the Length of the SAT Result in a Fatigue Effect?	117
11.8	How Do SAT Scores Relate to College Performance?	118
11.9	Performance Over Multiple Time Periods	122
11.10	Longer-Term Performance.....	126
11.11	Differential Validity for Subgroups	128
11.11.1	Gender	129
11.11.2	Race and Ethnicity	131
11.11.3	Students with Disabilities	132
11.11.4	Fatigue Effects	135
11.12	Summary of the MHSA SAT Component	135
CHAPTER 12.	VALIDITY OF THE MHSA SCIENCE COMPONENT	138
12.1	Questionnaire Data	139
12.1.1	Self-image	139
12.1.2	Attitude Towards Content Area	140
12.1.3	Match of Questions to What Is Learned in School.....	141
12.1.4	Difficulty of Assessment.....	141
REFERENCES	144
APPENDICES	152
APPENDIX A	2009–10 COMMITTEE MEMBERS	
APPENDIX B	MHSA READING ALIGNMENT REPORT	
APPENDIX C	MHSA MATHEMATICS ALIGNMENT REPORT	
APPENDIX D	POLICIES AND PROCEDURES	
APPENDIX E	PARTICIPATION RATES (SCIENCE)	
APPENDIX F	2010 NATIONAL TABLES	
APPENDIX G	INTERPRETING SCORES ON THE SAT	
APPENDIX H	ITEM-LEVEL CLASSICAL STATISTICS (SCIENCE)	
APPENDIX I	ITEM-LEVEL SCORE POINT DISTRIBUTIONS (SCIENCE)	
APPENDIX J	DIF RESULTS (SCIENCE)	
APPENDIX K	ITEM RESPONSE THEORY PARAMETERS (SCIENCE)	
APPENDIX L	TCCS AND TIFS (SCIENCE)	

APPENDIX M	LOOKUP TABLES
APPENDIX N	SCORE DISTRIBUTIONS
APPENDIX O	RELIABILITY
APPENDIX P	INTERRATER AGREEMENT (SCIENCE)
APPENDIX Q	SAMPLE REPORTS
APPENDIX R	INTERACTIVE REPORTS
APPENDIX S	DECISION RULES

Chapter 1. OVERVIEW OF THE MAINE HIGH SCHOOL ASSESSMENT

1.1 PURPOSE OF THE ASSESSMENT SYSTEM

The Maine High School Assessment (MHSA) is designed to measure student progress toward the achievement of the state standards contained in Maine’s system of *Learning Results*. 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 ready for college, career, and citizenship upon graduation.

From 1985 through 2005, grade 11 students took the state-developed Maine Educational Assessment (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 Maine Department of Education’s (MDOE) 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.

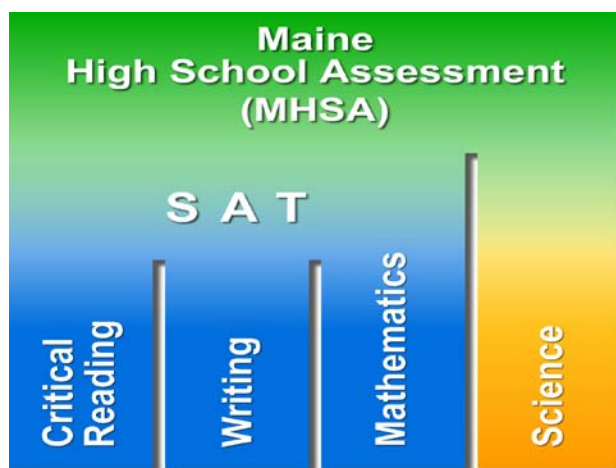
In 2009–2010, Maine adopted the New England Common Assessment Program (NECAP) content standards as part of the system of *Learning Results*, but retained the high school testing system. Alignment studies showed that there was sufficient alignment to the SAT mathematics section to allow the Math-A to be dropped. The same administration protocols and time lines were followed in 2009–2010: the science component was administered during a two-week window that ran from March 29 to April 9, 2010. 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 4. Details about the administration of the science components and the SAT were communicated to schools on an ongoing basis through informational letters, the MDOE Web site, and webinars. Additionally, workshops were held throughout the state on all aspects of accommodations, the registration process, SAT test center supervisor training, 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 two components of the 2010 MHSA (SAT 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.



1.2 PURPOSE OF THIS REPORT

The purpose of this report is to document the technical aspects of the 2009–2010 Maine High School Assessment (MHSA), one component of Maine’s Comprehensive Assessment System (MeCAS). Other components are the Personalized Alternate Assessment Portfolio, the New England Common Assessment Program, 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.

Chapter 2. TEST DESIGN AND DEVELOPMENT OF THE MHSA: SAT

The MHSA is intended to support good educational practice and is perceived as having an impact on instruction and curriculum. It features two components: The SAT and the MHSA Science test. Details on the content specifications and development of the SAT component are featured in this chapter; Chapter 3 covers content specifications and development of the Science component.

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.

2.1 THE MHSA: THE SAT 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 2009 administration of the SAT can be directly compared with scores from the May 2010 administration. The MHSA designates the May and June (makeup 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 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: SAT provides assessment opportunities for all students, regardless of their cognitive abilities, cultural backgrounds, or linguistic backgrounds.
2. Precisely defined constructs—The MHSA: SAT measures the constructs it is intended to measure and does not measure irrelevant material.
3. Accessible, non-biased items—The MHSA: SAT 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 MHSА: SAT 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—MHSА: SAT mathematics 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 MHSА: SAT 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 2010 forms required by the MHSА, like all forms of the SAT critical reading test, met the specifications presented in Table 2-1.

**Table 2-1. 2009–10 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. 2009–10 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.

Table 2-3. 2009–10 MHSА: 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)	

continued

¹ Described more fully in Chapter 8, equated delta is a transformation of 1–p, with a mean of 13 and a standard deviation of 4.

<i>Number and Percentage of Items by Delta Value</i>	
DV	N (%)
8	6 (9.0)
7	4 (6.0)
6	2 (3.0)
Total	67 (100)

Note: The equated delta distribution, mean, and standard deviation are 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 2010 forms required by the MHSA, like all forms of the SAT writing test, met the specifications presented in Table 2-4:

Table 2-4. 2009–10 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.

Figure 2-1. 2009–10 MHSА: Essay Scoring Guide

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 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.

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. 2009–10 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	

continued

<i>Writing Process Skill</i>	<i>Essay Prompt</i>	<i>Improving Paragraphs</i>
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. 2009–10 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
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. 2009–10 MHSA: SAT Writing Section Configuration

<i>Writing 1 (25 minutes)</i>	<i>Writing 2 (25 minutes)</i>	<i>Writing 3 (10 minutes)</i>
Essay	Items 1–11: Improving sentences (11)	Items 1–14: Improving sentences (14)
	Items 12–29: Identifying sentence errors (18)	
	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 section 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. 2009–10 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

The MHSA mathematics test consists of the traditional SAT mathematics. 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 May and June 2010 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. 2009–10 MHSA: SAT Mathematics Content Specifications

<i>Time Allotted: 70 minutes</i>	<i>Number</i>	<i>Percent of Test</i>
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 specifications listed. The four content areas specified in Table 2-9 are further defined in Table 2-10.

Table 2-10. 2009–10 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

continued

Geometry and Measurement

- Pythagorean Theorem and special properties of isosceles, equilateral, and right triangles
- Properties of parallel and perpendicular lines
- Coordinate geometry
- Geometric visualization
- Slope
- Similarity
- Transformations

Data Analysis, Statistics, and Probability

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

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

**Table 2-11. 2009–10 MHSAT: 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 4.

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. 2009–10 MHSAT: SAT
Mathematics Psychometric Specifications**

		<i>Item Type Difficulty</i>	
		MC	SPR
<i>Mean equated delta (SD)</i>		12.2 (3.2)	13.6–14.2 (3)
<i>Number of Items by Delta Value and Item Type</i>			
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

continued

<i>Number of Items by Delta Value and Item Type</i>			
MC		SPR	
12	5		
11	5		
10	4	10–11	2
9	3		
8	3	8–9	1
7	2		
6	2	<6–7	0
<6	1		
Total	44	Total	10

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

2.6 DEVELOPMENT

Each new form of the MHSA test 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. Experienced educators participate in the development process through the work of multiple committees. The current members of these committees can be found at www.collegeboard.com.

2.7 ITEM WRITING AND REVIEW

Test development specialists at Educational Testing Service (ETS) write the test items for the SAT. 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 4 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 PRETESTING THE ITEMS

Every item used in an operational form of the MHSA SAT 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 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.

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 6.

2.9 ANALYSIS OF PRETEST INFORMATION FOR THE MHSA: SAT

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. The statistical indices employed in analyzing and screening the MHSA SAT component follow.

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 inter-correlated, 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 non-scored 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 1-12, 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).

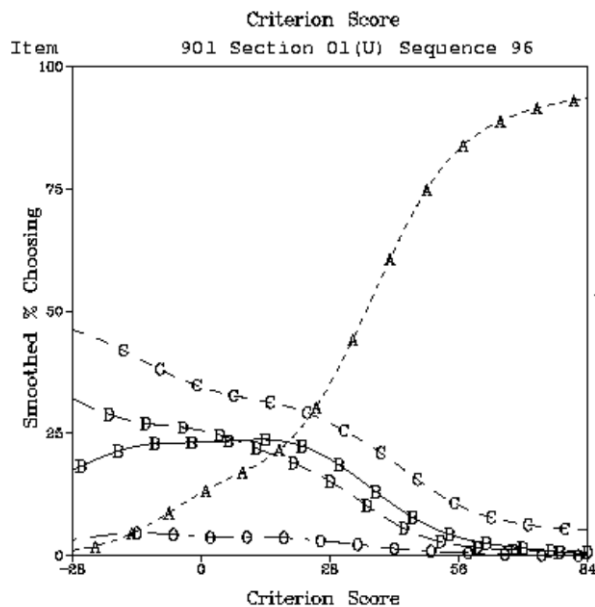
2.11 ITEM DISCRIMINATION AND ITEM/TEST RELATIONSHIP

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 the 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 MHSAs do 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. 2009–10 MHA: SAT 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 the 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 with 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-prettested 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 6.

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 is included that can occur in a variety of contexts (e.g., absolute value), 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 4.

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 in planning curricula, 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 2009–10 MHSA is one of the four released forms and will be available for Maine educators at no cost to inform teaching and learning of the NECAP Grade Expectations used in Maine. 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.

2.19 ALIGNMENT OF THE SAT TO THE NECAP STANDARDS

In 2009, Maine joined the NECAP consortium and the NECAP Grade Expectations were adopted into law. Alignment studies were conducted in August 2009 to compare the content of the SAT with the Grade

Expectations. The studies revealed that the alignment between the reading and mathematics Grade Expectations and those of the SAT critical reading test and mathematics test fully satisfied the criteria of the Webb alignment model.

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

2.19.1 Design of SAT Critical Reading

The 2009 NECAP Grade Expectations, covered by the SAT critical reading section, include the following:

1. Vocabulary Strategies and Breadth of Vocabulary
2. Initial Understanding of Literary Texts
3. Analysis and Interpretation of Literary Texts/Citing Evidence
4. Analysis and Interpretation of Literary Texts – Author’s Craft/Citing Evidence
5. Initial Understanding of Informational Text
6. Analysis and Interpretation of Informational Text/Citing Evidence

The number of items covering each performance indicator section of the reading standard is indicated in Table 2-13.

Table 2-13. 2009–10 MHSAs: Number of Items on the SAT Coded to NECAP Grade Expectations for Reading

<i>Reading Grade Expectation</i>	<i>SAT Critical Reading (Grade 11)</i>
Vocabulary Strategies and Breadth of Vocabulary	36
Initial Understanding of Literary Texts	21
Analysis and Interpretation of Literary Texts/Citing Evidence	29
Analysis and Interpretation of Literary Texts – Author’s Craft/Citing Evidence	17
Initial Understanding of Informational Text	27
Analysis and Interpretation of Informational Text/Citing Evidence	23

2.19.2 Design of SAT Mathematics

This section addresses only the SAT component of the MHSА Mathematics assessment. The 2009 NECAP Grade Expectations for Mathematics covered by the SAT mathematics section include the following:

1. Numbers and Operations
2. Geometry and Measurement
3. Functions and Algebra
4. Data, Statistics, and Probability

Table 2-14 displays the number of SAT items measuring each NECAP Grade Expectation. Refer to Appendix C for the Webb alignment study on the MHSА mathematics assessment.

Table 2-14. Number of Items on the SAT Coded to the NECAP Grade Expectations for Mathematics

<i>Mathematics Grade Expectation</i>	<i>SAT Mathematics (Grade 11)</i>
Numbers and Operations	28
Geometry and Measurement	16
Functions and Algebra	39
Data, Statistics, and Probability	7

Chapter 3. TEST DESIGN AND DEVELOPMENT OF THE MHSA: SCIENCE

3.1 TEST SPECIFICATIONS

3.1.1 Criterion Referenced Test

The MHSA contains a criterion referenced science test. Items on the science test are developed specifically for Maine and are directly linked to Maine’s science content standards. These content standards are the basis for the reporting categories and are used to help guide the development of test items.

3.1.2 Item Types

Maine educators and students are familiar with the types of items used in the assessment program. The types of items and their functions are described below:

- **Multiple-choice (MC)** items are used to provide breadth of coverage within a content area. Because they require no more than a minute for most students to answer, MC items make efficient use of limited testing time and allow for coverage of a wide range of knowledge and skills. There are four answer options for multiple-choice items in the MHSA Science test.
- **Constructed-response (CR)** items typically require students to use higher-order thinking skills—evaluation, analysis, summarization, and so on—to construct satisfactory responses. CR items take most students approximately 5 to 10 minutes to complete. Note that the use of released MHSA items to prepare students to respond to CR items is appropriate and encouraged.

3.1.3 Description of Test Design

The science test is structured using both *common* and *field-test* items. Common items are taken by all students. Student scores are based only on common items. Field-test items are divided among the forms of the test for each grade level. Each student takes only one form of the test and therefore answers a fraction of the field-test items. Field-test items are not identifiable to test takers and have a negligible impact on testing time. Because all students participate in the field test, it provides the minimum sample size (750–1,500 students per item) needed to produce reliable data that can be used to inform item selection for future tests.

3.2 SCIENCE TEST SPECIFICATIONS

3.2.1 Standards

The 2009–10 MHSA science test items are aligned to the content standards D: The Physical Setting and E: The Living Environment, which are Maine’s accountability standards and are described in the Science and Technology section of Maine’s *Learning Results: Parameters for Essential Instruction*. Content

specialists use the content standards, performance indicators, and descriptors 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.

3.2.2 Items Types

The science test includes multiple-choice (MC) and constructed-response (CR) items. Each MC item requires students to select the correct response from four choices. Each type of item is worth a specific number of points in the student's total science score, as shown in Table 3-1.

Table 3-1. 2009–10 MHSA: Science Item Types

<i>Item Type</i>	<i>Possible Score Points</i>
MC	-1/3, 0, or 1
CR	0, 1, 2, 3, or 4

MC = multiple-choice; CR = constructed-response

Fifty percent of the items were released from the 2009–10 MHSA science test. A practice test composed of released science items is available on the Maine Department of Education website: www.maine.gov/education/mhsa/science/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.

3.2.3 Test Design

Table 3-2 summarizes the numbers and types of items that were used to compute student scores on the 2009–10 MHSA science test. Additionally, each test form had eight MC matrix field-test items and one CR field-test item that did not affect student scores.

Table 3-2. 2009–10 MHSA: Science Items

Session 1	Session 2	TOTAL	
		MC	CR
16 MC, 2 CR	24 MC, 2 CR	40	4

MC = multiple-choice; CR = constructed-response

3.2.4 Blueprints

Table 3-3 shows the distribution of points across the science standards. For MHSA, D1 – D2 contained 1 CR item, D3 – D4 contained 1 CR item, and E1 – E5 contained 2 CR items.

