Massachusetts Department of
ELEMENTARY & SECONDARY
EDUCATION

Massachusetts
English Proficiency Assessment
MEPA

2009
Technical Report

This document was prepared by the
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 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

TABLE OF CONTENTS

Chapter 1. OVERVIEW OF THE MASSACHUSETTS ENGLISH PROFICIENCY ASSESSMENT

1.1 Overview and Purpose of the Assessment System (Defining the Population)

The Massachusetts English Proficiency Assessment (MEPA) measures the language skills of English language learner (ELL) students in the state. As required by the No Child Left Behind (NCLB) Act of 2001, the assessment gauges the proficiency of students who are new to Massachusetts schools as a baseline and then measures their progress toward acquiring English language proficiency. The tests are also used to meet state and federal accountability requirements and to help determine when a student is ready to be reclassified as no longer of limited English proficiency. The MEPA is based on *English Language Proficiency Benchmarks and Outcomes for English Language Learners* in the four areas of reading, writing, listening, and speaking.

Reading and writing are assessed via a written test, the MEPA-R/W. The test was developed for grade spans K–2, 3–4, 5–6, 7–8, and 9–12. For each content area (reading and writing) in grades 3–12, three sessions are available to cover the range of proficiencies. Students are assigned by their schools to two adjacent sessions. The K–2 assessment has two levels of differing complexity, and students are assigned to one of the two by their schools.

Listening comprehension and speaking skills are assessed via observation by trained administrators in the schools. This assessment, the Massachusetts English Language Assessment-Oral (MELA-O), makes use of a scoring matrix developed to be universal for all grade spans.

Scores on both the MEPA-R/W and the MELA-O are used to determine a student's overall MEPA score and to report a student's results in one of five performance levels—*Level 1*, *Level 2*, *Level 3*, *Level 4*, and *Level 5*—which describe a student's achievement on the MEPA.

While reclassification as formerly limited English proficient (FLEP) is not required for a student at a particular MEPA performance level, the Department of Elementary and Secondary Education (the Department) recommends reclassification for a student who scores at *Level 5*. In some instances, a student at *Level 4* may also be ready for reclassification if educators determine that he or she is ready to perform ordinary classroom work in English. In making such determinations, educators consider MEPA scores in combination with student scores on local reading, language, and other academic assessments; academic grades; MCAS scores; and the observations and recommendations of educators.

1.2 Description of This Report

The purpose of this document is to report on the technical details and characteristics of the 2009 MEPA tests, and to provide evidence of the validity and reliability of the results from those tests. The report contains information on test design and development, scoring, analysis, and reporting.

The report is organized in the same general order as the testing process itself, beginning with test design and development and ending with the reporting of results. Each step of the process is described in detail.

Chapter 2. CURRENT YEAR UPDATES

2.1 Fall 2008 Field Test

A major change in the 2009 MEPA program was the introduction of a common/matrix-sampled test design, in which the common items taken by every student in a grade span are the basis of the student's scores, while matrix-sampled questions are field tested for use as common items in future years. To address this change, a field test was conducted in fall 2008 in which the majority of limited English proficient (LEP) students participated. Each field test booklet included one reading session and one writing session, which were administered in two separate testing sessions, each being no longer than one class period. All levels of items (i.e., those from Sessions 1, 2, and 3) were represented so that participating students answered the full range of questions. This field test yielded common items for the spring 2009 tests in all grade spans.

2.2 Spring 2009 Online Pilot Test

In spring 2009, the Department and its testing contractor, Measured Progress of Dover, New Hampshire, worked collaboratively to conduct an online pilot test administration to validate the use of online testing in the varied technology infrastructures of schools throughout the state. The specific goals of the online pilot test were to

- obtain feedback from students regarding the ease of use and the overall experience of taking the test online;
- obtain feedback from test administrators regarding starting, monitoring, and completing the online test administration;
- obtain feedback from school administrators regarding feasibility of online testing implementation and ease of use;
- sample the functionality of the online test in a variety of school technology environments;
- gather information about how different technology-related factors affect the testing system;
- identify issues that might impact the successful implementation of online testing in spring 2010;
- identify and prioritize potential system enhancements for spring 2010 online testing.

The online pilot included 11 schools and approximately 500 students from across the state. Following test administration, students were asked to respond to a four-question survey designed to gauge their level of interest and understanding of the online test system. Teachers and administrators were also asked to provide details of their experience with the administrative portion of the system. A report detailing the results of the pilot study is available at www.mcasservicecenter.com/McasDefault.asp?ProgramID=14.

