MASSACHUSETTS DEPARTMENT OF
ELEMENTARY & SECONDARY EDUCATION

MCAS

MASSACHUSETTS
COMPREHENSIVE
ASSESSMENT
SYSTEM

2004–2008 MEPA Technical Report

Massachusetts Department of
ELEMENTARY & SECONDARY
EDUCATION

This document was prepared by
Massachusetts Department of Elementary and Secondary Education
Mitchell D. Chester, Ed.D.
Commissioner

The Massachusetts Department of Elementary and Secondary 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.
Inquiries regarding the Department's compliance with Title IX and other civil rights laws may be directed to the Human Resources Director, 75 Pleasant St., Malden, MA 02148-4906 781-338-6105.

Massachusetts Department of Elementary and Secondary Education
75 Pleasant Street, Malden, MA 02148-4906
Phone 781-338-3000 TTY: N.E.T. Relay 800-439-2370
www.doe.mass.edu

SIGILLUM REIPUBLICÆ MASSACHUSETTENSIS.

TABLE OF CONTENTS

Chapter 1. Overview of This Report

The Massachusetts English Proficiency Assessment (MEPA) is part of the Massachusetts Comprehensive Assessment System (MCAS), established by the Education Reform Act of 1993. The main purposes of the MEPA are to

- measure the current level of English language proficiency of limited English proficient (LEP) students and their progress toward proficiency over time
- identify LEP students who have achieved proficiency in English
- provide data that can be used to strengthen curriculum, instruction, and classroom assessment

The purpose of this technical report is to document the technical quality and characteristics of the 2004–2008 MEPA test program, and to present evidence of the validity and reliability of the intended uses of the MEPA test results.

MEPA items were field-tested during the 2003–2004 school year and the program became operational in 2004–2005. The *2005 MEPA Technical Report*, released in 2007, documented MEPA's first operational year. Since the first MEPA contract culminated with the 2007–2008 cycle of testing, the present technical report will primarily document the three operational years that have yet to be documented (2005–2006, 2006–2007, and 2007–2008), and also present information for the 2004–2005 year that might be useful to the reader for considering the validity of MEPA scores throughout the duration of the first contract. Thus, this report covers the MEPA administration from fall 2004 through spring 2008. The report may also serve as a guide for replicating and/or improving the assessment procedures in subsequent years.

Specific sections of the report discuss test development, test administration, item scoring, scaling and equating, standard setting, reporting of results, item analyses, and reliability. Each of these topics contributes important information toward establishing the validity of the assessment program. Note, however, that certain aspects of a comprehensive validity argument are not included in the report that might also be important to consider when drawing conclusions about validity (e.g., consequences that arise from MEPA scores at student, school, district, and state levels).

Although some parts of this technical report may be useful for laypersons, the intended audience is experts in psychometrics and educational research. The report assumes a working knowledge of measurement concepts, such as reliability and validity, and statistical concepts such as mean and correlation. In some places, the reader is presumed also to have basic familiarity with advanced topics in measurement and statistics.

Chapter 2. DESCRIPTION OF THE MEPA PROGRAM

Title III of the No Child Left Behind (NCLB) Act of 2001 requires that states annually measure the **performance** of limited English proficient (LEP) students in the domains of reading, writing, listening, and speaking, and their **progress** toward acquiring these skills in English. In addition, Chapter 386 of the Massachusetts Acts of 2002 (known as Question 2) requires English language learners in Massachusetts to participate in assessments of English language proficiency. The MEPA program complies with these federal and state assessment requirements. MEPA results are used to

- help determine the level of English proficiency of LEP students
- measure student, school, and district performance on meeting the state's learning standards as detailed in the *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (www.doe.mass.edu/ell/benchmark.pdf)
- improve student achievement and classroom instruction by providing diagnostic feedback regarding student acquisition of knowledge and skills

The Massachusetts Department of Elementary and Secondary Education defines an LEP student as "a student whose first language is a language other than English and who is not able to perform ordinary classroom work in English." All LEP students in grades K–12 educated with Massachusetts public funds participate in MEPA, including

- students enrolled in public and charter schools
- students enrolled in educational collaboratives
- students enrolled in private schools that receive public funding for special education (including approved and unapproved schools within and outside of Massachusetts)
- students who receive educational services in institutional settings
- custodial students of the Departments of Children and Families and Youth Services
- students with disabilities

The MEPA test consists of two separate assessments.

The MEPA-Reading/Writing (MEPA-R/W) is a written test that assesses reading and writing knowledge and skills. All LEP students in grades 3–12 were required to participate in the MEPA-R/W, which was developed for LEP students in four grade spans: 3–4, 5–6, 7–8, and 9–12. Students in grades K–2 were assessed locally in 2006–2008 using norm-referred tests such as the LAS-RW and IPT pending development of a MEPA-R/W test for these students. The separate reading and writing tests consisted of three test sessions, each of increasing language complexity. Each student participated in two sessions of both reading and writing. Schools made separate decisions about which two sessions a student would take and were instructed to consider the Proficiency Level Descriptors in the *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (June 2003) to evaluate which two sessions best matched the student's needs.

The Massachusetts English Language Assessment-Oral (MELA-O) is an observational assessment that evaluates listening (comprehension) and speaking (production) skills in English. All LEP students in grades K–12 were required to participate in the MELA-O. Qualified MELA-O trainers and/or administrators assessed LEP students' listening and speaking skills (sometimes called "indicators" in this report) by observing the students as they participated in everyday classroom activities using the MELA-O Scoring Matrix, found at the end of section 5.1.2. Schools were responsible for submitting MELA-O scores from this locally administered assessment to the testing contractor.

Performance on both the MEPA-R/W items and the MELA-O indicators were incorporated into a student’s overall MEPA scaled score as described in the later sections of this report. The two assessments were designed to measure the range of performance identified by the four MEPA performance levels: *Beginning*, *Early Intermediate*, *Intermediate*, and *Transitioning* (described in more detail in section 5.1 of this report).

Chapter 3. TEST DEVELOPMENT AND DESIGN

3.1 MELA-O Specifications

The Massachusetts English Language Assessment-Oral (MELA-O) is a classroom assessment tool designed to evaluate the English speaking (production) and listening (comprehension) skills of limited English proficient students. The assessment was developed in the early 1990s by the Massachusetts Department of Elementary and Secondary Education in collaboration with researchers at the Evaluation Assistance Center-East at George Washington University. The MELA-O is aligned to the speaking and listening skills identified in the Department's *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (June 2003).

The MELA-O is designed to be administered in a classroom setting where an LEP student can be observed performing academic tasks and participating in ordinary social interactions with other students and the teacher. A student's speaking and listening skills are observed over time by a Qualified MELA-O Trainer (QMT) or Qualified MELA-O Administrator (QMA). Based on his/her observation of the student, the QMT/QMA uses the MELA-O Scoring Matrix to assign indicator scores for listening and speaking, including the four speaking subdomains of fluency, vocabulary, pronunciation, and grammar.

3.2 MEPA-R/W Specifications

The MEPA-R/W is a custom-designed reading and writing test. Test items are aligned with standards in the *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (June 2003), which is based on the *Massachusetts English Language Arts Curriculum Framework* (June 2001). The assessment was developed by the Massachusetts Department of Elementary and Secondary Education in collaboration with Massachusetts educators and the Department's MEPA contractor, Measured Progress, Inc. of Dover, New Hampshire.

3.2.1 Items

3.2.1.1 Item Types

The MEPA-R/W used the following question formats to measure student learning.

- Multiple-choice questions (MC)
  - Students read a question and selected the correct answer from four options.
  - A correct answer was assigned a score of 1 point, and an incorrect answer was assigned a score of 0 points.

- Reading short-answer questions (SA2)
  - Students generated a response of one or more sentences to a question that referenced a paragraph or passage they had read.
  - The response received a score of 0–2 points, based on an item-specific scoring guide.

- Writing short-answer questions (SA1)
  - Students read a question and generated a brief response, usually one word or a short statement.

    - The response received a score of 0–1 point, based on an item-specific scoring guide.

- Sentence-writing questions (SW)
    - Students wrote one or more sentences in response to a graphic or prompt.
    - The response received a score of 0–2 points, based on an item-specific scoring guide.

- Reading open-response questions (OR)
    - Students read a passage and then answered a question by creating a written response of one or more paragraphs.
    - The response received a score of 0–4 points, based on an item-specific scoring guide.

- Writing-prompt questions (WP)
    - Students wrote a composition in response to a writing prompt.
    - The composition received a score of 0–4 points, based on a scoring guide.

3.2.1.2 Item Clarity

Items were reviewed and edited to ensure adherence to the *Standards of Educational and Psychological Testing* (1999) as well as to ensure uniform style in accordance with the *MCAS Style Guide* (based primarily upon the *Chicago Manual of Style*, 14th edition). In accordance with principles delineated in these publications, items were expected to use correct grammar, punctuation, usage, and spelling, and be written in clear, concise style.

3.2.1.3 Item Development and Content Accuracy

The MEPA-R/W did not have a common/matrix-sampled test design (such as that used on the MCAS tests); all items were developed during the first year of the MEPA contract. Items went through a rigorous process of field testing during spring 2004 question tryouts. Scoring guides, where applicable, were also subjected to rigorous internal checks for content accuracy. As described below, Assessment Development and Bias Review Committees and external content expert reviewers assisted the Department in ensuring the content accuracy of all test materials.

Assessment Development Committees (ADCs) reviewed test items and passages. ADCs are made up of Massachusetts educators who have expertise working with English language learners. The Bias Review Committee, also comprised of educators, reviewed test items and passages, both prior to and following field-testing, for potential bias (i.e., material that may disadvantage a student for reasons that are not relevant to the construct being measured). External content experts—specialists in English language acquisition, with a preference for expertise in second language learning—reviewed newly developed items for content accuracy as well as developmental appropriateness; two content experts working independently critiqued each item.

3.2.1.4 Developmental Appropriateness

Each MEPA-R/W item was designed to be developmentally appropriate for the grade span of the test on which it appeared, determined largely by the Massachusetts *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (June 2003). ADC members' judgments were strongly considered where an interpretation was required about the appropriateness of an item or the concept tested by it.

3.2.2 Operational Test Design

3.2.2.1 Construction Process

The process of form construction was informed by classical test statistics, where items were selected based on average difficulty (*p*-value) and discrimination indices. Forms were subsequently evaluated by examining test characteristic curves (TCCs), test information functions (TIFs), and standard errors, where the item response theory (IRT) functions were derived using the one-parameter logistic (1PL) model for the dichotomous items and the one-parameter partial credit model (PCM) for the polytomous items. Classical test theory (CTT) statistics for forms were also evaluated for the sake of completeness.

Four MEPA-R/W test forms (A, B, C, and D) were assembled for operational use (see Table 3-1). IRT and CTT results indicated that the test forms were similar.

Table 3-1. 2004–2008 MEPA: Test Forms and Administration Dates

Administration	Test Form
Fall 2004	A*
Spring 2005	B
Fall 2005	B
Spring 2006	C
Fall 2006	C
Spring 2007	D
Fall 2007	B
Spring 2008	C

*Form A was released following the Fall 2004 administration.

3.2.2.2 Test Sessions

The reading and writing components of the MEPA-R/W were administered separately, each in three sessions. Students participated in two sessions of reading and two sessions of writing based on their level of English proficiency. Schools decided which two sessions each student was to take, considering each component separately. Session 1 of each component was based largely on visual stimuli and contained limited text. Sessions 2 and 3 of each component included increasingly complex tasks in reading or writing, as applicable. Tables 3-2 and 3-3 show the test blueprints for each test session of the reading and writing components, respectively.

Table 3-2. 2004–2008 MEPA: Reading Test Blueprint

Outcomes Assessed	Session 1: 16 Points Largely Based on Visual Stimuli (very limited text)			Session 2: 16 Points 2 Short Reading Passages			Session 3: 16 Points 1 Medium and 1 Long Reading Passage		
	MC	SA2	OR	MC	SA2	OR	MC	SA2	OR
Vocabulary	4	1	0	4	0	0	4	0	0
Beginning to Read	6	0	0	0	0	0	0	0	0
Comprehension	0	2	0	4	2	0	3	1	1
Literary Elements/ Expository Text	0	0	0	2	1	0	3	0	0
Total Items by Type	10	3	0	10	3	0	10	1	1
Total Points	10	6	0	10	6	0	10	2	4

MC = multiple-choice, 1 point; SA2 = short-answer, 2 points; OR = open-response, 4 points
Total points possible for sessions 1 & 2: 32
Total points possible for sessions 2 & 3: 32

Table 3-3. 2004–2008 MEPA: Writing Test Blueprint

Outcomes Assessed	Session 1: 14 Points Largely Based on Visual Stimuli			Session 2: 14 Points 1 Writing Prompt			Session 3: 16 Points 3 Writing Prompts		
	SA1	SW	WP	MC	SW	WP	MC	SW	WP
Writing	6	3	0	0	3	1	0	0	3
Editing	0	1	0	4	0	0	4	0	0
Total Items by Type	6	4	0	4	3	1	4	0	3
Total Points	6	8	0	4	6	4	4	0	12

MC=multiple-choice, 1 point; SA1= short-answer,1 point; SW=sentence-writing, 2 points; WP=writing-prompt, 4 points
Total points possible for sessions 1 & 2: 28
Total points possible for sessions 2 & 3: 30

Each test session is briefly described below.

Session 1

- Reading: The student was asked to recognize and read simple words and phrases and comprehend short, simple reading passages.
- Writing: The student was asked to write simple words and sentences.

Session 2

- Reading: The student was asked to read and comprehend literary and informational text.
- Writing: The student was asked to write sentences and paragraphs in response to writing stimuli.

Session 3

- Reading: The student was asked to read and comprehend moderately difficult literary and informational text.
- Writing: The student was asked to edit grade-level text and write short compositions.

Chapter 4. TEST ADMINISTRATION

4.1 Requirements for Student Participation

Federal and state laws require that LEP students be assessed annually to measure their proficiency in reading, writing, listening, and speaking in English. The assessments must be administered to all students who are identified by their districts as LEP. The few exemptions from participation are listed in section 4.1.1.3.

Districts were required to have a procedure in place to assess the English proficiency of all students in grades K–12 whose home language is not English to determine if they are proficient in English. All students identified as LEP, regardless of their language support program, were required to be tested, even if a parent declined sheltered English immersion or any language support program for the student. Tests were administered to LEP students enrolled in public schools and to those educated with public funds placed in out-of-district programs.

4.1.1 LEP Students with Disabilities

Both state and federal law require the participation of students with disabilities in statewide testing programs. For the purposes of MEPA, students with disabilities had either an Individualized Education Program (IEP) provided under the Individuals with Disabilities Education Act or a plan provided under Section 504 of the Rehabilitation Act of 1973.

4.1.1.1 MELA-O

All LEP students with disabilities in grades K–12 were required to participate in MELA-O except for those students who were deaf or hard of hearing.

4.1.1.2 MEPA-R/W

When taking the MEPA-R/W, LEP students with disabilities in grades 3–12 were provided the same accommodations documented in their IEPs or 504 plans, except in the cases listed below.

Test accommodations are allowable changes in the routine conditions under which LEP students with disabilities take the MEPA-R/W tests. Accommodations were allowed in four areas: changes in timing or scheduling of the test; changes in test setting; changes in test presentation; and changes in how the student responded to questions. Because untimed test sessions were allowed for all students, additional time was not considered a test accommodation.

A list of frequently used accommodations was published annually, with guidelines for making accommodations decisions, in *Requirements for the Participation of Students with Disabilities in MCAS* (the 2008 publication is available at the Department's website at www.doe.mass.edu/mcas/participation/sped.doc). However, schools could contact the Department to discuss the use of other accommodations that did not appear on the published list. Accommodations were allowable as long as they did not alter the test itself, or provide coaching or assistance to the student during test administration. Out-of-level testing (i.e., taking the test at a grade span that was inappropriately matched to the student's actual grade) was also not permitted. Additional information regarding test accommodations can be found in the Department's publication *Requirements for the Participation of Students with Disabilities in MCAS*.

4.1.1.3 Exceptions to MEPA/R-W Testing

Students with disabilities who used the following accommodations in the classroom were **not** required to participate in the MEPA-R/W, unless another appropriate accommodation would allow them to participate:

- Braille
- Electronic text reader

In addition, the following LEP students with disabilities were **not** required to participate in the MEPA-R/W:

- students who required the MCAS Alternate Assessment
- students who were deaf or hard of hearing **and** required the signed administration of sessions 1 and 2 for the reading and/or writing tests (see section 4.2.2 for further information about the administration of sessions 1 and 2)

4.2 Schedule of MEPA Test Administration

The MEPA tests, MELA-O and MEPA-R/W, were administered twice during each school year, once in the fall and once in the spring. Students in grades K–12 took the MELA-O. Students in grades 3–12 also took the MEPA-R/W.

In the 2004–2005 school year, all LEP students were required to participate in both the fall and spring MEPA administrations. The fall MEPA administration established each student's baseline scores; the spring MEPA administration helped determine their progress in achieving proficiency in English. For students who enrolled that year in Massachusetts schools after the fall 2004 MEPA administration, the spring 2005 administration determined their baseline assessments.

In operational years 2005–2006, 2006–2007, and 2007–2008, all grade 3 LEP students and those LEP students newly enrolled in Massachusetts schools were required to participate in the respective fall MEPA administration to determine their baseline scores. Again, all LEP students were required to participate in each spring MEPA administration.

Table 4-1 shows the MEPA test administration dates for the period covered by this report.

Table 4-1. 2004–2008 MEPA: Test Administration Dates

Year	Fall Test Administration Period		Spring Test Administration Period	
	MELA-O	MEPA-R/W	MELA-O	MEPA-R.W
2004–2005	September 20–October 22	October 18–22	February 28–April 8	March 28–April 8
2005–2006	October 3–28	October 24–28	February 27–March 24	March 20–24
2006–2007	October 3–31	October 23–31	February 12–March 16	March 12–16
2007–2008	October 1–31	October 22–31	February 25–March 19	March 10–19

4.2.1 MELA-O Administration

The testing window for MELA-O was approximately one month long to allow sufficient time to observe LEP students engaging in a variety of classroom activities and determine appropriate scores in listening and speaking.