Table 3-3. 2009–10 MHSA: Science Distribution of Score Points

Science Standards	Score Points
D1 – D2 (Earth & Space)	11
D3 – D4 (Physical)	23
E1 – E5 (Life)	22
Total score points	56

3.2.5 Depth of Knowledge

Each item on the MHSA science test is assigned a depth-of-knowledge (DOK) level. The DOK level reflects the complexity of mental processing students use to answer an item. DOK is not synonymous with difficulty. Each of the four DOK levels is described below.

- **Level 1 (Recall):** This level requires the recall of information such as a fact, definition, term, or simple procedure. These items require students only to demonstrate a rote response, use a well-known formula, or follow a set procedure.
- **Level 2 (Skill/Concept):** This level requires mental processing beyond that of recalling or reproducing a response. These items require students to make some decisions about how to approach the item.

- **Level 3 (Strategic Thinking):** This level requires reasoning, planning, and using evidence. These items require students to handle more complexity and abstraction than items at the previous two levels.
- **Level 4 (Extended Thinking):** This level requires planning, investigating, and complex reasoning over an extended period of time. Students are required to make several connections within and across content areas. This level may require students to design and conduct experiments. Due to the nature of this level, there are no level 4 items on the MHSA.

It is important that the MHSA measures a range of depths of knowledge. Table 3-4 shows the distribution of points across the DOK levels used on the MHSA.

**Table 3-4. 2009–10 MHSA: Science
Distribution of Score Points across
Depth of Knowledge (DOK)**

<i>DOK Level</i>	<i>Points</i>
1	12
2	28
3	16
Total	56

3.2.6 Use of Calculators and Reference Sheets

Calculators are not used or needed when taking the science test. There are no science reference sheets.

3.3 TEST DEVELOPMENT PROCESS

3.3.1 Item Development

Items used on the science test are developed and customized specifically for use on the MHSA and are consistent with Maine content standards and performance indicators. A Measured Progress test developer works with the Maine state science specialist and Maine educators to verify the alignment of items to the appropriate Maine content standards.

The development process combines the expertise of the Measured Progress test developer, the Maine state science specialist, and committees of Maine educators to help ensure items meet the needs of the MHSA program. All items used on the common portions of the MHSA were reviewed by a committee of Maine content experts, by a committee of Maine bias experts, and by three external content experts.

3.3.2 Item Reviews at Measured Progress

Measured Progress has test developers with expertise in each of the major science content areas who review the newly developed items for:

- item integrity, including science content and structure, format, clarity, possible ambiguity, and single correct answer;
- appropriateness and quality of graphic;
- appropriateness of scoring guide descriptions and distinctions;
- that the item is assessing the intended content standard;
- completeness of associated item documentation (e.g., scoring guide, content codes, key, grade level, depth of knowledge and contract identified); and
- appropriateness for the designated grade level.

3.3.3 Item Reviews at State Level

A committee of Maine classroom teachers from across the state reviewed the items before field testing. Teacher participants are selected based on their content-area expertise and grade-level familiarity. The purpose of the 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?

3.3.4 Bias and Sensitivity Review

Bias review is an essential component of the development process. During the bias review process, items were reviewed by a committee of Maine educators. Items were examined for issues that might offend or dismay students, teachers, or parents. Including such groups in the development of assessment items and materials can avoid many unduly controversial issues, and unfounded concerns can be allayed before the test forms are produced.

3.3.5 External Expert Review

The test items were classified into three groups based on science content. Three science experts (1 in earth/space science, 1 in life science, 1 in physical science) reviewed the group of items corresponding to their area of expertise. The expert reviewers primarily evaluated each item for correct science content. For the multiple-choice items, the experts also indicated whether the keyed answer was correct and whether it was the only correct answer among the options given. The DOE state science specialist and Measured Progress test developers reviewed the experts' evaluations and made appropriate adjustments to the items as necessary.

3.3.6 Reviewing and Refining

Recommended changes from the Item Review and Bias and Sensitivity meetings, as well as the comments from the three external science experts, were reviewed and considered by the Maine state science specialist. The Measured Progress test developer made the edits that were approved by the Maine state science specialist.

3.3.7 Item Editing

Measured Progress editors then reviewed and edited the items to ensure adherence to style guidelines in the *Chicago Manual of Style, 15th ed.*, and to sound testing principles. These principles include the stipulations that items

- demonstrate correct grammar, punctuation, usage, and spelling;
- are written in a clear, concise style;
- contain unambiguous explanations that tell students what is required to attain a maximum score;
- are written at a reading level that allows students to demonstrate their knowledge of the subject matter being tested regardless of reading ability;
- exhibit high technical quality regarding psychometric characteristics;
- have appropriate answer options or score-point descriptors; and
- are free of potentially insensitive content.

3.3.8 Item Selection and Operational Test Assembly

The Measured Progress test developer met with the Maine state science specialist to select the common items. In preparation for the meeting, the test developer and a psychometrician at Measured Progress considered the following in selecting the set of items to propose for the common:

- **Content coverage/match to test design and blueprints.** The test designs and blueprints stipulate a specific number of multiple-choice and constructed-response items for each content area. Item selection for the embedded field test was based on the number of items in the existing pool of items that are eligible for the common.
- **Item difficulty and complexity.** Item statistics drawn from the data analysis of previously field-tested items were used to ensure similar levels of difficulty and complexity from year to year, as well as quality psychometric characteristics.
- **“Cueing” items.** Items were reviewed for any information that might “cue” or provide information that would help to answer another item.

At the meeting, the Maine state science specialist reviewed the proposed sets of items for the common and field test and made the final selection of items.

The test developer laid out the items into test forms according to the test specifications. During assembly of the test forms, the following criteria were considered:

- **Key patterns.** The sequence of keys (correct answers) was reviewed to ensure that their order appeared random.
- **Option balance.** Items were balanced across forms so that each form contained a roughly equivalent number of key options (*As*, *Bs*, *Cs*, and *Ds*).
- **Page fit.** Item placement was modified to ensure the best fit and arrangement of items on any given page.
- **Relationships among forms.** Although field-test items differ from form to form, these items must take up the same number of pages in all forms so that sessions 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 test, and the number of graphics.

3.3.9 Operational Test Draft Review

After the forms were laid out as they would appear in the final test booklets, the forms were again thoroughly reviewed by Measured Progress editors to ensure that the items appeared exactly as intended. Any changes made during test construction were reviewed and approved by the test developer. The Maine state science specialist then read the forms for final approval.

3.3.10 Alternative Presentations

The common test was translated into Braille by National Braille Press, a subcontractor that specializes in test materials for blind and visually impaired students. In addition, Form 1 was adapted into a large-print version.

Chapter 4. TEST ADMINISTRATION: SAT

The SAT component of the MHSA was offered to all Maine juniors or third-year students on May 1 and June 5, 2010. There were 14,530 students who registered for the May date and 342 for the June makeup date. Of those 14,872 students, 13,558 (91.2%) registered under standard conditions, 1,106 (6.8%) registered with College Board–approved accommodations, 298 (2%) preregistered with Maine Purposes Only (MPO) accommodations, and 730 (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.

4.1 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 preparatory material may be viewed at www.maine.gov/education/mhsa/studentrp.html and www.collegeboard.com.

4.2 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.

4.3 PHYSICAL SETTING

In order that testing takes 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.

4.4 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.

4.5 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.

4.6 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 4-1 shows the actual test directions for student-produced–response items.

Figure 4-1. 2009–10 MHSA: SAT Instructions for Student-Produced Responses

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.

Answer: $\frac{7}{12}$

Write answer in boxes.

Grid in result.

Fraction line

Answer: 2.5

Decimal point

Answer: 201

Either position is correct.

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 $7\frac{1}{2}$. (If $\frac{31}{10}$ 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:

4.7 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.

4.8 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 for the 2009–10 testing year required the cancellation of scores or the scheduling of makeup tests.

4.9 SUBGROUP PERFORMANCE

In accordance with NCLB legislation that subgroup performance be analyzed and reported, Tables O-3 to O-8 in Appendix O present 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 five examinees.

4.10 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 or other team of educators.

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 as appropriate.

Historically, about 10% of those taking the state-administered MHSA tests have qualified for testing accommodations. Nationally, approximately 1.9% of SAT test takers qualify for College Board–approved Services for Students with Disabilities (SSD) accommodations. In the 2009–10 administration, 8.8% of those taking the MHSA qualified for testing accommodations: 4.6% in reading, 4.5% in mathematics, and 4.6% in writing.

4.10.1 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.

7. **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.
8. **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 have been 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.

4.10.2 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.

4.10.3 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).

4.10.4 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 February 16, 2010, deadline was established for Maine high school junior eligibility form submissions. March 26, 2010, was the standard deadline for eligibility form submissions for the May 1, 2010 SAT.

4.10.5 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 25, 2010.

4.10.6 MHSA Accommodation Request and Approval Statistics

Table 4-1 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 2010 MHSA administration. It includes any approvals for students who chose to take the test under MPO conditions.

**Table 4-1. 2009–10 MHSA: Summary
of Accommodations for 2010 MHSA Administration**

Total number of accommodations requested for College Board approval	1987
Total number of accommodations approved by College Board	1633
Total number of students using College Board accommodations	287
Total number of students using MPO accommodations	208
Total number of students using accommodations	495
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.	
<i>MPO ACCOMMODATIONS</i>	
MPO accommodations for May 2009	
MT1–Extended time same day	108
MT2–Extended time over several days	63
MT3–Multiple or frequent breaks	83
MT4–Flexible test day or start time	16
MT5–Flexible ordering of test sections	9
MS1–School location other than classroom	14
MS2–Offsite location with school personnel	2
MP1–Individual testing	18
MP2–Small group testing	199
MP3–Human reader	94
MP4–Sign language (not for reading test)	4
MP5–Stand, move, pace during testing	2
MP6–Alternative/Assistive Technology to Communicate	1
MP7–Proctored by special education or ESL Title 1 personnel	29
MP8–Large Print	1
MP10–Bilingual dictionary	21
MP11–Translation into native language	20
MP12–"Sheltered English" content	15
MR1–Scribe/recording device for other than essay	22
MR3–Other assistive devices	1
MR4–Word processor	9
MR7–Bilingual dictionary	2
MR8–Verification directions understood	103
MO1–Accommodations based on test content	2
Side by Side Test Books	2
Supervisor Rewrites Illegible Answers	2
<i>COLLEGE BOARD ACCOMMODATIONS</i>	
Large print–photo enlarged to 14 point	4
Large print–20 point	4
Large block answer sheet	4
Braille test	1

continued

<i>COLLEGE BOARD ACCOMMODATIONS</i>	
Braille graphs and figures	1
Braille device for written responses	1
Reader	183
Cassette test version	11
Writer to record responses	82
Computer to record written responses	42
<i>COLLEGE BOARD ACCOMMODATION (continued)</i>	
Reading–50% extended time	617
Writing–50% extended time	618
Mathematical calculations–50% extended time	610
Listening–50% extended time	254
Reading–100% extended time	65
Writing–100% extended time	67
Mathematical calculations–100% extended time	63
Extra breaks	328
Written directions, bring sign language interpreter	3
Extended breaks	60
Snacks and/or fluids permitted	14
Preferential seating	15
Write answers in the test book	4
Separate location	16
School-based testing	27
Auditory amplification, including FM system	1
Test blood sugar level	3
Small group setting	574
Other assistance – College Board will confirm	2

*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.

4.11 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 5. TEST ADMINISTRATION: SCIENCE

As the contractor responsible for the administration of the science 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 conducting item analysis for production of student results.

The science test was administered at all Maine high schools during the testing window of March 29 to April 9, 2010. As indicated in the *Principal and Test Coordinator Manual*, principals and/or their designated MHSA coordinators were responsible for the proper administration of the science portion 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.

5.1 RESPONSIBILITY FOR ADMINISTRATION

To ensure the administration of the science test 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 manuals may be accessed at <http://www.maine.gov/education/mhsa/testmaterial.html>.

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 science test. 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 and verifying the number of test and student response booklets being returned.

5.2 PARTICIPATION REQUIREMENTS AND DOCUMENTATION

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

Eligibility for taking the science test with accommodations was determined during the registration process for the SAT conducted by the College Board. (Please see Chapter 4 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 science test 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 science test 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 these students were forwarded to Measured Progress so they would not be included in any reports. A summary of participation rates, both overall and by demographic categories, is provided in Appendix E.

5.3 TEST SECURITY

Maintaining test security is critical to the success of the MHSA. The *Principal/Test Coordinator Manual* and the *Test Administrator Manual* explain in detail all test security measures and test administration procedures. School personnel were informed that any concerns about breaches in test security were to be reported to the school's test coordinator and/or principal immediately. The test coordinator and/or principal were responsible for immediately reporting the concern to the District Superintendent and the State Assessment Director at the Department of Education. Test security was also strongly emphasized at test administration workshops. Principals were required to log onto a secure Web site to complete the *Principal's Certification of Proper Test Administration* form; they also had to provide the number of secure tests received from Measured Progress, the number of tests administered to students, and the number of secure test materials that they were returning to Measured Progress. Principals were instructed to submit the form by entering a unique password, which acted as their digital signature. By signing and submitting the form, the principal certified that the tests were administered according to the test administration procedures outlined in the *Principal/Test Coordinator Manual* and *Test Administrator Manual*, that the security of the tests was maintained, that no secure material was duplicated or in any way retained in the school, and that all test materials had been accounted for and returned to Measured Progress.

5.4 TEST AND ADMINISTRATION IRREGULARITIES

There were no test irregularities in the spring 2010 administration.

5.5 TEST ADMINISTRATION WINDOW

The test administration window was March 29 through April 9, 2010.

5.6 SERVICE CENTER

To provide additional support to schools before, during, and after testing, Measured Progress established the MeCAS Service Center. The support of the Service Center is essential to the successful administration of any statewide test program. It provides a centralized location to which individuals in the field can call using a toll free number to ask specific questions or report any problems they may be experiencing. Representatives are responsible for receiving, responding to, and tracking calls, then routing issues to the appropriate person(s) for resolution. All calls are logged into a database, which includes notes regarding the issue and resolution of each call. The Service Center was open to receive calls from 7:30 AM to 4:00 PM Monday through Friday beginning two weeks before the start of testing and ending two weeks after testing.

Chapter 6. SCORING: 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.⁴ 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.

6.1 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 associated with 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 form
- Seating charts
- Test Question Ambiguity/Error form

6.2 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. Answer sheets are held in a climate controlled environment

⁴ Chapter 9 describes how scores are transformed to the MHS scale of 1100 to 1180.

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 multi-sheet 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.

6.3 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.

9. 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.
10. 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.
11. 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.

6.4 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 8.

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.

6.5 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 the level expected for essays at the same score point for

other prompts. (For example, 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 2010 SAT essays required a third reading (Figure 6-1). For the Maine-specific population of students who received official score reports, the percentage of essays requiring a third reading was the same (Figure 6-2).

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 6-1. Differences in Reader Scores for National Sample in May and June 2010

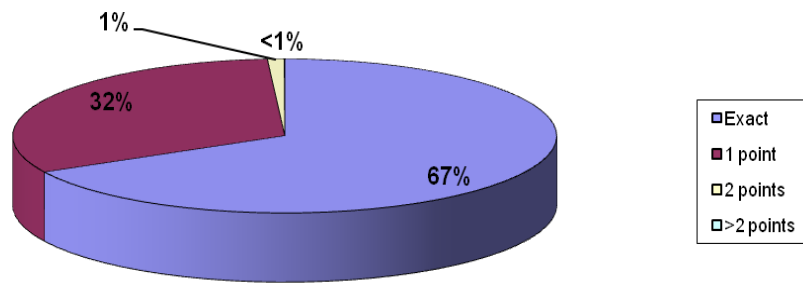
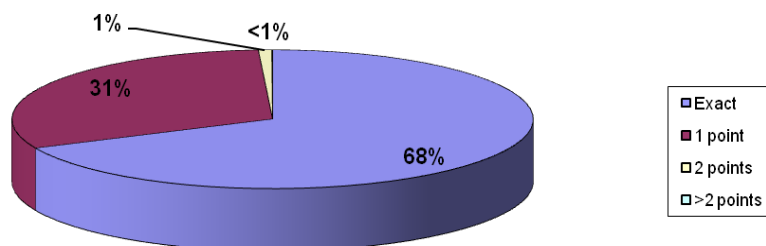


Figure 6-2. 2009–10 MHSAs: Differences in Reader Scores for Maine-Specific Sample* in May and June 2010



*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 2010 national administrations for all test takers is shown in Figure 6-3. The Maine-specific distributions for May and June 2010 are displayed in Figure 6-4. It should be noted that Figure 6-4 is based only upon students in Maine who received official College Board score reports for the May and June 2010 administrations.