Chapter 3. Test Development and Design

3.1 Item Types

This chapter relates specifically to the paper tests used for the MEPA-R/W. Custom test items, based on the Department's *English Language Proficiency Benchmarks and Outcomes for English Language Learners*, are written for these tests.

- Multiple-choice items appear in both reading and writing. Each multiple-choice item requires students to select a single best answer from four response options. These items are machine scored, and students receive a score of 1 for a correct response or a score of 0 for an incorrect or blank response.
- Short-answer items in reading require students to generate a brief response usually consisting of one or more sentences. Short-answer responses are scored by trained readers on a scale of 0 to 2 based on item-specific rubrics. The K–2 grade span also has short-answer items, although they are scored on a scale of 0 to 1.
- Open-response items in reading require students to generate a response of typically one or more paragraphs. Open-response items receive scores of 0 to 4 on the basis of item-specific rubrics and are scored by trained readers.
- Short-answer items in writing require students to generate a very brief response, typically one word or phrase. Trained readers assign scores of 0 or 1 based on item-specific rubrics.
- Sentence-writing items require students to respond to a graphic or prompt. Trained readers assign scores of 0 to 2 based on item-specific rubrics.
- Writing-prompt items require students to write compositions based on a prompt. Trained readers assign scores of 0 to 4 based on item-specific rubrics or scoring guides.

3.2 Operational Development Process

The MEPA-R/W development process is typical of large-scale assessments that utilize the common/matrix test design. This section of the technical report outlines the steps of the item development process.

Before item development begins, test developers select reading passages to be included in the tests for each grade span. The proposed passages are reviewed by the Department, assessment development committees, and a bias committee for quality, interest, grade appropriateness, freedom from bias or sensitivity, and content accuracy (see Appendix A for committee membership). These selections provide the basis for the passage-based items on the tests.

3.2.1 Item and Scoring Guide Development

In addition to aligning with the Department's *English Language Proficiency Benchmarks and Outcomes for English Language Learners*, test items are developed to be age and grade appropriate, engaging and of interest to students, and free of bias or sensitivity concerns. As will be described in greater detail, each item is rigorously reviewed by the contracted test development staff, assessment development committees, bias committee, and development staff at the Department. In addition, the answer key for each multiple-choice item is reviewed and verified. Distracter options are checked to ensure the presence of only one correct answer.

Scoring guides or rubrics are used for all item types, except multiple-choice, to indicate how to evaluate student responses and assign appropriate scores. Development of these guides begins in conjunction with the items' creation. The scoring guide or rubric for each item is drafted by the test developers and subsequently reviewed and edited as necessary throughout the development process.

3.2.2 Content Standards

The assessments are aligned to *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (see highlights in Appendix B or view the entire document at http://www.doe.mass.edu/ell/benchmark.pdf), which was created by the Department in June 2003, and based on the *Massachusetts English Language Arts Curriculum Framework* of June 2001. The benchmarks for reading address vocabulary and syntax in print, beginning to read in English, comprehension, literary elements and techniques, informational/expository text, and research. The writing benchmarks focus on pre-writing, writing, revising, editing, and media.

3.2.3 Internal Item Review

The lead test developers review all items to ensure that they

- are aligned to approved benchmarks and outcomes;
- are aligned to approved test blueprints;
- are appropriate for the skill level and vocabulary of the tested grade span;
- contain only essential information;
- have clear, correct, and understandable graphics, where applicable;
- meet universal design requirements;
- use contexts that would be familiar to students from diverse cultural and linguistic backgrounds;
- do not cue the correct answer to other items;
- do not repeat key wording in stems, distracters, or prompts in other items; and
- do not echo the wording of the text.

Items are also reviewed by proofreaders for style and mechanics and by content leads for quality, style, and variety.

3.2.4 External Item Review

In addition to internal review, each passage and item is evaluated by external assessment development committees. These committees are composed of educators from a diverse sampling of urban, suburban, and rural districts across the state, with experience in English language acquisition in the appropriate grade spans. These reviewers assess the passages and items for alignment to standards, content appropriateness, and age appropriateness.

3.2.5 Bias and Sensitivity Review

Following review by the assessment development committees, a separate committee of educators review each passage and item for potential bias or insensitivity using guidelines developed by the Department. The goal of this review is to ensure that the tests do not include language, symbols, or content that could be construed as potentially offensive, inappropriate, or negative.

