DOCUMENT RESUME

ED 459 184	TM 033 460
TITLE	MCAS Technical Report, 1999. Massachusetts Comprehensive Assessment System.
INSTITUTION	Massachusetts State Dept. of Education, Boston.
PUB DATE	2000-11-00
NOTE	112p.; Cover page varies. For the 1998 technical report, see ED 451 227.
PUB TYPE	Guides - Non-Classroom (055) -- Numerical/Quantitative Data (110)
EDRS PRICE	MF01/PC05 Plus Postage.
DESCRIPTORS	*Elementary School Students; Elementary Secondary Education; English; Language Arts; Mathematics; Psychometrics; Reliability; Sciences; *Scoring; *Secondary School Students; State Programs; *Test Construction; Test Interpretation; *Testing Programs; *Validity
IDENTIFIERS	Massachusetts; *Massachusetts Comprehensive Assessment System

ABSTRACT

This manual documents the technical aspects of the Massachusetts Comprehensive Assessment System (MCAS). In May 1999, Massachusetts public school students in grades 4, 8, and 10 participated in the second annual administration of the MCAS tests in English Language Arts, Mathematics, and Science & Technology. This report provides information about the technical quality of these assessments. It includes a description of the processes used to develop, administer, and score the tests and to analyze test results. The report may be used by educated laypersons, but the intended audience is experts in psychometrics and educational research. The report assumes working knowledge of measurement concepts such as reliability and validity and statistical concepts such as correlation and central tendency. The report contains these sections: (1) "Assessment Development"; (2) "Test Administration"; (3) "Development and Reporting of Scores"; (4) "Technical Characteristics"; and (5) "References." (Contains 6 figures 56 tables, and 26 references.) (SLD)

MASSACHUSETTS DEPARTMENT OF EDUCATION

MCAS

MASSACHUSETTS
COMPREHENSIVE
ASSESSMENT
SYSTEM

1999 MCAS Technical Report

November 2000

Massachusetts Department of Education

Massachusetts Department of Education

This document was prepared by the Massachusetts Department of Education
Dr. David P. Driscoll, Commissioner of Education

The Massachusetts Department of Education, an Affirmative Action employer, is committed to ensuring that all of its programs and facilities are accessible to all members of the public. We do not discriminate on the basis of age, color, disability, national origin, race, religion, sex, or sexual orientation.

This document printed on recycled paper

350 Main Street Malden, Massachusetts 02148-5023 #781-338-3000

TABLE OF CONTENTS

CHAPTER 1
BACKGROUND AND OVERVIEW

PURPOSE OF THIS MANUAL

The purpose of this technical manual is to document the technical aspect of the Massachusetts Comprehensive Assessment System (MCAS). In May 1999, students in grades 4, 8, and 10 participated in the second annual administration of the MCAS tests in English language arts, mathematics, and science and technology. Also administered to grades 8 and 10 students were the history and social science tests[1]. This report provides information about the technical quality of those assessments. This includes a description of the processes used to develop, administer, and score the tests and to analyze the test results. This report will serve as a guide for replicating and/or improving the procedures in subsequent years.

While some parts of this technical report may be used by educated laypersons, the intended audience is experts in psychometrics and educational research. The report assumes working knowledge of measurement concepts such as reliability and validity, and statistical concepts such as correlation and central tendency. For some chapters, the reader is presumed to have basic familiarity with advanced topics in measurement and statistics.

THE EDUCATION REFORM LAW OF MASSACHUSETTS OF 1993

The Massachusetts Comprehensive Assessment System (MCAS) was developed in response to the Education Reform Law of Massachusetts of 1993. Three sections of the reform act that are particularly relevant to the assessment program are excerpted and presented below.

> *The board shall direct the commissioner to institute a process to develop academic standards for the core subjects of mathematics, science and technology, history and social science, English, foreign languages and the arts. The standards shall cover grades kindergarten through twelve and shall clearly set forth the skills, competencies and knowledge expected to be possessed by all students at the conclusion of individual grades or clusters of grades. The standards shall be*

[1] Although a history and social science test was administered to grade 10 students the results were not reported in the two primary reporting media: performance levels and scaled scores. This decision was made because the school have not had the chance to implement the two-year world history curriculum assessed on the test.

> *formulated so as to set high expectations of student performance and to provide clear and specific examples that embody and reflect these high expectations, and shall be constructed with due regard to the work and recommendations of national organizations, to the best of similar efforts in other states, and to the level of skills, competencies and knowledge possessed by typical students in the most educationally advanced nations. The skills, competencies and knowledge set forth in the standards shall be expressed in terms which lend themselves to objective measurement, define the performance outcomes expected of both students directly entering the work force and of students pursuing higher education, and facilitate comparisons with students of other states and other nations.*
>
> *The "competency determinations" shall be based on the academic standards and curriculum frameworks for tenth graders in the areas of mathematics, science and technology, history and social science, and English, and shall represent a determination that a particular student has demonstrated mastery of a common core of skills, competencies and knowledge in these areas, as measured by the assessment instruments described in section one I. Satisfaction of the requirements of the competency determination shall be a condition for high school graduation. If the particular student's assessment results for the tenth grade do not demonstrate the required level of competency, the student shall have the right to participate in the assessments program the following year or years.*
>
> *... comprehensive diagnostic assessment of individual students shall be conducted at least in the fourth, eighth and tenth grades. Said diagnostic assessments shall identify academic achievement levels of all students in order to inform teachers, parents, administrators and the students themselves, as to individual academic performance. The board shall develop procedures for updating, improving or refining the assessment system. The assessment instruments shall be designed to avoid gender, cultural, ethnic or racial stereotypes and shall recognize sensitivity to different learning styles and impediments to learning. The system shall take into account on a nondiscriminatory basis the cultural and language diversity of students in the commonwealth and the particular circumstances of students with special needs. Said system shall comply with federal requirements for accommodating children with special needs. All potential English proficient students from language groups in which programs of transitional bilingual education are offered under chapter seventy-one A shall also be allowed opportunities for assessment of their performance in the language which best allows them to demonstrate educational achievement and mastery. For the purposes of this section, a "potential English proficient student" shall be defined as a student who is not able to perform ordinary class work in English; provided, however, that no student shall be allowed to be tested in a language other than English for longer than three consecutive years.*

CURRICULUM FRAMEWORKS

As required by the Educational Reform Act of 1993, the Massachusetts Department of Education developed and disseminated curriculum frameworks. These frameworks are intended to provide guidance for the reform of public education in Massachusetts by raising the standards and expectations of schools and students. The following four frameworks guided the development of MCAS test specifications (Massachusetts Department of Education, 1997a, 1997b, 1997c, 1997d):

- *English Language Arts Curriculum Framework*;
- *Mathematics Curriculum Framework: Achieving Mathematical Power*;
- *Science and Technology Curriculum Framework: Owning the Questions through Science and Technology*;
- *History and Social Science Curriculum Framework*.

English Language Arts

The English language arts standards are divided into four strands: language, literature, composition, and media. The framework also provides two suggested lists of authors, illustrators, and works.

Mathematics

The mathematics standards are divided into four content-based strands: number sense; patterns, relations, and functions; geometry and measurement; and statistics and probability. The framework also discusses four aspects of applying mathematical knowledge: problem solving, communication, reasoning, and connections.

Science and Technology

The science and technology standards are divided into four strands: inquiry; domains of science; technology; and science, technology, and human affairs. Domains of science is divided into three substrands: physical sciences, life sciences, and earth and space sciences. Technology is divided into two substrands: the design process and understanding and using technology.

History and Social Science

The history and social science standards are divided into four content-based strands: history, geography, economics, and civics and government. There are twenty learning standards related to these four learning strands.

PURPOSES OF THE MCAS

The statewide assessment program serves two main purposes. First, it is a tool for measuring the performance of individual students and schools against established state standards. Second, it is intended to improve classroom instruction by a) providing useful feedback about the quality of instruction and b) modeling effective assessment approaches that can be used in the classroom.

The Education Reform Act requires that, in addition to fulfilling local graduation requirements, students pass the state's grade 10 tests as a condition for receiving a high school diploma. The Massachusetts Board of Education has determined that this requirement will be applied for the first time to graduates of the Class of 2003. Students will be given multiple opportunities, if necessary, to pass the tests. The Board of Education has established that students in the class of 2003 will have to achieve a performance level of *Needs Improvement* or higher on the MCAS grade 10 English Language Arts and Mathematics tests.

Local educators should use results of the MCAS tests, together with results of local tests and assessments, to identify strengths and weaknesses in curriculum and instruction, and to determine the needs of individual students in order to serve them more effectively. In addition to MCAS results, local educators should make use of released MCAS test items, *The Massachusetts Comprehensive Assessment System Release of Spring 1999 Test Items (1999)*, and the *Test Item Analysis Report* (which contains student results for each of the questions provided in that year's release document). These resources can assist educators in developing and implementing instructional strategies designed to support the goal that all students attain the state's academic learning standards.

ORGANIZATION OF THIS MANUAL

The organization of this report is based on the conceptual flow of an assessment's life span; it begins with the initial test specification and addresses all the intermediate steps that lead to final score reporting. Section I covers the development of the MCAS tests. It consists of six chapters, covering general design issues, the specific designs of the English language arts, mathematics, science and technology, and history and social science assessments, and the test development process. Section II consists of one chapter describing the administration of the tests. Section III contains six chapters covering scoring, standard setting, equating, scaling, score reporting, and state results. Section IV presents three chapters addressing the technical characteristics of the tests. Topics covered include item analysis, reliability, and validity.

Because of the educational and political importance of high-stakes testing programs such as the MCAS, this technical report uses professional guidelines for evaluating and documenting the testing program, specifically the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 1985[2]) and the *Code of Fair Testing Practices in Education* (1988). The *Standards for Educational and Psychological Testing* covers technical standards for test development and evaluation, professional standards for test use, standards for particular applications (i.e., testing students of limited English proficiency and students with disabilities), and standards for administrative procedures (i.e., test administration, scoring and reporting, and protecting the rights of test takers). Table 1-1 shows the categories of standards from the *Standards for Educational and Psychological Testing* and shows where each category of standards is addressed in this technical manual.

[2] The 1985 standards were used because they were the latest editions when the test was developed.

Table 1-1
Location of Information Addressing Standards from
Standards for Educational and Psychological Testing

Standards		Location of Information
Technical Standards for Test Construction and Evaluation	Validity	Chapter 17
	Reliability and Errors of Measurement	Chapter 16
	Test Development and Revision	Chapters 2–7
	Scaling, Norming, Score Comparability, and Equating	Chapter 10-12 (Scaling and Equating, other topics not applicable)
	Test Publication: Technical Manuals and User's Guides	Chapters 1–17
Professional Standards for Test Use	General Principals of Test Use	Throughout technical manual
	Clinical Testing	Not applicable
	Educational and Psychological Testing in the Schools	Throughout technical manual
	Test Use in Counseling	Not applicable
	Employment Testing	Not applicable
	Professional and Occupational Licensure and Certification	Not applicable
	Program Evaluation	Not applicable for 1999 test
Standards for Particular Applications	Testing Linguistic Minorities	Chapter 8
	Testing People Who Have Handicapping Conditions	Chapter 8
Standards for Administrative Procedures	Test Administration, Scoring, and Reporting	Chapters 8, 9, 13
	Protecting the Rights of Test Takers	Not addressed in technical manual

The *Code of Fair Testing Practices in Education* covers developing appropriate tests, interpreting scores, striving for fairness, and informing test takers. Table 1-2 shows where each point covered by the *Code of Fair Testing Practices in Education* is addressed in this technical report (or where else the information is available).

Table 1-2
Location of Information Regarding Responsibilities for Test Developers in
Code of Fair Testing Practices in Education

Responsibility		Location of Information
Developing Appropriate Tests	Define what each test measures and what the test should be used for. Describe the populations for which the test is appropriate.	Chapters 1–6, 8; *MCAS Guides*
	Accurately represent the characteristics, usefulness, and limitations of each test for its intended purposes.	Chapter 2; *MCAS Guides*
	Explain relevant measurement concepts as necessary for clarity at the level of detail that is appropriate for the intended audiences.	Chapters 9-12, 15-17
	Describe the process of test development. Explain how the content and skills to be tested were selected.	Chapter 3–7
	Provide evidence that the test meets its intended purpose(s).	Chapters 2–6, 17
	Provide representative samples or complete copies of test questions, directions, answer sheets, manuals, and score reports to qualified users.	Chapter 13; *Release of Spring 1999 Test Items*
	Indicate the nature of the evidence obtained concerning the appropriateness of each test for groups of different racial, ethnic, or linguistic backgrounds who are likely to be tested.	Chapter 15
	Identify and publish any specialized skills needed to administer each test and to interpret scores correctly.	Not Applicable
Interpreting Scores	Provide timely and easily understood score reports that describe test performance clearly and accurately. Also explain the meaning and limitations of reported scores.	Chapter 13
	Describe the population(s) represented by any norms or comparison group(s), the dates the data were gathered, and the process used to select the samples of test takers.	Chapter 8
	Warn users to avoid specific, reasonably anticipated misuses of test scores.	
	Provide information that will help users follow reasonable procedures for setting passing scores when it is appropriate to use such scores with the test.	Chapters 10-12
	Provide information that will help users gather evidence to show that the test is meeting its intended purpose(s).	Chapters 2–6, 17

Table 1-2
Location of Information Regarding Responsibilities for Test Developers in
Code of Fair Testing Practices in Education

	Responsibility	Location of Information
Striving for Fairness	Review and revise test questions and related materials to avoid potentially insensitive content or language.	Chapter 7
	Investigate the performance of test takers of different races, genders, and ethnic backgrounds when samples of sufficient size are available. Enact procedures that help to ensure that differences in performance are related primarily to the skills under assessment rather than to irrelevant factors.	Chapters 7, 15
	When feasible, make appropriately modified forms of tests or administration procedures available for test takers with handicapping conditions. Warn test users of potential problems in using standard norms with modified tests or administration procedures that result in noncomparable scores.	Chapter 7
Informing Test Takers	When a test is optional, provide test takers or their parents/guardians with information to help them judge whether the test should be taken, or if an available alternative to the test should be used.	Not Applicable
	Provide test takers the information they need to be familiar with the coverage of the test, the types of question formats, the directions, and appropriate test-taking strategies. Strive to make such information equally available to all test takers.	Not covered in this manual[3]
	Provide test takers or their parents/guardians with information about rights test takers may have to obtain copies of tests and completed answer sheets, retake tests, have tests rescored, or cancel scores.	Not covered in this manual
	Tell test takers or their parents/guardians how long scores will be kept on file and indicate to whom and under what circumstances test scores will or will not be released.	Not covered in this manual
	Describe the procedures that test takers or their parents/guardians may use to register complaints and have problems resolved.	Not covered in this manual

Despite the many pages of tables, figures, and text in this manual, it is beyond the scope of this report to provide all available details about the MCAS. However, details that are pertinent to understanding the technical quality of the MCAS are included in the appendices or referenced in this manual.

[3] Much information is provided to teachers and administrators who are responsible for developing and implementing local curricula. Thus, responsibility for communicating in advance the coverage of the MCAS rests on schools. Nonetheless, the Department of Education makes information directly available to parents or guardians through the Internet and by working with the news media throughout the state.

SECTION I
ASSESSMENT DEVELOPMENT

CHAPTER 2
OVERVIEW OF TEST DESIGN

According to the *Standards of Educational and Psychological Testing* (1985, p. 9), the construct that a test is intended to measure should be embedded in a conceptual framework. This chapter discusses the conceptual framework that was used to design the MCAS assessments. The *Standards* (1985) also states (p. 25) that specifications used in constructing the test should be stated clearly. This chapter describes the specifications used for test construction. The MCAS test design has been explicated previously in two sets of documents: The *Curriculum Frameworks*, which present the learning standards intended to guide the development of local curriculum, and the *Guides to the Massachusetts Comprehensive Assessment System*, which describe what will be on the test. This chapter will summarize pertinent information from those two sets of materials and provide some additional detail.

CURRICULUM FRAMEWORKS

The Education Reform Law of Massachusetts stipulates that the MCAS be based on the *Curriculum Frameworks* for English language arts, mathematics, science and technology, and history and social science. The Department of Education convened committees of educators[4] from around the state to work with the Department to develop the learning standards based on the *Curriculum Frameworks.*

GUIDES TO THE MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM

To design the assessments, the *Curriculum Frameworks* were evaluated to determine for each subject area which dimensions could be adequately assessed in an on-demand paper-and-pencil test. The

[4] Members of different MCAS committees are listed in Appendix A.

product of this process was the *Guide to the Massachusetts Comprehensive Assessment System*[5] for each test (here called the *MCAS Guides).* The *MCAS Guides* provided the foundation for the test specifications that detail what each test will cover and emphasize, including the content strands (subject areas) and question types to be used in the MCAS.

ITEM TYPES

Every item type has its strengths and weaknesses. To ensure the strongest possible program for the May 1999 tests, each MCAS test used one or more of four different item types: multiple-choice, short answer, open response, and writing prompt.

Multiple-choice questions are highly efficient in terms of testing time, and thus allow for a breadth of content coverage. Multiple-choice questions, however, may be susceptible to guessing and, for tests requiring computation (much of mathematics and for some aspects of science) to back solving. That is, instead of using the intended solution strategy, students can insert each choice into the problem and rule out incorrect options, one by one. MCAS multiple-choice items were scored one point if correct and zero points if incorrect.

Short-answer questions require responses ranging from a few words or a number to several sentences. They are relatively immune to random guessing and back solving. For these reasons, MCAS used short-answer questions as part of the mathematics assessment. MCAS short-answer items were scored on a zero to one scale.

Open-response (extended-response) questions invite students to demonstrate not only their knowledge of facts and comprehension about a subject, but also how they can apply their

[5] Massachusetts Department of Education (1998b), *Guide to the Massachusetts Comprehensive Assessment System: English Language Arts*, Malden.

Massachusetts Department of Education (1998c), *Guide to the Massachusetts Comprehensive Assessment System: Mathematics.*

Massachuset ts Department of Education (1998d), *Guide to the Massachusetts Comprehensive Assessment System: Science and Technology.*

Massachusetts Department of Education (1998e), *Guide to the Massachusetts Comprehensive Assessment System: History and Social Science.*

knowledge. Open-response questions can take many forms, but they all require students to construct a detailed or descriptive answer (usually up to half a page long), and take between ten and fifteen minutes to complete. MCAS open-response questions were all scored on a zero to four scale.

MCAS writing prompts require students to write a composition , which is then evaluated for topic development and use of standard English conventions. Features of the MCAS writing prompts are described in Chapter 3 (in the section titled "Composition"), and scoring of the writing prompts is discussed in Chapter 9.

COMMON-MATRIX DESIGN

MCAS test questions are assigned to either the common or matrix-sampled portions of the tests. Common test questions are those that were identical in all twelve forms of the test at each grade level. Approximately eighty percent of the questions on any given test form were common questions. All individual student results are based exclusively on common questions; thus, the performance of every student at a grade level is based on identical questions. In addition, performance level results and average scaled scores for schools and districts are based exclusively on common questions.