Figure 6-3. National Distribution of SAT Essay Scores for May and June 2010

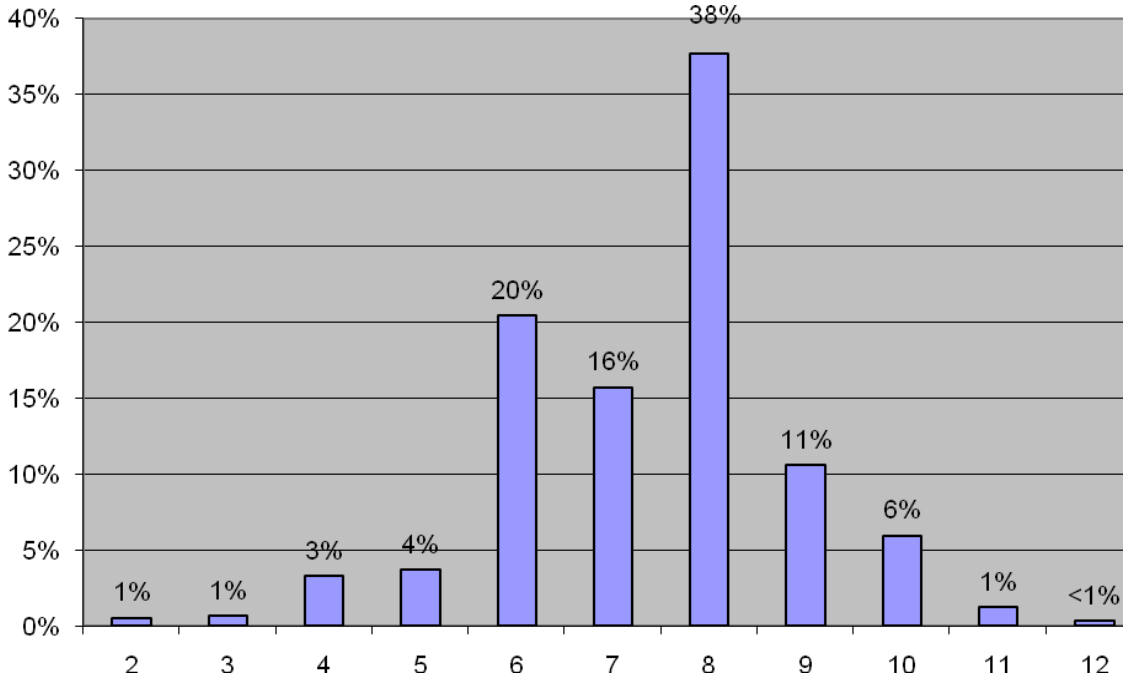
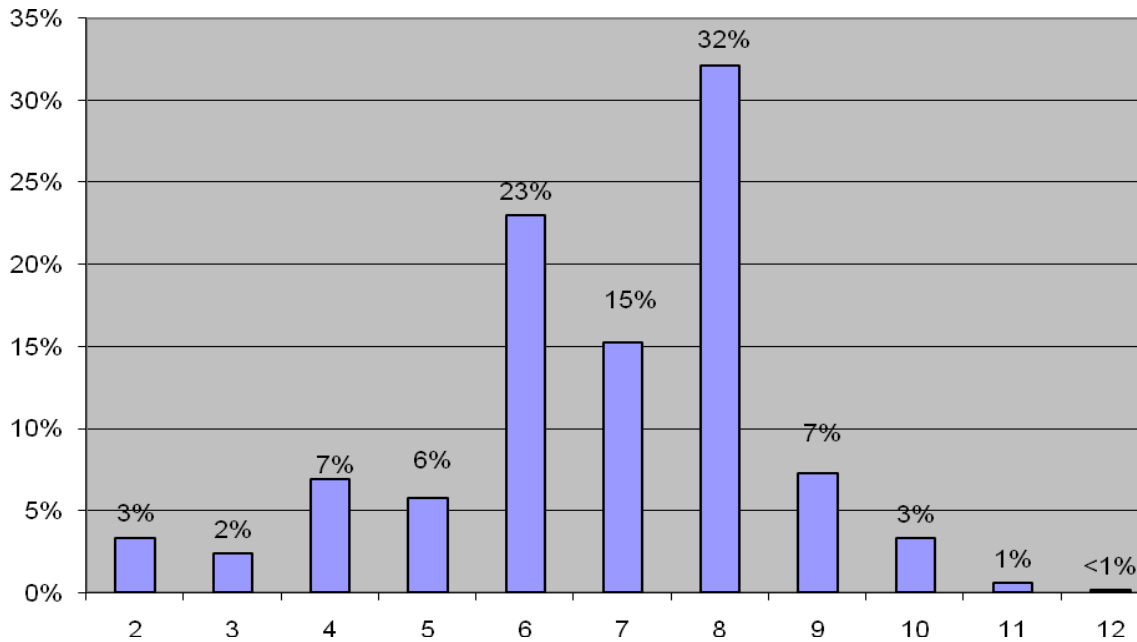


Figure 6-4. 2009–10 MHSAs: Maine-Specific Distribution* of SAT Essay Scores for May and June 2010



*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. 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 2009 College Board college-bound seniors cohort and the associated percentile ranks are shown in Table 6-1.

Table 6-1. 2009–10 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	14
790	99+	580	77	370	12
780	99	570	74	360	10
770	99	560	72	350	8
760	99	550	69	340	7
750	99	540	66	330	5
740	98	530	63	320	4
730	98	520	59	310	4
720	97	510	56	300	3
710	96	500	52	290	2
700	96	490	49	280	2
690	95	480	46	270	1
680	94	470	42	260	1
670	93	460	39	250	1
660	92	450	35	240	1
650	90	440	32	230	1
640	89	430	29	220	-1
630	87	420	25	210	-1
620	85	410	22	200	-
610	84	400	19	Mean	493
600	81	390	17	SD	111

SD = standard deviation

*Based on the 2009 College-Bound Seniors Cohort

As a point of reference, the SAT writing scores from the 2008 college-bound seniors cohort had a mean of 494 and a standard deviation of 110.

6.6 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.

6.7 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.

6.8 SUMMARY

The SAT component of the MHSAs 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.

Chapter 7. SCORING: SCIENCE

7.1 MACHINE-SCORED ITEMS

Multiple-choice item responses were compared to scoring keys using item analysis software. Correct answers were assigned a score of one point and incorrect answers were assigned -1/3 point. Student responses with multiple marks and blank responses were also assigned zero points.

The hardware elements of the scanners monitor themselves continuously for correct reads, and the software that drives these scanners also monitors correct data reads. Standard checks include recognition of a sheet that does not belong or is upside down or backwards and identification of critical data that are missing (e.g., a student ID number), test forms that are out of range or missing, and page or document sequence errors. When a problem is detected, the scanner stops and displays an error message directing the operator to investigate and to correct the situation.

7.2 PERSON-SCORED ITEMS

The images of student responses to constructed-response items were hand-scored through the Measured Progress electronic scoring system, iScore. Use of iScore minimizes the need for readers to physically handle answer booklets and related scoring materials. Student confidentiality was easily maintained, since all 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 at computer terminals accessed electronically scanned images of student responses. Readers evaluated each response and recorded each 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, imaged responses from a student’s entire booklet were always available for viewing when necessary, and the physical booklet was also available to the Chief Reader onsite. (Chief Reader and other scoring roles are described in Section 7.2.1 that follows.)

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

7.2.1 Scoring Location and Staff

Scoring Location

The iScore database, its operation, and its administrative controls are all based in Dover, New Hampshire. All 2009–10 MHSA Science test item responses were scored in Dover.

The iScore system monitored accuracy, reliability, and consistency across the scoring site. 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 the scoring site.

Staff Positions

The following staff members were involved with scoring the 2009–10 MHSA responses:

- The MHSA Scoring Project Manager, an employee of Measured Progress, was located in Dover, New Hampshire, and oversaw communication and coordination of scoring across the scoring site.
- The iScore Operational Manager and iScore administrators, employees of Measured Progress, were located in Dover, New Hampshire, and coordinated technical communication across the scoring site.
- A Chief Reader in the Science content area ensured consistency of scoring across the scoring site. Chief Readers also provided read-behind activities (defined in 7.2.7) for Quality Assurance Coordinators. Chief Readers are employees of Measured Progress.
- 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 the science content area. QACs provided read-behind activities (defined in 7.2.7) for Senior Readers at the scoring site. The ratio of QACs and Senior Readers to Readers was approximately 1:11.
- Senior Readers (SRs), selected from a pool of skilled and experienced Readers, provided read-behind activities (defined in 7.2.7) for the Readers at their scoring tables (2–12 Readers at each table). The ratio of QACs and SRs to Readers was approximately 1:11.
- Readers at the Dover, New Hampshire, scoring site scored operational and field-test MHSA 2009–10 student responses. Recruitment of Readers is described in Section 7.2.3.

7.2.2 Benchmarking Meetings

In preparation for implementing MHSA scoring guidelines, Measured Progress scoring staff prepared and facilitated benchmarking meetings held with the MHSA state science specialist representing their departments of education. The purpose of these meetings was to establish guidelines for scoring MHSA items during the current field-test scoring session and for future operational scoring sessions.

Several dozen student responses for each item that Chief Readers identified as illustrative midrange examples of the respective score points were selected. Chief Readers presented these responses to the MHSA

science content specialist during benchmarking meetings and worked collaboratively to finalize an authoritative set of score-point exemplars for each field-test item. As a matter of practice, these sets are included in the scoring training materials each time an item is administered.

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

7.2.3 Reader Recruitment and Qualifications

For scoring the 2009–10 MHSA, Measured Progress actively sought a diverse scoring pool. The broad range of Reader backgrounds typically includes scientists, editors, business professionals, authors, teachers, graduate school students, and retired educators. Demographic information about Readers (e.g., gender, race, educational background) was electronically captured for reporting.

Although a four-year college degree or higher was preferred, Readers were required to have successfully completed at least two years of college and to have demonstrated knowledge of the science content area. This permitted recruiting Readers currently enrolled in a college program, a sector of the population with relatively recent exposure to current classroom practices and trends in their fields. In all cases, potential Readers were required to submit documentation (e.g., resume and/or transcripts) of their qualifications.

Table 7-1 summarizes the qualifications of the 2009–10 MHSA scoring leadership and Readers.

Table 7-1. 2009–10 MHSA: Science Qualifications of Scoring Leadership and Readers—Spring Administration

<i>Scoring responsibility</i>	<i>Educational credentials</i>				<i>Total</i>
	<i>Doctorate</i>	<i>Master's</i>	<i>Bachelor's</i>	<i>Other</i>	
Scoring Leadership	0.0%	28.6%	57.1%	14.3%*	100.0%
Readers	13.6%	34.1%	50.0%	2.3%**	100.0%

Scoring Leadership = Chief Readers, QACs, and SRs

*1 QAC/SR had an Associate's degree

**1 Reader had an Associate's degree

Readers were either temporary Measured Progress employees or were secured through temporary employment agencies. All Readers were required to sign a nondisclosure/confidentiality agreement.

7.2.4 Methodology for Scoring Polytomous Items

Possible Score Points

The ranges of possible score points for the different polytomous items are shown in Table 7-2.

Table 7-2. 2009–10 MHSA: Science Possible Score Points for Polytomous Item Types

<i>Polytomous item type</i>	<i>Possible score point range</i>
Constructed-response	0–4
Non-Scorable Items	0

Non-Scorable 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*

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

Scoring Procedures

Scoring procedures for polytomous items included both single scoring and double-blind scoring. Single-scored items were scored by one Reader. Double-blind scored items were scored independently by two Readers, whose scores were tracked for “interrater agreement” (for further discussion of double-blind scoring and interrater agreement, see Section 7.2.7 and Appendix P).

7.2.5 Reader Training

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

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

Following the review of an item’s scoring guide, Readers reviewed the particular response set organized for that training: Anchor Sets, Training Sets, and Qualifying Sets. (These are defined below.)

During training, Readers could highlight or mark hard copies of the Anchor and Training Sets (as well as first Qualifying Sets after the qualification round), 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 for an item. This is a set approved by the science content specialist representing the MESA state department of education. Responses in Anchor Sets are 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 the MESA client and Measured Progress scoring services staff. Each contains one client-approved sample response per score point considered to be a midrange exemplar. The set includes a second sample response if there is more than one plausible way to illustrate the merits and intent of a score point.

Responses were read aloud to the room of Readers in descending score order. Announcing the true score of each anchor response, trainers facilitated group discussion of responses in relation to score point descriptions to help Readers internalize the typical characteristics of score points.

This Anchor Set continued to serve as a reference for Readers as they went on to calibration, scoring, and recalibration activities for that item.

Training Set

Next, Readers practiced applying the scoring guide and anchors to responses in the Training Set. The Training Set typically included 10 to 15 student responses designed to help establish both the full score point range and the range of possible responses within each score point. The Training Set often included unusual responses that were less clear or solid (e.g., shorter than normal, employing atypical approaches, simultaneously containing very low and very high attributes, and written in ways difficult to decipher). Responses in the Training Set were presented in randomized score point order.

After Readers independently read and scored a Training Set response, trainers polled Readers or used online training system reports to record their initial range of scores. Trainers then led group discussion of one or two responses, directing Reader attention to difficult scoring issues (e.g., the borderline between two score points). Throughout the process, trainers modeled how to discuss scores by referring to the anchor set and to scoring guides.

Qualifying Set

After the Training Set had been completed, Readers were required to score responses accurately and reliably in Qualifying Sets assembled for constructed-response items. The 10 responses in each Qualifying Set were selected from an array of responses that clearly illustrated the range of score points for that item as reviewed and approved by the state specialist. Hard copies of the responses were also made available to Readers after the qualification round so that they could make notes and refer back during the post-qualifying discussion.

To be eligible to live score one of the above items, Readers were required to demonstrate scoring accuracy rates of at least 80% exact agreement (i.e., to exactly match the predetermined score on at least 8 of the 10 responses) and at least 90% exact or adjacent agreement (i.e., to exactly match or be within one score point of the predetermined score on 9 or 10 of the 10 responses). In other words, Readers were allowed 1 discrepant score (i.e., 1 score of 10 that was more than one score point from the predetermined score), provided they had at least 8 exact scores.

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 the required accuracy rate on the second Qualifying Set, they were allowed to score operational responses.

Readers who did not achieve the required scoring accuracy rates on the second Qualifying Set were not allowed to score responses for that item. Instead, they either began training on a different item or were dismissed from scoring for that day.

7.2.6 Leadership Training

QACs and select SRs were trained in a separate training session immediately prior to Reader training. In addition to discussing the items and their responses, QAC and SR training included greater detail on the client's rationale behind the score points in order to better equip QACs and SRs to handle questions from regular Readers.

7.2.7 Monitoring of Scoring Quality Control

Readers were monitored for continued accuracy and consistency throughout the scoring process, using the following methods and tools (which are defined in this section):

- Embedded Committee-Reviewed Responses (CRRs)
- Read-Behind Procedures
- Double-Blind Scoring
- Recalibration Sets
- Scoring Reports

It should be noted that any Reader whose accuracy rate fell below the expected rate for a particular item and monitoring method was retrained on that item. Upon approval by the QAC or Chief Reader as appropriate (see below), the Reader was allowed to resume scoring. Readers who met or exceeded the expected accuracy rates continued scoring.