The MELA-O was to be administered only by an education professional who had been certified as a Qualified MELA-O Trainer (QMT) or a Qualified MELA-O Administrator (QMA). Training procedures are discussed further in section 5.1.2 of this report.

4.2.2 MEPA-R/W Administration

Each LEP student participated in only two of the three reading sessions and only two of the three writing sessions. Schools decided which two sessions each LEP student was to take, considering each component separately. Schools were to consider the proficiency level descriptors in the Massachusetts *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (June 2003)*,* as well as the following reading and writing skill levels, as they decided which two sessions in each component were appropriate for each LEP student:

- Sessions 1 and 2 assessed *Beginning* to *Early Intermediate* reading/writing skills.
- Sessions 2 and 3 assessed *Intermediate* to *Transitioning* reading/writing skills. Reading session 2 was composed of below-grade-level passages and items measuring reading comprehension; reading session 3 passages approached grade-level text, and session 3 items measured comprehension, inferential reading, and understanding of literary and expository text elements.

In making these decisions, schools were also instructed to review student scores on English proficiency assessments used by their districts and to consider observations by staff who worked closely with each student.

Chapter 5. SCORING

5.1 MELA-O Scoring

MELA-O scoring for each LEP student was completed at the school level. A Qualified MELA-O Administrator (QMA) or Qualified MELA-O Trainer (QMT) assigned scores based on observation of the student's classroom activities, and marked the student's scores on the student's individual MELA-O Scoring Matrix form (shown at the end of section 5.1.2).

Once assigned, MELA-O scores for grades K–2 LEP students were submitted electronically through the Department's security portal. MELA-O scores for LEP students in grades 3–12 were transcribed by QMAs/QMTs onto students' MEPA-R/W answer booklets, which were scanned by the testing contractor as described in section 5.2.1, and the MELA-O scores for these students were recorded for reporting at that time.

5.1.1 Methodology for Scoring the MELA-O

Each student received two MELA-O scores, one for listening (comprehension) and one for speaking (production). A single score ranging from 0 to 5 was assigned for listening. Speaking was scored in four separate subdomains: fluency, vocabulary, pronunciation, and grammar. Each subdomain was assigned a score of 0 to 5. The four subdomain scores were totaled to determine the student's overall speaking score, with a range of 0 to 20.

5.1.2 QMT/QMA Training

To become certified as a QMT, an education professional was required to participate in a specialized 12-hour training sponsored by the Department, and pass a Qualifying Test (described below) with a minimum score of 80% exact or adjacent scores. Using the QMT Training Manual, prospective QMTs were instructed in how to prepare and conduct training sessions for prospective QMAs.

To become certified as a QMA, an education professional was required to participate in a nine-hour training conducted by a QMT, and a minimum of one hour of practice classroom rating of students. Part of the nine hours of training included review and discussion of the MELA-O Training Tape, which consisted of 14 samples of different students engaged in speaking and listening activities in classrooms.

After training, participants took a Qualifying Test, which they were required to pass with a minimum score of 60% exact or adjacent scores. The Qualifying Test consisted of a videotape showing samples of six different students engaged in speaking and listening activities in classrooms. The test included students at varying levels of oral proficiency in elementary, middle, and high schools. Using the MELA-O Scoring Matrix (shown at the end of this section), each training participant recorded speaking and listening scores for each student on a Qualification Answer Sheet. This activity took approximately two hours.

Each participant's scores were transferred from his or her Qualification Answer Sheet to a Scoring Sheet by the QMT. In order to be "correct," each score was required to fall within the acceptable range noted on the Scoring Sheet.

- A score within the acceptable range for listening was awarded 1 point.
- A score within the acceptable range for speaking was awarded 1 point.

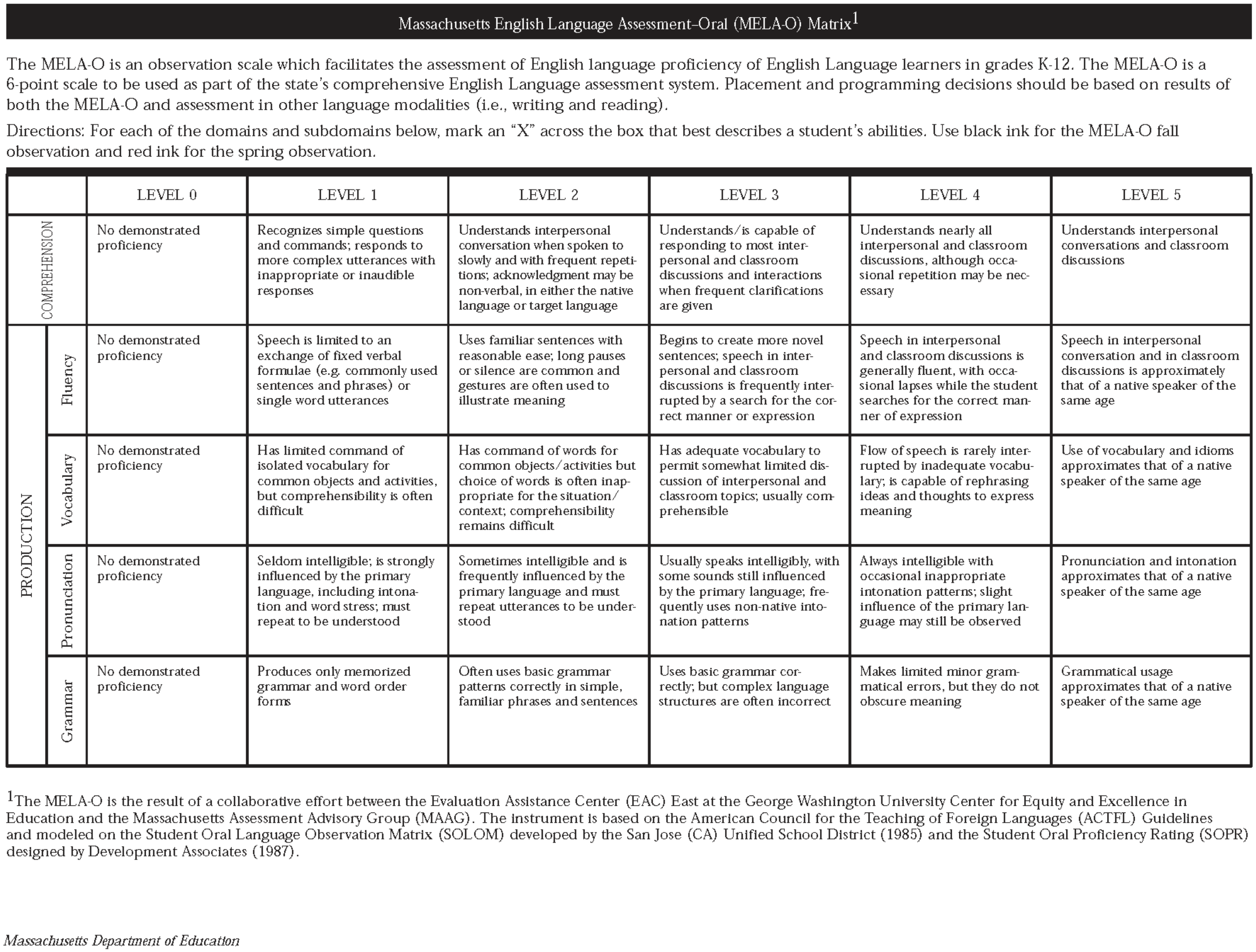
- These two scores were totaled to arrive at a sum for each sample of 0, 1, or 2 points.
- The minimum calibration passing score was 12 points total (a 60% minimum passing standard).
- Qualification Answer Sheets and Scoring Sheets for all passing participants were attached to their personal documentation and sent to the Massachusetts Department of Elementary and Secondary Education.

Participants who did not pass the Qualifying Test could spend additional time practicing scoring, using the MELA-O Scoring Matrix, the MELA-O Training Tape, and actual classroom observation. The QMT determined when to allow the participant additional opportunities to take the Qualifying Test.

Beginning in 2007, the qualification standards for certifying as a QMT or QMA were reset. The revised Qualifying Test provided 10 student samples; participants were required to assess the students across 5 matrix areas for a total of 50 possible scores. To re-qualify as a QMT, a participant was required to attain at least 35 exact scores with no more than two discrepant scores (those scores 2 or more points from exact), or 31–34 exact scores with no more than one discrepant score. To re-qualify as a QMA, a participant was required to attain at least 30 exact scores with no more than two discrepant scores, or 26–29 exact scores with no more than one discrepant score.

Two versions of the MELA-O Scoring Matrix are shown in Figures 5-1 and 5-2. The second version reflects a slight update that was operational for 2008 scoring.

Figure 5-1. 2004–2007 MEPA: MELA-O Scoring Matrix



Massachusetts English Language Assessment–Oral (MELA-O) Matrix[1]

The MELA-O is an observation scale which facilitates the assessment of English language proficiency of English Language learners in grades K-12. The MELA-O is a 6-point scale to be used as part of the state's comprehensive English Language assessment system. Placement and programming decisions should be based on results of both the MELA-O and assessment in other language modalities (i.e., writing and reading).

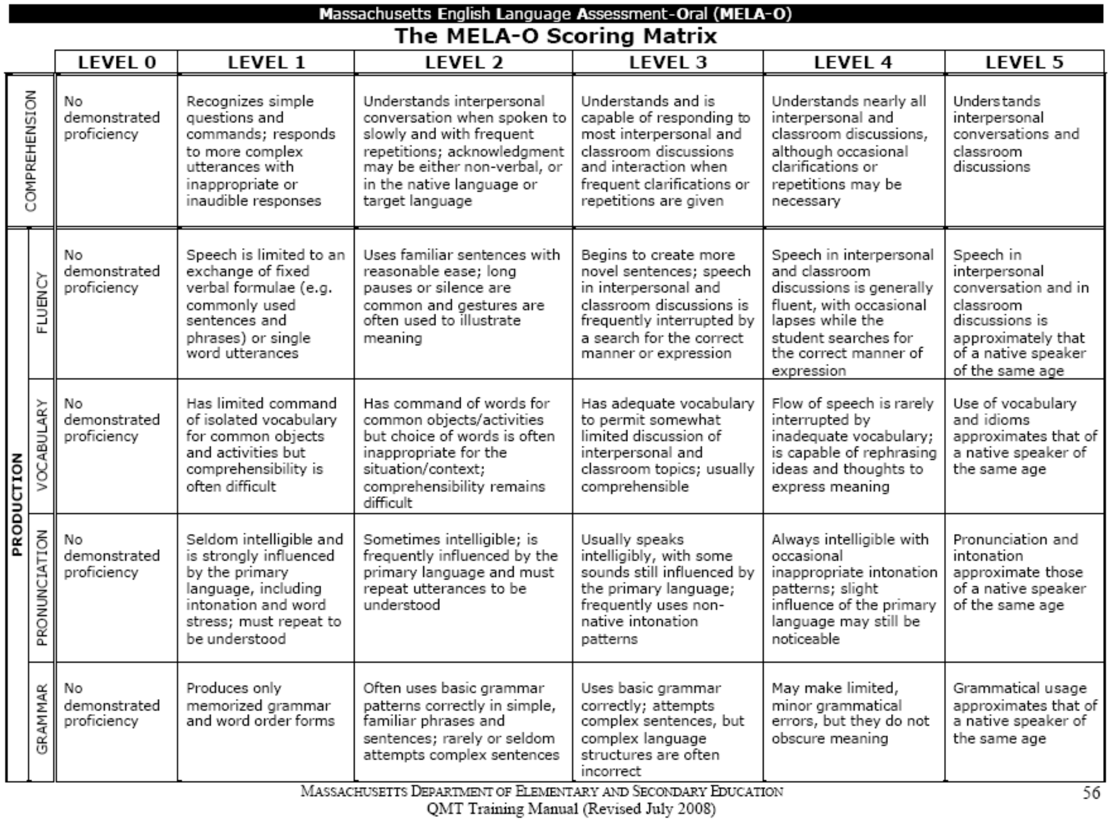
Directions: For each of the domains and subdomains below, mark an "X" across the box that best describes a student's abilities. Use black ink for the MELA-O fall observation and red ink for the spring observation.

		LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
COMPREHENSION		No demonstrated proficiency	Recognizes simple questions and commands; responds to more complex utterances with inappropriate or inaudible responses	Understands interpersonal conversation when spoken to slowly and with frequent repetitions; acknowledgment may be non-verbal, in either the native language or target language	Understands/is capable of responding to most interpersonal and classroom discussions and interactions when frequent clarifications are given	Understands nearly all interpersonal and classroom discussions, although occasional repetition may be necessary	Understands interpersonal conversations and classroom discussions
PRODUCTION	Fluency	No demonstrated proficiency	Speech is limited to an exchange of fixed verbal formulae (e.g. commonly used sentences and phrases) or single word utterances	Uses familiar sentences with reasonable ease; long pauses or silence are common and gestures are often used to illustrate meaning	Begins to create more novel sentences; speech in interpersonal and classroom discussions is frequently interrupted by a search for the correct manner or expression	Speech in interpersonal and classroom discussions is generally fluent, with occasional lapses while the student searches for the correct manner of expression	Speech in interpersonal conversation and in classroom discussions is approximately that of a native speaker of the same age
	Vocabulary	No demonstrated proficiency	Has limited command of isolated vocabulary for common objects and activities, but comprehensibility is often difficult	Has command of words for common objects/activities but choice of words is often inappropriate for the situation/context; comprehensibility remains difficult	Has adequate vocabulary to permit somewhat limited discussion of interpersonal and classroom topics; usually comprehensible	Flow of speech is rarely interrupted by inadequate vocabulary; is capable of rephrasing ideas and thoughts to express meaning	Use of vocabulary and idioms approximates that of a native speaker of the same age
	Pronunciation	No demonstrated proficiency	Seldom intelligible; is strongly influenced by the primary language, including intonation and word stress; must repeat to be understood	Sometimes intelligible and is frequently influenced by the primary language and must repeat utterances to be understood	Usually speaks intelligibly, with some sounds still influenced by the primary language; frequently uses non-native intonation patterns	Always intelligible with occasional inappropriate intonation patterns; slight influence of the primary language may still be observed	Pronunciation and intonation approximates that of a native speaker of the same age
	Grammar	No demonstrated proficiency	Produces only memorized grammar and word order forms	Often uses basic grammar patterns correctly in simple, familiar phrases and sentences	Uses basic grammar correctly; but complex language structures are often incorrect	Makes limited minor grammatical errors, but they do not obscure meaning	Grammatical usage approximates that of a native speaker of the same age

[1] The MELA-O is the result of a collaborative effort between the Evaluation Assistance Center (EAC) East at the George Washington University Center for Equity and Excellence in Education and the Massachusetts Assessment Advisory Group (MAAG). The instrument is based on the American Council for the Teaching of Foreign Languages (ACTFL) Guidelines and modeled on the Student Oral Language Observation Matrix (SOLOM) developed by the San Jose (CA) Unified School District (1985) and the Student Oral Proficiency Rating (SOPR) designed by Development Associates (1987).

Massachusetts Department of Education

Figure 5-2. 2008 MEPA: MELA-O Scoring Matrix



Massachusetts English Language Assessment-Oral (MELA-O)

The MELA-O Scoring Matrix

		LEVEL 0	LEVEL 1	LEVEL 2	LEVEL 3	LEVEL 4	LEVEL 5
COMPREHENSION		No demonstrated proficiency	Recognizes simple questions and commands; responds to more complex utterances with inappropriate or inaudible responses	Understands interpersonal conversation when spoken to slowly and with frequent repetitions; acknowledgment may be either non-verbal, or in the native language or target language	Understands and is capable of responding to most interpersonal and classroom discussions and interaction when frequent clarifications or repetitions are given	Understands nearly all interpersonal and classroom discussions, although occasional clarifications or repetitions may be necessary	Understands interpersonal conversations and classroom discussions
PRODUCTION	FLUENCY	No demonstrated proficiency	Speech is limited to an exchange of fixed verbal formulae (e.g. commonly used sentences and phrases) or single word utterances	Uses familiar sentences with reasonable ease; long pauses or silence are common and gestures are often used to illustrate meaning	Begins to create more novel sentences; speech in interpersonal and classroom discussions is frequently interrupted by a search for the correct manner or expression	Speech in interpersonal and classroom discussions is generally fluent, with occasional lapses while the student searches for the correct manner of expression	Speech in interpersonal conversation and in classroom discussions is approximately that of a native speaker of the same age
	VOCABULARY	No demonstrated proficiency	Has limited command of isolated vocabulary for common objects and activities but comprehensibility is often difficult	Has command of words for common objects/activities but choice of words is often inappropriate for the situation/context; comprehensibility remains difficult	Has adequate vocabulary to permit somewhat limited discussion of interpersonal and classroom topics; usually comprehensible	Flow of speech is rarely interrupted by inadequate vocabulary; is capable of rephrasing ideas and thoughts to express meaning	Use of vocabulary and idioms approximates that of a native speaker of the same age
	PRONUNCIATION	No demonstrated proficiency	Seldom intelligible and is strongly influenced by the primary language, including intonation and word stress; must repeat to be understood	Sometimes intelligible; is frequently influenced by the primary language and must repeat utterances to be understood	Usually speaks intelligibly, with some sounds still influenced by the primary language; frequently uses non-native intonation patterns	Always intelligible with occasional inappropriate intonation patterns; slight influence of the primary language may still be noticeable	Pronunciation and intonation approximate those of a native speaker of the same age
	GRAMMAR	No demonstrated proficiency	Produces only memorized grammar and word order forms	Often uses basic grammar patterns correctly in simple, familiar phrases and sentences; rarely or seldom attempts complex sentences	Uses basic grammar correctly; attempts complex sentences, but complex language structures are often incorrect	May make limited, minor grammatical errors, but they do not obscure meaning	Grammatical usage approximates that of a native speaker of the same age

MASSACHUSETTS DEPARTMENT OF ELEMENTARY AND SECONDARY EDUCATION 56
QMT Training Manual (Revised July 2008)

5.2 MEPA-R/W Scoring

5.2.1 Scanning of Answer Booklets

Once received by the testing contractor, each MEPA-R/W student answer booklet was scanned in its entirety into an electronic imaging system (iScore, a highly secure, server-to-server interface designed by Measured Progress). Student identification and demographic information, school information, and student answers to multiple-choice questions were converted to alphanumeric format; hand-written student responses were captured in digital image format (bitmaps).