The remaining twenty percent of the MCAS test questions in each test form were matrix-sampled questions, which differed across the twelve test forms at each grade level tested. Matrix-sampled questions serve three primary purposes. First, starting in 1999, they serve as the basis for equating tests from year to year. This allows for comparisons of performance at the school and district levels over time. Second, matrix-sampled questions, when combined with common questions, allow reporting in greater depth and detail for a broader range of the curriculum than is possible with common questions only. Results from the matrix-sampled questions and common questions are aggregated at the school and district levels to produce subject area subscores. Third, matrix-sampling allows for the field-testing of new items under operational testing conditions.

Common questions are publicly released following each year's test administration to inform local decisions about curriculum and instruction.[6] Released common questions are replaced each year with either questions from the previous year's matrix-sampled section.

The distribution of common and matrix-sampled questions for each grade level is shown in Table 2-1.

Table 2-1
May 1999 MCAS
Number of Test Questions in Each Content Area by Question Type and Function

Question Type: MC = Multiple- Choice, SA = Short Answer, OR = Open Response, WP = Writing Prompt

Grade	Question Function	Content Area									
		English Language Arts			Mathematics			Science & Technology		History & Social Sciences	
		MC	OR	WP	MC	SA	OR	MC	OR	MC	OR
4	Common	36	4	1	29	5	5	34	5	-	-
	Matrix	12	2	0	7	1	1	7	1	-	-
	Total	48	6	1	36	6	6	41	6	-	-
8	Common	36	4	1	29	5	5	34	5	34	5
	Matrix	12	2	0	7	1	1	7	1	7	1
	Total	48	6	1	36	6	6	41	6	41	6
10	Common	36	4	1	32	4	6	36	6	33	6
	Matrix	12	2	0	7	1	1	8	1	15	3
	Total	48	6	1	39	5	7	44	7	48	9

TEST SESSION STRUCTURE

Within each subject, test questions were organized in separate 45- or 60-minute sessions. The number of questions per session was based on estimated time spent on each type of question. For reading (language and literature), the length of the selection was also factored in. However, Department policy was to provide students with as much time as they could use productively (and without compromising schools' administration constraints). The amount of additional time per

[6] Massachusetts Department of Education (1999). *The Massachusetts Comprehensive Assessment System: Release of May 1999 Test Items.*

session that was generally considered reasonable ranged from five minutes to one-half hour. The number of sessions administered at each grade level in each subject area is shown in Table 2-2.

Table 2-2
May 1999 MCAS
Number Test Sessions*
Administered at Each Grade Level by Subject Area

Subject	Grade 4	Grade 8	Grade 10
English Language Arts – Composition	2	2	2
English Language Arts – Language and Literature	3	3	3
Mathematics	2	3	3
Science & Technology	2	3	3
History and Social Science	1**	3	3
All Subjects	10	14	14

*The recommended time per session for grades 8 and 10 was 45 minutes. The recommended time per session for grade 4 is 60 minutes with the exception of the English Language Arts – Composition sessions, which was 45 minutes per session.
**Question tryout.

MCAS 1999 tests were administered using three separate student booklets:

- English Language Arts Composition
- English Language Arts/Mathematics
- Science & Technology/History and Social Science

Each student used five separate answer booklets – one for each content area.

The English language arts test has one composition component only, administered in two consecutive 45-minute test sessions. In the first session, students were required to write a draft of a long composition in response to a writing prompt. In the second session, students revised the draft of their compositions to produce their final. There was one writing prompt administered for each grade. This prompt was administered to all students. The English language arts composition test was administered more than two weeks earlier than the other content areas.

The language and literature portions of the English language arts test contained reading passages followed by multiple-choice and open-response questions.

Mathematics tests in each grade level included multiple-choice, short-answer, and open-response questions. Both sessions of the grade 4 mathematics test contains short answer questions. For grades 8 and 10, short answer questions only appeared in the first session. For all sessions of grades 8 and 10 mathematics test, multiple-choice questions appear first in each session followed by short answer and/or open response questions. Each session of the grade 4 mathematics test starts with a series of multiple-choice questions followed by short answer and open response questions, then another series of multiple-choice questions followed by and open response question.

Science and technology sessions for all grades included multiple-choice and open-response questions only. Multiple-choice questions appeared first in each session, followed by open-response questions.

The grades 8 and 10 history and social science assessment are composed of multiple-choice and open response common and matrix items administered in three 45-minute sessions. In each session, the multiple-choice questions appeared first followed by open response questions.

CHAPTER 3
DESIGN OF THE ENGLISH LANGUAGE ARTS ASSESSMENT

LEARNING STANDARDS

Table 3-1 presents the English language arts learning standards from the *English Language Arts Curriculum Framework*.

Table 3-1
English Language Arts Learning Standards

Language Strand	1	Use agreed-upon rules for informal and formal discussions in small and large groups.
	2	Pose questions, listen to the ideas of others, and contribute their own information or ideas in group discussions and interviews in order to acquire new knowledge.
	3	Make oral presentations that demonstrate appropriate consideration of audience, purpose, and the information to be conveyed.
	4	Acquire and use correctly an advanced reading vocabulary of English words, identifying meanings through an understanding of word relationships.
	5	Identify, describe, and apply knowledge of the structure of the English language and standard English conventions for sentence structure, usage, punctuation, capitalization, and spelling.
	6	Describe and analyze how oral dialects differ from each other in English, how they differ from written standard English, and what role standard American English plays in informal and formal communication.
	7	Describe and analyze how the English language has developed and been influenced by other languages.

Table 3-1
English Language Arts Learning Standards

Literature Strand	8	Decode accurately and understand new words encountered in their reading materials, drawing on a variety of strategies as needed and then use these words accurately in speaking and writing.
	9	Identify the basic facts and essential ideas in what they have read, heard, or viewed.
	10	Demonstrate an understanding of the characteristics of different genres.
	11	Identify, analyze, and apply knowledge of theme in literature and provide evidence from the text to support their understanding.
	12	Identify, analyze, and apply knowledge of the structure and elements of fiction and provide evidence from the text to support their understanding.
	13	Identify, analyze, and apply knowledge of the structure, elements, and meaning of non-fiction or informational material and provide evidence from the text to support their meaning.
	14	Identify, analyze, and apply knowledge of the structure, elements, and theme of poetry and provide evidence from the text to support their understanding.
	15	Identify and analyze how an author's choice of words appeals to the senses, creates imagery, suggests mood, and sets tone.
	16	Compare and contrast similar myths and narratives from different cultures and geographic regions.
	17	Interpret the meaning of literary works, nonfiction, films, and media by using different critical lenses and analytic techniques.
	18	Plan and present effective dramatic readings, recitations, and performances that demonstrate appropriate consideration of audience and purpose.
Composition Strand	19	Write compositions with a clear focus, logically related ideas to develop it, and adequate supporting detail.
	20	Select and use appropriate genres, modes of reasoning, and speaking styles when writing for different audiences and rhetorical purposes.
	21	Improve organization, content, paragraph development, level of detail, style, tone, and word choice in revising their compositions.
	22	Use their knowledge of standard English conventions for sentence structure, usage, punctuation, capitalization, and spelling to edit their writing.
	23	Use self-generated questions, note-taking, summarizing, précis writing, and outlining to enhance learning when reading or writing.
	24	Use open-ended research questions, different sources of information, and appropriate research methods to gather information for their research projects.
	25	Develop and use rhetorical, logical, and stylistic criteria for assessing final versions of their compositions or research projects before presenting them to varied audiences.
Media Strand	26	Obtain information by using a variety of media and evaluate the quality of the information obtained.
	27	Explain how techniques used in electronic media modify traditional forms of discourse for different aesthetic and rhetorical purposes.
	28	Design and create coherent media productions with a clear focus, adequate detail, and consideration of audience and purpose.

CONTENT COVERAGE

The *Guide to the Massachusetts Comprehensive Assessment System: English Language Arts* identified the following standards to be assessed by the MCAS on-demand tests: language strand 4–7, literature strand 8–17, and composition strand 19–22.

ITEM TYPES

The *MCAS Guide* also presented the number of items by item type, component, and grade. Table 3-2 presents this information.

Table 3-2
May 1999 MCAS
English Language Arts
Distribution of Questions (Number per Student) by Component and Grade Level

Mode of Assessment	Language and Literature Component			Composition Component		
	Grade 4	Grade 8	Grade 10	Grade 4	Grade 8	Grade 10
Multiple-choice questions	48	48	48	0	0	0
Open-response questions*	6	6	6	0	0	0
Writing prompts	0	0	0	1	1	1

* Open-response questions assess learning standards from the literature strand only.

COMPOSITION

The composition component of the MCAS English language arts assessment included a long composition administered in two consecutive sessions totaling approximately 90 minutes

The long composition was structured to include some of the key elements of the writing process: drafting, revising, and finalizing. Consequently, the long composition was administered in two consecutive administration periods on the same school day, separated by a short break. In the first administration period, students prepared a first draft of their writing. Students were provided with space in the test booklet to generate and organize ideas and draft their writing. Following the break, students returned to revise and finalize their compositions during the second administration period.

The long composition prompt focused on a different writing mode at each grade: Grade 4, narrative; Grade 8; persuasive; and Grade 10, literary analysis.

READING SELECTIONS

MCAS selections are classified into one of two categories: literary, and non-narrative nonfiction. Table 3-3 describes these two genres.

Table 3-3
Genre of MCAS Selections

Literary	Non-Narrative, Nonfiction
● fiction ● poetry ● drama ● nonfiction ● essays ● biographies ● autobiographies	● instructions ● informational reports and articles ● letters ● interviews ● reviews ● essays ● speeches ● editorials ● critiques (emphasis on exposition in earlier grades, moving toward persuasive structures at higher grades)

Arguments can be made that some selections, especially essays or memoirs, can fit either category. When that happened, the Assessment Development Committee decided the classification on an individual basis.

In addition to selection genre, the *English Language Arts Curriculum Framework* (1997) provided two lists of suggested authors, illustrators, and works, referred to as its Appendix A and Appendix B. Its Appendix A was intended to reflect our "common literary and cultural heritage" and its Appendix B was planned to reflect "contemporary American and world literature." Table 3-4 presents the percent of selections broken down by genre and source (Appendix A, Appendix B, and other).

Table 3-4
Percent of Selections by Genre and Source

Grade	Literary			Non-Narrative NonFiction		
	Appendix A	Appendix B	Other	Appendix A	Appendix B	Other
4	25	13	12	0	0	50
8	30	15	15	0	0	40
10	30	15	15	5	15	20

DETAILED SPECIFICATIONS

Table 3-5 describes the exact number of items appearing in the 1999 MCAS English language arts assessment.

Table 3-5
May 1999 MCAS
Detailed Specifications for English Language Arts Assessment

(MC = Multiple-Choice; OR = Open-Response; WP = Writing Prompt)

Grade	Reporting Category	Common			Matrix (Total Across 12 Forms)		
		MC	OR	WP	MC	OR	WP
4	Language*	11	0	0	26	0	0
	Literature	25	4	0	118	24	0
	Composition	0	0	1	0	0	12
	Total	36	4	1	144	24	12
8	Language	7	0	0	21	0	0
	Literature	29	4	0	123	24	0
	Composition	0	0	1	0	0	12
	Total	36	4	1	144	24	12
10	Language	9	0	0	24	0	0
	Literature	27	4	0	120	24	0
	Composition	0	0	1	0	0	12
	Total	36	4	1	144	24	12

*In 1999, the grade 4 test included four "stand-alone" language items. These items appeared on the same pages as items associated with reading selections, but were not otherwise linked to the selections.

CHAPTER 4
DESIGN OF THE MATHEMATICS ASSESSMENT

LEARNING STANDARDS

The Massachusetts *Mathematics Curriculum Framework* (1996) presents four content strands: number sense; patterns, relations, and functions; geometry and measurement; and statistics and probability. These four content strands form the basis for mathematical problem solving, communication, reasoning, and connections.

Table 4-1 presents the mathematics content learning standards for pre-Kindergarten through grade 4, grades 5 through 8, and grades 9 and 10.

Table 4-1
Mathematics Learning Standards

	PreK–4	Grades 5–8	Grades 9 and 10
Number Sense	1. Number Sense and Numeration 2. Concepts of Whole Number Operations 3. Fractions and Decimals 4. Estimation 5. Whole Number Computation	1. Number and Number Relationships 2. Number Systems and Number Theory 3. Computation and Estimation	1. Discrete Mathematics 2. Mathematical Structure 3. Estimation
Patterns, Relations, and Functions	1. Patterns and Relationships 2. Algebra 3. Mathematical Structures	1. Patterns and Functions 2. Algebra	1. Algebra 2. Functions 3. Trigonometry
Geometry and Measurement	1. Geometry and Spatial Sense 2. Measurement	1. Geometry 2. Measurement	1. Geometry and Spatial Sense 2. Geometry from an Algebraic Perspective
Statistics and Probability	1. Statistics and Probability	1. Statistics 2. Probability	1. Statistics 2. Probability

CONTENT COVERAGE

The *Guide to the Massachusetts Comprehensive Assessment System: Mathematics* presented the approximate percentage of items for each content strand. Table 4-2 presents this information.

Table 4-2
Approximate Percent of Mathematics Test Questions by Content Strand

Content Strand	Grade 4	Grade 8	Grade 10
Number Sense	35	25	20
Patterns, Relations, and Functions	20	30	30
Geometry and Measurement	25	25	30
Statistics and Probability	20	20	20

MATHEMATICAL THINKING SKILLS

In addition to content knowledge, students are expected to demonstrate problem-solving and mathematical communication and reasoning skills, as well as skill at making connections between math content and its real-world application.[7] For the purposes of the MCAS assessment, these skills are grouped into three major areas: conceptual understanding, procedural knowledge, and problem solving.

Conceptual Understanding

Questions in this area assess student skills in labeling, verbalizing, and defining concepts; recognizing and generating examples and counter-examples; using models, diagrams, charts, and symbols to represent concepts; translating from one mode of representation to another; and comparing, contrasting, and integrating concepts.

Procedural Knowledge

Questions in this area assess student skills related to executing procedures and verifying results; explaining reasons for steps in procedures; recognizing correct and incorrect procedures; developing new procedures, or extending or modifying familiar ones; and recognizing situations in which a procedure is appropriate, necessary, or correctly applied.

Problem Solving

Questions in this area assess student skills in selecting appropriate mathematical concepts and procedures for both real-life and mathematical problem situations and appropriately applying these concepts and procedures; selecting and using appropriate problem-solving strategies; and verifying and generalizing solutions.

The *MCAS Guide* also addressed the distribution of test items by mathematical thinking skills. Table 4-3 presents this information for each grade level.

[7] The core concept of the Massachusetts *Mathematics Curriculum Framework* "is that students develop mathematical power through problem solving, communication, reasoning and [making] connections" (p. 1).

Table 4-3
Approximate Percent of Test Questions By Mathematical Thinking Skill

Mathematical Thinking Skill	Grade 4	Grade 8	Grade 10
Conceptual Understanding	40	30	30
Procedural Knowledge	40	25	25
Problem Solving	20	45	45

ITEM TYPES

Three types of mathematics questions were used at each grade level tested: multiple-choice, short answer, and open response. Short-answer questions require a brief response, usually a short statement or numeric solution to a computation or simple problem. Open-response questions require students to show their work in solving a problem and require responses in writing or in the form of a chart, table, diagram, or graph, as appropriate.

The approximate distribution of mathematics test questions by type for each grade level was presented in the *MCAS Guide* and is shown in Table 4-4.

Table 4-4
May 1999 MCAS
Approximate Distribution of Mathematics Questions by Type

Grade	Question Type	Number of Test Questions (per student test booklet)
4 and 8	Multiple-choice	36
	Short-answer	6
	Open-response	6
10	Multiple-choice	39
	Short-answer	5
	Open-response	7

DETAILED SPECIFICATIONS

Table 4-5 describes the exact number of items appearing in the 1999 MCAS mathematics assessment.

Table 4-5
May 1999 MCAS
Detailed Specifications for Mathematics Assessment

(MC = Multiple-Choice; SA = Short-Answer; OR = Open-Response)

Grade	Reporting Category	Common			Matrix (Total Across 12 Forms)		
		MC	SA	OR	MC	SA	OR
4	Number Sense	11	3	1	27	7	4
	Patterns, Numbers, and Relations	6	1	1	16	2	3
	Geometry and Measurement	5	1	2	23	3	3
	Statistics and Probability	7	0	1	18	0	2
	Total	29	5	5	84	12	12
8	Number Sense	8	2	1	19	9	2
	Patterns, Numbers, and Relations	6	2	2	29	2	3
	Geometry and Measurement	8	1	1	28	1	2
	Statistics and Probability	7	0	1	8	0	5
	Total	29	5	5	84	12	12
10	Number Sense	7	1	1	21	0	2
	Patterns, Numbers, and Relations	10	0	2	24	7	4
	Geometry and Measurement	9	1	2	23	4	4
	Statistics and Probability	6	2	1	16	1	2
	Total	32	4	6	84	12	12

CHAPTER 5
DESIGN OF THE SCIENCE AND TECHNOLOGY ASSESSMENT

LEARNING STANDARDS

The science and technology section of the MCAS is based on the learning standards described in the Massachusetts *Science & Technology Curriculum Framework* (1996). These learning standards were developed in collaboration with teachers, school and district administrators, scientists, technology experts, college faculty, parents, and representatives of business and community organizations across the state. The science and technology learning standards are too long to be included in this technical manual. The interested reader should refer to the Massachusetts *Science & Technology Curriculum Framework.*

Although science and technology are connected, they are not the same. Science, as stated in the Massachusetts *Science & Technology Curriculum Framework*, "involves the discovery of fundamental relationships that help explain the natural world" (p. 3). Technology, on the other hand, involves the creation of tools that expand people's capacity to solve problems and to use and control the natural and human-made environment.

The MCAS science and technology assessment is designed to assess two fundamental dimensions of learning: content knowledge and skills in using and applying science and technology.

CONTENT COVERAGE

Four major content strands identified by the *Science & Technology Curriculum Framework* serve as the foundation for the MCAS science and technology assessment and its reporting categories:

- Inquiry
- Domains of science:
- Physical sciences
- Life sciences
- Earth and space sciences
- Technology
- Science, technology, and human affairs

Table 5-2 shows the approximate distribution of MCAS science and technology questions by content strand and substrand for each grade level. For reporting purposes, MCAS questions are linked with the reporting category that most closely represents the standard(s) assessed.