3.2.6 Item Editing

Test developers keep detailed records of all of the recommendations from the meetings of the assessment development and bias committees. These recommendations are reviewed by the Department, and where appropriate, approved revisions are made to the items and graphics. The items are again reviewed by proofreaders and by the Department before being considered eligible to remain in the item pool.

3.2.7 Reviewing and Refining Items

After these final revisions, the items receive another test development review to make sure that no problems have been introduced in quality, style, or variety of item types, and that the items are consistent with Massachusetts Comprehensive Assessment System (MCAS) English Language Arts items. Items are also checked for cueing and echoing concerns and for alignment to benchmarks and outcomes across the range of the entire performance continuum.

3.2.8 Operational Test Assembly

Test developers assemble proposed sets of common items to be used in each grade span and of field test items to appear in each of the matrix forms of the spring administration. Items are chosen to match the item types and outcomes on the test blueprints, cover as many benchmarks as possible, and represent a range of difficulty within each outcome. Each proposed form is checked for possible cueing and echoing. Additional items in the pool are available as replacements if needed, and draft data are entered onto the test blueprints. Psychometric and research staff run test characteristic curves for the proposed common sets, which are then reviewed by the Department. Form-pulling meetings are held, during which the form selections are carefully analyzed, substitutions are made as needed, and new test characteristic curves are reviewed. The constructed forms then move into production.

3.2.9 Editing Drafts of Operational Tests

The constructed sets from the form-pulling meetings are reviewed by editorial staff and placed in test book format. Test developers and the Department review hard-copy drafts of each test form for layout consistency, consistency and accuracy of all non-item elements (i.e., headers, footers, and direction lines), and correct implementation of all form-pulling edits. The items are also checked for possible cueing of answers and for content accuracy. Several rounds of edits are made, with new hard-copy pages provided for Department and test development review after each round.

3.2.10 Alternative Presentation—Large Print

Large-print versions of the test are created from one form in each grade span, for use by visually impaired students. The final PDF files of the standard tests are printed in a larger format with no other modifications from the originals.

3.3 Guidelines for Test Designs and Blueprints

Items for the 2009 MEPA-R/W administration were designed according to *English Language Proficiency Benchmarks and Outcomes for English Language Learners* to assess language proficiency and to generate new items for testing in future years. This section provides further details of the test designs.

3.3.1 Selection Guidelines

Items selected for inclusion on the final forms of the 2009 MEPA tests met the following criteria, specified by the Department:

- Match to the framework
- Fulfillment of blueprint specifications
- Accuracy and clarity of item content
- Age and content appropriateness
- Achievement of a critical threshold for item statistics
- Freedom from bias
- Sensitivity to students with special needs

In addition, test characteristic curves, test information functions, and projected cut scores were reviewed for the common items. This ensured that the developed tests had appropriate measurement precision along the performance continuum.

3.3.2 Test Construction

Once 2009 form selection was completed, information related to test construction was finalized. Item types, outcomes, and correct keys (for multiple-choice items) were documented for each item on the test.

3.3.3 Field Test Design

After the fall 2008 field test, new field test items were embedded in the spring 2009 matrix forms to generate common items for future administrations. Eight matrix forms were tested in each grade span, with one exception. Because grade span 7–8 had significantly fewer students than the others, only six matrix forms were field tested for this group. The test blueprints show the number and type of field test items in each level or session. For the K–2 grade span, field test items were embedded in both levels. In grades 3–12, the field test items were embedded in Session 2, a session that all students took. Because the tests include items that are read aloud to the students, the spring matrix-sampled forms were distributed so that each school administered only one of the forms.

The fall 2009 test used the same common items as the spring test in a single form, but did not include any field test items.

3.3.4 Equating Design

An assessment program like the MEPA, which has multiple forms and levels (i.e., Sessions 1 and 2 versus Sessions 2 and 3) and which will continue to be administered in subsequent years, uses equating both “within year” and “across years.” The within-year equating is designed to place all item parameters for a given grade span onto a common measurement scale, whereas the across-year equating is designed to maintain the measurement scale from one year to the next. Both of these equating activities were ultimately done through a common item-linking technique that made use of item response theory (IRT) scaling.