MELA-O scores recorded on answer booklets for grades 3–12 students were also scanned and captured for reporting at this time.

5.2.2 Machine-Scored Items

Multiple-choice items were used in all sessions of the reading and writing tests. Student responses to these items were machine-scored by applying a scoring key to the captured responses. Correct answers were assigned a score of one point; incorrect answers were assigned a score of zero points. Blank responses and responses with multiple marks were also assigned zero points.

5.2.3 Hand-Scored Items

Student responses to open-response, short-answer, sentence-writing, and writing-prompt test items were individually read and evaluated by scorers employed by the testing contractor. Answer document images were sorted into item-specific groups for scoring purposes. A student's entire answer booklet was always available; however, for scoring purposes, scorers only reviewed response images one item at a time.

Measured Progress maintained strict security throughout the scoring process by using iScore, which ensured that access to student response images was restricted to scorers and those working in a scoring management capacity. District, school, and student names were not visible to scorers, thereby maintaining student confidentiality. Each student response, however, was linked through iScore to its original booklet number.

More information is provided below about the following aspects of hand-scoring:

- Scorer Recruitment and Qualification
- Methodology for Scoring Constructed-Response Items
- Training for Scoring Accuracy and Reliability
- Training of Scoring Leadership
- Operational Scoring Quality Control

5.2.3.1 Scorer Recruitment and Qualification

MEPA-R/W scorers were recruited and hired by the testing contractor. They comprised a diverse group of individuals with a wide range of backgrounds, ages, and experiences. Most scorers were quite experienced, having scored student responses for many other testing programs, and many had previously scored MEPA-R/W field-test responses.

All MEPA-R/W scorers completed at least two years of college; hiring preference was given to those with a four-year college degree. Potential scorers were required to submit documentation such as resumes and transcripts along with their applications. This documentation was carefully reviewed; if a potential scorer did not have at least two college credits with average or above-average grades in the specific content area to be scored, the scorer was eliminated from the applicant pool. Teachers and administrators (principals, guidance counselors, etc.) employed in Massachusetts schools were not eligible to score MEPA-R/W responses.

All scorers signed a non-disclosure/confidentiality agreement before being allowed to attend any of the training sessions.

5.2.3.2 Methodology for Scoring Constructed-Response Items

Scorers assigned scores based on item-specific scoring guides after receiving training on the items they were scoring.

Reading

Two types of constructed-response items were used on the MEPA reading test:

- short-answer (2 points)
- open-response (4 points)

Ten percent of the open-response items were double-blind scored, meaning they were scored independently by at least two different individuals.

Writing

The MEPA writing test used three types of constructed-response items:

- short-answer (1 point)
- sentence-writing (2 points)
- writing-prompt (4 points)

All writing constructed-response type items were double-blind scored.

5.2.3.3 Training for Scoring Accuracy and Reliability

Scorers were required to demonstrate the ability to score student responses accurately and consistently throughout the training, qualification, and scoring processes.

Chief Readers (CRs) employed by the testing contractor conducted scorer training. After introducing Measured Progress scoring staff and Massachusetts Department of Elementary and Secondary Education staff (if present), they presented an overview of the MEPA program that included MEPA's purposes and goals, unique features of the reading or writing tests, and a description of the testing population. This was followed by a general discussion about the confidentiality, security, and proprietary nature of testing and scoring materials, and about scoring procedures. After general guidelines about holistic scoring were shared, scorers began item-specific training.

Scorers thoroughly reviewed and discussed the Scoring Guide for the item they were to score. The Scoring Guide for each item included the item (or assignment) itself and a description of each score point and/or annotation. Scorers then carefully reviewed a large number of actual student responses from field test or previous test administrations that had been organized into three types of sets.

- Anchor Sets: Responses that were solid, exceptionally clear, typical examples of the score points; they were referred to throughout the training and scoring process as "true examples"
- Training Sets: Unusual, discussion-provoking responses (e.g., very high or low quality, short, exceptionally creative, disorganized) that further defined the score point by illustrating the range of responses typically encountered in operational scoring
- Qualifying Sets: Responses that were clear, typical examples of the score points

No scorer being trained was allowed to score live operational student responses until he or she achieved the minimum accuracy rate on a Qualifying Set. Each Qualifying Set consisted of 10 previously scored responses. The minimum accuracy rate was 70% exact matches on the pre-scored papers and 90% exact or adjacent agreement (i.e., only one non-matching paper but within a score point of being a match). For 1-point and 2-point items, any scorer who failed to meet the minimum standard was not allowed to score the item. For 4-point items, potential scorers who failed to meet this standard on the first Qualifying Set were retrained and subsequently scored a second Qualifying Set of 10 previously scored and approved responses. Potential scorers who failed to achieve the minimum accuracy rate for this second Qualifying Set were not allowed to score the item.

5.2.3.4 Training of Scoring Leadership

Scoring leadership, including Quality Assurance Coordinators (QACs) and Senior Readers (SRs), were trained prior to regular scorer training. Their training was identical to the scorer training described above, except that QACs and SRs were held to a higher minimum standard on Qualifying Sets: an accuracy rate of 80% exact and 90% exact or adjacent. If a potential QAC or SR did not achieve this minimum accuracy rate, he or she was not allowed to serve in a leadership role for that item. If they met the minimum accuracy rate for regular scorers, however, they could choose to act as a regular scorer for that item or to train for a leadership role in a different item or assignment.

5.2.3.5 Operational Scoring Quality Control

The scoring process was monitored by the Quality Assurance Coordinator and the Chief Reader. Chief Readers had the overall responsibility of ensuring that items were scored accurately, consistently, and according to approved scoring guidelines. There were separate Chief Readers for reading and writing.

The use of iScore enabled a constant measuring and monitoring of scorers for scoring accuracy and consistency; reading rates and total number of responses read were also monitored. During actual scoring of live operational student responses, scorers were required to maintain a daily scoring accuracy rate of 70% exact and 90% exact or adjacent, as measured by the following tools and techniques (each described in more detail below):

- embedded committee-reviewed responses
- read-behinds
- double-blinds
- computer-generated reports

There was a minimum scoring accuracy standard of 70% exact agreement and 90% exact or adjacent for embedded Committee-Reviewed Responses (CRRs), read-behinds, and double-blind scores. If a scorer fell below the minimum standard in any of these areas, iScore prevented further access to operational images and notified scoring leadership of the need for retraining. Scoring leadership determined whether or when a scorer was allowed to resume scoring. An individual scorer received only two opportunities to be retrained on a particular item. If a scorer fell below standard a third time, he or she was dismissed from scoring that item.

Embedded Committee-Reviewed Responses (CRRs)

Embedded CRRs are responses that were previously scored and whose scores had been reviewed and approved by Assistant Chief Readers or Chief Readers. Embedded CRRs were selected and loaded into the computerized scoring system for “blind” distribution to scorers. These responses looked identical to other live student responses. Therefore, during regular scoring, scorers did not know if a response was an embedded CRR or a live response. The Chief Readers had some flexibility in how they used embedded CRRs; some were pre-selected before scoring began and some were selected randomly during operational scoring. Some were released one at a time to scorers; some were released as an entire set of five or more responses. During the first full day of scoring, some items included 15 CRRs that were released at random points to ensure scorers were sufficiently calibrated at the beginning of scoring.

For some items, typically the 4-point items that had more potential for discrepant scores, 30 CRRs were available; scorers typically received 20 within the first 100 responses scored, and 10 additional responses within the next 100 responses scored.

Additionally, embedded CRRs were distributed throughout the scoring session so that they comprised roughly 2% of a scorer’s scores.

Read-Behinds

Scorers, especially those who needed retraining based on their CRR scores, were often monitored using read-behinds. The QAC and SR directed iScore to send a select number of responses, typically three at a time, and the scorer’s scores for them to a special queue accessible to the SR. Before viewing the scorer’s scores, the SR also scored the responses and recorded them. The system then compared the scores. Identical scores indicated that the individual scorer was calibrated to the state’s scoring guidelines. Differing scores indicated non-calibration, and the SR would have the opportunity to provide individualized scoring consultation to the scorer.

Double-Blinds

Double-blind scoring refers to responses that were scored independently by at least two different scorers who were unaware of each other’s scores.

Computer-Generated Reports

Scoring leadership utilized reports generated by iScore to ensure the following:

- overall accuracy, consistency, and reliability of scoring at the group level
- the availability of immediate, real-time individual scorer data, to allow early intervention that might have been necessary
- adherence to scoring schedules

Most reports were available to SRs and QACs at the scoring tables; other reports were only available to Chief Readers, Scoring Managers, and the Scoring Director.

The Department had full access to all reports; however, reports could be modified in such a way that scorers were identified by unique ID numbers rather than by name. The testing contractor typically provided the Department with the following reports:

- The Read-Behind Summary Report provided the total number of read-behind responses read by both a scorer and the Senior Reader/Quality Assurance Coordinator, noting the number and percentage of exact, adjacent, and discrepant scores.
- The Double-Blind Summary Report provided the total number of double-blind responses read by a scorer, noting the number and percentage of exact, adjacent, and discrepant scores.
- The Embedded CRR Summary provided for a scorer the total number of responses scored, the number of embedded CRR responses scored, and the number and percentage of exact, adjacent, and discrepant scores.

Chapter 6. EQUATING AND SCALING (COMPOSITE TEST LEVEL)

6.1 Equating

Both MEPA-R/W items and MELA-O indicators were analyzed through the use of Item Response Theory (IRT). Details on the IRT calibration are provided in 8.1.3. Item characteristics were also analyzed using standard classical test theory (CTT) methods (see section 8.1.1 and 8.1.2). Once IRT and CTT analyses were completed, four parallel MEPA-R/W test forms (A, B, C, and D) were assembled for operational use. Forms were administered as shown in Table 6-1.

Table 6-1. 2004–2008 MEPA:
Test Forms and Administration Dates

Administration	Test Form
Fall 2004	A*
Spring 2005	B
Fall 2005	B
Spring 2006	C
Fall 2006	C
Spring 2007	D
Fall 2007	B
Spring 2008	C

*Form A was released following the Fall 2004 administration.

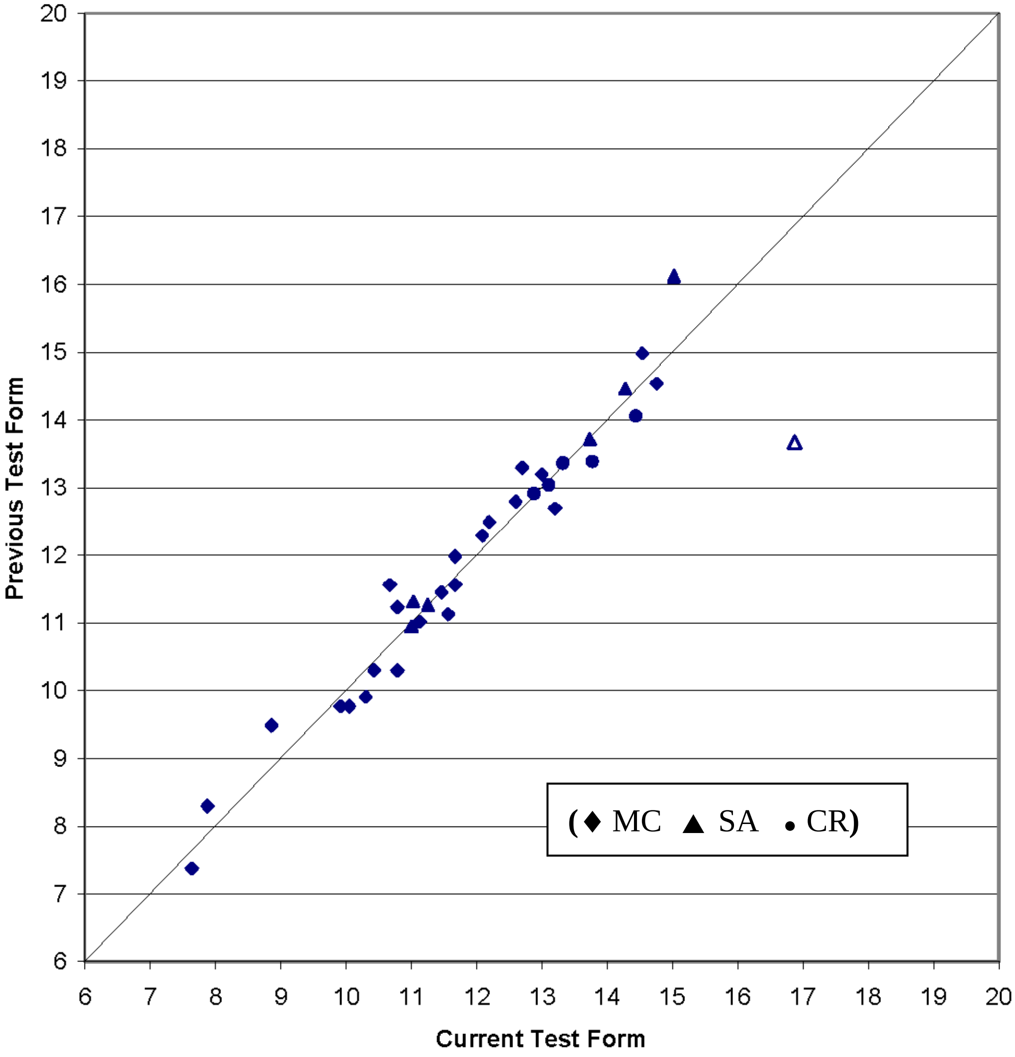
All operational MEPA-R/W items were originally calibrated to the IRT “base scale” of 2003–2004, when field tested. These calibrations determined IRT “pre-equated” parameters for all reading and writing items, with the exception of items that were modified after field testing. Any MEPA-R/W items that underwent modification were then calibrated onto the base scale using operational data the first time that the form in which they appeared went operational (along with the MELA-O indicators), fixing the parameters of unchanged MEPA-R/W items at their field-tested values. As a result, all MEPA-R/W items and MELA-O indicators are calibrated to the base scale.

The equating method described above is commonly known as the *anchor-test-nonequivalent-groups* design (Petersen, Kolen, & Hoover, 1989). The “anchor test,” in this case, was the set of items that remained unchanged from the field test. Note that the students who took the field test in 2003–2004 and those who took the operational test in any subsequent year were not equivalent groups. IRT is particularly useful in equating for nonequivalent groups (Allen & Yen, 1979), which is why the procedure was used for MEPA equating.

Prior to fixing the values of the parameters of the unchanged items, the items were evaluated for use as equating items using the delta method. Each item has two p-values, one for the field test and one for the operational test. (For open-response items, an adjusted p-value is used, calculated by taking the average item score and dividing by the maximum possible item score.) The p-values are transformed to the delta scale, which 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). The higher the delta value, the more difficult the item. The delta values were computed for evaluating potential equating items within grade spans (3–4, 5–6, 7–8, and 9–12).

Figure 6-1 illustrates how a delta plot is used to examine equating items. In the figure, different shapes identify different item types: ♦ for multiple-choice items, ▲ for short-answer items, and • for sentence-writing, open-response, and writing-prompt items. The perpendicular distance of each item to the regression line is computed. The unshaded shape in the illustration indicates the item with the greatest perpendicular distance from the regression line. Items that are not more than three standard deviations away from the regression line may be used as equating items. One item from Form B grade span 3–4, two items from Form C grade span 3–4, two items from Form C grade span 5–6, and one item from Form C grade span 9–12 were excluded from use as equating items as a result of the delta analyses. Tables showing the results of the delta analyses are provided in Appendix A; IRT item parameters for each grade span are provided in Appendix B.

Figure 6-1. 2004–2008 MEPA: Sample Delta Plot



6.2 Scaling

Overall scaled scores for MEPA ranged from 300 to 400. The scaled score cut points of 325 for the *Beginning/Early Intermediate* cut and 375 for the *Intermediate/Transitioning* cut were fixed across grade spans. The *Early Intermediate/Intermediate* scaled score cut point varied across grade spans depending on the location of the theta (θ) cut score established during MEPA standard setting in 2005, which is documented in the *2005 MEPA Technical Report.* Scaled score cut points are presented in Table 6-2.

Table 6-2. 2004–2008 MEPA: Scaled Score Cut Points

Grade Span	*Beginning/ Early Intermediate*	*Early Intermediate/ Intermediate*	*Intermediate/ Transitioning*
3–4	325	349	375
5–6	325	346	375
7–8	325	346	375
9–12	325	343	375

The scaled score (SS) for each student was calculated using the following formula:

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

where $\hat{\theta}$ is the student's estimated score on the theta scale.

The transformation line's slope, *m*, and intercept, *b*, were calculated as follows:

$$m = \frac{SS_3 - SS_1}{\theta_3 - \theta_1} = \frac{375 - 325}{\theta_3 - \theta_1}$$

$$b = SS_1 - m\theta_1$$

where SS_1 and SS_3 are the scaled score cuts, and θ_1 and θ_3 the theta cuts, between *Beginning/Early Intermediate*, and between *Intermediate/Transitioning*, respectively.

The transformation constants (slope and intercept) for each grade span are presented in Table 6-3.