Table 5-1
Approximate Distribution of Science and Technology Test Questions
By Content Strand and Substrand

Content Strand	Substrands	Grade 4	Grade 8	Grade 10
Inquiry	In accordance with the *Science & Technology Curriculum Framework* and assessment design, many questions that address other content strands will also be inquiry-based, and are therefore not limited to a specific percentage of questions.			
Domains of Science	Physical Sciences	25%	25%	25%
	Life Sciences	25%	25%	25%
	Earth and Space Sciences	25%	25%	25%
Technology	The Design Process	5%	5%	5%
	Understanding and Using Technology	15%	15%	15%
Science, Technology, and Human Affairs		5%	5%	5%

SKILLS IN USING AND APPLYING SCIENCE AND TECHNOLOGY

In addition to content knowledge, students will be expected to demonstrate various process skills fundamental to science and technology. Critical investigation and problem-solving skills include

- observation;
- hypothesis formulation and testing; and
- evaluation and use of evidence to propose, design, and test solutions.

For the purposes of the MCAS assessment, these scientific and technology-related process skills are grouped into three major areas: thinking skills, procedural skills, and application skills.

Thinking Skills

Questions in this area assess student understanding of concepts. In order to demonstrate thinking skills, students will be required, for example, to recognize, evaluate, analyze, and explain natural scientific and technological phenomena.

Procedural Skills

Questions in this area assess student knowledge and understanding of scientific and technological procedures.

Application Skills

Questions in this area assess student skill in selecting appropriate scientific and technological concepts and procedures and appropriately applying these concepts and procedures to solve real-life and theoretical problems.

ITEM TYPES

Two types of questions will be used at each grade level tested: multiple-choice and open response. The *Guide to the Massachusetts Comprehensive Assessment System: Science & Technology* presented the approximate number of items for each item type for each component in each grade. Table 5-2 presents this information.

Table 5-2
May 1999 MCAS
Approximate Distribution of Science & Technology Items by Type

Grade	Item Type	Number of Test Items (per student test booklet)
4 and 8	Multiple-choice	41
	Open response	6
10	Multiple-choice	44
	Open response	7

DETAILED SPECIFICATIONS

Table 5-3 describes the exact number of items appearing in the 1999 MCAS science and technology assessment. Note, technology and science, technology, and human affairs were collapsed and referred to as technology.

Table 5-3
May 1999 MCAS
Detailed Specifications for 1999 MCAS Science & Technology Assessment

Grade	Reporting Category	Common		Matrix (Total Across 12 Forms)	
		Multiple-Choice	Open-Response	Multiple-Choice	Open-Response
4	Inquiry	5	1	15	3
	Physical Sciences	8	1	17	2
	Life Sciences	7	1	17	3
	Earth & Space Sciences	7	1	18	2
	Technology	7	1	17	2
	Total	34	5	84	12
8	Inquiry	4	1	13	2
	Physical Sciences	7	1	19	1
	Life Sciences	7	1	16	2
	Earth & Space Sciences	8	1	18	3
	Technology	8	1	18	4
	Total	34	5	84	12
10	Inquiry	3	0	7	1
	Physical Sciences	8	2	24	3
	Life Sciences	9	1	24	3
	Earth & Space Sciences	8	1	24	3
	Technology	8	2	17	2
	Total	36	6	96	12

CHAPTER 6
DESIGN OF THE HISTORY AND SOCIAL SCIENCE

STUDY STRANDS AND LEARNING STANDARDS

The *History and Social Science Curriculum Framework* contains four (4) **Study Strands** and twenty (20) related **Learning Standards** to be assessed at grades 5, 8, and 10. Table 6-1 presents these Study Strands and related Learning Standards. (Note: The numbers preceding the Study Strands and Learning Standards are used as the basis for coding items on the History and Social Science assessment at all three grade levels.)

Table 6-1

Study Strands and Related Learning Standards

Study Strands	Learning Standards
1. History	1. Chronology and Cause
	2. Historical Understanding
	3. Research, Evidence, and Point of View
	4. Society, Diversity, Commonality, and the Individual
	5. Interdisciplinary Learning: Religion, Ethics, Philosophy, and Literature in History
	6. Interdisciplinary Learning: Natural Science, Mathematics, and Technology in History
2. Geography	7. Physical Spaces of the Earth
	8. Places and Regions of the World
	9. The Effects of Geography
	10. Human Alteration of Environments
3. Economics	11. Fundamental Economic Concepts
	12. Economic Reasoning
	13. American and Massachusetts History
	14. Today's Economy
	15. Theories of Economy
4. Civics and Government	16. Authority, Responsibility, and Power
	17. The Founding Documents
	18. Principles and Practices of American Government
	19. Citizenship
	20. Forms of Government

CONTENT COVERAGE

The *History and Social Science Curriculum Framework* groups the Study Strands and Learning Standards in a **Core Knowledge Era** format and places them within commonly recognized time periods in United States and World history. (Refer to pages 13 through 17 in the *Framework* for specific topics to be taught and assessed within each Core Knowledge Era.) Table 6-2 presents the Core Knowledge Eras for the United the Core Knowledge Eras for the World.

Table 6-2

Core Knowledge Eras

Core Knowledge Eras: The United States

1. Early America and Americans (Beginnings to 1650)
2. Settlements, Colonies, and Emerging American Identity (1600 to 1763)
3. The American Revolution: Creating a New Nation (1750 to 1815)
4. Expansion, Reform, and Economic Growth (1800 to 1861)
5. The Civil War and Reconstruction (1850 to 1877)
6. The Advent of Modern America (1865 to 1920)
7. The United States and Two World Wars (1914 to 1945)
8. The Contemporary United States (1945 to the Present)

Core Knowledge Eras: The World

1. Human Beginnings and Early Civilization (Prehistory to 1000 B.C.)
2. Classical Civilizations of the Ancient World (1000 B.C. to c. 500 A.D.)
3. Growth of Agricultural and Commercial Civilizations (500 to 1500 A.D.)
4. Emergence of a Global Age (1450 to 1750)
5. The Age of Revolutionary Change (1700 to 1914)
6. The World in the Era of Great Wars (1900 to 1945)
7. The World from 1945 to the Present

Content Coverage by Core Knowledge Eras

Table 6-3 presents the Core Knowledge Eras assessed on the 1999 examination.

Table 6-3

Core Knowledge Eras Assessed in 1999

Grade Level	Core Knowledge Eras Assessed
5 (Tryout)	World Core Knowledge Eras: 1. Human Beginnings and Early Civilizations to 1000 B.C. United States Core Knowledge Eras: 1. Early America and Americans (Beginnings to 16500) 2. Settlements, Colonies, and Emerging American Identity (1600 to 1763) 3. The American Revolution: Creating a New Nation (1750 to 1815); topics a. through g. only
8	**World Core Knowledge Eras:** 1. Human Beginnings and Early Civilizations (Prehistory to 1000 B.C.) 2. Classical Civilizations of the Ancient World (1000 B.C. to c. 500 A.D. 3. Growth of Agricultural and Commercial Civilizations (500 to 1500 A.D.); topics a. through c. only United States Core Knowledge Eras: 3. The American Revolution: Creating a New Nation (1750 to 1815); topics d. through h. only 4. Expansion, Reform, and Economic Growth (1800 to 1861) 5. The Civil War and Reconstruction (1850 to 1877)
10	World Core Knowledge Eras: 2. Classical Civilizations of the Ancient World (1000 B.C. to 500 A.D.); topics h. and i. only 3. Growth of Agricultural and Commercial Civilizations (500 to 1500 A.D.) 4. Emergence of a Global Age (1450 to 1750) 5. The Age of Revolutionary Change (1700 to 1914) 6. The World in the Era of Great Wars (1900 to 1945) 7. The World from 1945 to the Present

Detailed Specifications

Table 6-4, 6-5 and 6-6 details the number of items used on the 1999 History and Social Science assessment. (Note: Grade 5 is not shown because the 1999 assessment was a tryout and the Grade 5 test design is yet to be determined.)

Table 6-4
Number of Assessment Items by Session and Core Knowledge Era

CKE = Core Knowledge Era MC = Multiple-choice OR = Open Response

Grade 8

Session: CKE Assessed	Number of Items			
	Common		Matrix	
	MC	OR	MC	OR
Session 1: World CKE 1,2,3 a-c only	12	2	3	0
Session 2: U.S. CKE 3 d-f only, 4	11	2	2	0
Session 3: U.S. CKE 5	11	1	2	1
Total	34	5	7	1

Grade 10

Session: CKE Assessed	Number of Items			
	Common		Matrix	
	MC	OR	MC	OR
Session 1: World CKE 3 d- j, 4	11	2	5	1
Session 2: World CKE 5	11	2	5	1
Session 3: World CKE 6,7	11	2	5	1
Total	33	6	15	3

Table 6-5
Approximate Distribution of History & Social Sciences Items by Type

Grade	Item Type	Number of Test Items (per student test booklet)
8	Multiple-choice	41
	Open response	6
10	Multiple-choice	48
	Open response	9

Table 6-6
May 1999 MCAS
Detailed Specifications for History & Social Sciences Assessment

Grade	Reporting Category	Common		Matrix (Total Across 12 Forms)	
		Multiple-Choice	Open-Response	Multiple-Choice	Open-Response
8	History	22	3	19	5
	Geography	4	1	23	2
	Economics	4	1	20	2
	Civics	4	0	21	3
	Total	34	5	83	12
10	History	20	4	43	5
	Geography	5	0	10	4
	Economics	4	1	13	2
	Civics	4	1	14	7
	Total	33	6	80	18

CHAPTER 7
TEST DEVELOPMENT PROCESS

As described in the preceding chapters, MCAS tests were developed to meet a complex set of content and cognitive specifications. In addition, to provide accurate measurement across four performance categories, MCAS items needed to demonstrate acceptable statistical characteristics. To ensure an adequate selection of items to build final test forms, twice as many items were developed as were ultimately needed.

Given the large number of items required, a rigorous test development process was implemented. Table 7-1 presents the major steps in the MCAS test development process that followed the creation of test specifications. Additional information about each step is presented following the table.

Table 7-1
May 1999 MCAS
Major Steps in the Test Development Process

Step		When Occurred
1	Assessment Development Committee (ADC) item idea generation	March – April 1998
2	Item writing	March – July 1998
3	Internal item review	July – November 1998
4	Assessment Development Committee item review	July 1998
5	Item editing	September 1998 – January 1999
6	Item tryout form assembly	January - March 1998
7	Item tryout review	April 1998
8	Item tryout administration	May 17 – May 28, 1998
9	Item tryout scoring	June – July 1998
10	Item tryout data analysis	July 1998
11	Initial item selection	July - September 1998
12	Assessment Development Committee selection and editing of common and matrix items	November - December 1998
13	DOE-contractor review	January - February 1999
14	External bias and sensitivity review	January 1999
15	DOE-contractor bias and sensitivity resolution	January - March 1999
16	Operational test assembly	February – March 1999
17	Edit drafts of operational tests	February - March 1999
18	Braille translation	March 1999
19	Spanish translation	March 1999

ASSESSMENT DEVELOPMENT COMMITTEE (ADC) ITEM IDEA GENERATION

At the initial ADC meetings, specifications and designs were reviewed and item ideas were generated. Item ideas could range from broad-brush, "addition of two two-digit numbers with renaming (carrying) in a story problem" to targeted, "addition of two-digit numbers with renaming in a story problem that asks about the number of pieces of equipment in a park" to writing a complete draft item.

ITEM WRITING

Developers expanded upon the item ideas and edited the items for technical accuracy and adherence to sound testing practice.

INTERNAL ITEM REVIEW

- Lead or peer test developer within the content specialty reviewed the typed item, open-response scoring guide, and any reading selections and graphics.
- The content reviewer considered item "integrity," item content and structure, appropriateness to designated content area, item format, clarity, possible ambiguity, keyability, single "keyness," appropriateness and quality of reading selections and graphics, and appropriateness of scoring guide descriptions and distinctions (as correlated to the item and within the guide itself).
- The content reviewer also considered scorability and whether the scoring guide adequately addressed performance on the item.
- Fundamental questions for the content reviewer to ask included, but were not be limited to, the following:

  –What is the item asking?

  –Is the key the only possible key?

–Is the open-response item scorable as written (correct words used to elicit response defined by guide)?
–Is the wording of the scoring guide appropriate and parallel to the item wording?
–Is the item complete (e.g., with scoring guide, content codes, key, grade level, and contract identified)?
–Is the item appropriate for the designated grade level?

ASSESSMENT DEVELOPMENT COMMITTEE ITEM REVIEW

Item sets were brought to ADC meetings for review and revision.

ITEM EDITING

Editors reviewed and edited the items from the ADC item review to ensure uniform style (based on *The Chicago Manual of Style, 14th Edition*) and adherence to sound testing principals. These principals included that items

- were correct with regard to grammar, punctuation, usage, and spelling;
- were written in a clear, concise style;
- were unambiguous in explaining to students what is expected for a maximum score;
- were written at a reading level that prevents reading ability from interfering with the student demonstrating his or her knowledge of the tested subject matter;
- exhibited high technical quality regarding psychometric characteristics;
- had appropriate answer options or score-point descriptors; and
- raised no unnecessary sensitivity concerns.

ITEM TRYOUT FORM ASSEMBLY

Multiple test forms were created for English language arts, mathematics, and science and technology for each grade level (4, 8, and 10). Within each form, test questions were grouped by content (e.g., in order to form a more homogeneous criterion for item analysis, tryout forms were not built to be parallel). See section on final form assembly for more details of the test assembly process.

ITEM TRYOUT REVIEW

An editor reviewed the tryout forms. See section on final form review for more details of the review process.

ITEM TRYOUT ADMINISTRATION

The tryout was designed to mirror the administration of the operational assessment program. The test forms were spiraled so that each school would have some students taking each form and each form would be administered to a random sample of students. All students in grades 4, 8, and 10 in all schools in Massachusetts were required to participate in the tryout.

ITEM TRYOUT SCORING

Multiple-choice items were optically scanned. Open-response items were scored using a consensus-scoring model. That is, rather than developing a training pack with benchmark papers, a group of highly experienced scorers used scoring rubrics to guide discussion of student responses and came to mutually acceptable scores. Consensus scoring is less expensive and faster for small volumes of student papers.

ITEM TRYOUT DATA ANALYSIS

The following statistics were calculated for multiple-choice items: item difficulties (percent correct), item discriminations (point-biserial correlations), item quartile distributions (distribution of student responses or scores within each quartile of the criterion score distribution), and differential item functioning (DIF) statistics comparing males and females and white and black student responses.

The same statistics were calculated for short-answer questions, except there were insufficient students to calculate DIF statistics for white-black comparisons.

The same statistics were calculated for open-response items as were calculated for short-answer questions, except the Pearson product-moment correlation was used rather than the biserial correlation.

INITIAL ITEM SELECTION

Based on statistical information (see Table 7-2 for the format in which information was provided), comments from scorers, and professional judgment, test developers selected acceptable items to present to the ADCs. Note, not all item statistics were computed for item tryout items. For example, sample sizes were too small to calculate meaningful IRT statistics.

Table 7-2
Format of Item Statistics

Sample:	*A*			*Score Point*		*n*	*% of Total*	*% of 1st quartile*	*% of 2nd quartile*	*% of 3rd quartile*	*% of 4th quartile*	*Mean crit. score*
				OR	*MC*							
Criterion	*B*			*BL*	*BL*	*R*	*S*	*T*	*U*	*V*	*W*	*X*
				0	*A*							
Difficulty (Mn): C		*Discrimination (r): D*		*1*	*B*							
A: E	*c: F*	*b(01): G*	*b(12): H*	*2*	*C*							
Fit: K		*b(23): I*	*b(34): J*	*3*	*D*							
I(s12): L	*I(s23): M*	*I(s34): N*		*4*	*E*							
DIF(F-M): O	*DIF(B-W): P*	*DIF(H-W): Q*		*T*	*T*	*Y*						*Z*

A A description of the sample is entered here, such as: "1999 Massachusetts grade 4 item tryout sample for mathematics."

B The criterion measure used for biserial correlations and differential item functioning analyses is entered here, such as: "Form 12 Total Mathematics score."

C Classical item difficulty or item mean. For multiple-choice items this is equivalent to percent of students responding correctly (p-value); for open-response items this is equivalent to the average student item score.

D Classical item discrimination statistic. For multiple-choice items this is a corrected point-biserial correlation; for open-response items, this is a Pearson product-moment correlation (a corrected item-to-total score correlation).

E Item response theory item discrimination parameter.

F Item response theory lower asymptote (guessing) parameter (for the three-parameter logistic model). Used only for multiple-choice or other items where student guessing might lead to a correct answer.

G Item response theory difficulty parameter for differentiating scores of 0 and 1. There is one difficulty parameter for multiple-choice items, and one between each pair of consecutive score categories for open-response items.

H Item response theory difficulty parameter for differentiating scores of 1 and 2. This will be blank for multiple-choice items.

I Item response theory difficulty parameter for differentiating scores of 2 and 3. This will be blank for multiple-choice items.

J Item response theory difficulty parameter for differentiating scores of 3 and 4. This will be blank for multiple-choice items.

K Item response theory fit statistic, describing how well the IRT model fits the item's data.

L Amount of information item provides for differentiating between students at the first and second client-set performance standards. Requires that performance standards are already set. The sum of item information at these performance standard cut-points is directly related to the test's decision accuracy.

M Amount of information item provides for differentiating between students at the second and third client-set performance standards. Requires that performance standards are already set.

N Amount of information item provides for differentiating between students at the third and fourth client-set performance standards. Requires that performance standards are already set.

O Standardized difference between matched (by weighting to total group on criterion score) samples of male and female students. Significance of difference based on Mantel-Haenszel statistic and indicated by one asterisk (.01 level) or two asterisks (.001 level).

P Standardized difference between matched (by weighting to total group on criterion score) samples of white and black students.

Q Standardized difference between matched (by weighting to total group on criterion score) samples of white and Hispanic students.

R For open-response or multiple-choice items, the number of examinees who left this question blank. For open-response, the next five rows present the number of students with scores of 0, 1, 2, 3, and 4 respectively. More rows are added if there are additional score points. For multiple-choice items, those rows indicate the number of examinees who chose options A, B, C, D, and E, respectively.

S For each row in this column, the percent of examinees with each score (open-response) or who chose each option (multiple-choice) is indicated.

T Of those examinees scoring in the top quartile on the total criterion score, the percent whose response was blank. The next five rows present similar information for the other score points.

U Of those examinees scoring in the second quartile on the total criterion score, the percent whose response was blank. The next five rows present similar information for the other score points.

V Of those examinees scoring in the third quartile on the total criterion score, the percent whose response was blank. The next five rows present similar information for the other score points.

W Of those examinees scoring in the lowest quartile on the total criterion score, the percent whose response was blank.

X Mean total criterion score of those examinees whose score point was blank. For following rows, the mean criterion score is given for examinees achieving other score points. For multiple-choice items, this should be highest for the correct option. For open-response items, the means should be ordered for score points 0 to 4, and spread reasonably well.

Y Total sample size.

Z Sample mean on the criterion.

EXTERNAL BIAS AND SENSITIVITY REVIEW

A bias and sensitivity review committee of eighteen educators from around the state was convened for two three-day meetings to address potential bias and sensitivity issues. Bias is defined as question context or content that is irrelevant to the curriculum being assessed that affects test scores of an identifiable subgroup of students. Sensitivity refers to issues that are not related to the curriculum being assessed and might offend or distract students.