3.3.4.1 Within-Year Equating

For 2009 within-year equating, all test items were concurrently calibrated to an IRT scale for each grade span independently. For the K–2 grade span, four reading multiple-choice items and two writing short-answer items were tested in common between Levels A and B. For grade span 3–4 and above, the non-matrix items in Session 2 were common to all students in that grade span. Because the link between K–2 Levels A and B was smaller than what might typically be used in a within-year equating (i.e., less than 25 percent of the total number of points), only students in grade 1 were used in the calibration process. Details regarding the IRT models and the results of that analysis can be found in Chapter 7.

3.3.4.2 Across-Year Equating

Starting in 2010, across-year equating will be necessary in order to maintain the scale that was developed in 2009. Within the 2009 administration, items were field tested in both Levels A and B for grade span K–2 and in Session 2 for all other grade spans. These field test items were brought onto scale using the within-year equating method, and the 2010 test form will be constructed using these scaled field test items. That is, the majority of the items in Levels A and B for grade span K–2 and in Sessions 1, 2, and 3 for other grade spans will be pre-equated from the previous year's administration. A small number of items in Levels A and B for grade span K–2 and Sessions 1, 2, and 3 for other grade spans will need to be brought onto scale. These items, along with additional field test items, will be brought onto scale again using the within-year equating procedure. This process will then continue for the remainder of the program or until performance levels are revisited.

3.3.5 Test Booklet Design

The MEPA booklet layout is primarily governed by a style guide developed for the MCAS tests. Sessions start on the same page in each form per grade span. Most pages have a dual column layout with a center rule. When across-the-page items appear, they are typically placed at the top of a page. Whenever possible, items are situated so that they face their associated passage. Integrated test and answer booklets are used for the K–2 and 3–4 grade spans. Students in grades 5 to 12 use separate test and answer booklets.

3.4 Test Sessions

In grades 3 to 12, three sessions are available for reading and for writing. Students take two adjacent sessions, either Sessions 1 and 2 or Sessions 2 and 3. The Session 1 and 2 assessment is decidedly easier than the Session 2 and 3 assessment. This range in difficulty is meant to optimize the measurement of student performance across a wide measurement scale. Teachers, knowing how a student might perform, can select an assessment that is most appropriate for that student. In addition, a locator test is available to help teachers determine the appropriate sessions to be administered to each student. Other relevant factors in determining which sessions a student will participate in are the student's performance on local assessments, MCAS tests, and previous MEPA tests, as well as in the classroom. The locator tests for grades 3 to 12 include three passages and 13 to 14 multiple-choice items for reading and 12 multiple-choice items for writing.

The K–2 assessment has only one session for each content area, but two levels of each test are available. Students are assigned to either Level A or Level B and take the same level for both content areas. Schools are provided with a locator survey to help them inventory each student's skills and to aid in determining the appropriate level of the test to administer to each student. The locator survey is used in conjunction with local assessments, observations, and teacher judgment.

Chapter 4. TEST ADMINISTRATION

4.1 MEPA-R/W

The reading and writing portions of the 2009 assessment were administered as a traditional paper-and-pencil test. This section contains details of the test administration.

4.1.1 Responsibility for Administration

During the MEPA-R/W administration, school principals were responsible for the following:

- Enforcing test security
- Ensuring participation of all limited English proficient students at the appropriate grade level
- Providing accurate student information
- Coordinating the testing schedule
- Ensuring proper test administration
- Ensuring availability of accommodations

Principals were also responsible for designating test administrators who were fluent in English and, to the extent possible, were licensed classroom teachers working in the schools.

4.1.2 Administration Procedures

Administration procedures were explained in manuals provided to the principals and test administrators. Principals were responsible for ordering test materials for their schools and assigning space in which to administer the tests.

The K–2 Level A test was administered one to one or in small groups for most students, and administrators provided assistance to students in marking their responses, if needed. The Level A test was administered to larger groups of students (up to 15) only if the students were in grade 2 and if they all met more than half the skill requirements on the locator survey. The Level B test was administered in groups of up to 15 students, with smaller groups used for students who met less than half of the locator survey skill requirements. Test administrators monitored the correct placement of written responses for all K–2 students.

In the spring administration, all students within a school took the same form of the test. The fall tests had only one form per grade span, as only the common items were tested. The reading test was administered first followed by the writing test. The tests for all grade spans included items that administrators read aloud to students. The read-aloud scripts were provided in the *MEPA Test Administrator's Manual*.

4.1.3 Test Administration Window

The spring administration period for the MEPA-R/W was March 9 to 20, 2009, and the fall period was October 19 to 28, 2009.