Table 6-3. 2004–2008 MEPA: Transformation Constants for Composite MEPA Scores

Grade	Transformation Constants	
Span	Slope	Intercept
3–4	42.48	363.11
5–6	40.72	358.35
7–8	44.80	354.39
9–12	53.30	350.96

An estimated theta score ($\hat{\theta}$) was calculated for each student by translating his or her raw composite score to the corresponding θ score using the appropriate test characteristic curve (TCC). In deriving each student’s composite score, the treatment of MELA-O indicators was equivalent to that of MEPA-R/W reading and writing items. Note that the rubric and procedure for assigning MELA-O scores were common to each student; however, in the MEPA-R/W, for both reading and writing, some students took sessions 1 and 2 while others took sessions 2 and 3. Therefore, the IRT parameters for the MELA-O indicators and the MEPA-R/W reading and writing items were used together to calculate four TCCs for each administration of the MEPA (fall and spring), one for each possible combination of reading and writing sessions:

- reading and writing, sessions 1 and 2
- reading sessions 1 and 2, writing sessions 2 and 3
- reading sessions 2 and 3, writing sessions 1 and 2
- reading and writing, sessions 2 and 3

Appendix C provides tables showing each raw score and its corresponding theta and scaled scores for the overall composite scores for MEPA administrations between fall 2004 and spring 2008.

Appendix D displays TCCs and test information functions (TIFs) at the composite test level: four TCCs and four TIFs are provided for each grade span for MEPA administrations between fall 2004 and spring 2008. The TCCs show the expected (average) raw score corresponding to each θ value between -4.0 and 4.0 The TIFs display the amount of statistical information associated with each θ value. TIFs essentially depict test precision across the entire latent trait continuum.

6.2.1 Calculating Scaled Scores for Students with Extreme Low and High Scores

A so-called “dogleg” procedure was implemented at the bottom and top of the scaled score range (300 to 305 and 395 to 400, respectively) to ensure that the minimum and maximum raw scores translated to 300 and 400, respectively. The slope and intercept used to calculate the scaled scores for students whose estimated theta score ($\hat{\theta}$) corresponded to a scaled score of 305 or lower were calculated as follows:

$$m = \frac{305 - SS_{\min}}{\theta_{305} - \theta_{\min}} = \frac{305 - 300}{\theta_{305} - (-4.0)}$$

$$b = SS_{\min} - m\theta_{\min} = 300 - m(-4.0) = 300 + 4m$$

where θ_{305} is the θ value corresponding to a scaled score of 305, and the remaining terms are as defined above.

The scaled-score calculations for students at the extreme high end of the scaled score range (i.e., students whose estimated theta scores corresponded to a scaled score of 395 or higher) followed the same procedure, using a theta of θ_{395} instead of θ_{305}. Note that the transformation constants varied slightly depending on the administration (fall or spring) and the combination of sessions the student took. This is because the θ values corresponding to scaled scores of 305 and 395 varied somewhat across forms and sessions. The full set of transformation constants for extreme low and high scores is provided in Appendix E.

6.2.2 Composite Scaled Score Distributions

The composite scaled score distributions for each grade span for MEPA administrations between fall 2004 and spring 2008 are provided in Appendix F.

6.2.3 Scaled Score Error Band

In addition to the overall scaled score, an error band was also reported for each student. First, a raw score error band was calculated as follows:

$$UL_{raw} = RS + SE_{raw} \qquad \text{and} \qquad LL_{raw} = RS + SE_{raw}$$

where UL_{raw} and LL_{raw} are the upper and lower limit of the error band, respectively; RS is the student's raw score, and SE_{raw} is the standard error of measurement on the raw score scale.

SE_{raw} was calculated as follows (Lord & Novick, 1968):

$$SS_{raw} = \sqrt{\frac{RS(RS_{\max} - RS)}{RS_{\max} - 1}}$$

The maximum raw score varied depending on which combination of MEPA-R/W sessions the student took.

Once the raw score upper and lower limits were determined, they were translated into the corresponding values on the θ scale using the appropriate TCC. Finally, the θ scale upper and lower limits were scaled, using the appropriate slope and intercept terms, as described above. If either the upper or lower limit fell outside the scaled score range, it was truncated to the minimum or maximum scaled score value (300 or 400), as appropriate.

6.2.4 Reading and Writing Scaled Subscores

Because the total possible raw scores for MEPA-R/W reading and writing were different, and because the total possible raw score for writing varied depending on which sessions the student took, reading and writing raw scores were translated to a subscore scale that ranged from 1 to 30.

The reading scaled score (SS_r) was calculated as follows:

$$SS_R = m\hat{\theta}_R + b$$

where $\hat{\theta}_R$ is the student's estimated score on the theta scale for reading.

The slope and intercept were calculated as follows:

$$m = \frac{SS_{\max} - SS_{\min}}{\theta_{\max} - \theta_{\min}} = \frac{30 - 1}{4.0 - (-4.0)} = \frac{29}{8} = 3.625$$

$$b = SS_{\min} - m\theta_{\min} = 1 - 3.625(-4.0) = 15.5$$

The student's estimated reading theta score ($\hat{\theta}$) was obtained by translating his or her reading raw score to the corresponding θ value using the appropriate TCC, depending on which reading sessions the student took.

The process for determining the student's scaled score for writing was exactly the same as that described above for reading.

Tables showing the correspondence between reading and writing raw scores and their associated theta and scaled scores are provided in Appendix G.

Chapter 7. REPORTING OF RESULTS

MEPA results were reported in the form of performance levels and scaled scores for individual students, schools, districts, and the state. Students were assigned performance levels depending on the range within which their scaled scores fell, as determined through standard setting, described more fully below. MEPA results were provided via reports described in section 7.3. Sample reports are presented in Appendix H.

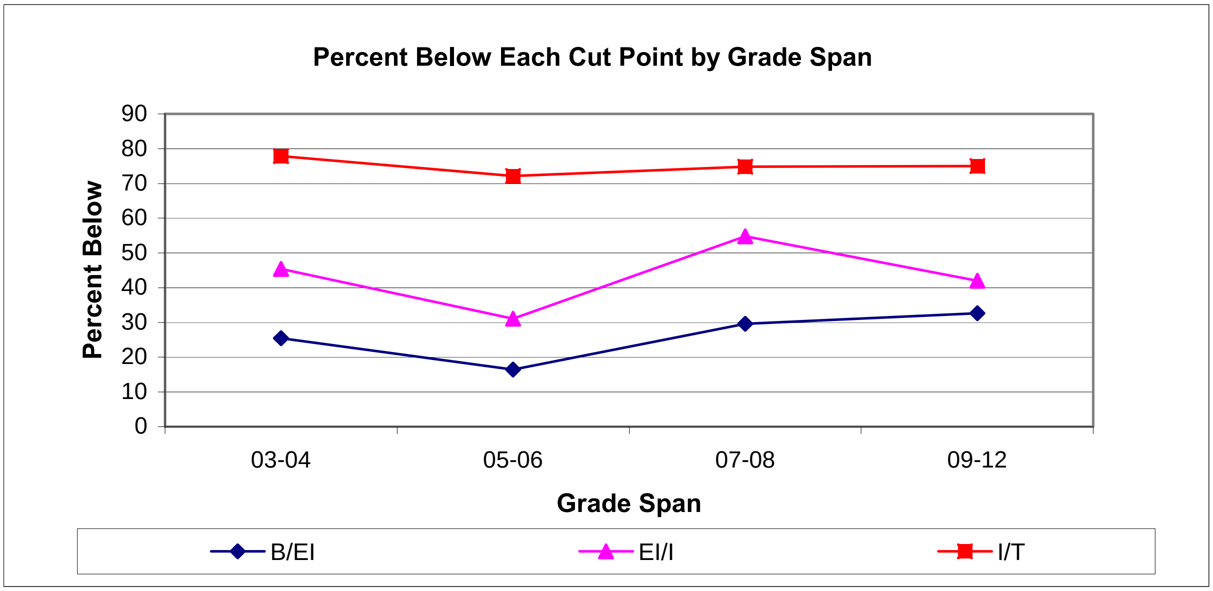
7.1 Standard Setting

Cut points for the MEPA were established at standard-setting meetings held February 2–4, 2005. Four panels were convened, one for each of the four grade spans (3–4, 5–6, 7–8, and 9–12). Using a modified version of the bookmark method, panelists recommended three cut points at each grade span: *Beginning/Early Intermediate, Early Intermediate/Intermediat*e, and *Intermediate/ Transitioning*. Panelists were first asked to familiarize themselves thoroughly with the assessment materials and performance level descriptors. They then went through three rounds of cut point placement; the final recommended cut points were the average placements from the third round.

One key component of the process was panelists' understanding of the format and logic of the ordered item booklet. The ordered item booklet displayed one item (or score category) per page, in ascending order of a standard IRT indicator of difficulty (see Appendix I for details). Like the MEPA-R/W items, MELA-O indicators had been calibrated via IRT; therefore, in the ordered item booklet, all MELA-O indicators were treated in the same manner as the MEPA-R/W constructed-response items. In particular, both MEPA-R/W items and MELA-O indicators appeared multiple times in the booklet, once per score point.

Once standard setting was complete, the results were evaluated to determine whether any adjustments needed to be made to the panelists' placements. Specifically, the percentage of students who would fall below each cut point was calculated based on the recommended cuts for each grade. Figure 7-1 shows that, while the cuts established for the *Intermediate/Transitioning* cut were fairly consistent across the four grade spans, there were some discrepancies for the other two cuts. In particular, the *Early Intermediate/Intermediate* cut for both grade span 5–6 and grade span 7–8, and the *Beginning/Early Intermediate* cut for grade span 5–6, showed some difference from cut points at the other grades.

Figure 7-1. 2004–2008 MEPA: Standard Setting Results



B/EI = *Beginning/Early Intermediate*, EI/I = *Early Intermediate/Intermediate,* I/T = *Intermediate/ Transitioning*

As a result of these discrepancies, smoothed cut points were also calculated by fitting a linear best-fit line to the lines shown above, then finding the theta cut value that corresponded to the smoothed percent-below value. Tables 7-1 through 7-4 show the original cut points, as recommended by the standard setting panelists, as well as the smoothed values.

Table 7-1. 2004–2008 MEPA: Standard Setting Results—Grade Span 3–4

Performance Level	Initial Cuts		Smoothed Cuts	
	Theta Cut	% in Category	Theta Cut	% in Category
Beginning		25.5		20.2
Early Intermediate	-0.727	20.0	-0.897	21.0
Intermediate	-0.269	32.5	-0.331	34.0
Transitioning	0.340	22.1	0.280	24.7

Table 7-2. 2004–2008 MEPA: Standard Setting Results—Grade Span 5–6

Performance Level	Initial Cuts		Smoothed Cuts	
	Theta Cut	% in Category	Theta Cut	% in Category
Beginning		16.4		24.1
Early Intermediate	-1.220	14.7	-0.819	18.2
Intermediate	-0.580	41.1	-0.299	32.8
Transitioning	0.343	27.8	0.409	24.8

Table 7-3. 2004–2008 MEPA: Standard Setting Results—Grade Span 7–8

Performance Level	Initial Cuts		Smoothed Cuts	
	Theta Cut	% in Category	Theta Cut	% in Category
Beginning		29.6		27.5
Early Intermediate	-0.582	25.1	-0.656	15.5
Intermediate	0.027	20.1	-0.194	31.2
Transitioning	0.472	25.2	0.460	25.8

Table 7-4. 2004–2008 MEPA: Standard Setting Results—Grade Span 9–12

Performance Level	Initial Cuts		Smoothed Cuts	
	Theta Cut	% in Category	Theta Cut	% in Category
Beginning		32.7		31.3
Early Intermediate	-0.450	9.3	-0.487	14.0
Intermediate	-0.205	33.0	-0.148	27.8
Transitioning	0.484	25.0	0.451	26.9

The final step in the standard setting process was to convene a panel to validate the smoothed cut points. The panel consisted of Department personnel and Measured Progress staff. Cut points for which the smoothed cut was more than one standard error of measurement different than the original cut were identified for validation. In addition, all four of the *Intermediate/Transitioning* cuts were identified for validation, since that cut is the most important for decision-making. In all, 8 of the 12 smoothed cuts were discussed by the panel. All cuts were found to be appropriate and consistent with the performance level descriptors. Therefore, the smoothed results were adopted as the final cut points for MEPA.

A complete report of the standard setting process is included as Appendix I.

7.2 Performance Level Descriptors

MEPA results were reported using four performance levels: *Beginning, Early Intermediate, Intermediate*, and *Transitioning*. The descriptors for each performance level are shown below.

- ***Beginning***. The student at this performance level is starting to develop the skills that will lead to effective communication in written and spoken English. A student performing at this level typically
  - recognizes simple written words and phrases
  - writes basic words or phrases, with frequent errors
  - speaks using basic words or phrases, with frequent errors
  - understands basic spoken vocabulary or phrases

- ***Early Intermediate***. The student at this performance level is developing skills that will lead to effective and complete communication in English. A student performing at this level typically
    - recognizes simple written words, phrases, and sentences, and reads and comprehends below-grade-level texts
    - writes short paragraphs with limited control of standard English conventions
    - speaks using common words and simple phrases; word choice is often inappropriate or incorrect
    - understands basic spoken vocabulary and phrases with frequent need for clarification

- ***Intermediate***. The student at this performance level demonstrates increasing skills in using and understanding English. Oral and written communication, although somewhat inconsistent, is solid and usually understandable. A student performing at this level typically
    - recognizes common written words and some academic words, and comprehends simple grade-level texts
    - writes short, simple compositions with partial control of standard English conventions
    - speaks using common words and phrases, and basic grammar and sentence structure; uses complex language structures but with occasional errors
    - understands most oral communication, with some need for clarification

- ***Transitioning***. The student at this performance level has achieved age-appropriate basic fluency in English, including reading, writing, listening, and speaking. A student performing at this level typically
    - recognizes most common and academic words, and reads and comprehends moderately difficult grade-level texts
    - writes short compositions demonstrating general control of standard English conventions
    - speaks using appropriate and correct words, phrases, and expressions, as well as basic and complex grammar and sentence structures
    - understands extended and prolonged oral communication, with little or no need for clarification

7.3 Student, School, and District Reports

Results for the 2004–2005, 2005–2006, 2006–2007, and 2007–2008 MEPA administrations for students in grades 3–12 were provided in the following reports:

- Spring MEPA Tests: Preliminary Participation Report
- Spring MEPA Tests: Preliminary Results by Year of Enrollment in U.S. (2004–2005, 2005-2006) or Massachusetts (2006–2007, 2007–2008) Schools
- Spring MEPA Tests: Roster of Student Results
- MEPA School and District Final Results
- Spring MEPA Parent/Guardian Report

The *Roster of Student Results* reports were generated and provided to schools and districts following the fall MEPA administration only. MEPA tests are intended to measure students' progress in acquiring fluency in English, and the fall MEPA administration is meant strictly to determine baseline scores. Complete MEPA results for both administrations were reported following the spring administration.[1]

Each report is briefly described below. Appendix H provides sample MEPA reports of results for the spring 2006 and spring 2008 MEPA administrations. Reports for the spring 2005 administration were identical (i.e., parallel) to those for spring 2006, and reports for the spring 2007 administration were identical to those for spring 2008; the spring 2005 and spring 2007 reports are therefore not provided. Reports are provided for grade span 3–4; reports for other grade spans are parallel and are therefore not provided. Additional interpretive information for these reports is provided in the Department's publication *Guide to Interpreting the MEPA Reports for Schools and Districts*.

7.3.1 Preliminary Reports

The following two reports were generated for each grade span (3–4, 5–6, 7–8, 9–12; no preliminary reports were generated for K–2) in a school:

- Spring MEPA Tests: Preliminary Participation Report (all years)
- Spring MEPA Tests: Preliminary Results by Year of Enrollment in U.S. (2004–2005, 2005–2006) or Massachusetts (2006–2007, 2007–2008) Schools

Each report is described below and in more detail in the *Guide to Interpreting the MEPA Reports for Schools and Districts*.

To ensure student confidentiality and to discourage generalizations about school performance based on very small student populations, a report was only generated for a grade span if more than 10 students in that grade span were tested in a school.

The data in these preliminary reports were generated based on the answer booklets received by the testing contractor following testing.[2] Copies of a school's preliminary reports were furnished to both the school and its district.

7.3.1.1 Spring MEPA Tests: Preliminary Participation Report

This report shows, for the school receiving the report, the following data for the grade span of the report:

- the number of students for whom answer booklets were received following testing; this number includes both students who were tested and those who did not participate
- the number of students who participated in testing
- the number of students who did not participate in testing in each category of non-participation (e.g., medically documented absence)
- the percentages of students who participated in each MEPA test, and in both MEPA tests

[1] For those students who participated in and had complete subcategory scores for both fall and spring MEPA testing, results were shown in the spring reports for both MEPA administrations. For a small number of students who participated in both MEPA administrations but whose results could not be linked through the students' State-Assigned Student Identification number (SASID), results were only reported for the MEPA administration linked to their SASIDs.

[2] Final participation results were based on whether answer booklets could be linked to students' SASID numbers; linked results were compared to Massachusetts' Student Information Management System (SIMS) LEP enrollment data to determine actual participation rates.

7.3.1.2 Spring MEPA Tests: Preliminary Results by Year of Enrollment in U.S. or Massachusetts Schools

This report shows, for the school receiving the report, student results for the grade span of the report in each of the following categories:

- the number and percentage of students for whom answer booklets were received following testing; this number includes both students who were tested and those who did not participate, and includes any student in grades 3–12 who took the MELA-O and/or the MEPA-R/W
- the overall average MEPA scaled score (only students who had complete scores in reading, writing, listening, and speaking were included in this calculation)
- the number and percentage of students in each performance level category (only students who had complete scores in reading, writing, listening, and speaking were included in this calculation)

For 2004–2005 and 2005–2006, results in this report were aggregated by the number of years students had been enrolled in United States schools for 1 year, 2 years, and 3 or more years. For 2006–2007 and 2007–2008, results were aggregated by the number of years students had been enrolled in Massachusetts schools for 1 year, 2 years, 3 years, 4 years, and 5 or more years.