SELECTION OF COMMON AND MATIRX ITEMS

Test developers presented item statistics to the Assessment Development Committees to assist in the Committees' recommendation for placement of items into the common and matrix portions of the test. The final decision for selections was made by the Department of Education with the assistance of the testing contractor.

OPERATIONAL TEST ASSEMBLY

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

- Content coverage/match to test design. The curriculum specialist completed an initial sort of items into sets based on a balance of content categories across sessions and forms, as well as a match to the test design (number of multiple-choice, short-answer, and open-response items).
- Item difficulty and complexity. Item statistics resulting from data analysis of previously tested items were used to assure similar levels of difficulty and complexity across forms.

- Visual balance. Item sets were reviewed to ensure that each reflected a similar length and "density" of selected items (e.g., length/complexity of reading selections, number of graphics).
- Option balance. Each item set was checked to verify that it contains a roughly equivalent number of key options (As, Bs, Cs, and Ds).
- Name balance. Item sets were reviewed to ensure diversity of names used.
- Bias. Each item set was reviewed to ensure fairness and balance based on gender, ethnicity, religion, socio-economic status, and other factors.
- Page fit. Item placement was modified to ensure the best fit and arrangement of items on any given page.
- Facing page issues. For multiple items that are associated with a single stimulus (graphic or reading selection), consideration was given to whether the group needs to begin on a left- or right-hand page, as well as to the nature and amount of material that needed to be on facing pages. These considerations serve to minimize the amount of "page flipping" required of the students.
- Relationships between forms. The set of "common" items must be placed identically in each version of the forms. Matrix-sampled item sets differ from form to form, but must take up the same number of pages in each form so that sessions and content areas begin on the same page in every form. Therefore, the number of pages needed for the longest form often drives the layout of each form.
- Visual appeal. The visual accessibility of each page of the form is always considered, including such aspects as the amount of "white space," the density of the text, and the number of graphics.

EDIT DRAFTS OF OPERATIONAL TESTS

Any changes that the test construction specialist makes are reviewed and approved by the test developer. Once a form is laid out in what is considered its final form, the form is read through to identify any final considerations, including the following:

- Editorial changes. All text is scrutinized for editorial accuracy, including consistency of instructional language, grammar, spelling, punctuation, and layout. Advanced Systems' publishing standards are based on *The Chicago Manual of Style, 14th Edition.*
- "Keying" items. Items are reviewed for any information that may "key" (or provide information that would help answer) another item. Decisions about moving keying items are based on the severity of the key-in and the placement of the items in relation to each other within the form.
- Key patterns. The final sequence of keys is reviewed to ensure that their order appears random (e.g., no recognizable pattern, no more than three of the same key in a row).

BRAILLE AND LARGE PRINT TESTS

One form of each of the Spring 1999 MCAS testes was translated into Braille by a subcontractor specializing in test materials for blind and visually-handicapped students. Additionally, one form of each of the spring 1999 MCAS tests was adapted into large print version.

SPANISH TRANSLATION

One form of the Spring 1999 MCAS mathematics, science and technology, and history and social science tests were adapted into Spanish. The Spanish version of the MCAS tests were presented in a bilingual format (Spanish/English) with identical test items presented on opposing pages: left-facing pages presented items in Spanish; right-facing pages presented identical items in English. This format was adopted based on field testing s Spanish adaptation and a bilingual format adaptation among Limited English Proficient (LEP) students in approximately 10 public schools.

In adapting a test to another language, a number of decisions have to be made. Depending on the nature of the original test, on the target language, and the intended examinee population, the adapted test may be very similar or quite different from the original. In this case, because intended examinees were known to come from different Hispanic countries, representing a variety of dialects rather than a single dialect, it was decided to use standard Spanish in the test, and to include certain dialectal variants as a gloss in brackets as needed. Because of the nature of the subjects being tested (math and science), and their link to the state standards, it was agreed ahead of time that the basic content of the tests should remain the same if possible.

There were a number of steps in the adaptation of MCAS for Spanish-speaking students. A preliminary review of the instruments showed that only two items needed to be replaced with items from other test forms in English. The two items identified in the review involved assumed knowledge of American culture. For example, one item assumed knowledge of how American football is played.

Another change that was made in the instruments involved translating English names to Spanish (James = Jaime), provided the names were easily translatable.

Two native speakers of Spanish were identified. Each was a professional translator with knowledge of item writing procedures and experience in test translation and test translation review. Each translator was a specialist in either math or science. The translator of the mathematics test had an undergraduate degree in mathematics from a university in Paraguay. The science translator had a degree in medical anthropology from a university in Colombia. Both had experience translating standardized tests, and had previously received instruction on item writing.

Both translators were oriented to the project. The orientation included information on the MCAS program and the most frequent countries of origin of examinees who would take the MCAS in Spanish. Subsequently, the translators began work on the first draft. Their first draft was reviewed by a senior translation specialist, who made initial decisions about how to handle wording common to both tests, such as that found in the instructions, headers, footers, item stems, etc. The senior translation specialist then sent each translator's work to the other with instructions that the translation be evaluated by comparing it line by line and item by item with the English version. The comments of each reviewer were reviewed, and then forwarded to the original translator with further observations or recommendations.

The DOE collected systematic feedback from teachers and students on the Spanish version following its administration. The feedback elicited from teachers concerning Spanish usage in the math and science tests showed that they felt the Spanish version accurately reflected the English original.

SECTION II
TEST ADMINISTRATION

CHAPTER 8
TEST ADMINISTRATION

RESPONSIBILITY FOR ADMINISTRATION

As indicated in the *Principal's Administration Manual* (Massachusetts Department of Education, 1999e), principals were responsible for the proper administration of the MCAS. Directors of charter schools, 766-approved private schools, institutional school programs, and educational collaboratives were responsible for the compliance with administration requirements in their school. Manuals and certification forms were used to ensure uniformity of administration procedures across schools.

PROCEDURES

Principals were instructed to read the *Principal's Administration Manual* thoroughly prior to testing and to be familiar with the instructions given in the *Test Administrator's Manual* (Massachusetts Department of Education, 1999f). The chapter "Conducting Test Administration" in the *Test Administrator's Manual* contains sections that detail the procedure to be followed for each test session. The chapter also contains the actual scripts "to be read aloud to students AS PRINTED during test administration" (p. 9). Another critical document produced and disseminated by the Department of Education was *The Massachusetts Comprehensive Assessment System: Requirements for Test Scheduling, Student Participation, and Test Security and Ethics* (Massachusetts Department of Education, 1999g).

ADMINISTRATOR TRAINING

In addition to the two administration manuals, the Massachusetts Department of Education, assisted by the testing contractor, conducted a series of administration workshops throughout the state in the month prior to the spring 1999 test administration.

TEST ADMINISTRATION SCHEDULE

MCAS testing materials were received in schools the week of April 19, 1999 for English language arts composition and May 10, 1999 for all other subject areas. The test administration window was from April 26-30, 1999 for English language arts composition and May 17 through June 2, 1999 for all other subject areas. The Department of Education supplied schools with sample test administration schedules for grades 4, 8, and 10. Table 8-1 presents the grade 10 sample test administration schedule.

Table 8-1
1999 Grade 10 Sample Test Administration Schedule

- Fourteen 45-minute test sessions, plus one 20–30 minute session for completion of student identification information, questionnaire, and an optional practice test
- Two 45-minute sessions per day maximum recommended
- Makeup sessions scheduled throughout the three weeks as necessary

May 1999

Monday	Tuesday	Wednesday	Thursday	Friday
17 Student Identification Questionnaire and Practice Test (30 min.)	18 English Language Arts English Language Arts	19 English Language Arts English Language Arts	20 English Language Arts	21 English Language Arts English Language Arts
24 Mathematics	25 Mathematics	26 Mathematics	27 Science & Technology Science & Technology	28 Science & Technology Science & Technology
31 History and Social Science Item Tryout History and Social Science Item Tryout	1	2	3	4

PARTICIPATION REQUIREMENTS

All public school students in grades 4, 8, and 10 were required to participate in the MCAS, per the Educational Reform Act of 1993, including students enrolled in charter schools, and students receiving publicly funded special education in 766-approved private schools, institutional schools, and collaboratives.

Students with Disabilities

Students with disabilities were defined as students with an Individualized Education Plan (IEP) or a plan of instructional accommodations provided under Section 504 of the Rehabilitation Act of 1973.

For such students, the IEP plan of the Section 504 team is required to consider the following questions in determining how a student will participate:

- Can this student take the tests under routine conditions?
- If the student is not able to take the tests under routine conditions, will he or she be able to take these tests if appropriate test accommodations are provided?
- If a student cannot take the tests, even with accommodations, what would be an appropriate alternative assessment to enable the student to demonstrate his or her knowledge of the standards contained in the curriculum frameworks?

Limited English Proficient Students

Limited English Proficient (LEP) students were defined as students who met any of the following conditions:

- were enrolled in a Transitional Bilingual Program;
- received English as a Second Language support;
- were not born in the United States and whose native language was a language other than English and who were currently not able to perform ordinary classroom work in English; or
- were born in the United States to non-English speaking parents and who were not currently able to perform ordinary classroom work in English.

LEP students were required to participate in the MCAS if they met either of the following criteria:

- student had been enrolled in school in the United States for more than three years; or
- student was in a Transitional Bilingual Education program or received English as a Second Language support and had been/would be recommended for regular education classes for the 1999–2000 school year.

Requirements for Spanish-Speaking LEP Students

Spanish-speaking LEP students who have completed three or more years of school in the United States were not eligible to take the Spanish language version of the MCAS; these students were required to take the English language version.

Spanish-speaking LEP students who do not yet have the fluency to participate in the English language version of the MCAS were required to participate in the Spanish language version of the mathematics and science and technology tests if they met all of the following criteria:

- had completed three or fewer years of school in the United States;
- were in a Transitional Bilingual Education program or received English as a Second Language support and were not to be recommended for regular education classes for the 1999–2000 school year; and
- possessed reading and writing skills in Spanish appropriate to their grade level.

Accommodations

The Massachusetts Department of Education published an extensive list of appropriate accommodations in *The Massachusetts Comprehensive Assessment System: Requirements for Test Scheduling, Student Participation, and Test Security and Ethics* (Massachusetts Department of Education, 1999g). Also, schools were directed to call the Department of Education to inquire about the use of accommodations not listed.

TEST SECURITY

Strict question and test security measures were implemented during all phases of development and production in order to maintain the fairness and integrity of the MCAS. To this end, each of the MCAS administration manuals contains a chapter on "Test Security and Ethics." In the chapter, it is stated

> *The quality and usefulness of the assessment data generated by MCAS depends, in large part, on uniformity of test administration and security of test materials. Valuable information about student achievement and curriculum effectiveness will be seriously compromised if test security is not strictly implemented and maintained (p. 5).*

The chapter includes sections on penalties, school/principal's responsibilities, and instructions to be given to students regarding the use of test materials. The school/principal's responsibilities include

- taking inventory of testing materials received by the school,

- monitoring the distribution and use of these materials, and
- ensuring the complete and error-free return of all materials.

ACCOUNTING FOR TEST MATERIALS

The administration manuals also contained explicit instructions for the handling of test booklets, answer documents, and other materials. Material tracking and verification forms were provided to principals and test administrators to help them account for test materials. Upon completion of testing, test administrators assembled the test materials for return to the principal. Used response documents were separated from unused ones and were packaged in special envelopes provided to schools. The school principal organized the testing materials, using the material verification form, to verify the return of all secure testing materials to the testing contractor.

Each principal received detailed instructions and a prepaid, pre-printed air-bill for returning test materials to the testing contractor. Principals were instructed to call the shipping contractor toll free when their materials were ready for pickup after testing. Shipped packages were completely and easily traceable. Personnel were able to track a particular package any time from date of pickup to date of delivery. A toll-free number was also provided to principals to provide notification of any problems or delays with pickup.

The outside of each box containing test materials was labeled by school and district. Upon receipt of each box, the labels were checked and the boxes were logged in. The resulting list was compared to a master distribution file on a daily basis. One week after the close of the testing window, a list of outstanding schools or missing boxes was produced, and applicable schools were contacted for discrepancy resolution.

Once boxes were scanned, they were placed on a holding skid (by grade) to be processed. In order to ensure accuracy, each person who checked materials worked with only one school at a time.

During log-in, staff opened boxes and reviewed administration forms. If any of the administration forms were missing, the school was contacted. A log-in supervisor used the principal's certification forms to enter into an electronic spreadsheet the following information:

- the number of materials sent to the school,
- the number of materials returned from the school, and
- the date the materials were logged into the spreadsheet.

In addition, the following information was entered into the spreadsheet and updated:

- the name of individual who logged in the materials,
- whether or not the school had a discrepancy and the date any discrepancy was sent to the school for resolution, and
- whether the school or the Department of Education has resolved the discrepancy.

The newly created spreadsheet was then compared to the master distribution file to determine if any discrepancies existed. If there was a difference between the number of materials sent to the school and the number received from the school, the discrepancy resolution process began.

Once the materials were accounted for, all demographic sheets were removed from the response booklets and placed under a school header pre-slugged with school name, school code, and the number of students in that school. This became the official file upon which school reports were based.

The used response booklets were processed by hand to check their general condition and to remove any unnecessary materials. Schools with materials that were returned with significant problems were reported to the school and the Department of Education. Efforts were made to correct gridding problems, and any missing or damaged headers were replaced.

About two percent of the total test forms were received from the schools in poor condition and could not be scanned. Unscannable forms were manually entered into the system. Large-print response booklets were also entered manually.

After the booklets were checked, they were oriented in one direction and boxed by school. The school header sheet was placed on the top of booklets in the box, which was then sent for scanning.

SECTION III
DEVELOPMENT AND REPORTING OF SCORES

CHAPTER 9
SCORING

Student answer booklets were scanned so that all information necessary to score responses and produce reports was captured and converted into an electronic format. This conversion included all student identification and demographic information, school information, multiple-choice data, and digital image clips of hand-written responses. This chapter summarizes the score processing procedures for the MCAS.

Multiple-choice questions were machine scored. All other questions were individually read and evaluated.

MACHINE-SCORED ITEMS

Student responses to multiple-choice were optically scanned. The scoring key was applied to the captured item responses. Correct answers were assigned a score of one point each; incorrect answers were assigned a score of zero points each. Multiple-choice questions were used for all content areas within English language arts, mathematics, science and technology, history and social science, except writing.

ITEMS SCORED BY READERS

Digital imaging and a computerized scoring system were used in the scoring process for all short-answer and open-response questions and short compositions. Digital imaging allowed electronic copies of student responses to be sent to readers who scored the responses. The computerized scoring system assigned student responses to readers. It provided maximum randomization of student work, ensuring that no one reader or group of readers scored multiple papers from the same school. It also provided continuous monitoring of the performance of readers, allowing leadership staff to re-score student responses and retrain readers when necessary. Scoring methods for each type of open-response question are cov ered in the following three subsections.

SCORING GUIDES FOR SHORT-ANSWER ITEMS

Short-answer questions, used in mathematics, were hand-scored by contractor staff. Correct answers were assigned a score of one point each; incorrect answers were assigned a score of zero points each. Most short-answer questions had a single correct numeric answer. In some cases, there were multiple acceptable answers (see Figure 9-1) or a range of correct answers (for example, correct answer: a number in the range of 356 to 358). Some short-answer questions were somewhat more complex to score. One example would be a question where the correct answer: is any set of 9 numbers with a range of 20, mean of 85, and median of 85; e.g., 75, 75, 75, 80, 85, 90, 95, 95. Figure 9-1 presents an example of a short-answer item with its scoring guide.

Figure 9-1
Example of a Short-Answer Item and Its Scoring Guide

Item	Write a RULE to find the next number in the pattern. 90, 87, 84, 81, ___	
Scoring guide	Score as correct:	Subtract 3 -3 minus 3

SCORING GUIDES FOR OPEN-RESPONSE ITEMS

Item-specific scoring guides were developed for each open-response item. Figure 9-2 presents an example of a scoring guide for an open-response item.

SCORING GUIDE FOR WRITING PROMPTS

Students were required to write one long composition in response to a writing prompt. The composition was assigned a score for topic/idea development (on a one to six scale) and a score for standard English conventions (on a one to four scale). Readers for the long compositions were composed of contractor scorers and teachers at three Massachusetts Writing Institutes. The *MCAS Writing Scoring Guide* in Figure 9-3 was used for scoring all compositions. In addition to the scores, "analytic annotations" were also used in reporting. These are comments on topic development, organization, details, language/style, sentences, grammar, and usage, and mechanics, as shown in Figure 9-3.

Figure 9-2
Example of an Open-Response Item and Its Scoring Guide

Item	To make a house handicapped accessible, a ramp is being constructed to the floor of the porch. The Americans with Disabilities Act requires that a ramp have an incline of no more than 5°. Assume that the maximum allowable angle is used and that the floor of the porch to which the ramp is constructed is 4 feet above the ground. (You may refer to the trigonometric table on your Mathematics Reference Sheet.) a. Draw and label a picture showing the ramp and porch. b. Based on the information above, how far is the end of the ramp from the porch? Show your work. c. Based on the information above, what is the length of the ramp? Show your work.
Scoring guide	Score 4 if The student scores 5 points Score 3 if The student scores 4 points Score 2 if The student scores 3 or 2 points Score 1 if The student scores 1 point Score 0 if Response is totally incorrect or irrelevant. Score Blank if No response Scoring information: Part a: 1 point for correct drawing of porch and ramp For drawing, the student must show right triangle with angle of 5° and 4' for length of vertical leg of right triangle opposite the 5° angle. Part b: 1 point for correct distance from porch = 45.71 feet 1 point for correct strategy displayed through work, e.g., tan 5° = 0,0875 = 4/x x = 4/0,0875 = 45.71 feet Note: Other correct approaches are acceptable.) Part c: 1 point for correct length of ramp = 45.9 feet 1 point for correct strategy displayed through work, e.g., $45.71^2 + 4^2$ = length of ramp ^{2} $(2089.4 + 16)^{.5}$ = length of ramp = 45.9 feet OR sin 5° = = 4/r r = 4/sin 5° r = 45.9 feet (or 45.87; 45.89) Some numbers in work may vary due to rounding, but answers should be correct to at least the nearest tenth of a foot. If rounding is to nearest foot, work must show ramp longer than horizontal distance before rounding. Note: If student reverses order of b and c, credit can be awarded as above, provided work/diagram shows student understands which length he/she found.