4.1.4 Participation Requirements and Documentation

All enrolled LEP students were required to participate in the spring 2009 MEPA-R/W administration, including, for the first time, students in the K–2 grade span. Participation in the fall 2009 MEPA-R/W was only required for LEP students enrolled in grades 1 to 12 who did not participate in the spring 2009 test. Kindergarteners did not participate in the fall MEPA-R/W testing. Exceptions were made to the participation requirements for both administrations for students with a medically documented absence, students who required accommodations that were not available (such as Braille), students who were deaf or hard of hearing, and students who required an alternate assessment due to significant disabilities (see Appendix C).

4.1.5 Documentation of Accommodations (Use and Appropriateness)

Prior to MEPA-R/W testing, Individualized Education Program (IEP) and 504 teams provided information to principals regarding the specific accommodation(s) students need in order to participate. These decisions were based on the needs of the individual students consistent with accommodations that they regularly use in the classroom. Examples of standard accommodations include scheduling of the test to meet the specific needs of the student, providing specific settings for test administration, altering the presentation of the test, and adjusting the method in which the student responds to test questions. IEP and 504 teams also could allow nonstandard accommodations if students met the specific criteria for each accommodation. Examples of nonstandard accommodations include having an administrator read the reading test aloud to the student or scribing the student's responses on the writing test. The principal indicated the use of specific accommodations by filling in the appropriate numbered accommodation bubbles on the student's answer booklet.

4.1.6 Administrator Training

Prior to the 2009 MEPA-R/W administration, workshops were held throughout the state to prepare principals or their designees for testing. Principals received manuals specifically written for the overall MEPA test administration. Principals held meetings with test administrators to review testing security, schedules, logistics, and materials. In addition, administrators were provided with manuals specific to each grade span that included general testing policies and tasks, as well as specific instructions and scripts for each test session.

4.1.7 Test Security

Test security was addressed extensively in the 2009 manuals provided to principals and test administrators. Principals were responsible for ensuring that all administrators and school personnel complied with the security requirements and instructions detailed in the manuals. All test materials were to be inventoried upon receipt, with any discrepancies reported immediately. Materials were then to be stored in secure locations until time for administration. Tracking charts were used to document the location of materials at all times when they were not in secure storage. Students received instructions about test security and were never to be unsupervised in the presence of test materials. Schools were responsible for returning all secure materials to the testing contractor, who worked with the Department to follow up on any discrepancies.

4.1.8 Test and Administration Irregularities

The 2009 manuals for principals and test administrators provided specific examples of security violations and contact information to be used in the event of a testing irregularity in their school. The manuals also included information on how to respond to other types of administration irregularities such as fire drills, power failures, and severe weather.

4.1.9 Service Center

Service center staff, trained on the logistical, programmatic, and grade span/content area-specific aspects of the 2009 MEPA program, were available to schools via a toll-free telephone number. The service center assisted schools in requesting additional testing materials and filling out forms, provided instructions about the delivery and return of MEPA materials, and answered questions about administration and reporting.

4.2 MELA-O

MELA-O is an observational assessment measuring students' proficiency in listening (comprehension) and speaking (production), as identified in the *English Language Proficiency Benchmarks and Outcomes for English Language Learners*. Within the category of production, four subdomains are evaluated: fluency, grammar, pronunciation, and vocabulary.

4.2.1 Responsibility for Administration

Students are assessed by trained and qualified MELA-O administrators (QMA) in the schools. Each district is responsible for ensuring that staff within the district have been trained to administer this assessment.

4.2.2 Administration Procedures

Students are primarily observed in a classroom setting by a QMA as they engage in normal academic interactions with the teacher or with other students. In order to attain an adequate sample of the students' language skills, the students may be observed on multiple occasions during the administration window. Students are rated (i.e., scored) on a 0–5 MELA-O scoring matrix in each category. Scores are reported to the Department.

4.2.3 Test Administration Window

The spring administration of the MELA-O took place from February 23 to March 20, 2009, and the fall MELA-O from October 1 to 28, 2009.

4.2.4 Participation Requirements and Documentation

LEP students in kindergarten through grade 12 were required to participate in the spring 2009 administration of the MELA-O listening and speaking test. The fall 2009 assessment was required for LEP students in grades 1 through 12 who did not participate in spring 2009. A very small number of students were not required to participate because they had an extended medically documented absence from school during the testing window or they had an IEP identifying them as deaf or hard or hearing.