7.3.2 Spring MEPA Tests: Roster of Student Results

This report provides to a school the MEPA results for each LEP student at that school. A separate *Roster of Student Results* report for each grade span was generated following each MEPA administration. Each LEP student enrolled at the school in the grade span of the report is listed alphabetically by last name, and his/her overall scaled score and performance level are shown, as well as his/her scaled subscores in reading, writing, listening, and speaking.[3] The *Roster of Student Results* shows results for up to three of the following MEPA test administrations: spring 2008, fall 2007, spring 2007, fall 2006, spring 2006, fall 2005, spring 2005, and fall 2004. If a student participated in more than one test administration, and his or her records from each administration were able to be matched based on student records from the Student Information Management System (SIMS), results for each administration were reported. If a student participated in only the spring administration for a given year, or if his or her records from previous administrations could not be matched based on SIMS, results from only the spring administration of that year were reported.

7.3.3 MEPA School and District Final Results

These reports were generated at the school level and the district level for grade spans 3–4, 5–6, 7–8, and 9–12; no report was generated for K–2. The school level report shows results for the relevant grade span at that school only; the district report shows results for that grade span from all schools in the district. Results were generated based on comparisons with SIMS LEP enrollment data.

The final results reports show data, for each grade span, in the following areas:

- comparative performance levels (all years)

[3] Since the number of possible points was the same for each student on the MELA-O, listening and speaking subscores were reported as raw scores. Because the total possible raw scores for MEPA-R/W reading and writing could vary, reading and writing subscores were reported as scaled scores. Further information on the scaling of these two subscores is provided in section 6.2.4 of this report.

- students in *Transitioning* (2005–2006, 2006–2007, 2007–2008)
- average score changes (all years)

To ensure student confidentiality and to discourage generalizations about school performance based on very small student populations, school and district final results reports were generated only if 10 or more students were represented in one of the following:

- the number of students enrolled and identified as LEP from October 1–March 1
- the number of students included in both fall and spring MEPA testing in a given school year

Additionally, final results in these reports only include results for students with complete scores in all four subcategories (reading, writing, listening, and speaking) for *both* the fall and spring MEPA administrations.

7.3.3.1 Comparative Performance Levels

Performance level comparisons were made between the following test administrations:

- students tested in the same grade span in two consecutive spring administrations (e.g., spring 2007 and spring 2008)
- students tested in the same grade span in the fall and spring administrations of the same school year (e.g., fall 2007 and spring 2008)

The following information is provided:

- the number of students at each performance level who participated and had complete subcategory scores for each MEPA administration
- the numbers and percentages of students whose performance level improved, maintained, or declined from fall to spring (when the total number of students in any performance level was less than 10, summary results for that performance level were not shown)

7.3.3.2 Students in Transitioning

The following information is provided:

- the number and percentage of students in the *Transitioning* performance level by the number of years in Massachusetts public schools

7.3.3.3 Average Score Changes

Because of the way MEPA tests were designed, comparisons between scaled scores from two different grade span tests were not necessarily valid. Results include:

- school (in school reports only), district, and state average scores for students who participated in and had complete subcategory scores for both MEPA administrations in the same grade span
- the differences in average scores between both MEPA administrations in the same grade span

7.3.4 Spring MEPA Parent/Guardian Report

This report shows students and their parents/guardians how the student performed in the MEPA administration(s) in which he/she participated. If a student participated in both the fall and the spring MEPA administrations, results were included for both administrations. If a student participated in both administrations but, for either fall or spring, was missing a score in any of the four scoring areas—reading, writing, listening, and speaking his or her results were not shown on the *Parent/Guardian Report* for the administration with the missing score. If a student participated in only one MEPA administration and had a missing score in one of the four scoring areas, no *Parent/Guardian Report* was generated.

A *Guide to the MEPA for Parents/Guardians* was provided with each *Parent/Guardian Report*, to assist parents/guardians and students in understanding and interpreting the results shown.

Shown on the top half of the *Parent/Guardian Report* results page are the student's overall MEPA scaled score and performance level, for the current year and up to two prior years if available. The score is also depicted graphically on a 300 to 400 scaled score range, surrounded by a standard error bar bracketing the student's expected score were he or she to take the test multiple times.

The bottom half of the results page gives two tools for comparing the student's scores to other criteria: a comparison of the student's score to the average *Transitioning* performance level score, and a comparison of the student's performance to the performance of students enrolled for various numbers of years in schools in the U.S. (2004–2005, 2006–2007) or in Massachusetts (2006–2007, 2007–2008). Each comparison is described below and in more detail in the *Guide to the MEPA for Parents/Guardians.*

7.3.4.1 Comparison to Transitioning Averages

Provided on each student's *Parent/Guardian Report* is a display comparing the student's performance to the average performance of a student *at or just above* the *Transitioning* performance level cut point (see sections 6.2 and 7.1 for information on cut points). MEPA results were sorted by overall scaled score least to greatest, and the scores of the first 500 students who received a scaled score of 375 or higher (the *Transitioning* cut point) were used to determine the average score for this display. Then the student's performance in each domain (reading, writing, listening, and speaking) is indicated as follows.

- If the student scored one or more standard deviations below the average, the "Below" box is checked.
- If the student scored between the average and one standard deviation below it, the "Approaching" box is checked.
- If the student scored at or above the average, the "At or Above" box is checked.

7.3.4.2 Comparison Relating to Number of Years of Enrollment

Also provided is a display showing statewide percentages of students at each performance level based on their numbers of years of enrollment in U.S. (2004–2005, 2005–2006) or in Massachusetts (2006–2007, 2007–2008) schools, with the student's performance level superimposed in the appropriate spot.

Chapter 8. STATISTICAL SUMMARIES

8.1 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* (1999) and the *Code of Fair Testing Practices in Education* (1988) include standards for identifying quality items. Items should assess only knowledge or skills that are identified as part of the domain being measured and should avoid assessing irrelevant factors. They should also be unambiguous and free of grammatical errors, potentially insensitive content or language, and other confounding characteristics. Further, questions must not unfairly disadvantage test takers from particular racial, ethnic, or gender groups.

Both qualitative and quantitative analyses were conducted to ensure that MEPA-R/W questions and MELA-O indicators meet these standards. Previous sections in this report have delineated various qualitative checks. This section of the report presents three categories of quantitative statistical evaluations: 1) difficulty indices, 2) item-test correlations, and 3) subgroup differences in item performance. Item response theory analyses are also discussed.

The results presented in this section are based on the fall 2004 through spring 2008 MEPA administrations. Throughout section 8.1, MELA-O indicators are included with constructed-response data.

8.1.1 Difficulty Indices and Item-Test Correlations

8.1.1.1 Difficulty Indices

All items 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 item. Multiple-choice items were scored dichotomously (correct versus incorrect), so for these items the difficulty index is simply the proportion of students who answered the item correctly. Constructed-response items were scored on a scale of either 0–2 or 0–4 points, and MELA-O indicators were scored on a scale of 0–5 points. By computing the difficulty index as the average proportion of points received, the indices for multiple-choice, constructed-response, and MELA-O indicators were placed on a similar scale; the index ranges from 0 to 1 regardless of the item type. Although this index is traditionally called a measure of difficulty, it is properly interpreted as an *easiness* index because larger values indicate easier items. An index of 0 indicates that no student received credit for the item, and an index of 1 indicates that every student received full credit for the item.

Items that were answered correctly by almost all students provide little information about differences in students’ performance, but they do indicate knowledge or skills that have been mastered by most students. Similarly, items that were correctly answered by very few students may indicate knowledge or skills that have not yet been mastered by most students, but such items provide little information about differences in students’ performance. In general, to provide best measurement, difficulty indices should range from near-chance performance (0.25 for four-option, multiple-choice items or essentially 0 for constructed-response items) to 0.90. Indices outside this range indicate items that were either too difficult or too easy for the target population.

Although difficulty is an important item characteristic, the relationship between performance on an item and performance on the whole test or a relevant test section may be more critical. An item that assesses relevant knowledge or skills should relate to other items that are purported to be measuring the same knowledge or skills.

8.1.1.2 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 MEPA-R/W multiple-choice items was the point-biserial correlation between item score and a criterion total score on the test. As such, the index ranges from –1 to 1, with the magnitude and sign of the index indicating the relationship's strength and direction, respectively. For constructed-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 0.3 to 0.6.

In general, discrimination indices are interpreted as indicating the degree to which high- and low-performing students responded differently on an item or, equivalently, the degree to which responses to an item help to differentiate between high- and low-performing students. From this perspective, indices near 1 indicate that high-performing students are more likely to answer the item correctly, indices near –1 indicate that low-performing students are more likely to answer the item correctly, and indices near 0 indicate that the item is equally likely to be answered correctly by high- and low-performing students.

Discrimination indices can be thought of as measures of how closely an item assesses the same knowledge and skills assessed by other items contributing to the criterion total score; that is, the discrimination index can be 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 2004–2008 MEPA, the criterion score for each item is the total score for all items.

8.1.1.3 Summary of Item Analysis Results

Summary statistics of the difficulty and discrimination indices for each item type are provided in Tables 8-1 through 8-4. In general, the item difficulty and discrimination indices are in acceptable and expected ranges. Very few items were answered correctly at near-chance rates; with the exception of the easier session 1 items, very few were answered correctly at near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing consistent constructs, and students who performed well on individual items tended to perform well overall. There were a small number of items with near-zero discrimination indices, but none was reliably negative. Occasionally, items with less desirable statistical characteristics need to be included in assessments to ensure that content is appropriately covered, but there were very few such cases in MEPA.

A comparison of indices across grade levels is complicated because these indices are population dependent. Direct comparisons would require that either the items or students were common across groups. Similarly, comparing the difficulty indices of multiple-choice and constructed-response items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that, in most cases, the difficulty indices for multiple-choice items are higher (indicating easier items) than the difficulty indices for constructed-response items. Similarly, the partial credit allowed for constructed-response items is advantageous in the computation of item-test correlations, so the discrimination indices for these items tend to be larger than the discrimination indices of other item types.

Table 8-1. 2004–2008 MEPA: Average Difficulty and Discrimination of Different Item Types for Composite Score for Grades 3–4

Administration	Statistics	Item Type		
		All	Multiple-Choice	Constructed-Response
Fall 2004	Difficulty	0.62 (0.15)	0.62 (0.16)	0.63 (0.14)
	Discrimination	0.49 (0.16)	0.39 (0.11)	0.62 (0.13)
	n	68	38	30
Spring 2005	Difficulty	0.71 (0.15)	0.73 (0.17)	0.69 (0.14)
	Discrimination	0.50 (0.15)	0.41 (0.12)	0.60 (0.12)
	n	68	38	30
Fall 2005	Difficulty	0.61 (0.16)	0.63 (0.17)	0.59 (0.15)
	Discrimination	0.51 (0.17)	0.41 (0.13)	0.63 (0.13)
	n	68	38	30
Spring 2006	Difficulty	0.71 (0.13)	0.71 (0.14)	0.72 (0.12)
	Discrimination	0.49 (0.16)	0.39 (0.11)	0.62 (0.12)
	n	68	38	30
Fall 2006	Difficulty	0.60 (0.15)	0.59 (0.15)	0.61 (0.14)
	Discrimination	0.50 (0.19)	0.38 (0.12)	0.67 (0.13)
	n	68	38	30
Spring 2007	Difficulty	0.72 (0.14)	0.74 (0.15)	0.70 (0.13)
	Discrimination	0.51 (0.16)	0.42 (0.11)	0.63 (0.12)
	n	68	38	30
Fall 2007	Difficulty	0.62 (0.17)	0.63 (0.17)	0.60 (0.16)
	Discrimination	0.50 (0.17)	0.40 (0.13)	0.62 (0.12)
	n	68	38	30
Spring 2008	Difficulty	0.72 (0.14)	0.71 (0.14)	0.73 (0.13)
	Discrimination	0.48 (0.16)	0.38 (0.11)	0.62 (0.12)
	n	68	38	30

**Table 8-2. 2004–2008 MEPA:
Average Difficulty and Discrimination of
Different Item Types for Composite Score for Grades 5–6**

Administration	Statistics	Item Type		
		All	Multiple-Choice	Constructed-Response
Fall 2004	Difficulty	0.70 (0.14)	0.74 (0.11)	0.65 (0.15)
	Discrimination	0.51 (0.16)	0.40 (0.09)	0.64 (0.12)
	n	68	38	30
Spring 2005	Difficulty	0.72 (0.13)	0.75 (0.10)	0.67 (0.14)
	Discrimination	0.53 (0.14)	0.44 (0.08)	0.65 (0.12)
	n	68	38	30
Fall 2005	Difficulty	0.54 (0.15)	0.61 (0.15)	0.45 (0.11)
	Discrimination	0.58 (0.17)	0.47 (0.09)	0.72 (0.12)
	n	68	38	30
Spring 2006	Difficulty	0.71 (0.13)	0.71 (0.13)	0.70 (0.14)
	Discrimination	0.49 (0.16)	0.38 (0.09)	0.63 (0.12)
	n	68	38	30
Fall 2006	Difficulty	0.55 (0.14)	0.59 (0.14)	0.49 (0.13)
	Discrimination	0.55 (0.17)	0.42 (0.1)	0.70 (0.11)
	n	68	38	30
Spring 2007	Difficulty	0.72 (0.13)	0.73 (0.13)	0.71 (0.12)
	Discrimination	0.50 (0.16)	0.39 (0.10)	0.64 (0.10)
	n	68	38	30
Fall 2007	Difficulty	0.56 (0.14)	0.62 (0.14)	0.47 (0.11)
	Discrimination	0.57 (0.16)	0.47 (0.09)	0.71 (0.12)
	n	68	38	30
Spring 2008	Difficulty	0.73 (0.14)	0.73 (0.14)	0.72 (0.14)
	Discrimination	0.47 (0.16)	0.36 (0.09)	0.6 (0.13)
	n	68	38	30

Table 8-3. 2004–2008 MEPA: Average Difficulty and Discrimination of Different Item Types for Composite Score for Grades 7–8

Administration	Statistics	Item Type		
		All	Multiple-Choice	Constructed-Response
Fall 2004	Difficulty	0.67 (0.15)	0.70 (0.14)	0.64 (0.15)
	Discrimination	0.50 (0.15)	0.40 (0.07)	0.63 (0.13)
	n	68	38	30
Spring 2005	Difficulty	0.69 (0.15)	0.72 (0.14)	0.65 (0.16)
	Discrimination	0.49 (0.17)	0.39 (0.11)	0.62 (0.13)
	n	68	38	30
Fall 2005	Difficulty	0.57 (0.18)	0.63 (0.18)	0.49 (0.13)
	Discrimination	0.52 (0.19)	0.40 (0.13)	0.66 (0.15)
	n	68	38	30
Spring 2006	Difficulty	0.70 (0.14)	0.73 (0.13)	0.65 (0.15)
	Discrimination	0.50 (0.15)	0.41 (0.09)	0.62 (0.13)
	n	68	38	30
Fall 2006	Difficulty	0.56 (0.17)	0.62 (0.15)	0.47 (0.15)
	Discrimination	0.54 (0.16)	0.44 (0.10)	0.67 (0.13)
	n	68	38	30
Spring 2007	Difficulty	0.67 (0.15)	0.69 (0.16)	0.65 (0.14)
	Discrimination	0.49 (0.16)	0.39 (0.11)	0.63 (0.12)
	n	68	38	30
Fall 2007	Difficulty	0.55 (0.17)	0.62 (0.18)	0.48 (0.12)
	Discrimination	0.52 (0.19)	0.40 (0.12)	0.67 (0.14)
	n	68	38	30
Spring 2008	Difficulty	0.70 (0.14)	0.74 (0.13)	0.66 (0.15)
	Discrimination	0.48 (0.15)	0.39 (0.10)	0.59 (0.13)
	n	68	38	30

Table 8-4. 2004–2008 MEPA:
Average Difficulty and Discrimination of
Different Item Types for Composite Score for Grades 9–12

Administration	Statistics	Item Type		
		All	Multiple-Choice	Constructed-Response
Fall 2004	Difficulty	0.65 (0.15)	0.65 (0.13)	0.65 (0.17)
	Discrimination	0.46 (0.16)	0.37 (0.09)	0.58 (0.14)
	n	68	38	30
Spring 2005	Difficulty	0.65 (0.14)	0.63 (0.14)	0.68 (0.13)
	Discrimination	0.45 (0.16)	0.34 (0.10)	0.60 (0.10)
	n	68	38	30
Fall 2005	Difficulty	0.55 (0.14)	0.56 (0.14)	0.54 (0.13)
	Discrimination	0.49 (0.17)	0.36 (0.10)	0.64 (0.11)
	n	68	38	30
Spring 2006	Difficulty	0.65 (0.15)	0.64 (0.14)	0.67 (0.16)
	Discrimination	0.45 (0.16)	0.35 (0.09)	0.59 (0.12)
	n	68	38	30
Fall 2006	Difficulty	0.56 (0.15)	0.58 (0.15)	0.53 (0.15)
	Discrimination	0.48 (0.18)	0.36 (0.1)	0.64 (0.13)
	n	68	38	30
Spring 2007	Difficulty	0.66 (0.15)	0.64 (0.15)	0.68 (0.14)
	Discrimination	0.44 (0.17)	0.33 (0.10)	0.57 (0.13)
	n	68	38	30
Fall 2007	Difficulty	0.57 (0.13)	0.58 (0.14)	0.57 (0.13)
	Discrimination	0.48 (0.18)	0.35 (0.10)	0.64 (0.12)
	n	68	38	30
Spring 2008	Difficulty	0.67 (0.15)	0.66 (0.15)	0.68 (0.16)
	Discrimination	0.41 (0.17)	0.3 (0.10)	0.56 (0.14)
	n	68	38	30

8.1.2 Subgroup Differences: Differential Item Functioning (DIF)

The *Code of Fair Testing Practices in Education* (1988) 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* (1999) includes similar guidelines.