Figure 9-3

MCAS WRITING SCORING GUIDE (LONG COMPOSITION)

Topic/Idea Development					
1	2	3	4	5	6
Little topic/idea development, organization, and/or details Little or no awareness of audience and/or task	Limited or weak topic/idea development, organization, and/or details Limited awareness of audience and/or task	Rudimentary topic/idea development and/or organization Basic supporting details Simplistic language	Moderate topic/idea development and organization Adequate, relevant details Some variety in language	Full topic/idea development Logical organization Strong details Appropriate use of language	Rich topic/idea development Careful and/or subtle organization Effective/rich use of language

Analytic Annotations			
		Commendations	Needs
Topic Development	The overall effect of the paper	TX effective and appropriate development of topic or ideas TY original development of topic or ideas	TJ appropriate response to task TK more development of the topic
Organization	The degree to which the response is ■ focused ■ clearly and logically ordered ■ clarified by paragraphs	OX evidence of planning OY consistently focused from beginning to end	OJ organization of ideas OK transitions among ideas
Details	The degree to which the response includes examples that develop the main points.	DX carefully chosen, relevant details DY develops details that support the topic	DJ more effective choice of details DK development of details beyond listing
Language/Style	The degree to which manipulation of language, including vocabulary, word choice, word combination, and sentence variety is effectively achieved.	LX word choice enhances meaning LY language creates distinctive voice, tone, or style	LI more variety/richness in word choice LJ more variety/richness in word choice LK more variety in sentence structure

Standard English Conventions			
1	2	3	4
Errors seriously interfere with communication AND Little control of sentence structure, grammar and usage, and mechanics	Errors interfere somewhat with communication and/or Too many errors relative to the length of the essay or complexity of sentence structure, grammar and usage and mechanics	Errors do not interfere with communication and/or Few errors relative to length of essay or complexity of sentence structure, grammar and usage and mechanics	Control of sentence structure, grammar and usage and mechanics (length and complexity of essay provide opportunity for student to show control of standard English conventions)

Analytic Annotations			
		Commendations	Needs
Sentences	The degree to which the response includes sentences that are correct in structure	SP correct sentence structure	SR correct sentence structure
Grammar and Usage	The degree to which the response demonstrates correct ■ use of standard grammatical rules of English ■ word usage and vocabulary	GUP correct application of grammatical rules GUQ control of vocabulary and word usage	GUR correct application of grammatical rules GUS greater attention to correct word usage
Mechanics	The degree to which the response demonstrates correct ■ spelling ■ capitalization ■ punctuation	MP control of mechanics aids clarity MQ correct mechanics in sophisticated construction	MR greater control of mechanics MS more careful proofreading

SELECTION OF SCORING STAFF

Scoring was led by a scoring director, scoring site managers (who managed the various scoring locations) and chief readers, curriculum specialists, who were responsible for managing the technical aspects of scoring. Chief readers were responsible for hiring quality assurance coordinators, overseeing the development of training materials, and ensuring training is implemented properly.

Chief readers worked with quality assurance coordinators and human resource specialists to hire qualified readers. For scoring of the MCAS, readers were required to have completed two years of college, but preferred to have earned a four-year college degree. In addition, readers were required to have an appropriate background for the discipline they scored. Applicant screening procedures included

- a formal, structured interview;
- reference checks; and
- a review of each returning reader's documented history on scoring projects similar to the MCAS to ensure that the contractor is not bringing any individual back to scoring who has not demonstrated successful work as a reader.

Table 9-4 summarizes the qualifications of the 1999 MCAS readers.

Table 9-4
Qualifications of 1999 MCAS Scorers

Scoring Responsibility		Educational Credentials				Teaching Experience	Total
		Doctorate	Masters	Bachelors	Other		
Leadership	N	4	28	18	1	47	51
	%	8%	55%	35%	2%	92%	100%
Readers	N	10	197	331	253	440	791
	%	1%	25%	42%	32%	56%	100%

There are two additional points to be made about scoring staff qualifications.

- Data do not include approximately 720 Massachusetts educators who scored a portion of the writing assessments as part of Department of Education-sponsored writing institutes; and,
- teaching experience ranged from one to thirty-two years.

READER TRAINING AND QUALIFICATION

For each item, quality assurance coordinators explained how the anchor pack papers exemplified the descriptors of the score points. After discussion of the anchor pack, readers attempted to score the training pack exemplars correctly. The quality assurance coordinators then reviewed the training pack and answered any questions readers had before actual scoring began. Subsequently, quality assurance coordinators monitored the scoring process and provided further training on any given item as warranted. Readers were required to maintain an acceptable scoring accuracy rate.

SCORING PROCESS

For short-answer and open-response questions, scoring was controlled by an electronic image scoring management system, which distributed digital images of student responses to readers. These responses were randomly assigned to readers. Thus, the probability is low that any reader would score more than one item from a particular student's response booklet. This procedure effectively minimized error variance due to reader sampling.

All readers had at their workstations a complete set of scoring materials (i.e., scoring guides, training packs) for each of the items. Quality assurance coordinators were available to advise and assist readers with their scoring efforts.

Quality assurance coordinators or other highly experienced scorers (verifiers) performed a series of read-behinds in which they scored responses previously scored by readers. Quality assurance coordinators used the agreement rates from these read-behinds to provide ongoing feedback to the readers.

Monitoring Scoring

The scoring management system tracked reader accuracy throughout the scoring process. After a reader scored a student response, the management system determined whether that response should also be scored by another reader, scored by a quality assurance coordinator or other scoring official, or routed for special attention[8]. Quality assurance coordinators and other scoring officials could get current reader accuracy reports and speed reports on-line at any time. Summary or detailed reports could be produced for any time period. Such capability served to ensure reliable and valid scoring.

The weighted averages of total (exact or adjacent) percent agreement of double-blind scores are reported in Table 9-5. Exact agreement is defined as both readers assigning the paper the same score); and adjacent agreement is defined as the two readers scores differing by one point. Up to 20% of the responses for each item received double-blind scores. The weighting was based on the number of responses that were rescored for each question. Note, these data may underestimate scorer accuracy. Blank respsonses were included in both the read-behind and double-blind rescoring. However, in many instances it was impossible for the reader to tell whether a mark on the image was written by the student or whether there was a crease in the paper, bleed-through from the other side of the page or dust on the image screen. Readers were instructed to score as zero any question for which the student had made a mark of any kind. Scores of zero and blank were counted as neither exact nor adjacent agreement, though the effect of blanks and zeroes on student scores was identical.

WRITING PROMPTS

Two different readers independently scored all compositions. If the two scores were not in exact or adjacent agreement, the two readers discussed and re-evaluated the composition to reach agreement on a score. By this method, the process of correcting inaccurate scores served as a way to prevent reader drift and provide continuous training. Samples of the scores assigned by readers to the compositions were regularly verified using the read-behind and double-blind methods to ensure the

[8] Student responses indicating possible child abuse or suicidal tendencies were flagged by readers for school attention.

quality of the scores. The final score for the compositions was the sum of the scores assigned by the two readers.

Table 9-5

1999 MCAS Double-Blind Total Agreement Rates

Subject		Grade 4	Grade 8	Grade 10
Language and Literature		95.4%	96.1%	97.4%
Mathematics	Short Answer	100%	100%	100%
	Open Response	96.1%	96.9%	97.9%
Science & Technology		94.1%	95.1%	95.6%
History and Social Science			95.7%	97.3%

CHAPTER 10
OVERVIEW OF DEVELOPING SCALED SCORES

The MCAS tests were designed to measure student performance against the learning standards contained in the *Curriculum Frameworks*. Consistent with this purpose, primary results on the MCAS tests are reported in terms of performance levels that describe student performance in relation to these established state standards. There are four performance levels:

- *Advanced:* Students at this level demonstrate a comprehensive and in-depth understanding of rigorous subject matter, and provide sophisticated solutions to complex problems.
- *Proficient:* Students at this level demonstrate a solid understanding of challenging subject matter and solve a wide variety of problems.
- *Needs Improvement:* Students at this level demonstrate a partial understanding of subject matter and solve some simple problems.
- *Failing:* Students at this level demonstrate a minimal understanding of subject matter and do not solve even simple problems.

Students received a separate performance level classification (based on total raw score) for each test. School and district level results were reported as the number and percentage of students who attained each performance level at each grade level tested.

In addition to performance levels, MCAS results are reported as scaled scores. Scaled scores in each content area range from 200 to 280. Scaled scores supplement the MCAS performance level results by providing information about the position of a student's results within a performance level. School- and district-level scaled scores are calculated by computing the average of student-level scaled scores.

The MCAS 1999 included tests in English Language Arts, Mathematics, and Science & Technology in their second annual administration for grades 4, 8, and 10. Also administered was the History and Social Science test for grade 8 in its first annual administration. Because the grade 8 History and Social Science test is different from the other tests in this sense, the process by which 1999 scaled scores in History and Social Science were developed was consistent with the process used for other content areas in 1998.

Scaled scores for the 1999 grade 8 History and Social Science test were developed in the same manner as 1998 scaled scores in other content areas. First, a standard setting process was implemented to determine the range of total raw scores that correspond to each performance level. Results of standard setting were used to determine the transformation of the raw scores to scaled scores. These steps for developing initial MCAS scaled scores in a content area were described in more detail in the *Massachusetts Comprehensive Assessment System 1998 Technical Report* (Massachusetts Department of Education, 1999, pp. 57-67).

To develop scaled scores for English Language Arts, Mathematics, and Science & Technology equating had to be performed. Equating is the process of converting test scores from different versions of the same test so that the resulting scores can be used interchangeably even though they are based on different sets of items. Equating allows for scores for MCAS 1998 and MCAS 1999 to be reported in the same scale.

The next two chapters provide details in developing scaled scores for the 1999 administration of MCAS. Chapter 11 describes the results of standard setting and details the conversion of raw scores to scaled scores for the grade 8 History and Social Science test. Chapter 12 describes how raw scores for English Language Arts, Mathematics, and Science & Technology were equated and translated to scaled scores.

CHAPTER 11
DEVELOPMENT OF SCORES:
GRADE 8 HISTORY AND SOCIAL SCIENCE

STANDARD SETTING

Standard setting is the process of determining the minimum, or threshold, score for each performance level. The multi-step process of setting standards for grade 8 History and Social Science began in February 1998, when the Massachusetts Board of Education adopted general descriptions for each of the four performance levels to be used in reporting. These general descriptions of Advanced, Proficient, Needs Improvement, and Failing (see Chapter 10) were the basis for all standard setting activities. Building on the general definitions, content specialists developed general performance level definitions for History and Social Science. Those descriptions, , were approved by the Board in June 1998 and were used in the standard-setting process.

The threshold scores for the grade 8 MCAS History and Social Science were set using the Body of Work (BoW) method. The hallmark of the BoW method is that panelists examine complete student response sets (student responses to multiple-choice questions and samples of actual student work on open-response questions) and match each student response set to one of the MCAS performance level categories. This is done in three major steps: 1) training/calibration, 2) range finding, and 3) pinpointing.

In August 1999, the Department of Education convened panels of Massachusetts citizens, both educators and non-educators, to participate in the standard-setting process for the MCAS. This process resulted in the identification of a minimum total test score (threshold score) for each performance level. The threshold scores were recommended to and accepted by the Board of Education. Details of the standard setting process are provided in the companion document *MCAS History and Social Science Standard Setting for Grade 8* (Massachusetts Department of Education, 2000).

Table 11-1 presents the final thresholds resulting from the standard setting. These thresholds were computed by applying the logistic regression technique on the classification data provided by panelists. The unit of analysis is each student's body of work. A separate regression analysis is done for each performance level threshold score. The standard error associated with each threshold score is also presented. Standard errors were estimated by applying the logistic regression technique separately to each panelist's data. Thus, for each threshold decision, there was a distribution of estimated thresholds. The standard error was estimated as the standard deviation of the distribution divided by the square root of the number of panelists.

Table 11-1
Grade 8 History and Social Science
Threshold (Minimum) Total Test Score For Each
Performance Category and Its Associated Standard Error
(Maximum Score on Test is 56)

Performance Category	Threshold	Standard Error
Advanced	46.37	.46
Proficient	38.83	.37
Needs Improvement	26.25	.26

TRANSLATING RAW SCORES TO SCALED SCORES (SCALING)

Students' raw scores, or total number of points, on the grade 8 History and Social Science test were translated to scaled scores using a process called scaling. Scaling simply converts raw points from one scale to another. Converting from raw scores to scaled scores does not change the rank ordering of students, give more weight to particular questions, or change students' performance level classifications.

Linear scaling parameters were determined so the minimum scaled score for Needs Improvement was 220, the minimum scaled score for Proficient was 240, and the minimum scaled score for Advanced was 260. This was done by solving two linear equations relating the raw threshold scores to these predetermined scaled score values. The resulting functions that translate raw scores (r) to scaled scores (S) are:

$$\begin{array}{ll} S = 1.59r + 177.25 & \text{if } r < 38.83, \text{and} \\ S = 2.65r + 136.19 & \text{if } r > 38.83 \end{array}$$

Note that the two linear equations correspond to either side of the proficient threshold. That is, the first equation yields scores lower than 240 and the second equation yields scores that are 240 or above.

After translations were applied, scores were rounded to the nearest even integer. Transformed scores below 200 were reported as 200. There were no transformed scores above 280. In any given year, test form difficulty and rounding might lead to some scaled scores between 200 and 280 not being obtainable. In the 1999 administration of the grade 8 History and Social Science test both 200 and 280 are obtainable. Raw score to scaled score conversion tables for all MCAS tests administered in 1999 are available in Appendix A of *Guide to Interpreting the 1999 MCAS Reports for Schools and Districts* (Massachusetts Department of Education, 1999).

CHAPTER 12
DEVELOPMENT OF SCORES:
ENGLISH LANGUAGE ARTS, MATHEMATICS, AND SCIENCE & TECHNOLOGY

Scaled scores for the 1999 MCAS English Language Arts, Mathematics, and Science & Technology were developed by equating the 1999 raw scores to the 1998 raw scores. Equating scores from alternate forms of a test adjusts for any difference in difficulty and allows for scores from the different forms to be comparable. Because the 1998 and 1999 versions of each test were developed from the same framework they may be considered alternate forms. Equating test scores from the 1998 and 1999 administration of each test makes it possible to report the results of the 1999 administration to be reported on the same scale that MCAS results were reported on the previous year. Equating simply converts raw points from MCAS 1999 to the MCAS 1998 raw score scale. The equated scores then are translated to scaled scores. The process of scaling does not change the rank ordering of students, give more weight to particular questions, or change students' performance level classifications.

Equating for MCAS used the *anchor-test-nonequivalent-groups* design with external anchor described by Petersen, Kolen, & Hoover (1993). The "anchor test" was a sub-set of matrix items that were included in both the 1998 and 1999 test administrations. These items are external to the test in that they do not contribute to the students' raw scores in either administration of the test. The groups of students who take each test in 1998 and 1999 are naturally-occurring groups and no assumption was made regarding their equivalence. Item Response Theory (IRT) is particularly useful in this type of equating (Allen & Yen, 1979).

Developing equated scores for the 1999 MCAS involved several steps. The first step was to construct the "anchor test;" that is, to determine the set of equating items. The second step was to calibrate the items in an IRT model. The IRT model used was a combination of the three-parameter logistic (3PL) model for multiple-choice items, the two-parameter logistic (2pl) model for short-answer items, and the graded response model (GRM) for the open-response items. The calibration was first performed on the 1998 data. The item parameters of the equating items resulting from this calibration were fixed for the calibration of the 1999 data. Fixing the parameters of the equating items ensures that the two forms of the test (1998 and 1999) are calibrated to the same scale of the trait being measured. Using test characteristics curves (TCC), raw scores from the 1999 MCAS were mapped or equated to raw scores on the 1998 MCAS. The equated scores were then translated to the 200 to 280 scale. The following sections detail this equating process.

DETERMINING THE SETS OF EQUATING ITEMS

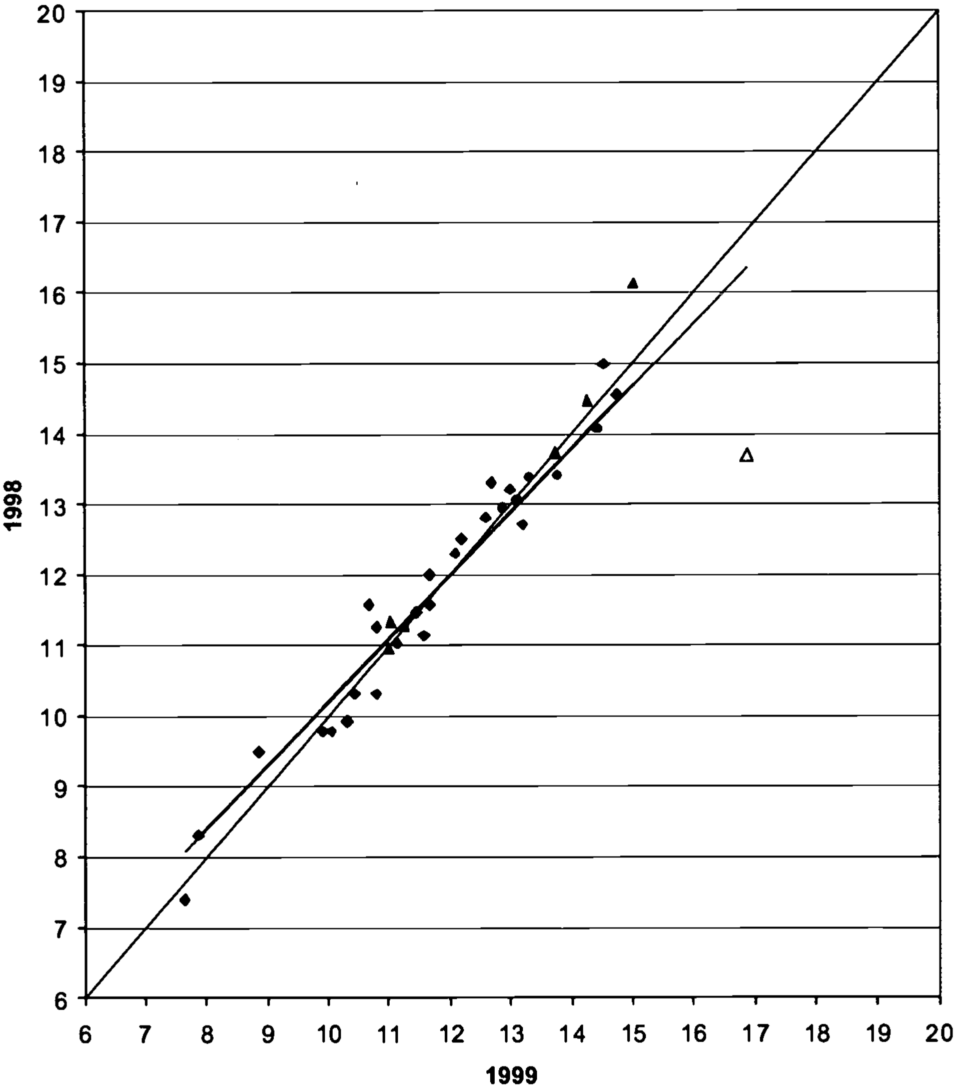
During the development stage of the 1999 MCAS tests, matrix items that were also administered in 1998 were identified as potential equating items. These items were designated based on the following guiding principles:

1. The average difficulty of the equating items should be about the same as the average difficulty of the 1998 test.
2. The total points from the equating items should be at least 20% of the total points on the test.
3. The position of each item in the 1999 form is about the same as its position in the 1998 form.
4. The distribution of the items across different relevant categories (i.e. items types and content areas) should be similar to that of the whole test.
5. There should not be any change in the item from one administration to the other.