4.2.5 Administrator Training

MELA-O administrators and MELA-O trainers (who may administer the test and also train new administrators) participate in a two-day training session and pass a qualifying test before they may begin administering the assessment. The qualifying test consists of a video showing clips of students in a classroom setting. The training participants use a scoring matrix to assign listening and speaking scores for each of the students. The students in the video clips had previously been assigned scores in each of the six sections of the matrix by a group of master trainers working in conjunction with the Department. These were deemed to be the correct scores against which the scores assigned by training participants are checked. A score assigned by a trainee that is more than two score points from the correct score is deemed discrepant. The minimum scores needed to qualify are:

- Qualified MELA-O trainers—35 correct scores out of 50 possible with no more than 2 discrepant scores, or 31 to 34 correct scores with no more than 1 discrepant score.
- Qualified MELA-O administrators—30 correct scores of 50 possible with no more than 2 discrepant scores, or 26 to 29 correct scores with no more than 1 discrepant score.

Trainers and administrators who had administered the MELA-O prior to 2007 were required to take part in a retraining session and to pass the qualifying test in order to continue assessing students after 2009 and/or training new administrators. The Department conducted six training and retraining sessions at multiple locations throughout 2009.

4.2.6 Service Center

As with the MEPA-R/W, the service center was available to provide schools and districts with information and to answer questions as needed.

Chapter 5. SCORING

5.1 MEPA-R/W

Once received by the testing contractor, each 2009 MEPA-R/W student answer booklet was scanned in its entirety into the electronic imaging system iScore. This highly secure, server-to-server interface was designed by Measured Progress.

Student identification and demographic information, school information, and student answers to multiple-choice questions were converted to alphanumeric format and were not visible to readers. Digitalized student responses to open-response, short-answer, sentence-writing, and writing-prompt test items were sorted, by grade level, into item-specific groups.

5.1.1 Machine-Scored Items

Multiple-choice items were used on 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 1 point; incorrect answers were assigned a score of 0 points. Blank responses and responses with multiple marks were also assigned 0 points.

5.1.2 Hand-Scored Items

Item-specific groups of responses were scored one at a time; readers scored one response at a time. Individual responses were linked through iScore to the original booklet number, so scoring leadership had access, if necessary, to a student’s entire answer booklet.

5.1.2.1 Scoring Locations and Staff

While the iScore database, its operation, and its administrative controls were all based in Dover, New Hampshire, 2009 MEPA-R/W responses were scored in two locations.

- Measured Progress Scoring Center, Longmont, Colorado
    - Grades 3–4 reading and writing
    - Grades 5–6 reading and writing
    - Grades 9–12 reading and writing
- Measured Progress Scoring Center, Dover, New Hampshire
    - Grades K–2 reading and writing
    - Grades 7–8 reading and writing

The following staff members were involved with scoring the 2009 MEPA-R/W responses.

- The MEPA-R/W scoring manager, located in Dover, oversaw communication and coordination of scoring across the two scoring sites.
- The iScore operations manager, located in Dover, coordinated technical communication across the two scoring sites.
- A scoring center manager provided logistical coordination for his or her scoring site.

- A chief reader in writing, reading, or reading/writing ensured consistency of benchmarking and scoring across all grade spans at the two scoring locations. Chief readers monitored and read behind both onsite and offsite quality assurance coordinators.
- Several quality assurance coordinators, selected from a pool of experienced senior readers, participated in benchmarking, training, scoring, and cleanup activities for specified content areas and grade levels. Quality assurance coordinators monitored and read behind senior readers.
- Senior readers, selected from a pool of skilled and experienced readers, monitored and read behind readers at their scoring tables. Each senior reader monitored 2 to 11 readers.

5.1.2.2 Benchmarking Meetings

Samples of student responses to field test items were read, scored, and discussed by scoring and Department staff at item-specific benchmarking meetings. All benchmarking meeting results were recorded and considered final upon Department signoff.

The primary goals of the field test benchmarking meetings were to

- revise, if necessary, an item's scoring guide;
- revise, if necessary, an item's scoring notes (elaborative notes about scoring that particular item that were listed, when needed, underneath the score point descriptions);
- assign official score points to as many of the sample responses as possible; and
- approve various individual and sets of responses (e.g., anchor, training) to be used to train field test scorers.