As part of the effort to identify such problems, MEPA items were evaluated in terms of differential item functioning (DIF) statistics. DIF procedures are designed to identify items for which subgroups of interest perform differently beyond the impact of differences in overall achievement. DIF indices indicate differential performance between two groups; however, the indices that categorize items as “low” or “high” DIF must not be interpreted as indisputable evidence of bias. Course-taking patterns, differences in group interests, or differences in school curricula can lead to differential performance. What must first be determined is whether the cause of this differential performance is construct-relevant. If differences in subgroup performance on an item can be plausibly attributed to construct-relevant factors, the item may be included in calculations of results.

The standardization DIF procedure (Dorans and Kulick, 1986) was used to evaluate differences among three MEPA subgroups: male/female, white/black, and white/Hispanic. This procedure calculates the average item performance for each subgroup at every total score. Then an overall average is calculated, weighting the total score distribution so it is the same for the reference and the focal group (e.g., male and female). The index ranges from –1 to 1 for multiple-choice items and is adjusted to the same scale for constructed-response items. Negative numbers indicate that the item was more difficult for female or non-white students. Dorans and Holland (1993) suggest that index values between –0.05 and 0.05 should be considered negligible. Dorans and Holland further state that items with values between –0.10 and –0.05 and between 0.05 and 0.10 (i.e., “low” DIF) should be inspected to ensure that no possible effect is overlooked, and that items with values less than –0.10 and greater than 0.10 (i.e., “high” DIF) are more unusual and should be examined very carefully.

Each MEPA item was categorized according to the guidelines adapted from Dorans and Holland (1993). Most MEPA items fell within the negligible range. Tables 8-5 to 8-8 show the number of items classified into each category separately by item type (multiple-choice versus constructed-response) for the following subgroup comparisons: male/female, white/black, and white/Hispanic. (Blank cells indicate comparisons for which there were insufficient numbers of students to compute reliable results.) Tables 8-9 to 8-12 show the number of items, by item type, that favor males or females in each of the three DIF categories.

Table 8-5. 2004–2008 MEPA: DIF Analysis by Session and Item Type for Grades 3–4

Administration	Session	Male/Female DIF Class									White/Black DIF Class									White/Hispanic DIF Class								
		All			MC			CR			All			MC			CR			All			MC			CR		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Fall 2004	1	23	0	0	10	0	0	13	0	0	22	1	0	10	0	0	12	1	0	19	4	0	8	2	0	11	2	0
	2	25	1	0	13	1	0	12	0	0	26	0	0	14	0	0	12	0	0	23	3	0	11	3	0	12	0	0
	3	13	6	0	10	4	0	3	2	0	15	4	0	11	3	0	4	1	0	18	1	0	13	1	0	5	0	0
Spring 2005	1	23	0	0	10	0	0	13	0	0	19	3	1	8	2	0	11	1	1	20	2	1	8	1	1	12	1	0
	2	25	1	0	13	1	0	12	0	0	25	1	0	13	1	0	12	0	0	26	0	0	14	0	0	12	0	0
	3	16	3	0	12	2	0	4	1	0	17	1	1	13	0	1	4	1	0	17	2	0	12	2	0	5	0	0
Fall 2005	1	22	1	0	10	0	0	12	1	0	15	3	0	7	0	0	8	3	0	20	3	0	8	2	0	12	1	0
	2	25	1	0	13	1	0	12	0	0	24	2	0	13	1	0	11	1	0	26	0	0	14	0	0	12	0	0
	3	19	0	0	14	0	0	5	0	0	0	0	0	0	0	0	0	0	0	18	1	0	13	1	0	5	0	0
Spring 2006	1	23	0	0	10	0	0	13	0	0	20	3	0	9	1	0	11	2	0	19	3	1	8	1	1	11	2	0
	2	25	1	0	13	1	0	12	0	0	20	5	1	10	3	1	10	2	0	23	3	0	11	3	0	12	0	0
	3	16	3	0	12	2	0	4	1	0	15	3	1	10	3	1	5	0	0	14	3	2	9	3	2	5	0	0
Fall 2006	1	22	1	0	10	0	0	12	1	0	17	6	0	9	1	0	8	5	0	19	4	0	8	2	0	11	2	0
	2	23	3	0	11	3	0	12	0	0	22	4	0	10	4	0	12	0	0	24	2	0	12	2	0	12	0	0
	3	16	3	0	12	2	0	4	1	0	14	4	1	9	4	1	5	0	0	13	5	1	8	5	1	5	0	0
Spring 2007	1	23	0	0	10	0	0	13	0	0	19	4	0	7	3	0	12	1	0	22	1	0	9	1	0	13	0	0
	2	26	0	0	14	0	0	12	0	0	26	0	0	14	0	0	12	0	0	25	1	0	13	1	0	12	0	0
	3	18	1	0	13	1	0	5	0	0	17	2	0	12	2	0	5	0	0	17	2	0	12	2	0	5	0	0
Fall 2007	1	21	2	0	9	1	0	12	1	0	18	4	1	8	2	0	10	2	1	19	4	0	9	1	0	10	3	0
	2	26	0	0	14	0	0	12	0	0	24	2	0	13	1	0	11	1	0	24	2	0	12	2	0	12	0	0
	3	18	1	0	13	1	0	5	0	0	17	2	0	12	2	0	5	0	0	18	1	0	13	1	0	5	0	0
Spring 2008	1	23	0	0	10	0	0	13	0	0	22	1	0	10	0	0	12	1	0	22	1	0	9	1	0	13	0	0
	2	24	2	0	12	2	0	12	0	0	25	1	0	13	1	0	12	0	0	24	2	0	12	2	0	12	0	0
	3	18	1	0	13	1	0	5	0	0	16	3	0	11	3	0	5	0	0	17	2	0	12	2	0	5	0	0

All = all items, MC = multiple-choice, CR = constructed-response. A = negligible DIF, B = low DIF, C = high DIF

Table 8-6. 2004–2008 MEPA: DIF Analysis by Session and Item Type for Grades 5–6

Administration	Session	Male/Female DIF Class									White/Black DIF Class									White/Hispanic DIF Class								
		All			MC			CR			All			MC			CR			All			MC			CR		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Fall 2004	1	23	0	0	10	0	0	13	0	0	19	3	1	7	2	1	12	1	0	19	4	0	8	2	0	11	2	0
	2	26	0	0	14	0	0	12	0	0	22	4	0	11	3	0	11	1	0	17	8	1	7	7	0	10	1	1
	3	19	0	0	14	0	0	5	0	0	14	5	0	10	4	0	4	1	0	18	1	0	13	1	0	5	0	0
Spring 2005	1	21	2	0	9	1	0	12	1	0	17	5	1	5	4	1	12	1	0	19	4	0	8	2	0	11	2	0
	2	25	1	0	14	0	0	11	1	0	23	2	1	11	2	1	12	0	0	22	3	1	10	3	1	12	0	0
	3	18	1	0	14	0	0	4	1	0	16	2	1	11	2	1	5	0	0	16	3	0	11	3	0	5	0	0
Fall 2005	1	22	1	0	9	1	0	13	0	0																		
	2	25	1	0	14	0	0	11	1	0																		
	3	0	0	0	0	0	0	0	0	0																		
Spring 2006	1	22	1	0	9	1	0	13	0	0	15	5	3	6	2	2	9	3	1	21	2	0	9	1	0	12	1	0
	2	23	3	0	13	1	0	10	2	0	23	3	0	11	3	0	12	0	0	24	1	1	12	1	1	12	0	0
	3	17	2	0	13	1	0	4	1	0	14	4	1	9	4	1	5	0	0	14	5	0	9	5	0	5	0	0
Fall 2006	1	18	4	1	7	2	1	11	2	0																		
	2	25	1	0	13	1	0	12	0	0																		
	3	0	0	0	0	0	0	0	0	0																		
Spring 2007	1	23	0	0	10	0	0	13	0	0	18	4	1	6	3	1	12	1	0	18	3	2	5	3	2	13	0	0
	2	22	3	1	12	1	1	10	2	0	22	3	1	11	2	1	11	1	0	21	4	1	9	4	1	12	0	0
	3	16	3	0	14	0	0	2	3	0	18	1	0	13	1	0	5	0	0	15	4	0	10	4	0	5	0	0
Fall 2007	1	22	1	0	9	1	0	13	0	0																		
	2	21	5	0	9	5	0	12	0	0																		
	3	0	0	0	0	0	0	0	0	0																		
Spring 2008	1	23	0	0	10	0	0	13	0	0	19	2	1	8	1	1	11	1	0	21	2	0	8	2	0	13	0	0
	2	24	2	0	13	1	0	11	1	0	22	4	0	10	4	0	12	0	0	23	3	0	11	3	0	12	0	0
	3	17	2	0	14	0	0	3	2	0	15	4	0	10	4	0	5	0	0	13	5	1	8	5	1	5	0	0

All = all items, MC = multiple-choice, CR = constructed-response. A = negligible DIF, B = low DIF, C = high DIF

Table 8-7. 2004–2008 MEPA: DIF Analysis by Session and Item Type for Grades 7–8

Administration	Session	Male/Female DIF Class									White/Black DIF Class									White/Hispanic DIF Class								
		All			MC			CR			All			MC			CR			All			MC			CR		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Fall 2004	1	22	1	0	10	0	0	12	1	0	16	7	0	8	2	0	8	5	0	20	1	2	10	0	0	10	1	2
	2	24	2	0	14	0	0	10	2	0	22	2	2	10	2	2	12	0	0	9	14	3	4	7	3	5	7	0
	3	17	2	0	12	2	0	5	0	0	18	1	0	13	1	0	5	0	0	14	4	1	10	3	1	4	1	0
Spring 2005	1	22	1	0	9	1	0	13	0	0	17	4	2	8	1	1	9	3	1	17	5	1	8	2	0	9	3	1
	2	22	4	0	11	3	0	11	1	0	17	9	0	8	6	0	9	3	0	18	7	1	7	6	1	11	1	0
	3	15	4	0	11	3	0	4	1	0	15	4	0	11	3	0	4	1	0	15	3	1	11	2	1	4	1	0
Fall 2005	1	19	4	0	8	2	0	11	2	0																		
	2	20	6	0	10	4	0	10	2	0																		
	3	0	0	0	0	0	0	0	0	0																		
Spring 2006	1	21	2	0	8	2	0	13	0	0	7	3	0	3	2	0	4	1	0	21	1	1	9	0	1	12	1	0
	2	24	2	0	13	1	0	11	1	0	18	4	4	7	3	4	11	1	0	19	5	2	8	4	2	11	1	0
	3	18	1	0	14	0	0	4	1	0	16	2	1	12	1	1	4	1	0	14	5	0	9	5	0	5	0	0
Fall 2006	1	19	4	0	7	3	0	12	1	0																		
	2	23	3	0	12	2	0	11	1	0																		
	3	0	0	0	0	0	0	0	0	0																		
Spring 2007	1	22	1	0	10	0	0	12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	25	1	0	14	0	0	11	1	0	20	6	0	9	5	0	11	1	0	20	4	2	8	4	2	12	0	0
	3	14	4	1	11	2	1	3	2	0	16	2	1	11	2	1	5	0	0	16	2	1	11	2	1	5	0	0
Fall 2007	1	22	1	0	9	1	0	13	0	0																		
	2	20	6	0	11	3	0	9	3	0																		
	3	0	0	0	0	0	0	0	0	0																		
Spring 2008	1	21	2	0	9	1	0	12	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	25	1	0	14	0	0	11	1	0	21	4	1	9	4	1	12	0	0	22	2	2	10	2	2	12	0	0
	3	18	1	0	13	1	0	5	0	0	14	3	2	9	3	2	5	0	0	13	6	0	8	6	0	5	0	0

All = all items, MC = multiple-choice, CR = constructed-response. A = negligible DIF, B = low DIF, C = high DIF

Table 8-8. 2004–2008 MEPA: DIF Analysis by Session and Item Type for Grades 9–12

Administration	Session	Male/Female DIF Class									White/Black DIF Class									White/Hispanic DIF Class								
		All			MC			CR			All			MC			CR			All			MC			CR		
		A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Fall 2004	1	22	1	0	9	1	0	13	0	0	16	6	1	6	3	1	10	3	0	20	2	1	8	1	1	12	1	0
	2	24	1	1	13	0	1	11	1	0	23	3	0	12	2	0	11	1	0	22	4	0	11	3	0	11	1	0
	3	15	4	0	13	1	0	2	3	0	16	2	1	11	2	1	5	0	0	18	0	1	13	0	1	5	0	0
Spring 2005	1	21	2	0	8	2	0	13	0	0	13	7	3	3	4	3	10	3	0	17	4	2	4	4	2	13	0	0
	2	23	3	0	13	1	0	10	2	0	19	6	1	8	5	1	11	1	0	23	3	0	11	3	0	12	0	0
	3	15	4	0	13	1	0	2	3	0	13	5	1	9	4	1	4	1	0	14	5	0	9	5	0	5	0	0
Fall 2005	1	20	3	0	7	3	0	13	0	0																		
	2	26	0	0	14	0	0	12	0	0																		
	3	12	7	0	9	5	0	3	2	0																		
Spring 2006	1	21	2	0	8	2	0	13	0	0	16	7	0	5	5	0	11	2	0	16	7	0	6	4	0	10	3	0
	2	24	2	0	13	1	0	11	1	0	19	6	1	7	6	1	12	0	0	21	4	1	10	3	1	11	1	0
	3	17	2	0	14	0	0	3	2	0	13	5	1	9	4	1	4	1	0	16	3	0	11	3	0	5	0	0
Fall 2006	1	20	2	1	7	2	1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	25	1	0	13	1	0	12	0	0	17	7	2	7	5	2	10	2	0	11	13	2	9	4	1	2	9	1
	3	14	5	0	12	2	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Spring 2007	1	22	1	0	9	1	0	13	0	0	15	6	2	4	5	1	11	1	1	16	5	2	5	3	2	11	2	0
	2	25	1	0	14	0	0	11	1	0	22	4	0	10	4	0	12	0	0	23	3	0	12	2	0	11	1	0
	3	15	4	0	13	1	0	2	3	0	15	2	2	11	1	2	4	1	0	15	3	1	10	3	1	5	0	0
Fall 2007	1	21	1	1	8	1	1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	2	24	2	0	13	1	0	11	1	0	7	0	1	3	0	1	4	0	0	14	9	3	7	5	2	7	4	1
	3	12	6	1	9	4	1	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Spring 2008	1	22	1	0	9	1	0	13	0	0	16	6	1	6	4	0	10	2	1	17	6	0	6	4	0	11	2	0
	2	24	2	0	13	1	0	11	1	0	22	3	1	10	3	1	12	0	0	21	4	1	9	4	1	12	0	0
	3	16	3	0	13	1	0	3	2	0	13	6	0	9	5	0	4	1	0	13	6	0	8	6	0	5	0	0

All = all items, MC = multiple-choice, CR = constructed-response. A = negligible DIF, B = low DIF, C = high DIF

Table 8-9. 2004–2008 MEPA:
DIF Categorization by Item Type: Grades 3–4

Administration	Item Type	Negligible DIF				Low DIF				High DIF			
		Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%
Fall 2004	MC	20	13	33	87	2	3	5	13	0	0	0	0
	CR	21	7	28	93	1	1	2	7	0	0	0	0
Spring 2005	MC	20	15	35	92	1	2	3	8	0	0	0	0
	CR	21	8	29	97	1	0	1	3	0	0	0	0
Fall 2005	MC	21	16	37	97	0	1	1	3	0	0	0	0
	CR	21	8	29	97	1	0	1	3	0	0	0	0
Spring 2006	MC	20	15	35	92	0	3	3	8	0	0	0	0
	CR	24	5	29	97	1	0	1	3	0	0	0	0
Fall 2006	MC	12	21	33	87	1	4	5	13	0	0	0	0
	CR	25	3	28	93	2	0	2	7	0	0	0	0
Spring 2007	MC	21	16	37	97	0	1	1	3	0	0	0	0
	CR	22	8	30	100	0	0	0	0	0	0	0	0
Fall 2007	MC	22	14	36	95	0	2	2	5	0	0	0	0
	CR	22	7	29	97	1	0	1	3	0	0	0	0
Spring 2008	MC	19	16	35	92	1	2	3	8	0	0	0	0
	CR	22	8	30	100	0	0	0	0	0	0	0	0

MC = multiple-choice, CR = constructed-response

Table 8-10. 2004–2008 MEPA:
DIF Categorization by Item Type: Grades 5–6

Administration	Item Type	Negligible DIF				Low DIF				High DIF			
		Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%
Fall 2004	MC	22	16	38	100	0	0	0	0	0	0	0	0
	CR	19	11	30	100	0	0	0	0	0	0	0	0
Spring 2005	MC	19	18	37	97	0	1	1	3	0	0	0	0
	CR	20	7	27	90	3	0	3	10	0	0	0	0
Fall 2005	MC	17	6	23	61	0	1	1	3	0	0	0	0
	CR	15	9	24	80	1	0	1	3	0	0	0	0
Spring 2006	MC	14	21	35	92	1	2	3	8	0	0	0	0
	CR	21	6	27	90	3	0	3	10	0	0	0	0
Fall 2006	MC	10	10	20	53	1	2	3	8	1	0	1	3
	CR	12	11	23	77	1	1	2	7	0	0	0	0
Spring 2007	MC	23	13	36	95	0	1	1	3	0	1	1	3
	CR	20	5	25	83	5	0	5	17	0	0	0	0
Fall 2007	MC	8	10	18	47	4	2	6	16	0	0	0	0
	CR	18	7	25	83	0	0	0	0	0	0	0	0
Spring 2008	MC	26	11	37	97	0	1	1	3	0	0	0	0
	CR	18	9	27	90	3	0	3	10	0	0	0	0