To determine the final set of equating items for each MCAS test a differential item functioning (DIF) approach using the delta plot method was applied. The p-values of each multiple-choice and short answer item were transformed to the delta metric. Each item has two p-values, one for each test administration. The delta scale is an inverse normal transformation of percentage correct to a linear scale with a mean of 13 and standard deviation of 4 (Holland & Wainer, 1993). A high delta value indicates a difficult item. For open response items, adjusted p-values, the average score divided by the maximum possible score (4), were transformed to the delta metric. The delta values computed for the potential equating items were plotted for each subject (English Language Arts, Mathematics, and Science & Technology) in each grade level (4, 8, 10).

Figures 12-1 is an example of delta plot for equating items. The graph shown is for grade 4 Mathematics. (Delta plots for other MCAS tests can be found in Appendix B.) The dark diagonal line is the regression line and the light diagonal line is the identity line. Different shapes were used to identify different item types: ? for multiple choice items; ? for short answer items; and, ? for open response items. The perpendicular distance of each item to the regression line was computed. The unshaded shape indicates the item with the greatest perpendicular distance from the regression line. Items that were not more than three standard deviations away from the regression line were used as equating items.

Figure 12-1
Sample Delta Plot
Grade 4 Mathematics
(? MC ? SA ? OR)



Of all the potential equating items, only one item was not used – a short answer mathematics item for grade 4. This item, represented by the un-shaded triangle in Figure 12-1, was more than four standard deviations away from the regression line. Table 12.1 presents the number of equating items used to calibrate the May 1999 MCAS items.

Table 12-1
Number and Percentage of Equating Items

Grade	Subject	Total Number of Items	Number of Equating Items				Total Number of Points	Points from Equating Items (%)
			MC	SA	OR	All (%)		
4	English Language Arts	208	36	-	7	43 (21%)	311	64 (21%)
	Mathematics	147	24	6	5	35 (21%)	198	77 (39%)
	Science & Technology	135	29	-	5	34 (22%)	186	50 (27%)
8	English Language Arts	209	45	-	9	54 (26%)	312	124 (40%)
	Mathematics	147	24	4	3	31 (21%)	198	60 (30%)
	Science & Technology	135	23	-	6	29 (21%)	186	48 (26%)
10	English Language Arts	208	60	-	10	70 (34%)	312	130 (42%)
	Mathematics	150	20	3	5	28 (19%)	204	55 (27%)
	Science & Technology	150	36	-	4	40 (27%)	204	49 (24%)

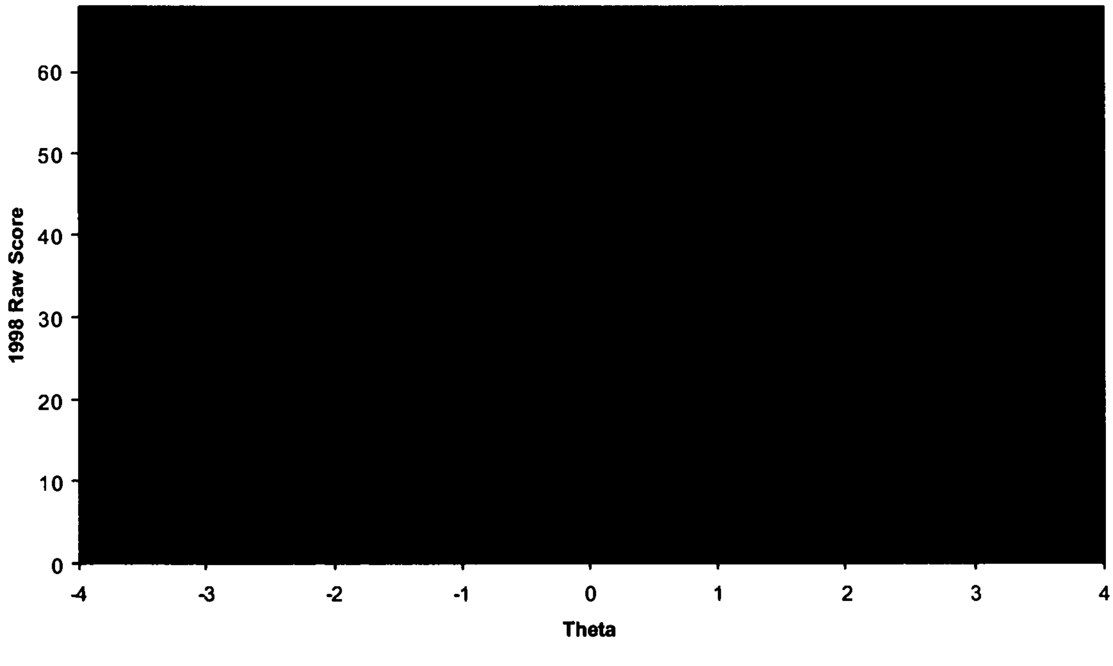
ITEM CALIBRATIONS

IRT calibration was performed on the common and matrix items from the 1998 MCAS tests using a combination of IRT models specific to item types (i.e., 3PL for multiple-choice, 2PL for short-answer, and GRM for open response). Each of these models expresses examinees tendencies to achieve certain scores on the items contributing to a scale as a function of a parameter that is not directly observed and commonly referred to as θ. Using the current version of PARSCALE, item parameters were estimated based on those models. From the parameter estimates, a test characteristic curve (TCC) was obtained using common items only – the same set of items on which individual student scores for the 1998 MCAS tests were based. Through this TCC, each raw score on the test can be mapped to a unique value of θ. An example of a TCCs is shown in Figure 12-2. Item parameters for the common items are included in Appendix C. Within each grade level and subject combination, the items are listed in order that they appear in *The Massachusetts Comprehensive Assessment System: Release of May 1998 Test Items* (Massachusetts Department of Education, 1998) within item type. Summary statistics of item parameters are included in Chapter 15.

Figure 12-2

Sample Test Characteristic Curve

Grade 4 English Language Arts



An IRT calibration was also performed on the 1999 MCAS student response data. This data set included responses to 1999 MCAS common and matrix items. So that the 1999 MCAS tests9 would be calibrated to the same θ scale as the 1998 tests, IRT parameters for the equating items were not estimated for this calibration. Instead, they were fixed to the estimated values resulting from the calibration of the 1998 MCAS data.

Parameters for common items are also available in Appendix C. Within each grade level and subject combination, the items are listed in order that they appear in *The Massachusetts Comprehensive Assessment System: Release of Spring 1999 Test Items* (Massachusetts Department of Education, 1999) by item type.

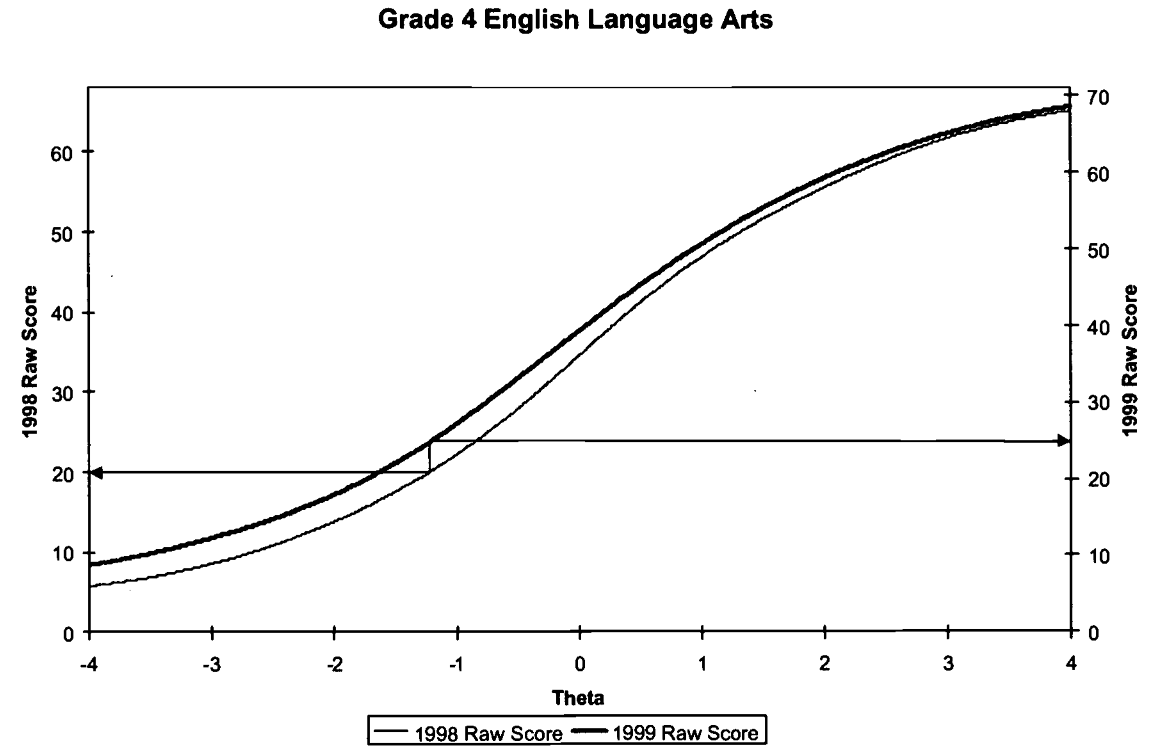
The item parameter estimates for the common items were used to obtain the TCC for 1999 MCAS tests. Using this TCC, each raw score was be mapped to a θ value.

During the item calibration stage, it was discovered that the c-parameters for multiple-choice items o the 1999 English Language Arts and Science & Technology test were converging to values around zero. The c-parameter, also referred to as pseudo-chance level parameter, is incorporated into the IRT model to take into account performance at the low end of the θ scale on multiple-choice items where guessing is a factor (Hambleton, Swaminathan, & Rogers, 1991). Having c-parameter values at zero means that items are fitting the 2PL model. Because all other multiple-choice items in different content areas were calibrated using the 3PL model, a decision was made to fix the c-parameters for English Language Arts and Science & Technology to 0.23. This value is a little lower than the random chance probability of selecting the correct choice. Fixing c-parameters typically assumes values that are smaller than the value that would result if examinees guessed randomly on the items (Lord, 1980).

EQUATED SCORES

Because the TCCs for the 1998 and 1999 MCAS were on the same θ metric, for each value of θ there is a corresponding raw score for each of the 1998 MCAS and 1999 MCAS common item sets. Thus, for each subject and grade combination, each MCAS 1999 raw score can be mapped to a MCAS 1998 raw score. For example, using the TCCs in Figure 12-3 (ELA Grade 4) a raw score of 25 in MCAS 1999 maps to a raw score of 20 in MCAS 1998. (Similar graphs for other subjects and grades are in Appendix C.) This mapping is referred to as IRT true-score equating (Lord, 1980) using fixed-b method to maintain a consistent θ metric.

Figure 12-3
Finding Equated Scores



SCALED SCORES

After raw scores from MCAS 1999 are mapped to MCAS 1998 raw scores (i.e., equated scores), these scores are translated to scaled scores. The functions that translate raw scores to scaled scores are:

$$
\begin{array}{ll}
S = m_1 r + b_1 & \text{if } r < P, \text{ and} \\
S = m_2 r + b_2 & \text{if } r > P
\end{array}
$$

where S is the scaled score, r is the raw score, and P is the proficient threshold. The values of the ms, the bs, and the Ps are shown in Table 12-1. These scaling constants are based on the results of standard setting processes implemented for English Language Arts, Mathematics and Science and

Technology in August 1998. A through discussion of the processes is found in *Massachusetts Comprehensive Assessment System 1998 Technical Report* (Massachusetts Department of Education, 1999; pp. 57-65).

Table 12-2
Transformation Constants Used to Compute Scaled Scores

Grade	Subject Area	Transformation Constants				Proficient Threshold *(P)*
		m_1	b_1	m_2	b_2	
4	English Language Arts	0.88	198.10	1.55	167.00	46.46
	Mathematics	1.48	192.10	2.44	161.55	31.70
	Science and Technology	1.70	188.23	2.07	177.15	29.81
8	English Language Arts	1.45	179.76	1.20	189.95	41.00
	Mathematics	2.00	174.09	1.96	175.17	32.50
	Science and Technology	2.71	158.95	1.97	180.76	29.52
10	English Language Arts	1.38	168.15	1.30	171.89	51.49
	Mathematics	1.89	174.01	1.78	177.85	34.39
	Science and Technology	1.55	185.30	1.65	181.63	34.61

After the transformation constants were applied, scores were rounded to the nearest even integer. Transformed scores below 200 were reported as 200; transformed scores above 280 were reported as 280. A more through discussion of the scaling functions are found in the *Massachusetts Comprehensive Assessment System 1998 Technical Report* (; Massachusetts Department of Education, 1999; pp.63-64).

Going back to the example in Figure 12-2, the MCAS 1999 raw score of 25 mapped to the MCAS 1998 raw score of 20 will result in a scaled score of 216. Table 12-1 presents the equated and scaled scores for each raw score of MCAS 1999 grade 4 English Language Arts. Similar tables for all other subjects and grade levels are in Appendix D.

In any given year, test form difficulty and rounding might lead to some scaled scores between 200 and 280 not being obtainable. Table 12-2 reports the highest and lowest attainable scores in MCAS 1999 (including grade 8 History and Social Science).

Table 12-3
May 1999 MCAS
Minimum and Maximum Obtainable Scores

Grade	Subject Area	Raw Score		Scaled Score	
		Minimum	Maximum	Minimum	Maximum
4	English Language Arts	0	71	200	270
	Mathematics	0	54	200	280
	Science and Technology	0	54	200	280
8	English Language Arts	0	72	200	268
	Mathematics	0	54	200	274
	Science & Technology	0	54	200	280
	History and Social Science	0	54	200	280
10	English Language Arts	0	72	200	280
	Mathematics	0	60	200	280
	Science and Technology	0	60	200	280

Table 12-4
Conversion of Raw Scores to Scaled Scores: Grade 4 English Language Arts

1999 Raw Score	1998 Raw Score	Scaled Score	1999 Raw Score	1998 Raw Score	Scaled Score
71	67	270	35	30	224
70	66	270	34	29	224
69	65	268	33	28	222
68	65	268	32	27	222
67	64	266	31	26	220
66	63	264	30	25	220
65	62	264	29	24	220
64	61	262	28	23	218
63	59	258	27	22	218
62	58	256	26	21	216
61	57	256	25	20	216
60	56	254	24	19	214
59	55	252	23	18	214
58	54	250	22	17	214
57	53	250	21	17	214
56	52	248	20	16	212
55	51	246	19	15	212
54	50	244	18	14	210
53	49	242	17	13	210
52	48	242	16	12	208
51	47	240	15	11	208
50	46	238	14	10	206
49	45	238	13	9	206
48	44	236	12	8	204
47	43	236	11	7	204
46	42	236	10	7	204
45	41	234	9	6	204
44	40	234	8	3	200
43	39	232	7	2	200
42	37	230	6	1	200
41	36	230	5	0	200
40	35	228	4	0	200
39	34	228	3	0	200
38	33	228	2	0	200
37	32	226	1	0	200
36	31	226	0	0	200

CHAPTER 13
SCORE REPORTING

Table 13-1 lists the primary MCAS reports.

Table 13-1
Primary MCAS Reports

1.	*Student Report for Parents/Guardians*
2.	*Student Labels*
3.	*School Test Item Analysis Report*
4.	*District Test Item Analysis Report*
5.	*School Report*
6.	*District Report*
7.	*1999 Statewide Summary of District Performance on the Massachusetts Comprehensive Assessment System (MCAS)*
8.	*MCAS Student Results CD*
10.	*MCAS School and District Results CD*
11.	*Report of 1999 Statewide Results: The Massachusetts Comprehensive Assessment System (MCAS)*

STUDENT REPORT FOR PARENTS/GUARDIANS

Student reports show the scaled score for each subject area, as well as a score band that indicates the standard error of measurement surrounding each score. Performance level definitions are provided so that parents/guardians will understand how to interpret the scaled scores. Specific comments are provided about the student's writing performance. Information is also provided to show how the student's performance compared to the average scores from the student's school, district, and state. An overview of test content is provided, along with a cautionary statement about interpreting scores and guidelines for parents/guardians for helping their children improve. The report also indicates that the child's school should be contacted if there are any questions about the child's report.

The Department of Education provides additional documentation, *Understanding Your MCAS 1999 Student Report for Parents/Guardians* (Massachusetts Department of Education, 1999), which explains in detail how to interpret student reports. This interpretive manual is available in English, Cape Verdean, Chinese, Haitian, Khmer, Portuguese, Russian, Spanish, and Vietnamese. In addition, while all student reports were printed in English, report shells were available in the aforementioned languages to aid parents and guardians in interpreting their child's report.

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 files, if desired. Student labels included results of item analyses.

SCHOOL AND DISTRICT TEST ITEM ANALYSIS REPORT

The *Test Item Analysis Report* shows the answers that each student gave on the multiple-choice questions, as well as his/her score on each open-response question. The report also summarizes overall performance at the school, district, and state levels for each of the question types.

Each school receives a separate *Test Item Analysis Report* for each subject area and grade. The report is designed to be used in conjunction with the publication *The Massachusetts Comprehensive Assessment System: Release of May 1999 Test Items* (Massachusetts Department of Education, 1999), which contains all common test questions. When the report and the publication are used together, educators are provided with a detailed picture of student performance. The *Guide to Interpreting the 1999 MCAS School and District Reports* (Massachusetts Department of Education, 1999) also explains the *Test Item Analysis Report* in detail.

SCHOOL AND DISTRICT REPORTS

The school, district, and union reports are intended for administrators and other interested parties. The school report includes performance level definitions, scaled score intervals, student status definitions, and information about how summary statistics are affected by students not tested; all of which are intended to help the reader interpret the report. The school report provides all results for the school, the district, and the entire state. The results provided are

- the number of students tested by student status (regular, students with disabilities, and limited English proficient students) for all subject areas combined and separately for each subject area,
- the percentage of students in each performance level by subject area,

- the distribution of scaled scores by subject area,
- the number of students in each performance level by subject area and student status,
- subject area subscores by subject subarea and by question type,
- three-year comparisons of school results, and
- average subject score by number of years in the school or district.

The district report is the same as the school report, except that it does not include the school-level data and the three-year comparisons are by district rather than by school. The *Guide to Interpreting the 1998 MCAS School and District Reports* (Massachusetts Department of Education, 1999) explains the school and district reports in detail.

1999 STATEWIDE SUMMARY OF DISTRICT PERFORMANCE ON THE MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM (MCAS)

The *1999 Statewide Summary of District Performance on the Massachusetts Comprehensive Assessment System (MCAS)* (Massachusetts Department of Education, 1999) summarizes performance of all districts in the state, providing a page of information for each.