Items with score point ranges of 0–2, 0–3, and 0–4 were benchmarked with multiple examples of each score point. Items with score point ranges of 0–1 (correct/not correct) were not formally submitted to the benchmarking meeting unless there were questions about how a particular response should be scored. If clarifications were needed for these, examples of correct and not correct score point responses were chosen as exemplars for the readers.

5.1.2.3 Reader Recruitment and Qualifications

2009 MEPA-R/W readers were primarily obtained through the services of a temporary employment agency. They represented a wide range of backgrounds, ages, and experiences. Most readers were highly experienced, having scored student responses for a number of other testing programs, and many had previously scored MCAS and MEPA-R/W responses.

All MEPA-R/W readers had successfully completed at least two years of college; hiring preference was given to those with a four-year college degree. Potential readers were required to submit applications and documentation such as resumes and transcripts. This documentation was carefully reviewed. If a potential reader did not clearly demonstrate knowledge of reading, writing, or English, or have at least two college courses with average or above-average grades in these subjects, the potential scorer was eliminated from the applicant pool. Teachers, tutors, and administrators (principals, guidance counselors, etc.) currently under contract or employed by or in Massachusetts schools, or anyone under 18 years of age, are ineligible to score MEPA-R/W responses.

5.1.2.4 Methodology for Scoring Polytomous Items

The 2009 MEPA-R/W contained polytomous items (including short-answer items) requiring students to generate a brief response, with scores of 0–1 or 0–2 assigned, and open-response items requiring a longer or more complex response, with scores of 0–3 or 0–4 assigned.

In addition to the option of assigning a score point between 0 and 4, depending on the item, readers could designate a response as one of the following:

- Blank—The written response form was completely blank (no graphite).
- Unreadable—The text on the computer screen was too faint to see accurately.
- Wrong Location—The response seemed to be a legitimate answer to a different question.

Responses initially marked “Unreadable” or “Wrong Location” were resolved by readers and iScore staff by matching all responses with the correct item and/or pulling the actual test booklet to look at the student’s original work.

Table 5-1 presents a K–2 Level B writing open-response scoring guide, one of the many different MEPA-R/W scoring guides used in 2009. The task associated with this scoring guide asked students to look at three pictures and then write a story with a beginning, middle, and end.

Table 5-1. 2009 MEPA: 3 point Open-response Item Scoring Guide—Writing K–2 Level B

Score	*Description*
3	The response is a **thoroughly** accurate depiction of the objects and events shown in the graphics. ● The response matches the progression of events in all three graphics. ● Sentences are complete (with at least a subject and verb); some may be complex. ● The response forms a well connected beginning, middle, and end, and clearly expresses the story depicted in the three graphics. ● There are few or no phonetic spellings and/or only minor errors in conventions that do not interfere with communication.
2	The response is a **partially** accurate or general depiction of the objects and events shown in the graphics. ● The response matches the progression of events in all three graphics. ● Sentences may or may not be complete. ● The response forms a beginning, middle, and end, and generally expresses the story depicted in the graphics. ● Spelling may be phonetic, but words are recognizable. Errors in spelling and conventions begin to interfere with communication.
1	The response is a **minimally** accurate or vague depiction of the objects and events shown in the graphics. ● The response may or may not match the progression of events in all three graphics. ● Most sentences are not complete. ● The response may or may not express the full story depicted in the graphics. ● Some words may be phonetically/visually recognizable; errors in conventions may seriously interfere with communication.
0	The response is irrelevant or is written in a language other than English.
Blank	No response.

Scoring Note: Holistic scoring allows for a range within each score point. One bullet alone does not define a score point. In holistic terms, a 3 response will be thorough, a 2 response will be partial, and a 1 response will be minimal. Response must show a plausible interpretation of events in sequence.

In addition to the scores or notations previously listed, readers may have also flagged a response as "Crisis"; these responses were sent to scoring leadership and the Department for immediate attention.

A response may be flagged as a crisis paper if it indicates

- perceived, credible desire to harm self or others;
- perceived, credible, and unresolved instances of mental, physical, and/or sexual abuse;
- presence of dark thoughts or serious depression;
- sexual knowledge well outside of the student's developmental age;
- ongoing, unresolved misuse of legal/illegal substances (including alcohol);
- knowledge of or participation in real, unresolved criminal activity;
- direct or indirect request for adult intervention/assistance (e.g., crisis pregnancy, doubt about how to handle a serious problem at home).