MC = multiple-choice, CR = constructed-response

Table 8-11. 2004–2008 MEPA:
DIF Categorization by Item Type: Grades 7–8

Administration	Item Type	Negligible DIF				Low DIF				High DIF			
		Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%
Fall 2004	MC	24	12	36	95	0	2	2	5	0	0	0	0
	CR	15	12	27	90	3	0	3	10	0	0	0	0
Spring 2005	MC	13	18	31	82	1	6	7	18	0	0	0	0
	CR	19	9	28	93	2	0	2	7	0	0	0	0
Fall 2005	MC	10	8	18	47	2	4	6	16	0	0	0	0
	CR	12	9	21	70	4	0	4	13	0	0	0	0
Spring 2006	MC	21	14	35	92	0	3	3	8	0	0	0	0
	CR	21	7	28	93	2	0	2	7	0	0	0	0
Fall 2006	MC	11	8	19	50	2	3	5	13	0	0	0	0
	CR	9	14	23	77	1	1	2	7	0	0	0	0
Spring 2007	MC	19	16	35	92	0	2	2	5	0	1	1	3
	CR	15	11	26	87	4	0	4	13	0	0	0	0
Fall 2007	MC	8	12	20	53	0	4	4	11	0	0	0	0
	CR	10	12	22	73	3	0	3	10	0	0	0	0
Spring 2008	MC	18	18	36	95	0	2	2	5	0	0	0	0
	CR	20	8	28	93	2	0	2	7	0	0	0	0

MC = multiple-choice, CR = constructed-response

Table 8-12. 2004–2008 MEPA:
DIF Categorization by Item Type: Grades 9–12

Administration	Item Type	Negligible DIF				Low DIF				High DIF			
		Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%	Favor Female	Favor Male	*n*	%
Fall 2004	MC	17	18	35	92	0	2	2	5	0	1	1	3
	CR	19	7	26	87	4	0	4	13	0	0	0	0
Spring 2005	MC	19	15	34	89	1	3	4	11	0	0	0	0
	CR	18	7	25	83	5	0	5	17	0	0	0	0
Fall 2005	MC	17	13	30	79	2	6	8	21	0	0	0	0
	CR	16	12	28	93	2	0	2	7	0	0	0	0
Spring 2006	MC	18	17	35	92	2	1	3	8	0	0	0	0
	CR	17	10	27	90	3	0	3	10	0	0	0	0
Fall 2006	MC	14	18	32	84	1	4	5	13	0	1	1	3
	CR	19	8	27	90	3	0	3	10	0	0	0	0
Spring 2007	MC	25	11	36	95	0	2	2	5	0	0	0	0
	CR	16	10	26	87	4	0	4	13	0	0	0	0
Fall 2007	MC	17	13	30	79	4	2	6	16	0	2	2	5
	CR	15	12	27	90	3	0	3	10	0	0	0	0
Spring 2008	MC	21	14	35	92	1	2	3	8	0	0	0	0
	CR	21	6	27	90	3	0	3	10	0	0	0	0

MC = multiple-choice, CR = constructed-response

8.1.3 Item Response Theory Analyses

All MEPA-R/W items and MELA-O indicators were calibrated using item response theory (IRT) methodology. IRT uses mathematical models to define a relationship between an unobserved measure of a student's knowledge or level of preparedness, usually referred to as theta (θ), and the probability (p) of getting a dichotomous item correct or of getting a particular score on a polytomous item. In IRT, it is assumed that all items are independent measures of the same construct (i.e., the same θ).

There are several commonly used IRT models to specify the relationship between θ and p (Hambleton and van der Linden, 1997; Hambleton and Swaminathan, 1985). The generalized partial credit model (GPCM) was employed for MELA-O indicators and polytomous MEPA-R/W items, and can be defined as:

$$P_{ijk}\left(k \middle| \theta_i, \zeta_j\right) = \frac{\exp \sum_{v=0}^{k} \left[Da_j\left(\theta_i - b_j + d_v\right)\right]}{\sum_{c=1}^{m} \exp \sum_{v=1}^{c} \left[Da_j\left(\theta_i - b_j + d_v\right)\right]}$$

where
k represents an observed category score,
θ represents student ability for student *i*,
ζ represents the set of estimated item parameters for item *j*,
i indexes the student,
j indexes the item,
v indexes response category,
m represents total number of response categories,
a represents item discrimination,
b represents item difficulty,
d represents a category step parameter, and
D is a normalizing constant equal to approximately 1.701.

In the case of MEPA, the a_j term in the above equation is equal to 1.0 for all items. The one-parameter logistic (1PL) model was employed for dichotomous MEPA-R/W items. For these items, the above equation reduces to the following:

$$P_j\left(\theta_i\right) = \frac{\exp\left(\theta_i - b_j\right)}{1 + \exp\left(\theta_i - b_j\right)}$$

The process of determining the specific mathematical relationship between θ and p is referred to as item calibration. Once items are calibrated, they are defined by a set of parameters which specify a non-linear, monotonically increasing relationship between θ and p. Once the item parameters are known, the $\hat{\theta}$ for each student can be calculated. In IRT, $\hat{\theta}$ is considered to be an estimate of the student's true score and has some characteristics that may make its use preferable to the use of raw scores in rank ordering students. Parscale Version 4.1 was used to complete the IRT analyses. For more information about item calibration and $\hat{\theta}$ determination, the reader is referred to Lord and Novick (1968) or Hambleton and Swaminathan (1985).

8.2 Assessment Reliability

Although each individual item's performance is an important focus for evaluation, a complete evaluation of an assessment must also address the way that items 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 with perfect accuracy: some students will receive scores that underestimate their true level of knowledge, and other students will receive scores that over estimate their true level of knowledge. Items that function well together produce assessments that have less measurement error (i.e., errors made should be few 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 items into two groups and then correlate students' scores on the two half-tests. This is known as a split-half estimate of reliability. If the two half-test scores correlate highly, items on the two half-tests must be measuring very similar knowledge or skills. This is evidence that the items complement one another and function well as a group. This also suggests that measurement error will be minimal.

In the determination of assessment reliability for MEPA, MELA-O speaking and listening indicators were treated in the same manner as MEPA-R/W reading and writing test items. Throughout section 8.2, MELA-O indicators have been included with constructed-response data.

8.2.1 Reliability and Standard Errors of Measurement

8.2.1.1 Cronbach's α Coefficient

The split-half method requires the psychometrician to select which items 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. Cronbach's α coefficient is an estimate of the average of all possible split-half reliability coefficients.

Cronbach's α coefficient is computed using the following formula:

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

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

Table 8-13 presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement for each MEPA administration and grade span.

Table 8-13. 2004–2008 MEPA: Reliabilities, Standard Errors of Measurement, and Descriptive Statistics for Composite Scores

Grade Span	Administration	Sessions*	*n*	Points	Min	Max	Mean	S.D.	Rel.	S.E.M.
3–4	Fall 2004	1/2	4615	85	0	85	48.22	18.32	0.95	4.11
		2/3	3279	87	7	84	56.60	11.43	0.90	3.62
	Spring 2005	1/2	3856	85	0	84	52.91	17.96	0.95	4.02
		2/3	4850	87	5	85	63.07	10.57	0.89	3.51
	Fall 2005	1/2	2362	85	0	84	49.34	18.17	0.96	3.63
		2/3	2208	87	10	83	54.43	11.81	0.91	3.54
	Spring 2006	1/2	1994	85	0	85	54.02	17.50	0.95	3.91
		2/3	3089	87	8	87	60.86	10.33	0.89	3.43
	Fall 2006	1/2	2380	85	0	83	49.45	19.12	0.96	3.82
		2/3	2614	87	9	84	53.96	11.02	0.89	3.65
	Spring 2007	1/2	1914	85	0	83	53.36	18.84	0.96	3.77
		2/3	3413	87	5	83	60.46	11.05	0.91	3.31
	Fall 2007	1/2	2368	85	0	83	48.58	18.51	0.95	4.14
		2/3	2693	87	5	83	54.03	11.87	0.91	3.56
	Spring 2008	1/2	1905	85	3	83	53.59	17.62	0.95	3.94
		2/3	3431	87	10	85	59.80	10.44	0.89	3.46
5–6	Fall 2004	1/2	2684	85	0	84	48.23	19.34	0.96	4.06
		2/3	3210	87	3	87	60.87	11.10	0.90	3.56
	Spring 2005	1/2	2334	85	0	84	50.96	19.69	0.96	3.94
		2/3	4128	87	4	87	63.01	11.05	0.90	3.50
	Fall 2005	1/2	436	85	0	82	34.37	21.20	0.96	4.24
		2/3	133	87	5	85	60.32	13.57	0.93	3.59
	Spring 2006	1/2	1198	85	0	82	52.01	17.62	0.95	3.94
		2/3	2128	87	1	86	62.07	10.77	0.89	3.57
	Fall 2006	1/2	389	85	0	77	34.68	20.45	0.96	4.09
		2/3	143	87	9	84	56.87	14.13	0.92	4.00
	Spring 2007	1/2	1187	85	1	84	51.89	18.89	0.95	4.22
		2/3	2554	87	7	87	62.81	11.13	0.90	3.52
	Fall 2007	1/2	417	85	0	81	35.93	21.37	0.96	4.27
		2/3	168	87	7	85	58.80	14.47	0.93	3.83
	Spring 2008	1/2	1106	85	4	83	52.53	17.64	0.95	3.94
		2/3	2780	87	8	85	62.79	10.41	0.88	3.61
7–8	Fall 2004	1/2	2249	85	0	84	47.19	18.43	0.95	4.20
		2/3	2841	87	0	84	59.16	12.48	0.92	3.63
	Spring 2005	1/2	2056	85	0	83	46.51	17.39	0.94	4.26
		2/3	3677	87	12	85	62.01	11.69	0.90	3.70
	Fall 2005	1/2	379	85	1	82	34.83	18.93	0.95	4.23
		2/3	112	87	29	80	58.43	11.05	0.90	3.50
	Spring 2006	1/2	1000	85	0	81	46.16	16.72	0.94	4.10
		2/3	1938	87	10	87	61.98	11.61	0.91	3.48
	Fall 2006	1/2	422	85	0	83	33.68	18.27	0.95	4.08
		2/3	167	87	5	83	55.73	15.67	0.94	3.84
	Spring 2007	1/2	1000	85	2	82	46.42	17.52	0.94	4.29
		2/3	1911	87	10	85	59.36	12.78	0.91	3.83
	Fall 2007	1/2	428	85	1	80	33.59	18.04	0.95	4.03
		2/3	163	87	8	81	54.63	14.66	0.93	3.88
	Spring 2008	1/2	993	85	6	84	47.70	16.54	0.94	4.05
		2/3	1877	87	12	85	61.34	11.62	0.90	3.67

(cont'd)

Grade Span	Administration	Sessions*	*n*	Points	Min	Max	Mean	S.D.	Rel.	S.E.M.
9–12	Fall 2004	1/2	3494	85	0	85	46.47	16.29	0.94	4.00
		2/3	5253	87	0	84	56.73	12.63	0.91	3.79
	Spring 2005	1/2	3466	85	0	84	48.77	16.27	0.94	3.98
		2/3	5828	87	8	86	58.87	11.52	0.89	3.82
	Fall 2005	1/2	717	85	1	81	34.71	15.67	0.94	3.84
		2/3	185	87	7	81	53.51	13.82	0.91	4.15
	Spring 2006	1/2	1768	85	0	79	44.52	15.07	0.93	3.99
		2/3	1753	87	5	82	56.81	12.40	0.90	3.92
	Fall 2006	1/2	695	85	0	83	33.57	16.24	0.94	3.98
		2/3	210	87	8	79	53.06	14.52	0.92	4.11
	Spring 2007	1/2	1672	85	0	82	45.34	15.83	0.93	4.19
		2/3	1579	87	4	84	58.25	12.12	0.89	4.02
	Fall 2007	1/2	660	85	0	77	34.74	17.27	0.95	3.86
		2/3	185	87	24	84	55.31	11.95	0.90	3.78
	Spring 2008	1/2	1621	85	0	83	43.83	15.61	0.93	4.13
		2/3	1612	87	7	84	57.45	11.01	0.88	3.81

S.D. = standard deviation, Rel. = reliability, S.E.M. = standard error of measurement.
* Because of the small number of students who took sessions 1 and 2 for one subject and sessions 2 and 3 for the other, only students who took the same combination of sessions for both reading and writing are included in these calculations.

NOTE: In 2004–2005 school year, all LEP students were required to participate in both MEPA administrations; the fall administration established the students' baseline. For students who enrolled after the fall administration, the spring 2005 administration determined their baseline assessments. For operational years 2005–2006 through 2007–2008, all third-grade LEP students and those LEP students newly enrolled in Massachusetts schools were required to participate in the respective fall MEPA administration to determine their baseline, and all LEP students were required to participate in each spring MEPA administration.

As described previously, the standard error of measurement of each test was taken into consideration when reporting individual student scores. These standard errors were computed at each raw score level and used to report error bands around the associated scaled scores (see section 5.2 for details).

8.2.1.2 Stratified Coefficient α

According to Feldt and Brennan (1989), a prescribed distribution of items over categories (such as different item types) indicates the presumption that at least a small, but important, degree of unique variance is associated with the categories. In contrast, Cronbach's coefficient α is built upon the assumption that there are no such local or clustered dependencies. A stratified version of coefficient α corrects for this problem:

$$\alpha_{strat} = 1 - \frac{\sum_{j=1}^{k} \sigma_{x_j}^2 (1 - \alpha_j)}{\sigma_x^2}$$

where
j indexes the subtests or categories,
$\sigma_{x_j}^2$ represents the variance of each of the k individual subtests or categories,
α_j is the unstratified Cronbach's α coefficient for each subtest, and
σ_x^2 represents the total test variance.

Stratified coefficient α was calculated separately for each grade span, administration, and combination of sessions taken. The stratification results provided in Tables 8-14 through 8-17 are based on item type.

Table 8-14. 2004–2008 MEPA: Reliability Statistics Overall, by Item Type, and Stratified for Grades 3–4

Administration	Sessions	α	α MC	*n* MC	α CR	*n* CR	Stratified α
Fall 2004	1/2	0.95	0.83	24	0.95	25	0.96
	2/3	0.90	0.85	28	0.85	17	0.91
Spring 2005	1/2	0.95	0.87	24	0.94	25	0.96
	2/3	0.89	0.83	28	0.85	17	0.90
Fall 2005	1/2	0.96	0.87	24	0.95	25	0.96
	2/3	0.91	0.85	28	0.87	17	0.92
Spring 2006	1/2	0.95	0.86	24	0.95	25	0.96
	2/3	0.89	0.81	28	0.84	17	0.89
Fall 2006	1/2	0.96	0.71	24	0.93	25	0.99
	2/3	0.89	0.66	28	0.73	17	0.96
Spring 2007	1/2	0.96	0.74	24	0.90	25	0.99
	2/3	0.91	0.71	28	0.77	17	0.97
Fall 2007	1/2	0.95	0.70	24	0.90	25	0.99
	2/3	0.91	0.70	28	0.71	17	0.97
Spring 2008	1/2	0.95	0.68	24	0.90	25	0.99
	2/3	0.89	0.69	28	0.74	17	0.96

MC = multiple-choice, CR = constructed-response

Table 8-15. 2004–2008 MEPA: Reliability Statistics Overall, by Item Type, and Stratified for Grades 5–6

Administration	Sessions	α	α MC	*n* MC	α CR	*n* CR	Stratified α
Fall 2004	1/2	0.96	0.87	24	0.95	25	0.96
	2/3	0.90	0.81	28	0.87	17	0.91
Spring 2005	1/2	0.96	0.88	24	0.95	25	0.96
	2/3	0.90	0.86	28	0.85	17	0.91
Fall 2005	1/2	0.96	0.88	24	0.96	25	0.97
	2/3	0.93	0.88	28	0.92	17	0.94
Spring 2006	1/2	0.95	0.85	24	0.95	25	0.96
	2/3	0.89	0.81	28	0.85	17	0.90
Fall 2006	1/2	0.96	0.71	24	0.92	25	0.99
	2/3	0.92	0.77	28	0.82	17	0.97
Spring 2007	1/2	0.95	0.72	24	0.91	25	0.99
	2/3	0.90	0.72	28	0.79	17	0.96
Fall 2007	1/2	0.96	0.78	24	0.93	25	0.99
	2/3	0.93	0.78	28	0.80	17	0.98
Spring 2008	1/2	0.95	0.61	24	0.89	25	0.99
	2/3	0.88	0.72	28	0.73	17	0.95

MC = multiple-choice, CR = constructed-response

Table 8-16. 2004–2008 MEPA: Reliability Statistics Overall, by Item Type, and Stratified for Grades 7–8

Administration	Sessions	α	α MC	*n* MC	α CR	*n* CR	Stratified α
Fall 2004	1/2	0.95	0.83	24	0.94	25	0.95
	2/3	0.92	0.85	28	0.88	17	0.92
Spring 2005	1/2	0.94	0.81	24	0.94	25	0.95
	2/3	0.90	0.84	28	0.86	17	0.91
Fall 2005	1/2	0.95	0.81	24	0.95	25	0.96
	2/3	0.90	0.83	28	0.86	17	0.91
Spring 2006	1/2	0.94	0.86	24	0.94	25	0.95
	2/3	0.91	0.86	28	0.87	17	0.92
Fall 2006	1/2	0.95	0.72	24	0.90	25	0.99
	2/3	0.94	0.81	28	0.84	17	0.98
Spring 2007	1/2	0.94	0.73	24	0.88	25	0.98
	2/3	0.91	0.73	28	0.81	17	0.97
Fall 2007	1/2	0.95	0.66	24	0.90	25	0.99
	2/3	0.93	0.79	28	0.82	17	0.98
Spring 2008	1/2	0.94	0.63	24	0.85	25	0.98
	2/3	0.90	0.78	28	0.79	17	0.97

MC = multiple-choice, CR = constructed-response

Table 8-17. 2004–2008 MEPA: Reliability Statistics Overall, by Item Type, and Stratified for Grades 9–12

Administration	Sessions	α	α MC	*n* MC	α CR	*n* CR	Stratified α
Fall 2004	1/2	0.94	0.83	24	0.93	25	0.94
	2/3	0.91	0.83	28	0.88	17	0.92
Spring 2005	1/2	0.94	0.83	24	0.93	25	0.94
	2/3	0.89	0.78	28	0.86	17	0.90
Fall 2005	1/2	0.94	0.83	24	0.94	25	0.95
	2/3	0.91	0.81	28	0.90	17	0.92
Spring 2006	1/2	0.93	0.80	24	0.93	25	0.94
	2/3	0.90	0.82	28	0.87	17	0.91
Fall 2006	1/2	0.94	0.62	24	0.88	25	0.98
	2/3	0.92	0.74	28	0.85	17	0.98
Spring 2007	1/2	0.93	0.61	24	0.84	25	0.98
	2/3	0.89	0.66	28	0.81	17	0.96
Fall 2007	1/2	0.95	0.73	24	0.89	25	0.98
	2/3	0.90	0.62	28	0.83	17	0.97
Spring 2008	1/2	0.93	0.54	24	0.82	25	0.98
	2/3	0.88	0.64	28	0.79	17	0.96

MC = multiple-choice, CR = constructed-response

8.2.2 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 level descriptors were defined and students were classified into performance levels, empirical analyses were conducted to determine the statistical accuracy and consistency of the classifications.