MCAS STUDENT RESULTS CD

The student results CD is an electronic version of the *Test Item Analysis Report*. Districts were provided with a CD containing student data for each school in the district.

MCAS SCHOOL AND DISTRICT RESULTS CD

The *MCAS School and District Results CD* is an electronic version of the *1999 Statewide Summary of District Performance on the Massachusetts Comprehensive Assessment System (MCAS).*

REPORT OF 1999 STATEWIDE RESULTS: THE MASSACHUSETTS COMPREHENSIVE ASSESSMENT SYSTEM (MCAS)

The *Report of 1999 Statewide Results: The Massachusetts Comprehensive Assessment System (MCAS)* (Massachusetts Department of Education, 1999) presented statewide participation rates, performance levels, and scaled score results.

CHAPTER 14
STATE RESULTS

This chapter presents key participation and performance results from the May 1999 MCAS administration.

Table 14-1

Students Tested on the MCAS Tests of Spring 1999[1]

Grade Level	Enrolled[2]	Percent Tested in English Language Arts[3]	Percent Tested in Mathematics	Percent Tested in Science & Technology	Percent Tested in History & Social Science
Grade 4	78,841	96.1	97.3	97.3	--
Grade 8	73,021	95.3	96.5	96.2	96.0
Grade 10	63,183	92.6	94.1	93.8	--
Total	215,045	94.8	96.1	95.9	

[1] This includes regular education students, students with disabilities and limited English proficient students

[2] Enrollment figures presented here are based on information on the Mathematics test at grades four, eight and ten. Because MCAS tests in each content area were processed independently, enrollment figures vary slightly across content areas.

[3] Percentages of students tested in English Language Arts are underestimated due to special circumstances involved in the processing of results from the English Language Arts tests. Because the *Writing* and *Language* and *Literature* portions of the test were administered at different times, it was necessary to match student results from the two portions of the test. Approximately one percent of the students at grades 4 and 8 and two percent at grade 10 could not be matched and were counted as two students who were not tested rather than as one student who was tested. This resulted in an overestimate of the number of students enrolled and an underestimate of the number and percentage of students tested.

Table 14-2

Regular Education Students Tested on the MCAS Tests of Spring 1999

Grade Level	Enrolled[1]	Percent Tested in English Language Arts[2]	Percent Tested in Mathematics	Percent Tested in Science & Technology	Percent Tested in History & Social Science
Grade 4	63,658	98.8	99.4	99.4	--
Grade 8	60,169	98.3	98.6	98.4	98.2
Grade 10	53,042	96.2	96.7	96.4	--
Total	176,869	97.8	98.3	98.2	

[1] Enrollment figures presented here are based on information on the Mathematics test at grades four, eight and ten. Because MCAS tests in each content area were processed independently, enrollment figures vary slightly across content areas.

[2] Percentages of students tested in English Language Arts are underestimated due to special circumstances involved in the processing of results from the English Language Arts tests. Because the *Writing* and *Language* and *Literature* portions of the test were administered at different times, it was necessary to match student results from the two portions of the test. Approximately one percent of the students at grades 4 and 8 and two percent at grade 10 could not be matched and were counted as two students who were not tested rather than as one student who was tested. This resulted in an overestimate of the number of students enrolled and an underestimate of the number and percentage of students tested.

Table 14-3

Students With Disabilities Tested on the MCAS Tests of Spring 1999

Grade Level	Enrolled[1]	Percent Tested in English Language Arts[2]	Percent Tested in Mathematics	Percent Tested in Science & Technology	Percent Tested in History & Social Science
Grade 4	13,011	91.8	93.5	93.7	--
Grade 8	11,543	90.3	92.1	91.6	91.3
Grade 10	8,559	87.8	89.6	89.7	--
Total	33,113	90.2	92.0	91.9	

[1] Enrollment figures presented here are based on information on the Mathematics test at grades four, eight and ten. Because MCAS tests in each content area were processed independently, enrollment figures vary slightly across content areas.

[2] Percentages of students tested in English Language Arts are underestimated due to special circumstances involved in the processing of results from the English Language Arts tests. Because the *Writing* and *Language* and *Literature* portions of the test were administered at different times, it was necessary to match student results from the two portions of the test. Approximately one percent of the students at grades 4 and 8 and two percent at grade 10 could not be matched and were counted as two students who were not tested rather than as one student who was tested. This resulted in an overestimate of the number of students enrolled and an underestimate of the number and percentage of students tested.

Table 14-4

Limited English Proficient Students Tested on the MCAS Tests of Spring 1999

Grade Level	Enrolled[1]	Percent Tested in English Language Arts[2]	Percent Tested in Mathematics[3]	Percent Tested in Science & Technology	Percent Tested in History & Social Science
Grade 4	2,172	65.9	56.6	54.9	--
Grade 8	1,309	40.6	40.0	39.1	36.9
Grade 10	1,582	29.5	30.8	29.0	--
Total	5,063	49.1	44.2	42.6	

[1] Enrollment figures presented here are based on information on the Mathematics test at grades four, eight and ten. Because MCAS tests in each content area were processed independently, enrollment figures vary slightly across content areas.

[2] Percentages of students tested in English Language Arts are underestimated due to special circumstances involved in the processing of results from the English Language Arts tests. Because the *Writing* and *Language* and *Literature* portions of the test were administered at different times, it was necessary to match student results from the two portions of the test. Approximately one percent of the students at grades 4 and 8 and two percent at grade 10 could not be matched and were counted as two students who were not tested rather than as one student who was tested. This resulted in an overestimate of the number of students enrolled and an underestimate of the number and percentage of students tested.

[3] Percentages of students tested in Mathematics, Science & Technology, and History and Social Science may be underestimated. It appears that several hundred students with limited English proficiency were tested but not classified as LEP students on these tests.

Table 14-5

1999 Statewide MCAS Performance Level Results by Student Status: Grade 4

(percentage of students at each performance level)[1]

Content Area	Student Status	Scaled Scores	Performance Level				
			Advanced	Proficient	Needs Improvement	Failing (Tested)	Failing (Absent)
English Language Arts	All	231	0	21	67	12	0
	Regular	234	1	25	69	5	0
	S w/ Disabilities	222	0	3	60	37	0
	LEP	222	0	3	53	43	0
Mathematics	All	235	12	24	44	19	0
	Regular	237	15	27	45	14	0
	S w/ Disabilities	224	3	10	44	42	0
	LEP	218	1	5	34	61	0
Science & Technology	All	240	10	46	36	8	0
	Regular	242	11	50	33	5	0
	S w/ Disabilities	231	3	27	50	20	0
	LEP	220	0	7	45	48	0

[1] Percentages may not total 100 percent due to rounding. For the purpose of computing school, district, and state results, students who were absent without a medically documented excuse from any subject area MCAS test were assigned the minimum scaled score of 200 and a performance level of *Failing* for that subject area.

Table 14-6

1999 Statewide MCAS Performance Level Results by Student Status: Grade 8

(percentage of students at each performance level)[1]

Content Area	Student Status	Scaled Scores	Performance Level				
			Advanced	Proficient	Needs Improvement	Failing (Tested)	Failing (Absent)
English Language Arts	All	238	3	53	31	12	0
	Regular	241	4	61	29	6	0
	S w/ Disabilities	224	0	16	42	41	1
	LEP	221	0	14	39	47	1
Mathematics	All	226	6	22	31	39	1
	Regular	229	7	26	34	32	1
	S w/ Disabilities	211	1	5	18	75	1
	LEP	207	1	3	8	87	0
Science & Technology	All	224	5	23	27	44	1
	Regular	227	6	26	29	38	1
	S w/ Disabilities	210	1	6	15	77	2
	LEP	204	0	2	7	91	0
History and Social Science	All	221	1	10	40	47	1
	Regular	223	1	12	45	41	1
	S w/ Disabilities	210	0	2	17	80	2
	LEP	206	0	0	9	91	0

[1] Percentages may not total 100 percent due to rounding. For the purpose of computing school, district, and state results, students who were absent without a medically documented excuse from any subject area MCAS test were assigned the minimum scaled score of 200 and a performance level of *Failing* for that subject area.

Table 14-7

1999 Statewide MCAS Performance Level Results by Student Status: Grade 10

(*percentage of students at each performance level*)[1]

Content Area	Student Status	Scaled Scores	Performance Level				
			Advanced	Proficient	Needs Improvement	Failing (Tested)	Failing (Absent)
English Language Arts	All	229	4	30	34	31	1
	Regular	232	5	35	36	23	1
	S w/ Disabilities	212	0	6	21	71	2
	LEP	213	0	6	25	66	3
Mathematics	All	222	9	15	23	50	3
	Regular	225	10	17	26	44	3
	S w/ Disabilities	206	1	3	9	84	3
	LEP	203	0	1	4	92	4
Science & Technology	All	226	3	21	39	34	3
	Regular	228	3	24	41	29	3
	S w/ Disabilities	213	0	5	23	69	3
	LEP	208	0	1	13	80	6

[1] Percentages may not total 100 percent due to rounding. For the purpose of computing school, district, and state results, students who were absent without a medically documented excuse from any subject area MCAS test were assigned the minimum scaled score of 200 and a performance level of *Failing* for that subject area.

SECTION IV
TECHNICAL CHARACTERISTICS

CHAPTER 15
ITEM ANALYSES

As noted in Brown (1983), "a test is only as good as the items it contains." A complete evaluation of a test's quality must include an evaluation of each question. Both the *Standards for Educational and Psychological Testing* and the *Code of Fair Testing Practices in Education* include standards for identifying quality questions. Questions should assess only knowledge or skills that are under assessment and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses are conducted to ensure that MCAS questions meet these standards. Previous sections in this report have delineated the qualitative checks on question quality. The current chapter focuses on more quantitative evaluations. The statistical evaluations are presented in three sections: 1) difficulty indices, 2) item-test correlations, and 3) subgroup differences in item performance. The results presented in this chapter are based on the statewide administration of MCAS in Spring of 1999. About 78,000 grade 4 students, 73,000 grade 8 students, and 63,000 grade 10 students participated in the assessment.

DIFFICULTY INDICES

All multiple-choice, short-answer, and open-response questions were evaluated in terms of difficulty and relationship to overall score according to standard classical test theory practice. Difficulty was measured by averaging the proportion of points received across all students who received the question. Multiple-choice and short-answer questions were scored dichotomously (correct v. incorrect), so for these questions, the difficulty index is simply the proportion of students who correctly answered the question. Open-response questions allowed for scores between zero and four. By computing the difficulty index as the average proportion of points received, the indices for multiple-choice, short-answer, and open-response questions are placed on a similar scale; the index ranges from zero to one regardless of the question type. Although this index is traditionally described as a measure of difficulty (as it is described here), it is properly interpreted as an easiness index because larger values indicate easier questions. An index of zero indicates that no student received credit for the question, and an index of one indicates that every student received full credit for the question.

ITEM-TEST CORRELATIONS

Within classical test theory, these relationships are assessed using correlation coefficients that are typically described as either item-test correlations or, more commonly, discrimination indices. The discrimination index used to analyze MCAS multiple-choice items and zero- or one-scored short-answer items was the point-biserial correlation between item score and a criterion total score on the test.

For open-response items, item discrimination indices were based on the Pearson product-moment correlation. The theoretical range of these statistics is also from –1 to 1, with a typical range from .3 to .6.

Discrimination indices can be thought of as measures of how closely a question assesses the same knowledge and skills assessed by other questions contributing to the criterion total score; that is, the discrimination index can be interpreted as a measure of construct consistency. In light of this interpretation, the selection of an appropriate criterion total score is crucial to the interpretation of the discrimination index. For the MCAS, appropriate criterion scores were selected based on item type and function (common or matrix). The selected criterion scores are provided in Table 15-1. For example, the criterion score for common open-response and short-answer items was the total score on all common multiple-choice, open-response, and short-answer items.

Table 15-1
Criterion Score Used in Computing the Discrimination Index
For Each Item Type and Function

Item Type	Item Function	Scores Included in the Total			
		MC Common	MC Matrix	OR & SA Common	OR & SA Matrix
Multiple-Choice (MC)	Common	✓			
	Matrix	✓	✓		
Open Response (OR) and Short Answer (SA)	Common	✓		✓	
	Matrix	✓	✓	✓	✓
Writing Prompt (WP)	Common	✓		✓	
	Matrix				

For the writing prompt, the reading score was used as the criterion.

SUMMARY OF ITEM ANALYSIS RESULTS

Frequency distributions and summary statistics of the difficulty and discrimination indices for each question are provided in Appendix E and Table 15-2. Appendix E provides distribution information of item difficulty and discrimination by test form while Table 15-2 provides separate distribution information for common and matrix multiple-choice questions.

ITEM RESPONSE THEORY (IRT) PARAMETER ESTIMATES

For equating (see Chapter 12) test items from the 1998 and 1999 administrations of MCAS were calibrated using IRT models. The IRT models used for calibration are the three-parameter logistic (3PL) model for multiple-choice items, two-parameter logistic (2PL) model for short answer items, and the graded response model (GRM) for open response items. The parameters estimated for the 3PL model are discrimination (*a*), difficulty (*b*), and the pseudo-chance level (*c*) parameters. For the 2PL model, only the discrimination and difficulty parameters were estimated. Threshold parameters (d_1, d_2, d_3, and d_4) were estimated for the open response items in addition to the discrimination and difficulty parameters. The computer program PARSCALE (Muraki, 1997) was used for all the IRT calibrations. As mentioned in Chapter 12, *c*-parameters for grade 10 English languages arts and science & technology 1999 were not estimated. Instead their values were fixed to .22. This is also true for some other multiple-choice items. Please see Appendix C for the full lists of item parameter estimates and their respective standard error. Note that a standard error value of zero indicates that the particular parameter was not estimated. Tables 15-3 and 15-4 present the summary statistics of item parameter estimates for multiple-choice, short answer, and open response items for 1998 and 1999 MCAS administrations, respectively.

Writing prompts were calibrated using the GRM. Each prompt was calibrated as two items: one for idea development (with score levels 2 to 12) and one for English convention (with score levels 2-8). The parameter estimates are found in Appendix C.

Table 15-2
Average Difficulty and Discrimination of Different Question Types
For Each Subject and Grade

Grade	Questions		Reading			Mathematics			Science & Technology			History and Social Science		
			n	Diff	Disc	*n*	Diff	Disc	*n*	Diff	Disc	*n*	Diff	Disc
4	MC	All	180	0.65	0.40	133	0.63	0.39	118	0.63	0.33			
		Common	36	0.64	0.37	29	0.66	0.39	34	0.65	0.33			
		Matrix	144	0.65	0.40	84	0.62	0.39	84	0.62	0.33			
	Short Answer		-	-	-	17	0.52	0.38	-	-	-			
	Open Response		28	0.49	0.54	17	0.45	0.59	17	0.40	0.49			
8	MC	All	180	0.67	0.40	113	0.55	0.40	118	0.59	0.34	118	0.55	0.32
		Common	36	0.68	0.39	29	0.56	0.41	34	0.63	0.38	34	0.57	0.35
		Matrix	144	0.66	0.41	84	0.54	0.40	84	0.58	0.33	84	0.55	0.31
	Short Answer		-	-	-	17	0.48	0.52	-	-	-	-	-	-
	Open Response		28	0.49	0.64	17	0.37	0.66	17	0.41	0.59	17	0.34	0.60
10	MC	All	180	0.62	0.41	116	0.43	0.37	132	0.52	0.34	213	0.46	0.34
		Common	36	0.63	0.39	32	0.52	0.39	36	0.59	0.38	33	0.48	0.33
		Matrix	144	0.62	0.41	84	0.40	0.36	96	0.49	0.33	180	0.46	0.34
	Short Answer		-	-	-	16	0.31	0.50	-	-	-	-	-	-
	Open Response		28	0.39	0.69	18	0.28	0.68	18	0.24	0.54	42	0.17	0.58

Table 15-3
Averages of Parameter Estimates of Common Items in MCAS 1998

Grade	Subject	Item Type	n	*a*	*b*	*c*	d_1	d_2	d_3	d_4
4	English Language Arts	Multiple-Choice	28	.90	-.14	.15	-	-	-	-
		Open Response	5	.86	.40	-	2.11	.92	-.86	-2.17
	Mathematics	Multiple-Choice	21	.80	-.24	.13	-	-	-	-
		Short Answer	5	.58	-.13	-	-	-	-	-
		Open Response	6	.83	.05	-	1.76	.35	-.63	-1.48
	Science & Technology	Multiple-Choice	26	.68	-1.03	.09	-	-	-	-
		Open Response	6	.65	-.44	-	2.91	1.21	-.84	-3.27
8	English Language Arts	Multiple-Choice	28	.61	-1.05	.06	-	-	-	-
		Open Response	5	.96	-.04	-	2.25	1.01	-.92	-2.35
	Mathematics	Multiple-Choice	21	.95	-.28	.12	-	-	-	-
		Short Answer	5	1.06	-.37	-	-	-	-	-
		Open Response	6	1.22	.12	-	1.10	.55	-.42	-1.22
	Science & Technology	Multiple-Choice	25	.67	-.18	.07	-	-	-	-
		Open Response	6	1.05	.59	-	2.13	.60	-.84	-2.02
10	English Language Arts	Multiple-Choice	32	.89	-.10	**.22**	-	-	-	-
		Open Response	8	1.32	.62	-	1.71	.82	-.66	-1.87
	Mathematics	Multiple-Choice	27	.86	-.16	.07	-	-	-	-
		Short Answer	5	1.08	.39	-	-	-	-	-
		Open Response	8	1.39	.66	-	1.11	.42	-.44	-1.09
	Science & Technology	Multiple-Choice	30	.91	.26	**.22**	-	-	-	-
		Open Response	8	1.23	1.29	-	1.69	.59	-.58	-1.70

Table 15-4
Averages of Parameter Estimates of Common Items in MCAS 1999

Grade	Subject	Item Type	n	*a*	*b*	*c*	d_1	d_2	d_3	d_4
4	English Language Arts	Multiple-Choice	35	.79	-.31	.16	-	-	-	-
		Open Response	4	.77	.34	-	2.89	.84	-1.00	-2.74
	Mathematics	Multiple-Choice	29	.77	-.59	.11	-	-	-	-
		Short Answer	5	.58	-.28	-	-	-	-	-
		Open Response	5	.81	.31	-	1.66	.53	-.58	-1.61
	Science & Technology	Multiple-Choice	34	.64	-.70	.15	-	-	-	-
		Open Response	5	.62	.78	-	2.86	.93	-.92	-2.88
8	English Language Arts	Multiple-Choice	36	.66	-1.08	.08	-	-	-	-
		Open Response	4	1.03	-.31	-	2.32	.86	-.84	-2.35
	Mathematics	Multiple-Choice	29	1.15	-.03	.16	-	-	-	-
		Short Answer	5	.98	-.32	-	-	-	-	-
		Open Response	5	1.23	.31	-	1.07	.35	-.34	-1.08
	Science & Technology	Multiple-Choice	34	.65	-.71	.09	-	-	-	-
		Open Response	5	.83	.77	-	2.37	.90	-.74	-2.53
10	English Language Arts	Multiple-Choice	36	.83	-.20	**.22**	-	-	-	-
		Open Response	5	1.31	.36	-	1.85	.88	.71	-2.02
	Mathematics	Multiple-Choice	32	.92	.03	.12	-	-	-	-
		Short Answer	4	.82	.38	-	-	-	-	-
		Open Response	6	1.34	.72	-	1.20	.26	-.49	-.97
	Science & Technology	Multiple-Choice	36	.92	.19	**.22**	-	-	-	-
		Open Response	6	1.03	1.65	-	1.48	.56	-.49	-1.55

SUBGROUP DIFFERENCES IN QUESTION PERFORMANCE

The *Code of Fair Testing Practices in Education* explicitly states that subgroup differences in performance should be examined when sample sizes permit, and actions should be taken to make certain that differences in performance are due to construct-relevant, rather than irrelevant, factors. The *Standards for Educational and Psychological Testing* includes similar guidelines. As part of the effort to identify such problems, MCAS questions were evaluated in terms of differential item functioning (DIF) statistics.