Furthermore, the accuracy rate required of a Reader to *qualify* to score responses live was higher than that required to *continue* to score responses live. The reason for the difference is that an “exact score” in double-blind scoring requires that *two* Readers choose the same score for a response (in other words, the score is dependent on peer agreement), whereas an “exact score” in qualification requires only that a *single* Reader match a score pre-established by scoring leadership. The use of multiple monitoring techniques is critical toward monitoring reader accuracy during the process of live scoring.

Embedded Committee-Reviewed Responses (CRRs)

Committee-reviewed responses (CRRs) are previously scored responses that are loaded (“embedded”) by scoring leadership into iScore and distributed “blindly” to Readers during scoring. Embedded CRRs may be chosen either before or during scoring and are inserted into the scoring queue so that they appear the same as all other live student responses.

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

Any Reader who fell below the required scoring accuracy rate was retrained before being allowed by the QAC to continue scoring. Once allowed to resume scoring, scoring leadership carefully monitored these Readers by increasing the number of read-behinds (defined in the next section).

Embedded CRRs were employed for all constructed-response items.

Read-Behind Procedures

Read-behind scoring refers to scoring leadership (usually a SR) scoring a response after a Reader has already scored the response. The practice was applied to all open-ended item types.

Responses placed into the read-behind queue were randomly selected by scoring leadership; Readers were not aware which of their responses would be reviewed by their SR. The iScore system allowed one, two, or three responses per Reader to be placed into the read-behind queue at a time.

The SR entered his or her score into iScore before being allowed to see the Reader’s score. The SR then compared the two scores, and the score of record (i.e., the reported score) was determined as follows:

- If there was exact agreement between the scores, no action was necessary; the regular Reader’s score remained.
- If the scores were adjacent (i.e., differed by 1 point), the SR’s score became the score of record. (A significant number of adjacent scores for a Reader triggered an individual scoring consultation with the SR, after which the QAC determined whether or when the Reader could resume scoring.)

- If the scores were discrepant (i.e., differed by more than 1 point), the SR’s score became the score of record. (This triggered an individual consultation with the SR, after which the QAC determined whether or when the reader could resume scoring on that item.)

Table 7-3 illustrates how scores were resolved by read-behind.

Table 7-3. 2009–10 MHSAs: Science Examples of Read-Behind Scoring Resolutions

<i>Reader score</i>	<i>QAC/SR score</i>	<i>Score of record</i>
4	4	4
4	3	3*
4	2	2*

* QAC/SR’s score.

SRs were tasked with conducting, on average, five read-behinds per Reader throughout each half-scoring day; however, SRs conducted a proportionally greater number of read-behinds for Readers who seemed to be struggling to maintain, or who fell below, accuracy standards.

In addition to regular read-behinds, scoring leadership could choose to do read-behinds on any Reader at any point during the scoring process to gain an immediate, real-time “snapshot” of a Reader’s accuracy.

Double-Blind Scoring

Double-blind scoring refers to two Readers independently scoring a response without knowing whether or not the response was to be double-blind scored. The practice was applied to all open-ended item types. Table 7-4 shows by which method(s) both common and equating open-ended item responses for each operational test were scored.

Table 7-4. 2009–10 MHSAs: Science Frequency of Double-Blind Scoring

<i>Grade</i>	<i>Content area</i>	<i>Responses double-blind scored</i>
HS	Science	10%
HS	Unreadable responses	100%
HS	Blank responses	100%

If there was a discrepancy (a difference greater than 1 score point) between double-blind scores, the response was placed into an arbitration queue. Arbitration responses were reviewed by scoring leadership (SR or QAC) without knowledge of the two Readers’ scores. Scoring leadership assigned the final score.

Appendix P provides the MHSAs 2009–10 percentages of agreement between Readers for each common item.

Scoring leadership consulted individually with any Reader whose scoring rate fell below the required accuracy rate, and the QAC determined whether or when the reader could resume scoring on that item. Once the Reader was allowed to resume scoring, scoring leadership carefully monitored the Reader’s accuracy by increasing the number of read-behinds.

Recalibration Sets

To determine whether Readers were still calibrated to the scoring standard, Readers were required to take an online Recalibration Set at the start and midpoint of the shift of their resumption of scoring.

Each Recalibration 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 of five responses of the first Recalibration Set, or exact on two or fewer, were not permitted to score on that item that day and were either assigned to a different item or dismissed for the day.
- Readers who were discrepant on only one of 5 responses of the first Recalibration Set, and/or exact on three, were retrained by their SR by discussing the Recalibration Set responses in terms of the score point descriptions and the original Anchor Set. After this retraining, such Readers began scoring operational responses under the proviso that the Reader's scores for that day and that item would be kept only if the Reader was exact on all 5 of 5 responses of the second Recalibration Set administered at the shift midpoint. The QAC determined whether or when these Readers had received enough retraining to resume scoring operational responses. Scoring leadership also carefully monitored the accuracy of such Readers by significantly increasing the number of their read-behinds.
- Readers who were not discrepant on any response of the first Recalibration Set, and exact on at least four, were allowed to begin scoring operational responses immediately, under the proviso that this Recalibration performance would be combined with that of the second Recalibration Set administered at the shift midpoint.

The results of both Recalibration Sets were combined with the expectation that Readers would have achieved an overall 80 percent-exact and 90 percent-adjacent standard for that item for that day.

The Scoring Project Manager voided all scores posted on that item for that day by Readers who did not meet the accuracy requirement. Responses associated with voided scores were reset and redistributed to Readers with demonstrated accuracy for that item.

Recalibration Sets were employed for all constructed-response items.

7.2.8 Reports Generated During Scoring

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. Due to the complexity of scoring a large-scale assessment project such as that for MHSA, computer-generated reports were necessary to ensure that

- overall group-level accuracy, consistency, and reliability of scoring were maintained at acceptable levels
- immediate, real-time individual Reader data were available to allow early 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 number and percentages of exact, adjacent, and discrepant scores with the SR/QAC. Scoring leadership could choose to generate this report by choosing options (such as “Today,” “Past Week,” and “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., sent back out to the floor to be rescored by other Readers). The benefit of this report is that it can reveal the degree to which an individual Reader agrees with their QAC or SR on how best to score live responses.
- **The Double-Blind Summary** showed the total number of double-scored responses of each Reader, and noted the number and percentages of exact, adjacent, and discrepant scores with second Readers. This report was used in conjunction with other reports to determine whether a Reader’s scores should be voided (i.e., sent back out to the floor to be rescored by other Readers). The benefit of this report is that it can reveal the degree to which Readers are in agreement with each other about how best to score live responses.
- **The Accuracy Summary** combined read-behind and double-blind data, showing the total number for the Readers, their accuracy rates, and their score-point distributions.
- **The Embedded CRR Summary** showed, for each Reader (by item or across all items), the total number of responses scored, the number of embedded CRRs scored, and the numbers and percentages of exact, adjacent, and discrepant scores with the Chief Reader. This report was used in conjunction with other reports to determine whether a Reader’s scores should be voided (i.e., sent back out to the floor to be rescored by other Readers). The benefit of this report is that it can reveal the degree to which an individual Reader agrees with their Chief Reader on how to best score live responses. Also, since embedded CRRs are administered during the first hours of scoring, this report can provide an early illustration of agreement between Readers and Chief Readers.
- **The Qualification Statistics Summary** listed each Reader by name and ID number, identified which Qualifying Set(s) they did and did not take and, for the ones taken, their pass rate. In addition to the pass rates of individuals, the report also showed numbers of Readers passing or

failing a particular Qualifying Set. The QAC could use this report to determine how Readers within their scoring group performed on specific Qualifying Sets.

- **The Summary Statistics Report** showed the total number of student responses for an item, and identified, for the time at which the report was generated, the following:
 - the number of single and double-blind scorings that had been performed
 - the number of single and double-blind scorings yet to be performed

Chapter 8. PSYCHOMETRIC TOPICS: 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 critical reading section provides a strong focus on the construct of reading, 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 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.

8.1 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 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, and Hammond, 2005).

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.

8.2 SAT STATISTICAL CHARACTERISTICS

The statistical characteristics of the SAT, based on the two forms administered in May and June 2010, 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 F. Tables F-1 through F-3 provide summaries of the scores for examinees participating in SAT testing in May and June 2010 by section for each form. Tables F-4 through F-6 present the rounded scaled score conversion tables by section for each SAT form.

8.3 RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

8.3.1 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 F-7 of Appendix F). 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 8-1. These values range from 0.79 to 0.92 for mathematics, 0.81 to 0.94 for critical reading, and 0.70 to 0.90 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, students who are economically disadvantaged, and students with limited English proficiency (beyond the first year). Subgroup reliabilities range from 0.85 to 0.93 in mathematics, from 0.90 to 0.94 in critical reading, and from 0.84 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 8-2. Average SAT scores and standard deviations on the raw score scale for Maine students are reported in Table 8-3.

8.3.2 Standard Errors of Measurement

The standard error of measurement (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 8-1 and 2-2 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 1.8 to 3.4 for mathematics, 2.2 to 4.1 for critical reading, and 2.0 to 3.7 for writing. Form 2 reliabilities and SEMs were not provided for the Maine-specific sample due to small sample size.

Conditional SEMs (i.e., SEMs at each scaled-score point) are provided in the raw score to scaled score lookup tables, which are presented in Appendix M. These are the actual tables that were used to determine student scaled scores, error bands, and achievement levels.

Table 8-1. 2009–10 MHTA: SAT Reliability Coefficients and SEMs* for Sections of the MHTA**

			<i>Form 1—May 2010</i>	
			<i>N=12,730</i>	
<i>Test Section</i>			<i>Reliability</i>	<i>SEM</i>
Math 1	Dressel-KR20	Raw	0.79	2.1
Math 2	Dressel-KR20	Raw	0.81	1.8
Math 3	Dressel-KR20	Raw	0.80	1.9
Total MHTA mathematics	Alpha	Raw	0.92	2.9
	Var. components	Raw	0.92	3.3
Critical Reading 1	Dressel-KR20	Raw	0.84	2.4
Critical Reading 2	Dressel-KR20	Raw	0.85	2.5
Critical Reading 3	Dressel-KR20	Raw	0.81	2.2
Total critical reading	Alpha	Raw	0.94	3.4
	Var. components	Raw	0.94	4.0
Writing 1	Dressel-KR20	Raw	0.87	3.1
Writing 2	Dressel-KR20	Raw	0.70	2.0
Total writing MC	Alpha	Raw	0.90	3.0
	Var. components	Raw	0.90	3.7

* See Appendix G 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 8-2. 2009–10 MHSA: SAT Reliability Coefficients and SEMs for Sections of the MHSA*

		Form 1—May 2010				
Test Section	Subgroup	N	KR-20		Variance Components	
			Reliability	SEM	Reliability	SEM
Total MHSA mathematics	Male	6,354	0.93	2.9	0.93	3.4
	Female	6,376	0.92	2.9	0.92	3.3
	Students with disabilities	1,090	0.86	2.8	0.85	3.4
	Economically disadvantaged	3,799	0.89	2.9	0.89	3.4
Total critical reading	Male	6,354	0.94	3.4	0.94	4.1
	Female	6,376	0.94	3.3	0.94	4.0
	Students with disabilities	1,090	0.90	3.4	0.90	4.1
	Economically disadvantaged	3,799	0.92	3.4	0.92	4.1
Total writing MC	Male	6,354	0.91	3.0	0.90	3.7
	Female	6,376	0.90	3.0	0.90	3.6
	Students with disabilities	1,090	0.84	3.0	0.84	3.6
	Economically disadvantaged	3,799	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 8-3. 2009–10 MHSA: SAT Raw Score Summary Statistics for Total Group and Subgroups

Form 1 May 2010		Mathematics			Critical Reading			Writing		
		N	Mean	SD	N	Mean	SD	N	Mean	SD
Gender	Male	6,354	21.3	12.6	6,354	25.8	16.5	6,354	18.5	11.8
	Female	6,376	19.8	11.4	6,376	26.7	16.0	6,376	21.0	11.5
	All	12,730	20.5	12.0	12,730	26.2	16.3	12,730	19.8	11.7
Students with disabilities	Yes	1,090	9.4	8.9	1,090	12.0	13.1	1,090	8.5	9.1
	No	11,640	21.6	11.8	11,640	27.5	15.9	11,640	20.8	11.3
	All	12,730	20.5	12.0	12,730	26.2	16.3	12,730	19.8	11.7
Economically disadvantaged	Yes	3,799	15.5	10.2	3,799	20.0	14.6	3,799	15.1	10.3
	No	8,931	22.7	12.1	8,931	28.9	16.2	8,931	21.7	11.7
	All	12,730	20.5	12.0	12,730	26.2	16.3	12,730	19.8	11.7
Limited English Proficiency	Currently receiving LEP	149	11.7	11.0	149	8.7	9.7	149	8.6	7.1
	Formerly received LEP	41	19.4	9.4	41	23.7	10.7	41	19.3	6.7
	No LEP	12,540	20.6	12.0	12,540	26.4	16.2	12,540	19.9	11.7

SD = standard deviation

Appendix N contains scaled-score distribution graphs showing the relative and cumulative percentages of students at each scaled score. The total number (N) of students tested is also given, from which the number of students assigned each scaled score can be derived.

Appendix N also shows, in Table N-1, achievement-level distributions for each of the last three administrations.

Table 8-4 below shows the MHSAs scaled-score ranges that correspond to each achievement level.

**Table 8-4. 2009–10 MHSAs:
SAT Range of Scores for Each Achievement Level**

<i>Content Area</i>	<i>Substantially Below Proficient</i>	<i>Partially Proficient</i>	<i>Proficient</i>	<i>Proficient with Distinction</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

8.4 CLASSIFICATION ACCURACY AND CONSISTENCY OF MHSAs: SAT CUT SCORES

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston and Lewis, 1995). After the achievement levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the MHSAs, students are classified into one of four achievement levels: Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction. This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2009–10 MHSAs because it is easily adaptable to all types of testing formats, including mixed-format tests.

The accuracy and consistency estimates reported below make use of “true scores” in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their “true” classifications.

For the 2009–10 MHSAs, after various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy was created where cell $[i, j]$ represented the estimated proportion of students whose true score fell into classification i (where $i = 1$ to 4) and observed score into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments per Livingston and Lewis (1995), a new four-by-four contingency table was created and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell $[i, j]$ of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where $i = 1$ to 4) and whose observed score on the second form would fall into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

Another way to measure consistency is to use Cohen's (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{(\text{Observed agreement}) - (\text{Chance agreement})}{1 - (\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}},$$

where

$C_{i.}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on the first hypothetical parallel form of the test;

$C_{.i}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on the second hypothetical parallel form of the test;

C_{ii} is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

8.4.1 Accuracy and Consistency

Results of the accuracy and consistency analyses described above are provided in Table 8-5. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, for Mathematics, the conditional accuracy value is 0.85 for Substantially Below Proficient. This figure indicates that among the students whose true scores placed them in this classification, 85% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.80 indicates that 80% of students with observed scores in the Substantially Below Proficient level would be expected to score in this classification again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, in testing done for NCLB accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the Partially Proficient/Proficient threshold is of greatest interest. For the 2009–10 MHSA, Table 8-6 provides accuracy

and consistency estimates at each cutpoint as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Note that, as with other methods of evaluating reliability, DAC statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 8-5 and 8-6 should be interpreted with caution.