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.

8.2.2.1 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.

Calculating Accuracy

Following Livingston and Lewis (1995), which can be used for both multiple-choice and constructed-response items, the true-score distribution for the MEPA 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 (described in Livingston and Lewis, 1995), a 4 × 4 accuracy contingency table is created for each content area test and grade level. The cells in the table show the proportions of students who were classified into each performance category by their actual (or observed) scores on the MEPA (i.e., observed status) and by their “true scores” (i.e., “true status”).

8.2.2.2 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 items 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 MEPA. 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 MEPA because their technique can be used with both constructed-response and multiple-choice items.

Calculating Consistency

Contingency Table Construction. To estimate consistency (i.e., the proportions of students classified into exactly the same categories by two forms of the test), 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 4 × 4 consistency contingency table is created for each test and grade level to show the proportions of students who are classified into each performance category by the actual test and by another (hypothetical) parallel test form.

Kappa. Another way to measure consistency is to use Cohen’s (1960) coefficient κ (kappa), which assesses the proportion of consistent classification after removing the proportion that would be expected to be consistent 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. Because κ is corrected for chance, the values of κ are lower than other consistency estimates.

8.2.3 Results of Accuracy, Consistency, and Kappa Analyses

Summaries of the MEPA accuracy and consistency analyses are provided in Tables 8-18 through 8-41.

The first section of each table shows overall accuracy and consistency indices as well as kappa.

The second section of each table shows accuracy and consistency values, conditional upon performance level. In each case, the denominator is the number of students who were actually placed into a given performance level. For example, the conditional accuracy value is 0.7343 for the *Intermediate* category for grade span 3–4 for the fall 2004 administration. This indicates that, of the students whose actual scores placed them in the *Intermediate* category, 73.43% of them would be expected to be in the *Intermediate* category if they were categorized according to their true score. Similarly, the corresponding consistency value of .6441 indicates that 64.41% of that same group of students would be expected to score in the *Intermediate* category if a second, parallel test form were used.

The third section of the summary tables shows information at each of the cut points. For certain tests, concern may be greatest regarding decisions made about a particular threshold. For example, if a college gave credit to students who achieved an Advanced Placement test score of 4 or 5, but not 1, 2, or 3, one might be interested in the accuracy of the dichotomous decision for below 4 versus 4 or above. The values in Tables 8-18 through 8-41 indicate the accuracy and consistency of the dichotomous decisions either above or below the associated cut point. False positive and false negative accuracy rates are also provided; these values are estimates of the proportion of students who were categorized above the cut when their true score would place them below the cut, and vice versa.

Table 8-18. 2005 Fall MEPA:
Accuracy and Consistency Summary for Grades 3–4

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.807		0.737	0.646
				Accuracy	**Consistency**
Indices Conditional on Performance Level		*Beginning*		0.856	0.815
		Early Intermediate		0.710	0.615
		Intermediate		0.724	0.636
		Transitioning		0.924	0.852
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.944	0.031	0.025	0.922
	EI : I	0.931	0.042	0.028	0.904
	I : T	0.932	0.046	0.022	0.907

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-19. 2006 Spring MEPA:
Accuracy and Consistency Summary for Grades 3–4

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.837	0.779	0.641

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.800	0.724
	Early Intermediate	0.689	0.591
	Intermediate	0.702	0.617
	Transitioning	0.947	0.903

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.971	0.015	0.015	0.959
	EI : I	0.945	0.031	0.024	0.923
	I : T	0.922	0.050	0.028	0.893

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-20. 2006 Fall MEPA:
Accuracy and Consistency Summary for Grades 3–4

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.891	0.847	0.795

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.919	0.891
	Early Intermediate	0.843	0.785
	Intermediate	0.854	0.801
	Transitioning	0.947	0.910

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.969	0.017	0.015	0.956
	EI : I	0.960	0.022	0.018	0.944
	I : T	0.962	0.023	0.015	0.947

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-21. 2007 Spring MEPA:
Accuracy and Consistency Summary for Grades 3–4

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.912	0.877	0.803

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.899	0.862
	Early Intermediate	0.826	0.761
	Intermediate	0.835	0.779
	Transitioning	0.969	0.949

		Accuracy			Consistency
		Accuracy	***False Positives***	***False Negatives***	
Indices at Cut Points	*B : EI*	0.982	0.009	0.009	0.975
	EI : I	0.971	0.016	0.014	0.959
	I : T	0.960	0.023	0.017	0.944

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-22. 2007 Fall MEPA:
Accuracy and Consistency Summary for Grades 3–4

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.891		0.847	0.794
				Accuracy	**Consistency**
Indices Conditional on Performance Level	*Beginning*			0.919	0.892
	Early Intermediate			0.838	0.778
	Intermediate			0.849	0.794
	Transitioning			0.949	0.912
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.968	0.017	0.015	0.956
	EI : I	0.961	0.022	0.018	0.944
	I : T	0.962	0.023	0.015	0.947

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-23. 2008 Spring MEPA:
Accuracy and Consistency Summary for Grades 3–4

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.915		0.881	0.808
				Accuracy	**Consistency**
Indices Conditional on Performance Level	*Beginning*			0.893	0.850
	Early Intermediate			0.841	0.782
	Intermediate			0.851	0.800
	Transitioning			0.968	0.947
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.985	0.008	0.008	0.979
	EI : I	0.971	0.015	0.013	0.960
	I : T	0.959	0.024	0.017	0.943

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-24. 2005 Fall MEPA:
Accuracy and Consistency Summary for Grades 5–6

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.873		0.829	0.720
				Accuracy	**Consistency**
Indices Conditional on Performance Level	*Beginning*			0.945	0.939
	Early Intermediate			0.628	0.508
	Intermediate			0.708	0.610
	Transitioning			0.927	0.857
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.950	0.032	0.018	0.931
	EI : I	0.959	0.026	0.015	0.943
	I : T	0.963	0.025	0.012	0.950

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-25. 2006 Spring MEPA:
Accuracy and Consistency Summary for Grades 5–6

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.812	0.747	0.629

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.821	0.766
	Early Intermediate	0.633	0.528
	Intermediate	0.712	0.631
	Transitioning	0.938	0.880

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.956	0.024	0.020	0.939
	EI : I	0.936	0.038	0.027	0.912
	I : T	0.919	0.054	0.027	0.890

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-26. 2006 Fall MEPA:
Accuracy and Consistency Summary for Grades 5–6

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.904	0.866	0.793

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.959	0.950
	Early Intermediate	0.759	0.664
	Intermediate	0.818	0.750
	Transitioning	0.939	0.889

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.963	0.022	0.015	0.948
	EI : I	0.968	0.019	0.013	0.955
	I : T	0.973	0.017	0.010	0.962

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-27. 2007 Spring MEPA:
Accuracy and Consistency Summary for Grades 5–6

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.900	0.860	0.797

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.908	0.876
	Early Intermediate	0.804	0.731
	Intermediate	0.854	0.804
	Transitioning	0.961	0.934

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.976	0.013	0.012	0.966
	EI : I	0.966	0.019	0.016	0.952
	I : T	0.958	0.025	0.017	0.942

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-28. 2007 Fall MEPA:
Accuracy and Consistency Summary for Grades 5–6**

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.936	0.909	0.856

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.975	0.968
	Early Intermediate	0.819	0.744
	Intermediate	0.864	0.811
	Transitioning	0.957	0.926

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.975	0.014	0.011	0.965
	EI : I	0.979	0.012	0.009	0.970
	I : T	0.982	0.011	0.008	0.974

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-29. 2008 Spring MEPA:
Accuracy and Consistency Summary for Grades 5–6**

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.886	0.840	0.758

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.880	0.836
	Early Intermediate	0.766	0.681
	Intermediate	0.823	0.766
	Transitioning	0.958	0.927

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.975	0.013	0.012	0.965
	EI : I	0.962	0.021	0.017	0.946
	I : T	0.949	0.031	0.020	0.928

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-30. 2005 Fall MEPA:
Accuracy and Consistency Summary for Grades 7–8**

	Accuracy	Consistency	Kappa (κ)
Overall Indices	0.850	0.800	0.673

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.924	0.921
	Early Intermediate	0.579	0.444
	Intermediate	0.665	0.560
	Transitioning	0.919	0.838

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.936	0.045	0.019	0.909
	EI : I	0.955	0.030	0.015	0.938
	I : T	0.958	0.029	0.013	0.943

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-31. 2006 Spring MEPA:
Accuracy and Consistency Summary for Grades 7–8**

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.821		0.761	0.655
				Accuracy	**Consistency**
Indices Conditional on Performance Level		*Beginning*		0.848	0.810
		Early Intermediate		0.616	0.508
		Intermediate		0.702	0.617
		Transitioning		0.946	0.894
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.951	0.028	0.021	0.932
	EI : I	0.940	0.036	0.024	0.917
	I : T	0.929	0.048	0.023	0.904

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-32. 2006 Fall MEPA:
Accuracy and Consistency Summary for Grades 7–8**

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.936		0.910	0.854
				Accuracy	**Consistency**
Indices Conditional on Performance Level		*Beginning*		0.974	0.968
		Early Intermediate		0.829	0.758
		Intermediate		0.874	0.824
		Transitioning		0.954	0.920
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.974	0.015	0.011	0.963
	EI : I	0.979	0.012	0.009	0.970
	I : T	0.984	0.010	0.007	0.977

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-33. 2007 Spring MEPA:
Accuracy and Consistency Summary for Grades 7–8**

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.893		0.851	0.793
				Accuracy	**Consistency**
Indices Conditional on Performance Level		*Beginning*		0.919	0.895
		Early Intermediate		0.785	0.703
		Intermediate		0.840	0.784
		Transitioning		0.959	0.929
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.970	0.017	0.014	0.958
	EI : I	0.964	0.020	0.016	0.949
	I : T	0.960	0.025	0.016	0.944

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-34. 2007 Fall MEPA:
Accuracy and Consistency Summary for Grades 7–8**

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.938		0.913	0.857
				Accuracy	Consistency
Indices Conditional on Performance Level		*Beginning*		0.975	0.969
		Early Intermediate		0.831	0.761
		Intermediate		0.875	0.826
		Transitioning		0.954	0.921
		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.975	0.014	0.011	0.965
	EI : I	0.980	0.012	0.009	0.971
	I : T	0.984	0.010	0.006	0.978

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-35. 2008 Spring MEPA:
Accuracy and Consistency Summary for Grades 7–8**

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.895		0.853	0.793
				Accuracy	Consistency
Indices Conditional on Performance Level		*Beginning*		0.917	0.891
		Early Intermediate		0.786	0.706
		Intermediate		0.840	0.785
		Transitioning		0.960	0.931
		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.971	0.016	0.013	0.960
	EI : I	0.965	0.020	0.016	0.950
	I : T	0.959	0.025	0.016	0.943

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

**Table 8-36. 2005 Fall MEPA:
Accuracy and Consistency Summary for Grades 9–12**

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.845		0.793	0.668
				Accuracy	Consistency
Indices Conditional on Performance Level		*Beginning*		0.924	0.918
		Early Intermediate		0.556	0.434
		Intermediate		0.717	0.619
		Transitioning		0.916	0.823
		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.934	0.044	0.022	0.908
	EI : I	0.948	0.035	0.018	0.927
	I : T	0.960	0.028	0.011	0.946

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-37. 2006 Spring MEPA:
Accuracy and Consistency Summary for Grades 9–12

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.792		0.792	0.726
				Accuracy	**Consistency**
Indices Conditional on Performance Level	*Beginning*			0.842	0.806
	Early Intermediate			0.551	0.444
	Intermediate			0.712	0.630
	Transitioning			0.930	0.858
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.940	0.035	0.025	0.917
	EI : I	0.928	0.044	0.028	0.900
	I : T	0.922	0.055	0.024	0.894

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-38. 2006 Fall MEPA:
Accuracy and Consistency Summary for Grades 9–12

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.910		0.874	0.801
				Accuracy	**Consistency**
Indices Conditional on Performance Level	*Beginning*			0.961	0.954
	Early Intermediate			0.718	0.609
	Intermediate			0.831	0.767
	Transitioning			0.945	0.901
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.964	0.021	0.014	0.950
	EI : I	0.970	0.018	0.012	0.958
	I : T	0.975	0.016	0.009	0.965

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-39. 2007 Spring MEPA:
Accuracy and Consistency Summary for Grades 9–12

		Accuracy		**Consistency**	**Kappa (κ)**
Overall Indices		0.868		0.818	0.747
				Accuracy	**Consistency**
Indices Conditional on Performance Level	*Beginning*			0.903	0.875
	Early Intermediate			0.696	0.591
	Intermediate			0.818	0.758
	Transitioning			0.951	0.911
		Accuracy			**Consistency**
		Accuracy	**False Positives**	**False Negatives**	
Indices at Cut Points	*B : EI*	0.962	0.021	0.017	0.946
	EI : I	0.955	0.026	0.019	0.938
	I : T	0.951	0.031	0.018	0.932

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-40. 2007 Fall MEPA:
Accuracy and Consistency Summary for Grades 9–12

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.899		0.859	0.792

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.954	0.943
	Early Intermediate	0.736	0.634
	Intermediate	0.844	0.785
	Transitioning	0.941	0.892

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.961	0.023	0.016	0.945
	EI : I	0.965	0.020	0.014	0.951
	I : T	0.973	0.017	0.010	0.962

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

Table 8-41. 2008 Spring MEPA:
Accuracy and Consistency Summary for Grades 9–12

		Accuracy		Consistency	Kappa (κ)
Overall Indices		0.867		0.817	0.744

		Accuracy	Consistency
Indices Conditional on Performance Level	*Beginning*	0.896	0.865
	Early Intermediate	0.705	0.602
	Intermediate	0.825	0.768
	Transitioning	0.949	0.909

		Accuracy			Consistency
		Accuracy	False Positives	False Negatives	
Indices at Cut Points	*B : EI*	0.963	0.021	0.017	0.948
	EI : I	0.955	0.026	0.020	0.937
	I : T	0.950	0.032	0.019	0.931

B = Beginning, EI = Early Intermediate, I = Intermediate, T = Transitioning

REFERENCES

Publications by the Massachusetts Department of Elementary and Secondary Education (formerly known as Massachusetts Department of Education)

2005 MEPA Technical Report.
http://www.doe.mass.edu/mcas/mepa/?section=results

English Language Proficiency Benchmarks and Outcomes for English Language Learners.
http://www.doe.mass.edu/ell/benchmark.pdf

Guide to Interpreting the MEPA Reports for Schools and Districts.
http://www.doe.mass.edu/mcas/mepa/?section=results

Guide to the MEPA for Parents/Guardians.
http://www.doe.mass.edu/mcas/mepa/?section=results

Massachusetts English Language Arts Curriculum Framework.
http://www.doe.mass.edu/frameworks/ela/0601.pdf

Qualified MELA-O Trainer (QMT) Training Manual.
http://www.doe.mass.edu/mcas/mepa/qmt_manual.pdf

Requirements for the Participation of Students with Limited English Proficiency in MCAS and MEPA.
http://www.doe.mass.edu/mcas/participation/lep.pdf

Other Publications Referenced in This Report

Allen, M. J., & W. M. Yen. (1979). *Introduction to Measurement Theory.* Belmont, CA: Waveland Press, Inc.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 1999. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

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

Dorans, N. J., and E. Kulick. 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.

Feldt, L. S., and R. L. Brennan. 1989. Reliability. In R. L. Linn (ed.), *Educational Measurement.* 3rd ed., pp. 105–146.

Hambleton, R. K., and H. Swaminathan. 1985. *Item response theory: Principles and applications*. Boston: Kluwer Academic Publishers.

Hambleton, R. K., and W. J. van der Linden, eds. 1997. *Handbook of modern item response theory.* New York: Springer-Verlag.

Holland, P. W., and H. Wainer, eds. 1993. *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Joint Committee on Testing Practices. 1988. *Code of fair testing practices in education*. Washington, DC: National Council on Measurement in Education.

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

Lord, F. M., and M. R. Novick. 1968. *Statistical theories of mental tests scores.* Reading, MA: Addison-Wesley Publishing Company, Inc.

Petersen, N. S., M. J. Kolen, and H. D. Hoover. 1989. Scaling, norming, and equating. In R. L. (ed.), *Educational measurement*. 3rd ed., pp. 221–262. New York: Macmillan Publishing Company.

APPENDICES