DIF procedures are designed to identify questions for which subgroups of interest perform differently beyond the impact of differences in overall achievement. For the MCAS, the standardization DIF procedure (Dorans and Kulick, 1986) was employed to evaluate three subgroup pairs: male v. female, white v. black, and white v. Hispanic[9]. This procedure calculates the difference in item performance for groups of students matched for achievement on the total test. That is, the average item performance is calculated for students at every total score, then an overall average is calculated weighting the total score distribution so it is the same for the two groups.

The index ranges from –1 to 1 for multiple-choice and short-answer questions and is adjusted to the same scale (by dividing by four) for open-response questions. Negative numbers indicate that the question was more difficult for female, black, or Hispanic students. Positive numbers indicate that the question was easier for female, black, or Hispanic students.

Dorans and Holland (1993) suggested that index values between –0.05 and 0.05 should be considered negligible for dichotomously scored questions (such as MCAS multiple-choice and short-answer questions). Most MCAS multiple-choice and short-answer questions fall within this range. Dorans and Holland further stated that dichotomously scored questions 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 questions with values outside the [–0.10, 0.10] range (i.e., "high" DIF) are more unusual and should be examined very carefully. These standards can be

[9] The Mantel-Haenszel procedure was also used to determ ine DIF during the test development process. Items with statistically significant DIF were flagged and indicated in the statistical information presented to the Bias and Sensitivity Review Committee.

applied to open-response questions by accounting for the larger range of possible index values and scaling appropriately. That is, values of the DIF index can range from –4.0 to 4.0, so the corresponding ranges are between –0.2 and 0.2 for negligible difference, between –0.4 and –0.2 and between 0.2 and 0.4 for "low" DIF and outside [-0.4, 0.4] for "high" DIF.

DIF indices indicate differential performance between two groups. That differential performance may or may not be indicative of bias in the test. Course-taking patterns, group differences in interests, or differences in school curricula can lead to DIF. If subgroup differences in performance are related to construct-relevant factors, the questions should be considered for inclusion on a test.

Each question was categorized according to the guidelines adapted from Dorans and Holland (1993). Tables 15-5, 15-6, and 15-7 provide the number of questions in each of the three DIF categories for male-female, white-black, and white-Hispanic comparisons. The counts in these tables include all items on the 1999 MCAS tests, including newly-developed field-tested items.

Table 15-5
Number of Questions in Each Male-Female DIF Category:

Grade	DIF Level	English Language Arts		Mathematics			Science & Technology		History and Social Science	
		MC	OR	MC	SA	OR	MC	OR	MC	OR
4	Negligible	153	27	97	15	15	94	17		
	Low	21	1	15	2	2	23	0		
	High	6	0	1	0	0	1	0		
8	Negligible	150	20	90	16	14	85	12	89	11
	Low	24	8	20	1	3	26	5	25	6
	High	6	0	3	0	0	7	0	4	0
10	Negligible	144	27	95	16	14	90	14	153	37
	Low	29	1	19	0	3	32	4	55	5
	High	7	0	2	0	1	10	0	5	0

Table 15-6
Number of Questions in Each White-Black DIF Category:

Grade	DIF Level	English Language Arts		Mathematics			Science & Technology		History and Social Science	
		MC	OR	MC	SA	OR	MC	OR	MC	OR
4	Negligible	153	28	90	16	16	99	16		
	Low	27	0	23	1	1	17	1		
	High	0	0	0	0	0	2	0		
8	Negligible	140	28	100	15	17	96	17	89	16
	Low	35	0	12	2	0	21	0	26	1
	High	5	0	1	0	0	1	0	3	0
10	Negligible	134	28	101	16	18	106	18	169	41
	Low	38	0	12	0	0	23	0	43	1
	High	8	0	3	0	0	3	0	1	0

Table 15-7
Number of Questions in Each White-Hispanic DIF Category:

Grade	DIF Level	English Language Arts		Mathematics			Science & Technology		History and Social Science	
		MC	OR	MC	SA	OR	MC	OR	MC	OR
4	Negligible	152	27	87	11	15	94	17		
	Low	26	1	23	5	2	20	0		
	High	2	0	3	1	0	4	0		
8	Negligible	135	28	99	13	16	89	17	89	17
	Low	42	0	12	4	1	27	0	27	0
	High	3	0	2	0	0	2	0	2	0
10	Negligible	129	28	105	16	18	103	18	162	41
	Low	40	0	10	0	0	26	0	49	1
	High	11	0	1	0	0	3	0	2	0

CHAPTER 16
RELIABILITY

Although an individual question's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way that questions function together and complement one another. Any measurement includes some amount of measurement error; that is, no measurement can be perfectly accurate. This is true of academic assessments—no assessment can measure students perfectly accurately; some students will receive scores that underestimate their true ability, and other students will receive scores that overestimate their true ability. Questions that function well together produce assessments that have less measurement error; that is, the errors made should be small on average. Such assessments are described as reliable.

There are a number of ways to estimate an assessment's reliability. One approach is to split all test questions into two groups and then correlate students' scores on the two half tests. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, questions on the two half tests have to be measuring very similar knowledge or skills. This is evidence that the questions complement one another and function well as a group. This also suggests that measurement error will be minimal.

The split-half method requires the psychometrician to select which questions contribute to each half-test score. This decision may have an impact on the resulting correlation. Cronbach (1951) provided a statistic that avoids this concern about the split-half method.

RELIABILITY AND STANDARD ERRORS OF MEASUREMENT

Table 16-1 presents descriptive statistics, Cronbach's a coefficient, and raw and scaled score standard errors of measurement for each subject area (English language arts, mathematics, and science and technology), separately for each grade level. The item analysis sample excludes students who did not take one or more sections of the subject.

Note, two scaled-score standard errors of measurement are presented: one for scaled scores below 240 and one for scaled scores of 240 and above. This is because different slopes are used in the linear transformation to scaled scores at these two different parts of the scaled score range.

Table 16-1
Reliabilities, Standard Errors of Measurement and Descriptive Statistics

Grade	Subject	n	Raw Score						Scaled Score	
									<240	>=240
			Min.	Max.	Mean	S.D.	Rel.	S.E.M.	S.E.M.	S.E.M.
4	English Language Arts	76,114	4	70	41.7	10.62	.88	3.68	2.82	1.92
	Mathematics	76,981	0	54	30.6	10.71	.89	3.55	3.21	3.36
	Science & Technology	76,977	0	52	29.9	8.58	.85	3.32	3.61	3.13
8	English Language Arts	70,156	4	72	45.6	11.87	.90	3.75	3.45	1.87
	Mathematics	71,238	0	54	24.3	11.03	.91	3.31	3.71	2.53
	Science & Technology	71,221	0	53	25.5	8.92	.90	2.82	4.10	2.51
	History and Social Science	71,182	0	53	19.0	6.89	.85	2.67	4.50	2.58
10	English Language Arts	59,769	4	72	42.4	13.45	.91	4.04	3.95	2.29
	Mathematics	61,201	0	60	24.1	12.00	.92	3.39	3.35	3.54
	Science & Technology	61,143	0	57	23.9	9.98	.90	3.16	3.80	2.49

RELIABILITY OF PERFORMANCE LEVEL CATEGORIZATION

All test scores contain measurement error; thus classifications based on test scores are also subject to measurement error. After the performance levels were specified and students were classified into those levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications.

Accuracy

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

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 questions if two complete, parallel, forms of the test are given to the same group of students. This is usually impractical, especially on lengthy tests such as the MCAS. To overcome this issue, techniques have been developed to estimate both accuracy and consistency of classification decisions based on a single administration of a test. The technique developed by Livingston and Lewis (1995) was used for the MCAS because their technique can be used with both constructed-response and multiple-choice questions.

Calculating Accuracy

All of the accuracy and consistency estimation techniques described below make use of the concept of "true scores" in the sense of classical test theory. A true score is the score that would be obtained on a test that had no measurement error. It is a theoretical concept that cannot be observed, although it can be estimated. Following Livingston and Lewis (1995), the true-score distribution for the MCAS was estimated using a four-parameter beta distribution, which is a flexible model that allows for extreme degrees of skewness in test scores.

In the Livingston and Lewis method, the estimated "true scores" are used to classify students into their "true" performance category, which is labeled "true status." After various technical adjustments (which are described in Livingston and Lewis, 1995), a 4 × 4 contingency table is created for each test and grade level. The cells in the table are the proportion of students who were classified into each performance category by the actual (or observed) scores on the MCAS (i.e., observed status) and by the "true scores" (i.e., "true status"). As an example, Table 16-2 shows the accuracy contingency table for fourth-grade English language arts. The accuracy contingency tables for all grades and subjects are provided in Appendix F (under step 5). Additional steps in the analysis are also shown in Appendix F.

Table 16-2
Accuracy Contingency Table for Grade 4 English Language Arts

True Status	Observed Status			
	Failing	Needs Improvement	Proficient	Advanced
Failing	**.09**	.02	.00	.00
Needs Improvement	.03	**.61**	.05	.00
Proficient	.00	.04	**.16**	.00
Advanced	.00	.00	.00	**.00**

Proportions on the diagonal (in bold) indicate exact agreement between the observed status and "true status." If the test were perfectly accurate, all of the off-diagonal cells would be zero. Accuracy is the sum of the diagonal (i.e., the proportion of exact agreement across the four performance levels). In Table 14-2, the diagonal sums to .86, indicating that 86 percent of the students were classified into exactly the same performance categories by their observed scores and their "true scores."

Kappa

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 classification that would be expected by chance. Cohen's κ can be used to estimate the classification consistency of a test from two parallel forms of the test. The second form in this case was the one estimated using the Livingston and Lewis (1995) method. Cohen's κ is shown in Table 14-3 (on page 85). Because κ is corrected for chance, the values of κ are lower than the other consistency estimates in Table 16-3.

Calculating Consistency

To estimate consistency, the "true scores" are used to estimate the distribution of classifications on an independent, parallel test form. After statistical adjustments (see Livingston and Lewis, 1995), a new 4 × 4 contingency table is created for each test and grade level that shows the proportions of students who were classified into each performance category by the actual test and by another (hypothetical) parallel test form. Consistency, which is the proportion of students classified into exactly the same categories by the two forms of the test, is the sum of the diagonal for the new contingency table. The consistency contingency tables are shown under step 7 in Appendix F.

Results of Accuracy, Consistency, and Kappa Analyses

The accuracy, consistency, and kappa indices for all grades and subjects are summarized in Table 16-3.

Table 16-3
Estimates of Accuracy and Consistency of Performance Level Classification

Grade	Subject	Accuracy	Consistency	Kappa (κ)
4	English Language Arts	.86	.80	.60
	Mathematics	.77	.68	.54
	Science & Technology	.77	.68	.51
8	English Language Arts	.77	.73	.57
	Mathematics	.79	.71	.58
	Science & Technology	.78	.70	.56
	History and Social Science	.80	.72	.53
10	English Language Arts	.79	.70	.57
	Mathematics	.81	.74	.58
	Science & Technology	.81	.73	.59

Another way of evaluating accuracy is to estimate the probability of students being classified as being in a particular performance-level category, given that their "true status" was that same category. For example, what is the probability that students who are really Proficient (based on their theoretical "true score") will be classified as Proficient based on their MCAS scores? Table 16-4 shows these estimated probabilities.

Table 16-4
Estimated Probability of Being Classified at a Proficiency Level
Given that the "True Status" is that Level

Grade	Subject	Failing	Needs Improvement	Proficient	Advanced
4	English Language Arts	.80	.89	.79	.77
	Mathematics	.79	.79	.67	.87
	Science & Technology	.79	.71	.82	.79
8	English Language Arts	.85	.61	.94	.60
	Mathematics	.91	.74	.69	.67
	Science & Technology	.92	.68	.67	.53
	History and Social Science	.90	.74	.66	.66
10	English Language Arts	.89	.74	.78	.53
	Mathematics	.92	.65	.65	.85
	Science & Technology	.89	.76	.76	.70

For certain decisions, concern may be highest regarding decisions made about a particular threshold. For example, if a college gave credit to students who achieved an Advanced Placement test score of four or five, but not one, two, or three, one might be interested in the accuracy of the dichotomous decision, below four versus four or above. Table 14-5 reports accuracy and consistency for various dichotomous categorizations on the MCAS.

Table 16-5
Accuracy and Consistency of Dichotomous Categorizations

Grade	Subject	Accuracy			Consistency		
		F/NI	NI/P	P/A	F/NI	NI/P	P/A
4	English Language Arts	.94	.92	.996	.92	.88	.99
	Mathematics	.91	.90	.95	.88	.86	.94
	Science & Technology	.94	.87	.95	.92	.82	.93
8	English Language Arts	.88	.91	.98	.89	.88	.96
	Mathematics	.92	.92	.95	.89	.89	.94
	Science & Technology	.92	.91	.95	.88	.88	.93
	History and Social Science	.87	.93	.99	.83	.91	.99
10	English Language Arts	.92	.91	.96	.89	.87	.94
	Mathematics	.91	.93	.97	.87	.90	.95
	Science & Technology	.91	.92	.98	.87	.88	.97

CHAPTER 17
VALIDITY

As noted in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1995, p. 9), "validity is the most important consideration in test evaluation." Validity refers to whether specific inferences made from test scores are appropriate, meaningful, and useful. There are several types of validity-related evidence that can be used to support appropriate, meaningful, and useful inferences based on test scores.

CONTENT-RELATED EVIDENCE

As noted in the *Standards* (p. 10), evidence of test validity begins with test development and continues throughout the entire testing process. Chapters 2 through 5 provide extensive evidence regarding the alignment between the content of the MCAS and the Massachusetts *Curriculum Frameworks*.

RELATIONSHIP BETWEEN MCAS SCORES AND SCORES ON OTHER TESTS

The *1999 MCAS Technical Manual* described two studies, Gong (1999) and Thacker and Hoffman (1999), that correlated MCAS scores with scores on SAT-9 and MAT-7. In addition, these studies examined subgroup differences (e.g., gender and racial/ethnic) between performance on the MCAS and the two standardized norm-referenced tests. Additional discussion of the relationship between performance on the MCAS tests and other tests such as the ITBS and NAEP was presented in the *1998 MCAS Technical Summary.*

A statewide sample of grade 8 students completing the 1999 MCAS test in Science & Technology also participated in the spring 1999 administration of the TIMSS test. Results from the TIMSS test are expected to be available in the winter of 2001. At that time, a detailed analysis of the relationship between student performance on the MCAS and TIMSS test will be conduced.

SECTION V
REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2000). *Standards for educational and psychological testing.* Washington, DC: APA.

Brown, F. G. (1983). *Principles of educational and psychological testing* (3rd Edition). Fort Worth: Holt, Rinehart and Winston.

Code of Fair Testing Practices in Education. (1988) Washington, DC: Joint Committee on Testing Practices. (Mailing Address: Joint Committee on Testing Practices, American Psychological Association, 750 First Avenue, NE, Washington, DC 20002-4242.)

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement,* 20, 37–46.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297–334.

Dorans, N. J., & Holland, P.W. (1993). DIF detection and description. In P.W. Holland & H. Wainer (Eds.) *Differential item functioning* (pp. 35–66). Hillsdale, NJ: Lawrence Erlbaum Associates.

Dorans, NJ, & 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.

Gong, Brian (1999). Relationships between student performance on the MCAS (Massachusetts Comprehensive Assessment System) and other tests—collaborating district A, grades 4 and 10. Prepared for the Massachusetts Department of Education. Dover, NH. The National Center for the Improvement of Educational Assessment, Inc.

Livingston, S.A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement,* 32, 179–197.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum.

Massachusetts Department of Education (1997a), *English Language Arts Curriculum Framework*, Malden, MA

Massachusetts Department of Education (1997b), *Mathematics Curriculum Framework: Achieving Mathematical Power*, Malden, MA

Massachusetts Department of Education (1997c), *Science and Technology Curriculum Framework: Owning the Questions Through Science and Technology*, Malden, MA

Massachusetts Department of Education (1997d), *History and Social Science Curriculum Framework*, Malden, MA

Massachusetts Department of Education (1998a). *The Massachusetts Comprehensive Assessment System: Release of May 1998 Test Items*, Malden, MA

Massachusetts Department of Education (1998b), *Guide to the Massachusetts Comprehensive Assessment System: English Language Arts*, Malden, MA

Massachusetts Department of Education (1998c), *Guide to the Massachusetts Comprehensive Assessment System: Mathematics,* Malden, MA

Massachusetts Department of Education (1998d), *Guide to the Massachusetts Comprehensive Assessment System: Science and Technology*, Malden, Mass.

Massachusetts Department of Education (1999a), *Principals Administration Manual,* Malden, MA

Massachusetts Department of Education (1999b), *Test Administrators Manuals,* Malden, MA

Massachusetts Department of Education (1999c), *The Massachusetts Comprehensive Assessment System: Requirements for Test Scheduling, Student Participation, and Test Security and Ethics*, Malden, MA

Massachusetts Department of Education (1999d), *Understanding Your MCAS 1999 Student Report for Parents/Guardians*, Malden, MA

Massachusetts Department of Education (1999e), *Guide to Interpreting the 1999 MCAS School and District Reports*, Malden, MA

Massachusetts Department of Education, (1999f), *1999 Statewide Summary of District Performance on the Massachusetts Comprehensive Assessment System (MCAS)*, Malden, MA

Massachusetts Department of Education, (1999g), *Report of 1999 Statewide Results: The Massachusetts Comprehensive Assessment System (MCAS)*, Malden, MA

Thacker, Arthur A. and Hoffman, R. Gene (1999). Relationship between MCAS and SAT-9 for one district in Massachusetts. (Report No. FR-WATSD-99-05). Radcliff, KY.: HumRRO.