Table 8-5. 2009-10 MHSA: SAT Summary of Decision Accuracy (and Consistency) Results by Subject—Conditional on Cutpoint

Subject	Grade	Substantially Below/Partially			Partially/Proficient			Proficient/Proficient with Distinction		
		Accuracy (consistency)	False positive	False negative	Accuracy (consistency)	False positive	False negative	Accuracy (consistency)	False positive	False negative
Reading	11	0.94 (0.92)	0.03	0.03	0.92 (0.89)	0.04	0.03	0.96 (0.95)	0.02	0.01
Writing	11	0.92 (0.89)	0.04	0.04	0.90 (0.86)	0.06	0.04	0.96 (0.95)	0.03	0.01
Mathematics	11	0.93 (0.90)	0.04	0.04	0.91 (0.88)	0.05	0.04	0.98 (0.97)	0.02	0.01

Table 8-6. 2009-10 MHSA: SAT Summary of Decision Accuracy (and Consistency) Results by Subject—Overall and Conditional on Performance Level

Subject	Grade	Overall	Kappa	Conditional on level			
				Substantially Below Proficient	Partially Proficient	Proficient	Proficient with Distinction
Reading	11	0.83 (0.76)	0.66	0.87 (0.82)	0.75 (0.66)	0.85 (0.80)	0.87 (0.76)
Writing	11	0.78 (0.70)	0.58	0.83 (0.77)	0.69 (0.60)	0.82 (0.75)	0.84 (0.67)
Mathematics	11	0.82 (0.75)	0.62	0.85 (0.80)	0.68 (0.58)	0.88 (0.82)	0.84 (0.67)

8.5 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 8-7 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 8-7 provides the speededness data for the state of Maine. The critical reading portion is unspeeded with the exception of Section 2, which is speeded but only by a small margin with not all items reached by 80% of examinees, slightly less examinees reaching 75% of the items than in critical reading sections 1 and 3, and variance indices of 0.11 and 0.15 for May and June respectively. However, note that the results for June 2010 should be interpreted with caution given the small sample size (180 examinees) 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 indices indicate a lack of speededness. One exception is section 2 in May 2010, where the variance index of 0.16 indicates the section is at the bottom of the moderate speededness range as proposed by Swineford (1974). The writing portion is unspeeded for both May and June 2010. Completion rate data for the national SAT population are provided in Appendix F, Table F-8.

**Table 8-7. 2009–10 MHSAs: Maine Completion
Rate Statistics for Sections of the College Board SAT**

<i>Form</i>	1	2	1	2	1	2
Administration	05/10	06/10	05/10	06/10	05/10	06/10
Sample size*	11,573	180	11,573	180	11,573	180
	<i>Critical Reading 1</i>		<i>Mathematics 1</i>		<i>Writing 1</i>	
% completing section	83.4	82.2	48.6	60.6	82.8	76.7
% completing 75%	98.6	99.4	97.6	98.3	100.0	100.0
Number of items reached by 80%	24	24	19	19	35	34
Mean not reached	0.5	0.4	0.9	0.7	0.4	0.5
SD not reached	1.5	1.2	1.4	1.3	1	1.2
NR variance/score variance	0.07	0.04	0.09	0.08	0.01	0.02
Number of items	24	24	20	20	35	35
	<i>Critical Reading 2</i>		<i>Mathematics 2</i>		<i>Writing 2</i>	
% completing section	78.7	77.8	49.8	38.9	90.5	87.8
% completing 75%	95.9	96.1	96.3	98.9	98.3	99.4
Number of items reached by 80%	23	23	15	16	14	14
Mean not reached	0.9	0.8	1.2	1.3	0.2	0.2
SD not reached	2.1	2.1	1.6	1.2	0.8	0.6
NR variance/score variance	0.11	0.15	0.16	0.10	0.05	0.03
Number of items	24	24	18	18	14	14
	<i>Critical Reading 3</i>		<i>Mathematics 3</i>			
% completing section	82.9	86.7	74	63.9		
% completing 75%	98.9	97.8	96.6	97.8		
Number of items reached by 80%	19	19	15	14		
Mean not reached	0.3	0.3	0.7	0.8		
SD not reached	1.1	1.0	1.6	1.4		
NR variance/score variance	0.05	0.04	0.14	0.12		
Number of items	19	19	16	16		

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.

8.6 ITEM STATISTICS

8.6.1 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 compared to those who attempted to answer the item. For the SAT, p -values are converted onto a standard scale called the delta index.

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

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 2010 were 11.4 (2.4), which exactly meets the specification. For mathematics and writing, the mean deltas for the two forms administered in May and June 2010 are also very close to the specifications, differing by no more than 0.20, with the exception of the standard deviations for the Math student-produced–response items. Table 8-8 summarizes the mean equated delta and standard deviation for each content area by form for students testing on the MHSA in Maine only.

Table 8-8. 2009–10 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	Form 1	Form 2
			May 2010 12,730 Equated Delta	June 2010 193 Equated Delta
Mathematics MC	N	44	44	44
	Mean	12.2	12.2	12.1
	SD	3.2	3.2	3.0
Mathematics SPR	N	10	10	10
	Mean	13.6-14.2	14.2	14.3
	SD	3.0	2.6	3.6
Total critical reading	N	67	67	67
	Mean	11.4	11.4	11.4
	SD	2.4	2.4	2.4
Total writing	N	49	49	49
	Mean	10.1	10.1	10.2
	SD	2.5	2.3	2.4

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.

8.6.2 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 being 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 8-9 presents the biserial coefficients for the May and June 2010 forms of the SAT for students taking the MHSA in Maine only. ‘

**Table 8-9. 2009–10 MHSAs: Maine Summary
Statistics for Biserial Coefficients* for Mathematics,
Critical Reading, and Writing Sections of the College Board SAT**

<i>Content Area</i>			<i>Form 1</i>	<i>Form 2</i>
			<i>May 2010</i> 12,730	<i>June 2010</i> 193
Mathematics MC	N		44	44
	Mean	.53-.57	0.53	0.46
	SD		0.12	0.20
Mathematics SPR	N		10	9
	Not Computed ^b			1
	Mean	.60-.70	0.63	0.58
Total critical reading	SD		0.14	0.14
	N		67	65
	Not Computed ^b			2
Total writing	Mean	.49-.53	0.54	0.47
	SD		0.11	0.15
	N		49	49
Total writing	Mean	.49-.53	0.49	0.44
	SD		0.09	0.16
	N		49	49

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

^bAn *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.

8.7 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 and 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 and 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 value 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; therefore, 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 2010 SAT test forms for the national population are listed in Tables F-11 and F-12 of Appendix F. Table 8-10 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. 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 African American students exceeds the criteria of 200 students, some caution should be used in the interpretation of these results as well. A potential second impact of the small sample sizes is that more items may have been classified with C-DIF; however, that did not occur in the May 2010 data. In fact, no C-DIF items were identified in any comparison.

Table 8-10. 2009–10 MHSAs: Maine Differential Item Functioning (DIF) Summary Form: 1 Administration: 5/10

<i>Category of Maximum Absolute DIF Value for All Comparisons</i>				<i>Female</i> N=6,376	<i>African American</i> N=270
				<i>Male</i> N=6,354	<i>White</i> N=12,056
<i>Content Area</i>	<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	5	9.3	3	3
	A	44	81.5	48	49
	-B	5	9.3	3	2
	-C	0	0.0	0	0
	Total	54	100	54	54
Total critical reading	+C	0	0.0	0	0
	+B	1	1.5	1	0
	A	59	88.1	63	63
	-B	7	10.4	3	4
	-C	0	0.0	0	0
	Total	67	100	67	67
Total writing	+C	0	0.0	0	0
	+B	2	4.1	0	2
	A	44	89.8	48	45
	-B	3	6.1	1	2
	-C	0	0.0	0	0
	Total	49	100	49	49

8.8 SUMMARY

The scores reported for SAT test takers must be accurate and 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 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 G or visit www.collegeboard.com/prod_downloads/sat/sat-program-handbook.pdf.

Chapter 9. PSYCHOMETRIC TOPICS: SCIENCE TEST

9.1 CLASSICAL ITEM ANALYSIS

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 item. Both *Standards for Educational and Psychological Testing* (AERA et al., 1999) and *Code of Fair Testing Practices in Education* (2004) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors. Items should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. In addition, items must not unfairly disadvantage students in particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that MHSA science items meet these standards. Qualitative analyses are described in earlier chapters of this report; this chapter focuses on quantitative evaluations. Statistical evaluations are presented in four parts: 1) difficulty indices, 2) item-test correlations, 3) differential item functioning (DIF) statistics, and 4) dimensionality analyses. The item analyses presented here are based on the statewide administration of the MHSA science test in spring 2010.

Note that, to facilitate interpretability of the calculated statistics, formula scoring of multiple-choice items was not implemented for purposes of calculating classical difficulty and discrimination indices or DIF statistics.

9.1.1 Classical Difficulty and Discrimination Indices

All multiple-choice and constructed-response items are evaluated in terms of item difficulty according to standard classical test theory practices. Difficulty is defined as the average proportion of points achieved on an item and is measured by obtaining the average score on an item and dividing it by the maximum possible score for the item. For purposes of calculating classical item statistics, the multiple-choice items were scored dichotomously (i.e., without formula scoring); therefore, for these items, the difficulty index is simply the proportion of students who correctly answered the item. Constructed-response items are scored polytomously, meaning that a student can achieve a score of 0, 1, 2, 3, or 4. By computing the difficulty index as the average proportion of points achieved, the indices for the different item types are placed on a similar scale, ranging from 0.0 to 1.0 regardless of the item type. Although this index is traditionally described as a measure of difficulty, it is properly interpreted as an *easiness* index, because larger values indicate easier items. An index of 0.0 indicates that all students received no credit for the item, and an index of 1.0 indicates that all students received full credit for the item.

Items that are answered correctly by almost all students provide little information about differences in student abilities, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that are correctly answered by very few students provide little information about differences in student

abilities, but may indicate knowledge or skills that have not yet been mastered by most students. In general, to provide the best measurement, difficulty indices should range from near-chance performance (0.25 for four-option multiple-choice items or essentially zero for constructed-response items) to 0.90, with the majority of items generally falling between 0.4 and 0.7. However, on a standards-referenced assessment such as the MHSA science test, it may be appropriate to include some items with very low or very high item difficulty values to ensure sufficient content coverage.

A desirable characteristic of an item is for higher-ability students to perform better on the item than lower-ability students do. The correlation between student performance on a single item and total test score is a commonly used measure of this characteristic of the item. Within classical test theory, the 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 constructed-response items, the item-discrimination index used was the Pearson product-moment correlation. For the multiple-choice items, formula scoring was not implemented for purposes of calculating classical item statistics, so the item discrimination index used was the point-biserial correlation. The theoretical range of these statistics is -1.0 to 1.0, with a typical observed range from 0.2 to 0.6.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score. That is, the discrimination index can be thought of as a measure of construct consistency.

A summary of the item difficulty and item discrimination statistics is presented in Table 9-1. Note that the statistics are presented for all items as well as by item type (multiple-choice and constructed-response). The mean difficulty and discrimination values shown in the table are within generally acceptable and expected ranges.

Table 9-1. 2009–10 MHSA: Science Summary of Item Difficulty and Discrimination Statistics

<i>Item type</i>	<i>Number of items</i>	<i>p-Value</i>		<i>Discrimination</i>	
		<i>Mean</i>	<i>Standard deviation</i>	<i>Mean</i>	<i>Standard deviation</i>
ALL	44	0.52	0.18	0.36	0.1
CR	4	0.38	0.09	0.58	0.05
MC	40	0.53	0.18	0.34	0.08

Comparing the difficulty indices of multiple-choice items and constructed-response items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that the difficulty indices for multiple-choice items tend to be higher (indicating that students performed better on these items) than the difficulty indices for constructed-response items. Similarly, discrimination indices for the four-point constructed-response items were larger than those for the dichotomous items because of the greater variability of the former (i.e., the partial credit these items allow) and the tendency for correlation coefficients to be higher given greater variances of the correlates.

In addition to the item difficulty and discrimination summaries presented above, item-level classical statistics and item-level score-point distributions were also calculated. Item-level classical statistics are provided in Appendix H; item difficulty and discrimination values are presented for each item. The item difficulty and discrimination indices are within generally 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. There were a small number of items with low discrimination indices, but none were negative. While it is not inappropriate to include items with low discrimination values or with very high or very low item difficulty values to ensure that content is appropriately covered, there were very few such cases on the MHSAs science test. Item-level score-point distributions are provided for constructed-response science items in Appendix I; for each science item, the percentage of students who received each score point is presented.

9.1.2 Differential Item Functioning

Code of Fair Testing Practices in Education (2004) explicitly states that subgroup differences in performance should be examined when sample sizes permit and that actions should be taken to ensure that differences in performance are because of 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, MHSAs science items were evaluated in terms of differential item functioning (DIF) statistics.

For the MHSAs science test, the standardization DIF procedure (Dorans and Kulick, 1986) was employed to evaluate subgroup differences. The standardization DIF procedure is designed to identify items for which subgroups of interest perform differently, beyond the impact of differences in overall achievement. The DIF procedure calculates the difference in item performance for two groups of students (at a time) matched for achievement on the total test. Specifically, average item performance is calculated for students at every total score. Then an overall average is calculated, weighting the total score distribution so that it is the same for the two groups.

When differential performance between two groups occurs on an item (i.e., a DIF index in the “low” or “high” categories, explained below), it may or may not be indicative of item bias. Course-taking patterns or differences in school curricula can lead to DIF, but for construct-relevant reasons. On the other hand, if subgroup differences in performance could be traced to differential experience (such as geographical living conditions or access to technology), the inclusion of such items should be reconsidered.

Computed DIF indices have a theoretical range from -1.0 to 1.0 for multiple-choice items, and the index is adjusted to the same scale for constructed-response items; here, again, formula scoring was not applied to the items for purposes of calculating DIF statistics. Dorans and Holland (1993) suggested that index values between -0.05 and 0.05 should be considered negligible. The preponderance of MHSAs science 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.

For the 2009–10 MHSA, six subgroup comparisons were evaluated for DIF:

- Male versus female
- No Disability versus Disability
- Not Economically Disadvantaged versus Economically Disadvantaged
- Non-LEP versus LEP
- White (non-Hispanic) versus Asian
- White (non-Hispanic) versus Black or African American

The table in Appendix J presents the number of items classified as either “low” or “high” DIF, overall and by group favored.

9.1.3 Dimensionality Analysis

The MHSA science test was designed to measure and report a single score on science achievement using a unidimensional scale from 1100 to 1180. Thus, this test is said to measure a single dimension, and the term “unidimensional” is used to describe such a test.

Because the high school science test was 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 science dimension that all the items have in common. 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 Item Response Theory (IRT) models that are used for calibrating, linking, scaling, and equating the 2009-10 MHSA science test forms.

The purpose of dimensionality analysis is to investigate whether violations of the assumption of test unidimensionality are statistically detectable and, if so, (a) the degree to which unidimensionality is violated and (b) the nature of the multidimensionality. Findings from dimensionality analyses performed on the 2009-10 MHSA science test are reported below. (Note: Only common items were analyzed since they are used for score reporting.)

Dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, and Gao, 2001) and DETECT (Zhang and 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 item response theory. Parametric techniques, such as 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 item characteristic curves.

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 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. Non-zero conditional covariances are essentially violations of the principle of local independence, and 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 cross-validation sample. Then 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 cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items displays local dependence, conditioning on total score on the non-clustered 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 displays local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal 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 cross-validation sample (if a DIMTEST exploratory analysis has been conducted, one could use the same training and cross-validation samples as were used with DIMTEST, 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 cross-validation sample data to average the conditional covariances. Within-cluster conditional covariances are summed, and from this sum the between-cluster conditional covariances are subtracted. This 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 cross-validation 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 and DETECT were applied to the 2009-10 MHSA science test. The data were first split into a training sample and a cross-validation sample. Because the total sample size was just over 14,000 student examinees, the training sample and cross-validation sample each had slightly more than 7,000 students.

DIMTEST was then applied to the MHSA science test. Because of the very large sample size of this test, DIMTEST would be sensitive even to quite small violations of unidimensionality; and the null hypothesis was rejected with a p-value less than 0.00005. The occurrence of statistical rejection of the null hypothesis was not surprising because strict unidimensionality is an idealization that rarely holds exactly for a given dataset. Thus, it was important to use DETECT to estimate the effect size of the violation of local independence found by DIMTEST.

Next, a DETECT analysis was conducted on the MHSA science test. This resulted in a DETECT statistic of 0.15, 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 0.45, 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 65.3%.

The clusters reported by DETECT were investigated and they indicated some tendency for the four-point open-response (OR4) items to cluster separately from the multiple-choice (MC) items. Specifically, there was one cluster in which OR4 items accounted for about 70% of the points in the cluster, whereas on the test as a whole the OR4 items only accounted for about 30% of the total points. The OR4 items in this cluster had strong positive conditional covariances with each other, but they also had many positive conditional covariances with MC items (instead of the strong negative conditional covariances you would expect if they were a strongly distinct dimension).

Finally, we note that these results are very similar to results that occurred in the analyses of the MHSA science tests in both 2007–08 and in 2008–09. In particular, for both years, rejection of the DIMTEST null hypothesis of unidimensionality occurred with a p-value less than 0.00005. The DETECT effect sizes for the two years were 0.23 and 0.17, respectively, indicating weak and very weak multidimensionality, respectively.

Taken together, the DIMTEST and DETECT results for the science test indicate that the test has very small, though detectable, violations of unidimensional local independence, and that these violations seem primarily related to the two item types used on the test. Thus, although these results indicate some definite multidimensionality, it is very weak in magnitude. Therefore, no changes in test design or scoring for the science test seem to be warranted in regard to multidimensionality. In particular, the dimensionality analysis results support the application of unidimensional IRT to the MHSA science test for purposes of calibrating,

linking, scaling, and equating. Indeed, the results support using unidimensional IRT to place the MHSA science items onto a single-score scale for reporting purposes.

9.2 IRT SCALING AND EQUATING

The MHSA science test uses a pre-equating model in which items are calibrated using IRT and placed on scale at the time of field testing. These item parameters are then used to assemble test forms that meet content blueprints and psychometric quality criteria. The sections below describe the procedures used to calibrate the MHSA items and to calculate scaled scores and achievement levels used for reporting.

9.2.1 Item Response Theory

As mentioned above, all MHSA science items were calibrated using IRT. IRT uses mathematical models to define a relationship between an unobserved measure of student performance, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item. In IRT, it is assumed that all items are independent measures of the same construct (i.e., of the same θ). Another way to think of θ is as a mathematical representation of the latent trait of interest. Several common IRT models are used to specify the relationship between θ and p (Hambleton and van der Linden, 1997; Hambleton and Swaminathan, 1985). The process of determining the specific mathematical relationship between θ and p is called item calibration. After items are calibrated, they are defined by a set of parameters that specify a nonlinear, monotonically increasing relationship between θ and p . Once the item parameters are known, an estimate of θ for each student can be calculated. This estimate, $\hat{\theta}$, is considered to be an estimate of the student's true score or a general representation of student performance. It has characteristics that may be preferable to those of raw scores for equating purposes.

Because of the use of formula scoring we use a polytomous IRT model for all items. The MC items are scored such that an incorrect response is given a score of -0.33, an omit is given a score of 0, and a correct answer is given a score of 1 (i.e., formula scoring). The student response records are initially coded as 0, 1, or 2, and the integer scoring function is modified from 0, 1, and 2 to -0.33, 0, and 1 after the IRT calibration and equating process is complete. Thus, the response category probability values as estimated during IRT calibration are multiplied by their respective value from the modified scoring function.

In the graded response model (GRM) for polytomous items, an item is scored in $k + 1$ graded categories that can be viewed as a set of k dichotomies. At each point of dichotomization (i.e., at each threshold), a two-parameter model can be used. This implies that a polytomous item with $k + 1$ categories can be characterized by k item category threshold curves (ICTC) of the two-parameter logistic form:

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

where
i indexes the items,
j indexes students,
k indexes threshold,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

After computing *k* ICTCs in the GRM, *k* + 1 item category characteristic curves (ICCCs) 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 Item Characteristic Curve (ICC) for polytomous items is computed as a weighted sum of ICCCs, where each ICCC is weighted by a score assigned to a corresponding category.

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

For more information about item calibration and determination, the reader is referred to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

9.2.2 Item Response Results

The tables in Appendix K give the IRT item parameters of all common items on the 2009–10 MHSA tests. In addition, Appendix L shows graphs of the test characteristic curves (TCCs) and test information functions (TIFs), which are defined below.

TCCs display the expected (average) raw score associated with each θ_j value between -4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 10.1, the expected raw score at a given value of θ_j is

$$E(X | \theta_j) = \sum_{i=1}^n P_i(1|\theta_j),$$

where
 i indexes the items (and n is the number of items contributing to the raw score),
 j indexes students (here, θ_j runs from -4.0 to 4.0), and
 $E(X | \theta_j)$ is the expected raw score for a student of ability θ_j .

The expected raw score monotonically increases with θ_j , consistent with the notion that students of high ability tend to earn higher raw scores than do students of low ability. Most TCCs are “S-shaped”: flatter at the ends of the distribution and steeper in the middle.

The TIF displays the amount of statistical information that the test provides at each value of θ_j . Information functions depict test precision across the entire latent trait continuum. There is an inverse relationship between the information of a test and its standard error of measurement (SEM). For long tests, the SEM at a given θ_j is approximately equal to the inverse of the square root of the statistical information at θ_j (Hambleton, Swaminathan, and Rogers, 1991), as follows:

$$SEM(\theta_j) = \frac{1}{\sqrt{I(\theta_j)}}$$

Compared to the tails, TIFs are often higher near the middle of the θ distribution where most students are located and where most items are sensitive by design.

9.2.3 Achievement Standards

MHSA standards to establish science achievement-level cut scores were set in May 2009. The standard-setting meeting and results were discussed in the 2009 technical report and standard-setting report provided at that time. The theta-metric cut scores that emerged from the standard-setting meeting will remain fixed throughout the assessment program unless standards are reset for any reason.

9.2.4 Scaled Scores

9.2.4.1 Description of Scale

Because the θ scale used in IRT calibrations is not readily understood by most stakeholders, reporting scales were developed for the MHSA tests. The reporting scale is a simple linear transformation of the underlying θ scale used in the IRT calibrations. Scaled scores range from 1100 to 1180; the Partially

Proficient/Proficient cut was set at 1142 and the Proficient/Proficient with Distinction cut was set at 1162. (At the student level, scaled scores were reported as even numbers only.)

By providing information that is more specific about the position of a student's results, scaled scores supplement achievement-level scores. School and SAU-level scaled scores are calculated by computing the average of student-level scaled scores. Students' raw scores (i.e., total number of points) on the MHSA tests were translated to scaled scores using the data analytic process known as *scaling*. Scaling simply converts from one scale to another. In the same way that a given temperature can be expressed on either Fahrenheit or Celsius scales, or the same distance can be expressed in either miles or kilometers, student scores on the 2009–10 MHSA tests can be expressed in raw or scaled scores.

It is important to note that converting from raw scores to scaled scores does not change students' achievement-level classifications. Given the relative simplicity of raw scores, it is fair to question why scaled scores are reported instead of raw scores. First, because multiple-choice items are formula scored, fractional and negative total raw scores are possible, making them undesirable for use in score reporting. In addition, scaled scores make consistent the reporting of results across years. Due to the fact that different sets of items make up each year's test form, raw cut scores may vary slightly from year to year, but the scaled cut scores remain the same. It is this uniformity across scaled scores that facilitates the understanding of student performance.

9.2.4.2 Calculations

The scaled scores are obtained by a simple translation of ability estimates ($\hat{\theta}$) using the linear relationship between threshold values on the θ metric and their equivalent values on the scaled-score metric. Students' ability estimates are based on their raw scores and are found by mapping through the TCC. Scaled scores are calculated using the linear equation

$$SS = m\hat{\theta} + b$$

where
 m is the slope and
 b is the intercept

The linear transformation is determined by fixing the 1142 and 1162 values. Table 9-2 presents the scaled-score cuts (i.e., the minimum scaled score for getting into the next achievement level). It is important to repeat that the values in Table 9-2 do not change from year to year, because the cut scores along the θ scale do not change unless standards are reset. Also, in a given year it may not be possible to attain a particular scaled score, but the scaled-score cuts will remain the same.

Table 9-2. 2009–10 MHSA: Science Scaled Score Cuts and Minimum and Maximum Scores

Minimum	Scaled Score Cuts			Maximum
	SBP/PP	PP/P	P/PWD	
1100	1134	1142	1162	1180
SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient With Distinction				

Table 9-3 shows the cut scores on θ and the slope and intercept terms used to calculate the scaled scores. Note that the values in Table 9-3 will not change unless the standards are reset.

Table 9-3. 2009–10 MHSA: Science Cut Scores (on θ Metric), Intercept, and Slope

θ Cuts			Transformation Constants	
SBP/PP	PP/P	P/PWD	Slope	Intercept
-0.3318	0.3616	2.3362	10.12863	1138.337
SBP = Substantially Below Proficient; PP = Partially Proficient; P = Proficient; PWD = Proficient With Distinction				

Appendix M contains raw score to scaled-score lookup tables for this year and last year. These are the actual tables that were used to determine student scaled scores, error bands, and achievement levels. The SEMs reported in the lookup tables are *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.

9.2.4.3 Score Distributions

Appendix N contains scaled-score distribution graphs showing the relative and cumulative percentages of students at each scaled score. Appendix N also shows, in Table N-1, achievement-level distributions. Because standards for the MHSA science assessment were set in 2009, results are shown for the 2008–09 and 2009–10 administrations.

9.3 RELIABILITY

Although an individual item’s performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way items function together and complement one another. Tests that function well provide a dependable assessment of the student’s level of ability. Unfortunately, no test can do this perfectly. A variety of factors can contribute to a given student’s score being either higher or lower than his or her true ability. For example, a student may misread an item, or mistakenly fill in the wrong bubble when he or she knew the answer. Collectively, extraneous factors that impact a student’s score are referred to as measurement error. Any assessment includes some amount of measurement error; that is, no measurement is perfect. This is true of all academic assessments—some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability.

When tests have a high amount of measurement error, student scores are very unstable. Students with high ability may get low scores or vice versa. Consequently, one cannot reliably measure a student’s true level of ability with such a test. Assessments that have less measurement error (i.e., errors made are small on average and student scores on such a test will consistently represent their ability) are described as reliable.

There are a number of ways to estimate an assessment’s reliability. One possible approach is to give the same test to the same students at two different points in time. If students receive the same scores on each test, then the extraneous factors affecting performance are small and the test is reliable. (This is referred to as “test-retest reliability.”) A potential problem with this approach is that students may remember items from the first administration or may have gained (or lost) knowledge or skills in the interim between the two administrations. A solution to the “remembering items” problem is to give a different, but parallel test at the second administration. If student scores on each test correlate highly, the test is considered reliable. (This is known as “alternate forms reliability,” because an alternate form of the test is used in each administration.) This approach, however, does not address the problem that students may have gained (or lost) knowledge or skills in the interim between the two administrations. In addition, the practical challenges of developing and administering parallel forms generally preclude the use of parallel forms reliability indices. One way to address the latter two problems is to split the test in half and then correlate students’ scores on the two half-tests; this in effect treats each half-test as a complete test. By doing this, the problems associated with an intervening time interval and with creating and administering two parallel forms of the test are alleviated. This is known as a “split-half estimate of reliability.” If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests 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 into halves will result in a different correlation. Another problem with the split-half method of calculating reliability 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, α (alpha), which eliminates the problem of the split-half method by comparing individual item variances to total test variance. Cronbach’s α was used to assess the reliability of the 2009–10 MHSA:

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

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.

9.3.4 Reliability and Standard Errors of Measurement

Table 9-4 presents descriptive statistics, Cronbach’s α coefficient, and raw score standard errors of measurement (SEMs) for the 2009–10 MHSA science assessment.

Table 9-4. 2009–10 MHSA: Science Raw Score Descriptive Statistics, Cronbach’s Alpha, and Standard Errors of Measurement (SEM)

Grade	Number of students	Raw score			Alpha	SEM
		Maximum	Mean	Standard deviation		
11	14,026	56	21.75	12.31	0.89	4.15

9.3.5 Subgroup Reliability

The reliability coefficients presented in the previous section were based on the overall population of students who took the 2009–10 MHSA science test. Appendix O presents reliabilities for various subgroups of interest. Subgroup Cronbach’s α ’s were calculated using the formula defined above based only on the members of the subgroup in question in the computations; values are only calculated for subgroups with 10 or more students.

For several reasons, the results of this section should be interpreted with caution. First, inherent differences between grades and content areas preclude making valid inferences about the quality of a test based on statistical comparisons with other tests. Second, reliabilities are dependent not only on the measurement properties of a test, but on the statistical distribution of the studied subgroup. For example, it can be readily seen in Appendix O that subgroup sample sizes may vary considerably, which results in natural variation in reliability coefficients. Alternatively, α , which is a type of correlation coefficient, may be artificially depressed for subgroups with little variability (Draper and Smith, 1998). Third, there is no industry standard to interpret the strength of a reliability coefficient, and this is particularly true when the population of interest is a single subgroup.

9.3.6 Subcategory Reliability

Of even more interest are reliabilities for the science reporting subcategories within MHSA, described in Chapter 3. Cronbach’s α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Appendix O. Once again as expected, because they are based on a subset of items rather than the full test, computed

subcategory reliabilities were lower (sometimes substantially so) than were overall test reliabilities, and interpretations should take this into account. The subcategory reliabilities were lower than those based on the total test and approximately to the degree one would expect based on classical test theory. Qualitative differences between subtests once again preclude valid inferences about the quality of the full test based on statistical comparisons among subtests.

9.3.7 Interrater Consistency

Chapter 7 of this report describes in detail the processes that were implemented to monitor the quality of the hand-scoring of student responses for science constructed-response items. One of these processes was double-blind scoring: approximately 10% of student responses were randomly selected and scored independently by two different scorers. Results of the double-blind scoring were used during the scoring process to identify scorers that required retraining or other intervention and are presented here as evidence of the reliability of the MHSA science test. A summary of the interrater consistency results is presented in Table 9-5 below. Results in the table are collapsed across the hand-scored items. The table shows the number of score categories, the number of included scores, the percent of exact agreement, the percent of adjacent agreement, the correlation between the first two sets of scores, and the percent of responses that required a third score. This same information is provided at the item level in Appendix P. These interrater consistency statistics are the result of the processes implemented to ensure valid and reliable hand-scoring of items as described in detail in Chapter 7.

Table 9-5. 2009–10 MHSA: Science Summary of Interrater Consistency Statistics Collapsed Across Items

<i>Grade</i>	<i>Number of score categories</i>	<i>Number of included scores</i>	<i>Percent exact</i>	<i>Percent adjacent</i>	<i>Correlation</i>	<i>Percent of third scores</i>
11	5	5422	65.14	31.89	0.83	2.91

9.3.8 Reliability of Achievement-Level Categorization

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston and Lewis, 1995). After the achievement levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical decision accuracy and consistency (DAC) of the classifications. For the MHSA science test, students are classified into one of four achievement levels: Substantially Below Proficient, Partially Proficient, Proficient, or Proficient with Distinction. This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated, because errorless test scores do not exist. Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2009–10 MHSa science test because it is easily adaptable to all types of testing formats, including mixed-format tests.

The accuracy and consistency estimates reported below make use of “true scores” in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated. In the Livingston and Lewis method, estimated true scores are used to categorize students into their “true” classifications.

For the 2009–10 MHSa science test, after various technical adjustments (described in Livingston and Lewis, 1995), a four-by-four contingency table of accuracy was created where cell $[i, j]$ represented the estimated proportion of students whose true score fell into classification i (where $i = 1$ to 4) and observed score into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students whose true and observed classifications matched) signified overall accuracy.

To calculate consistency, true scores were used to estimate the joint distribution of classifications on two independent, parallel test forms. Following statistical adjustments per Livingston and Lewis (1995), a new four-by-four contingency table was created and populated by the proportion of students who would be categorized into each combination of classifications according to the two (hypothetical) parallel test forms. Cell $[i, j]$ of this table represented the estimated proportion of students whose observed score on the first form would fall into classification i (where $i = 1$ to 4) and whose observed score on the second form would fall into classification j (where $j = 1$ to 4). The sum of the diagonal entries (i.e., the proportion of students categorized by the two forms into exactly the same classification) signified overall consistency.

Another way to measure consistency is to use Cohen’s (1960) coefficient κ (kappa), which assesses the proportion of consistent classifications after removing the proportion of consistent classifications that would be expected by chance. It is calculated using the following formula:

$$\kappa = \frac{(\text{Observed agreement}) - (\text{Chance agreement})}{1 - (\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.}C_{.i}}{1 - \sum_i C_{i.}C_{.i}},$$

where

$C_{i.}$ is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on the first hypothetical parallel form of the test;

C_i is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on the second hypothetical parallel form of the test;
 C_{ii} is the proportion of students whose observed achievement level would be Level i (where $i = 1 - 4$) on both hypothetical parallel forms of the test.

Because κ is corrected for chance, its values are lower than other consistency estimates.

9.3.8.2 Accuracy and Consistency

The accuracy and consistency analyses described above are provided in Tables 9-6 and 9-7. The table includes overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.85 for Substantially Below Proficient. This figure indicates that among the students whose true scores placed them in this classification, 85% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.79 indicates that 79% of students with observed scores in the Substantially Below Proficient level would be expected to score in this classification again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, in testing done for NCLB accountability purposes, the primary concern is distinguishing between students who are proficient and those who are not yet proficient. In this case, the accuracy of the Partially Proficient/Proficient threshold is of greatest interest. For the 2009–10 MHSA science test, Table 9-6 provides accuracy and consistency estimates at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

The above indices are derived from Livingston and Lewis's (1995) method of estimating the accuracy and consistency of classifications. It should be noted that Livingston and Lewis discuss two versions of the accuracy and consistency tables. A standard version performs calculations for forms parallel to the form taken. An "adjusted" version adjusts the results of one form to match the observed score distribution obtained in the data. The tables use the standard version for two reasons: (1) this "unadjusted" version can be considered a smoothing of the data, thereby decreasing the variability of the results; and (2) for results dealing with the consistency of two parallel forms, the unadjusted tables are symmetrical, indicating that the two parallel forms have the same statistical properties. This second reason is consistent with the notion of forms that are parallel; that is, it is more intuitive and interpretable for two parallel forms to have the same statistical distribution.

Note that, as with other methods of evaluating reliability, DAC statistics calculated based on small groups can be expected to be lower than those calculated based on larger groups. For this reason, the values presented in Tables 9-6 and 9-7 should be interpreted with caution.

Table 9-6. 2009-10 MHSAs: Science Summary of Decision Accuracy (and Consistency) Results by Subject and Grade—Conditional on Cutpoint

Subject	Grade	Substantially Below/Partially			Partially/Proficient			Proficient/Proficient with Distinction		
		Accuracy (consistency)	False positive	False negative	Accuracy (consistency)	False positive	False negative	Accuracy (consistency)	False positive	False negative
Science	11	0.91 (0.87)	0.05	0.05	0.89 (0.85)	0.06	0.05	0.98 (0.97)	0.01	0

Table 9-7. 2009-10 MHSAs: Science Summary of Decision Accuracy (and Consistency) Results by Subject and Grade—Overall and Conditional on Performance Level

Subject	Grade	Overall	Kappa	Conditional on level			
				Substantially Below Proficient	Partially Proficient	Proficient	Proficient with Distinction
Science	11	0.79 (0.72)	0.57	0.85 (0.79)	0.52 (0.42)	0.87 (0.81)	0.80 (0.57)

Chapter 10. MHSa SCORE REPORTING

All students who participate in the MHSa receive score reports that contain Maine-specific scores on the SAT and science tests. Those students who take the SAT under college-reportable conditions (i.e., without Maine purposes only [MPO] accommodations) also receive SAT score reports directly from the College Board.

10.1 PRIMARY REPORTS

The primary reports for the 2009–10 MHSa are listed below.

- Individual Student Report for Parents/Guardians
- Student Results Label
- Interactive Reporting
- 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 Q.

10.2 INDIVIDUAL 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 back side provides a complete picture of an individual student's performance on the MHSa, 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.

10.3 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.

10.4 INTERACTIVE REPORTING

There are four interactive reports that were available: Item Analysis Report, Achievement Level Summary, Released Items Summary Data, and Longitudinal Data. Each of these interactive reports is described in the following sections. Sample interactive reports are provided in Appendix R. To access these four interactive reports, the user clicked the interactive tab on the home page of the system and selected the report desired from the drop down menu. Next, the user applied basic filtering options, such as the name of the SAU or school and the grade level test, to open the specific report. At this point, the user had the option of printing the report for the entire grade level or applying advanced filtering options to select a subgroup of students to analyze. Advanced filtering options include gender, ethnicity, limited English proficient (LEP), IEP, and SES. All interactive reports, with the exception of the Longitudinal Data Report, allowed the user to provide a custom title for the report.

10.4.1 Item Analysis Report

The Item Analysis Report provides a roster of all students in a school and provides performance on the items that are released to the public. The student names and identification numbers are listed as row headers down the left side of the report.

For each student, multiple-choice items are marked either with a plus sign (+), indicating that the student chose the correct multiple-choice response, or a letter (from A to D), indicating the incorrect response chosen by the student. For constructed-response items, the number of points earned is shown. All responses to released items are shown in the report, regardless of the student's participation status. The columns on the right side of the report show the Total Test results, broken into several categories. Content Strand Points Earned columns show points earned by the student in each content area subcategory relative to total possible points. A Total Points Earned column is a summary of all points earned and total possible points in the content area. The last two columns show the student's scaled score and achievement level. Students reported as Not Tested are given a code in the achievement level column to indicate the reason the student did not test. It is important to note that not all items used to compute student scores are included in this report; only released items are included. At the bottom of the report, the average percentage correct for each multiple-choice item and average scores for the short-answer and constructed-response items are shown for the school, SAU, and state. When advanced filtering criteria are applied by the user, the School and SAU Percent Correct/Average Score rows at the bottom of the report are blanked out and only the Group row and the State

row for the group selected will contain data. This report can be saved, printed, or exported as a PDF, XLS, or CSV file.

The Item Analysis Roster is confidential and should be kept secure within the school and SAU. FERPA requires that access to individual student results be restricted to the student, the student's parents/guardians, and authorized school personnel.

10.4.2 Achievement Level Summary

The Achievement Level Summary provides a visual display of the percentages of students in each achievement level for a selected grade. The four achievement levels are represented by various colors in a pie chart. A separate table is also included below the chart that shows the number and percentage of students in each achievement level. This report can be saved, printed, or exported as a PDF or JPG file.

10.4.3 Item Analysis Data

The Released Items Summary Data report is a school-level report that provides a summary of student responses to the released items for a selected grade. The report is divided into two sections by item type (multiple-choice and open-response). For multiple-choice items, the total number/percent of students who answered the item correctly and the number of students who chose each incorrect option or provided an invalid response are reported. An invalid response on a multiple-choice item is defined as "the item was left blank" or "the student selected more than one option for the item." For open-response items, point value and average score for the item are reported. Users are also able to view the actual released items within this report. If a user clicks on a particular magnifying glass icon next to a released item number, a pop-up box will open displaying the released item.

10.4.4 Longitudinal Data Report

The Longitudinal Data Report is a confidential student-level report that provides individual student performance data for multiple test administrations. The state-assigned student identification number is used to link students across test administrations. Student performance on future test administrations will be included on this report over time. This report can be saved, printed, or exported as a PDF file for a single student or for all students within a group.

10.5 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. 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
- Limited English Proficiency 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

10.6 DECISION RULES

To ensure that reported results for the 2009–10 MHSA were 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 MHSA 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 MHSA can be found in Appendix S.

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.

10.7 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 MHSA 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 the 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 MHSA 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.

Chapter 11. VALIDITY RESEARCH ON THE MHSA SAT COMPONENT

This chapter seeks to bring together a wide range of validity evidence regarding the MHSA SAT Component 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 purpose of this chapter is to provide some of the more recent evidence supporting the interpretation of SAT scores. 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.

11.1 CONSTRUCT VALIDITY

The SAT is described variously as “a measure of the critical thinking skills you will need for academic success in college,”⁵ or as an assessment of “student reasoning based on knowledge and skills developed by the student in school coursework.”⁶ 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 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).

⁵ College Board. (2004). SAT preparation booklet 2004–2005 for the new SAT. (New York: Author, 2004), 3.

⁶ *About the new SAT*, retrieved from www.collegeboard.com on January 21, 2005.

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.

11.2 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—known as 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 11.4).

11.3 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, “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.⁷ 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 content from a third-year high school mathematics course, such as exponential growth, absolute value, and functional notation. The new test also places greater emphasis on other topics, such 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 2-12. 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.

⁷ College Board, *2005 College Bound Seniors: Total Group Profile Report* (New York: Author, 2005), Table 3-1.

11.4 WRITING

Although the College Board has previously offered tests of writing⁸ 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.

11.5 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, 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

⁸ 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).

11.6 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 give an advantage to 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.

11.7 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 data, indicating no significant changes in performance trends between the two tests.

11.8 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, and Ramist, 1990; Hezlett, Kuncel, Vey, Ahart, Ones, Campbell, and Camara, 2001; Young and 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 college 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 and Novick, 1968) and shrinkage (Rozeboom, 1978) were applied.

The relationship of each of the predictors with FGPA and ECGPA is shown in Table 11-1. The values in the table represent the weighted-average validity coefficients across all of the participating institutions.

Table 11-1. 2009–10 MHS: SAT Component—Weighted Average Correlations for Predictors with FGPA and ECGPA

<i>Predictor</i>	<i>FGPA</i>			<i>ECGPA</i>		
	<i>N</i>	<i>Corrected</i>	<i>Uncorrected</i>	<i>N</i>	<i>Corrected</i>	<i>Uncorrected</i>
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 11-2.

Table 11-2. 2009-10 MHS: SAT Component—Weighted Average Incremental Validity Results Across Institutions for Predicting First Year GPA (Model A)

	<i>Step</i>	<i>Adjusted</i>		<i>Unadjusted</i>	
		<i>R</i>	ΔR	<i>R</i>	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	0.02
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	0.02

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 11-2, 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 11-3.

Table 11-3. 2009-10 MHSAs: SAT Component—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 et al. (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 11-4.

Table 11-4. 2009-10 MHSA: SAT Component—Correlations of Test Scores and Student Self-Reports with College Writing Performance

<i>Predictor Variables</i>		<i>Total Writing Performance</i>			<i>Average Writing Performance</i>		
		<i>Total Sample</i> <i>N = 112</i>	<i>Females</i> <i>N = 69</i>	<i>Males</i> <i>N = 43</i>	<i>Total Sample</i> <i>N = 154</i>	<i>Females</i> <i>N = 93</i>	<i>Males</i> <i>N = 61</i>
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, and 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 SAT II 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 scores in the prediction of total writing performance and average writing performance are shown in Table 11-5.

Table 11-5. 2009-10 MHSA: SAT Component—Predictive Contributions of SAT-Verbal and SAT II: Writing Test

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

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

Note: Figures for *R* in parentheses corrected for shrinkage. Source: Breland, Kubota, and 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.

11.9 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.

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 11-6. 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, more than one-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 11-6. 2009-10 MHSA: SAT Component—Percentage of Students by Academic Intensity, HSGPA, SAT Score, and Level of College Selectivity

		<i>College Selectivity Level</i>				
		<i>Total</i>	<i>1 (low)</i>	<i>2</i>	<i>3</i>	<i>4 (high)</i>
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, and 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 11-7 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 11-7. 2009-10 MHSAs: SAT Component—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</i>	<i>N</i>	<i>Percent</i>	<i>Percent</i>
				<i>CGPA</i> ≥ 3.5	<i>CGPA</i> ≥ 2.5	<i>CGPA</i> ≥ 3.5	<i>CGPA</i> ≥ 2.5
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

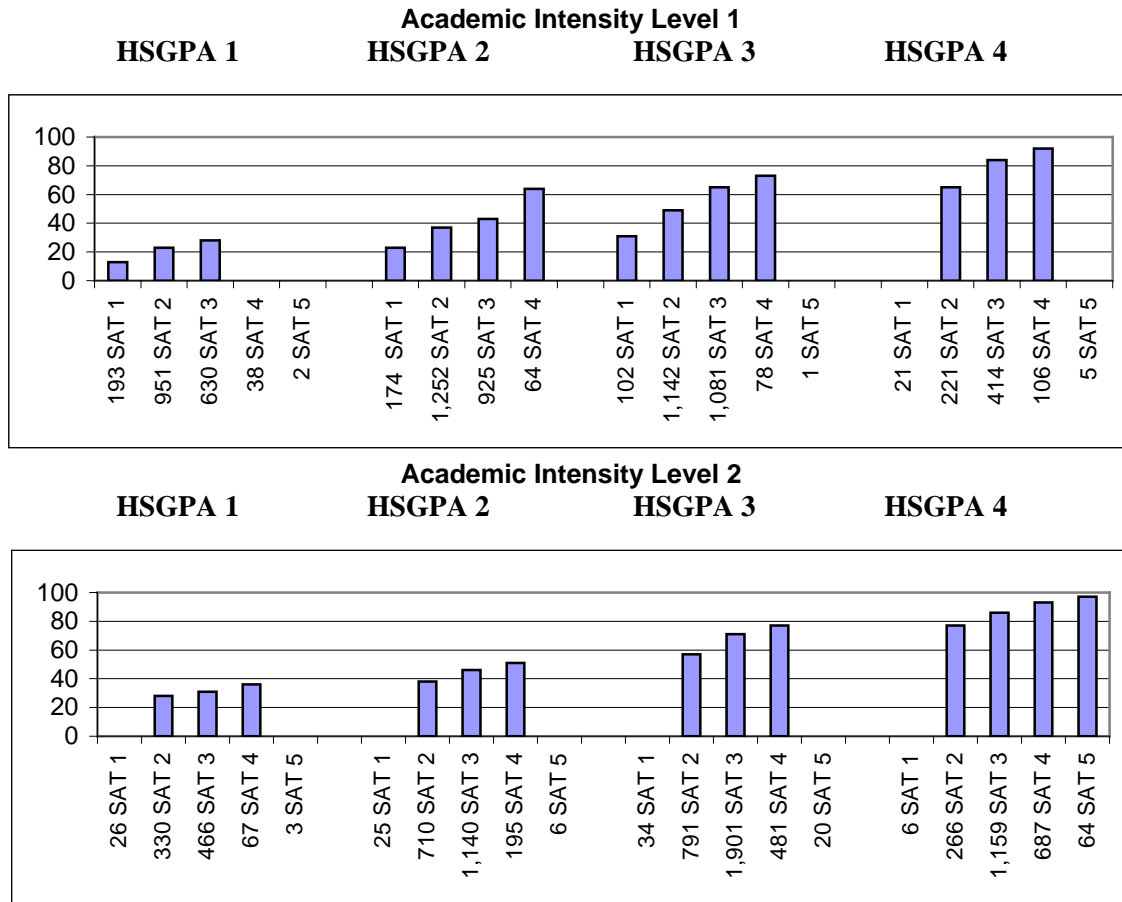
Source: Bridgeman, Pollack, and Burton (2004) Table 5

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 11-1 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).⁹ 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.

⁹ Academic intensity Level 3 is not included because of the small number of students in this category in Level 1 schools.

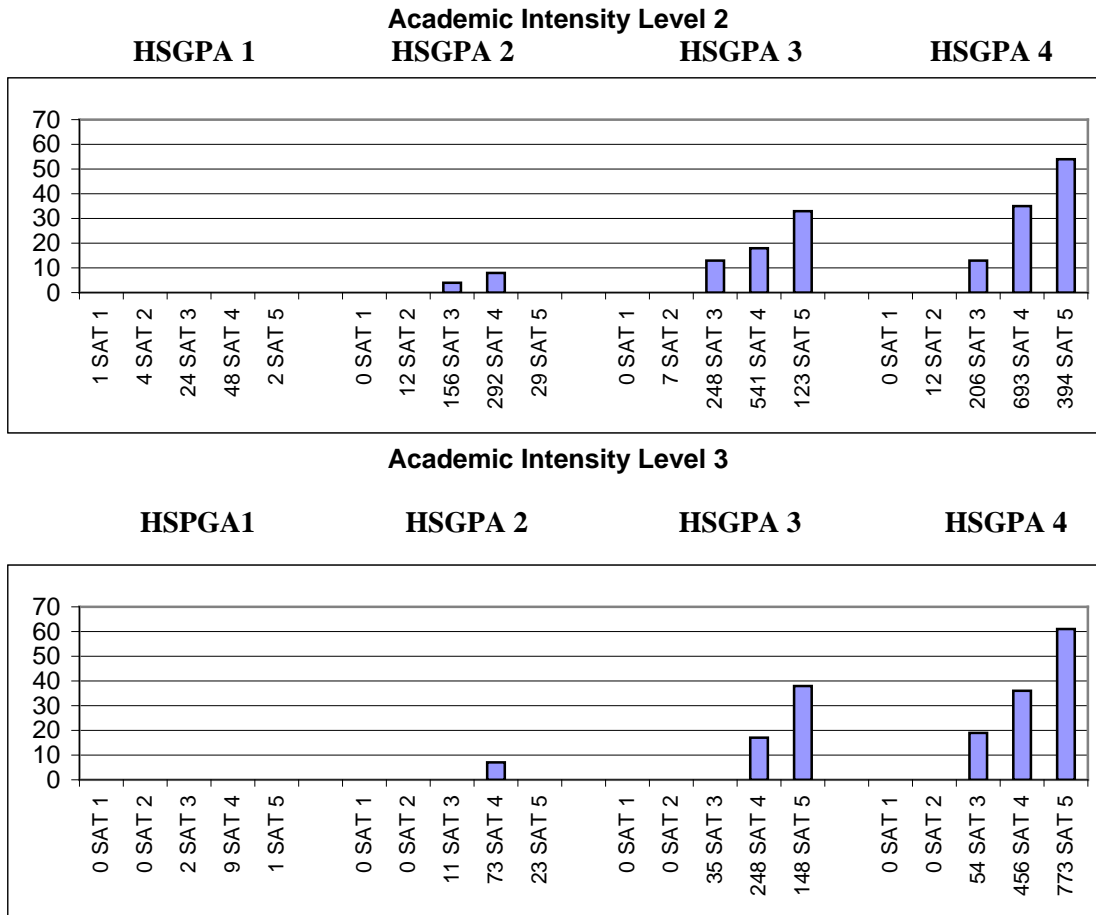
Figure 11-1. Level 1 Colleges—Percentage of Freshmen with CGPAs of 2.5 or Higher by the First Two Levels of Academic Intensity, HSGPA, and SAT Score



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, and Burton (2004), Figure 2 and Table A1.

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 11-2 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.

Figure 11-2. Level 4 Colleges—Percentage of Students with CGPAs of 3.5 or Higher by Highest Two Levels of Academic Intensity, HSGPA, and SAT Score



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, and Burton (2004), Figure 6 and Table A2.

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.

11.10 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 11-8. 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 11-8. 2009-10 MHSA: SAT Component—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 and 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, and 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.

11.11 DIFFERENTIAL VALIDITY FOR SUBGROUPS

A considerable amount of research in the last fifteen years has examined the question of whether 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, Lewis, and McCamley (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 11-9. 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 11-9. 2009-10 MHSAs: SAT Component—Effectiveness by Student Group Correlation with FGPA

	<i>Gender</i>			<i>Ethnic Group</i>				
	<i>All Students</i>	<i>Male</i>	<i>Female</i>	<i>Native American</i>	<i>Asian American</i>	<i>African American</i>	<i>Hispanic</i>	<i>White</i>
<i>N</i>	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

11.11.1 Gender

Drawn from the Ramist et al. (1994) study, Table 11-9 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 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 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.

11.11.2 Race and Ethnicity

In the Ramist, Lewis, and McCamley-Jenkins (1994) study reported in Table 11-9, 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.

11.11.3 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, and 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, and 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 and Schneider, 2000) and account for the largest percentage of college freshmen with disabilities (Cahalan, Mandinach, and 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 and 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.

11.11.4 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.

11.12 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.

Chapter 12. VALIDITY OF THE MHSA SCIENCE COMPONENT

Because interpretations of test scores, and not a test itself, are evaluated for validity, the purpose of the *2009–10 MHSA Technical Report* is to describe several technical aspects of the MHSA in support of score interpretations (AERA et al., 1999). Each chapter contributes an important component in the investigation of score validation: test development and design; test administration; scoring, scaling, and equating; item analyses; reliability; and score reporting.

The MHSA science test, as described in Chapters 3 and 5, was written and aligned in its entirety to *Maine’s Learning Results* science accountability standards. MHSA science results are intended to facilitate inferences about student achievement on the 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 constructing a validity argument. The evidence around test content, response processes, internal structure, relationship to other variables, and consequences of testing speaks to different *aspects* of validity but are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

Evidence on test-content validity is meant to determine how well the assessment tasks represent the curriculum and standards for each content area and grade level. Content validation is informed by the item development process, including how the test blueprints and test items align to the curriculum and standards. Viewed through the lens provided by the standards, evidence based on test content was extensively described in Chapter 3. Item alignment with accountability standards; 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 questions are aligned by Maine educators to the 2007 *Maine’s Learning Results* (MLRs), and undergo several rounds of review for content fidelity and appropriateness. Items are presented to students in multiple formats (constructed-response and multiple-choice). Finally, tests are administered according to state-mandated standardized procedures, with allowable accommodations.

Chapter 5 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 7 added to the body of content validation evidence.

Evidence based on internal structure is presented in great detail in the discussions of item analyses, reliability, and scaling and equating in Chapter 9. Technical characteristics of the internal structure of the

assessments are presented in terms of classical item statistics (item difficulty, item-test correlation), differential item functioning analyses, dimensionality analyses, reliability, standard errors of measurement, and item response theory parameters and procedures. In general, 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 most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall.

Evidence based on the consequences of testing is addressed in the scaled score information in Chapter 9 and the reporting information in Chapter 10. Each of these chapters speaks to the efforts undertaken to promote accurate and clear information provided to the public regarding test scores. Scaled scores offer the advantage of simplifying the reporting of results across subsequent years. Achievement levels provide users with reference points for mastery, which is another useful and simple way to interpret scores. Several different standard reports are provided to stakeholders.

12.1 QUESTIONNAIRE DATA

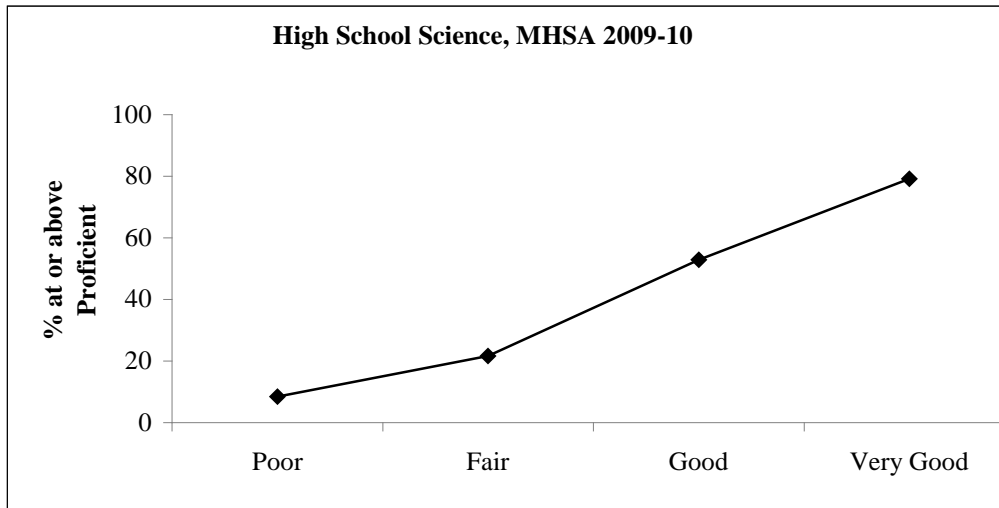
External validity of the MHSA is conveyed by the relationship of test scores and situational variables such as self-image, attitude toward content matter, and match of test questions to what is learned in school. These situational variables were all based on student questionnaire data collected during the administration of the MHSA. Note that no inferential statistics are included; however, because the numbers of students are quite large, differences in average scores may be statistically significant.

12.1.1 Self-image

Examinees were asked how they would rate themselves as a student in science. Figure 12-1 indicates a strong positive relationship between self image as a student and MHSA scores.

Question: Which of the following best describes how you rate yourself as a student in science?

Figure 12-1.

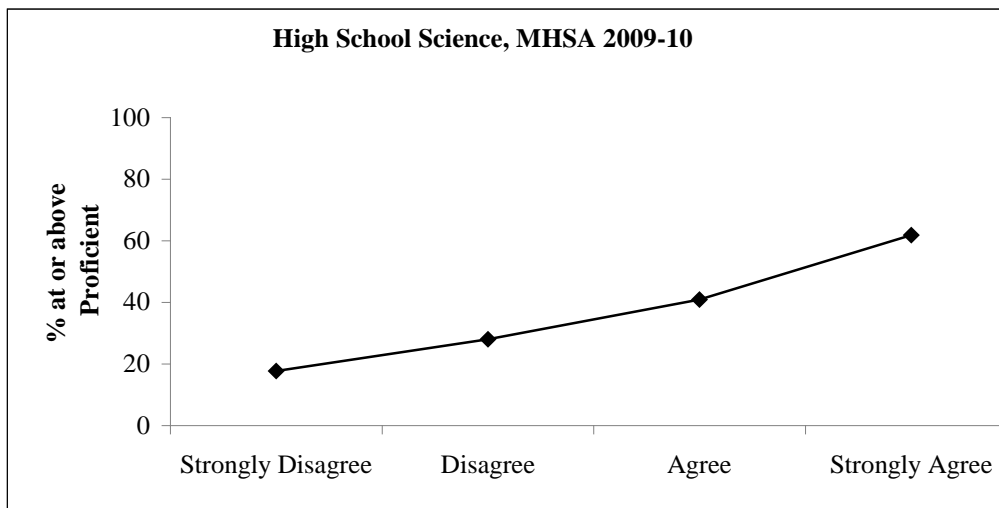


12.1.2 Attitude Towards Content Area

Students were asked how they felt about the statement “My knowledge of science will be useful to me as an adult.” Figure 12-2 indicates that students’ attitudes toward science are related positively to MHSA scores.

Question: My knowledge of science will be useful to me as an adult.

Figure 12-2.

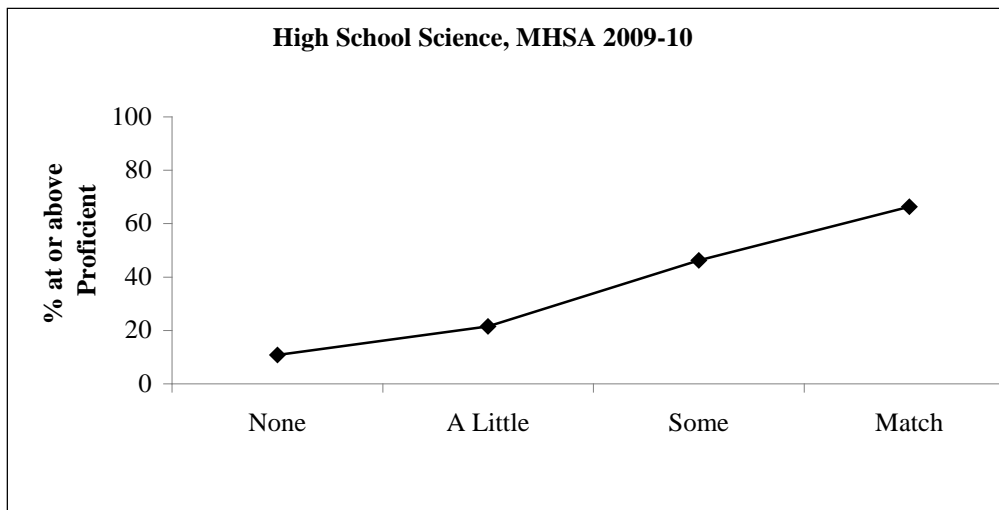


12.1.3 Match of Questions to What Is Learned in School

Students were asked how well the questions on the MHSA test matched what they had learned in school about science. Figure 12-3 indicates that there is a positive relationship between how well students feel the questions match what they've learned in science and MHSA scores.

Question: How well do the questions that you have just been given on this MHSA test match what you have learned in school about science?

Figure 12-3.

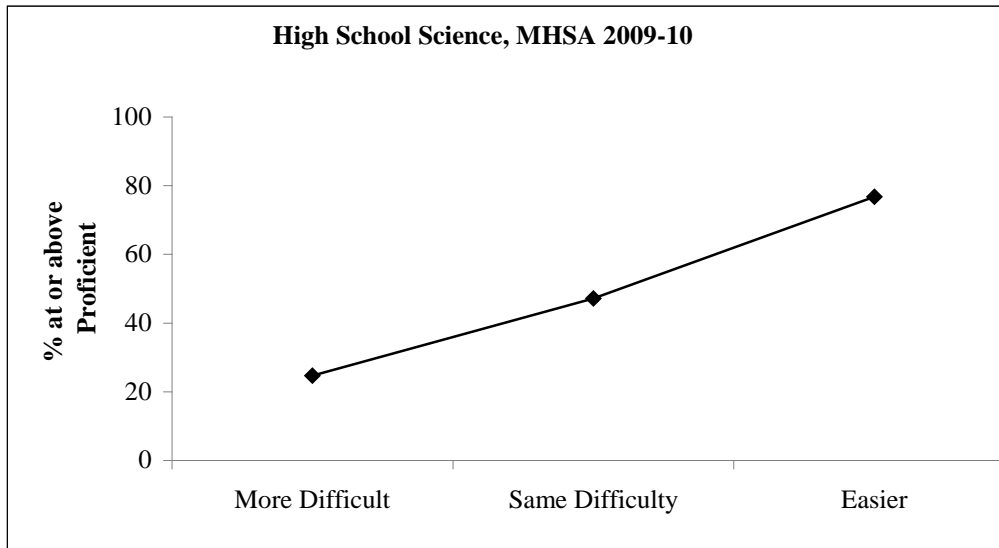


12.1.4 Difficulty of Assessment

Students were asked how difficult they found the test. Figure 12-4 indicates that there is a strong *negative* relationship between how difficult the students felt the items were and overall MHSA Science scores (i.e., students who found the test more difficult received lower scores than students who found the test easier).

Question: How difficult was this science test?

Figure 12-4.



The evidence presented in this report supports inferences of student achievement on the content represented in *Maine's Learning Results* and grade-level expectations for science for the purposes of program and instructional improvement and as a component of school accountability.

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 No. 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).
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 37-46.
- 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., & 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., & 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.
- Dorans, Liu & Hammond (2000). – referred to in text on page 78, but not listed here in reference list. Please add.
- 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*.
- Draper and Smith, 1998 - referenced in text on page 108, but not listed here. Please add.
- 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.
- Hambleton, R. K., Swaminathan, H., & Rogers, J. H. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage Publications, Inc.
- Hambleton, R. K., & van der Linden, W. J. (1997). *Handbook of modern item response theory*. New York, NY: Springer-Verlag.
- 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, 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.
- 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.
- 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, 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.