Student responses were either single scored, in which each response was scored only once, or double-blind scored, in which each response was independently read and scored by two separate readers. For each 2009 MEPA-R/W item, at least 10 percent of the responses were randomly double-blind scored; neither reader knew it had been scored before or what score it had been given. A double-blind response with discrepant scores between the two readers (i.e., a difference greater than 1 if there were 3 or more score points) was sent to the arbitration queue and read by a senior reader or quality assurance coordinator.

Above and beyond the 10 percent double-blind scoring, senior readers, at random points throughout the scoring shift, performed read behinds on each of the readers at their table. This process involved senior readers viewing responses recently scored by a particular reader and, without knowing the reader's score, assigning their own score to that same response.

Tables 5-2 and 5-3 outline the resolution rules for instances when the two read-behind or two double-blind scores were not identical (i.e., adjacent or discrepant).

Table 5-2. 2009 MEPA: Resolution Chart for Read-Behind Scoring*

Reader 1	*Reader 2*	*QAC/SR read behind*	*Final*
4	-	4	4
4	-	3	3
4	-	2	2
0	-	1	1

* In all cases, the quality assurance coordinator/senior reader score is the final score of record.

Table 5-3. 2009 MEPA: Resolution Chart for Double-Blind Scoring*

Reader 1	*Reader 2*	*QAC/SR resolution*	*Final*
4	4	-	4
4	3	-	4
3	4	-	4
4	2	3	3
4	1	2	2
3	1	1	1

* If reader scores are identical or adjacent, the highest score is used as the final score. If reader scores are neither identical nor adjacent, the resolution score is used as the final, reported score.

5.1.2.5 Reader Training

Chief readers were responsible for ensuring that scoring leadership and readers scored consistently, fairly, and only according to the approved scoring guidelines. Chief readers started the training process with an overview of the MEPA-R/W; this general orientation included the purpose and goal of the testing program and any unique features of the test and the testing population.

Scoring materials were carefully compiled and checked for consistency and accuracy. The time, order, and manner in which the materials were presented to readers were standardized to ensure readers had the same training experience and, as much as possible, the same environment for each item, content area, and grade level at each scoring location.

Depending on availability of technology, the trainer may have had an opportunity to choose between several possible modes of delivery. In some cases, chief readers and quality assurance coordinators were able to deliver the training via a headset with a microphone, with all readers listening through headphones. This was the preferred method if there were simultaneous training sessions happening in the same room at the same time, or if there was a very large number of readers, as the electronic amplification helped to ensure all readers could hear without strain. Some items were trained via a remote location; that is, the chief reader was communicating directly with readers, even though he or she was physically in one room or scoring location and readers were sitting at their computers in a separate room or different scoring location. Direct interaction between reader and trainer continued uninterrupted, either via instant messaging and two-way audio communication devices or onsite training supervisors.

After the general orientation, the trainer thoroughly reviewed and discussed the scoring guide for the item to be scored, which consisted of the item itself, the scoring rubric, and any item-specific scoring notes. All scoring guides had previously been approved by the Department and were used without any additions or deletions.

Before assigning scores to operational student responses, prospective readers carefully reviewed up to four different sets of actual student responses, some of which had been used to train readers when the item was a matrix field test item.

- Anchor set—Responses that were solid, exceptionally clear, typical examples of the score points, referred to throughout the training and scoring process as “true examples.”
- Training set—Unusual, discussion-provoking responses (e.g., very high/low/short, exceptionally creative, disorganized) that further defined the score point range by illustrating the range of responses typically encountered in operational scoring.
- Ranking set—One clear example of each mid-range score point distributed to readers in mixed or scrambled score point order. At the appropriate time during training, readers rank ordered them according to their true score points.
- Qualifying set—Readers of 3- and 4-point items were given a test of 10 responses that were clear, typical examples of each of the score points as a way to determine if they were able to score according to the Department-approved scoring rubric.

Meeting or surpassing the minimum acceptable standard on an item’s qualifying set was an absolute requirement for scoring student responses to that item. An individual reader had to attain a scoring accuracy rate of 70 percent exact and 90 percent exact and adjacent agreement (at least 7 out of the 10 were exact score matches and either 0 or 1 discrepant) on either of two potential qualifying sets.

Scoring leadership (quality assurance coordinators and senior readers) had to meet or surpass the higher qualification standard of at least 80 percent exact and 90 percent exact and adjacent.

5.1.2.6 Monitoring of Scoring Quality Control and Consistency

When MEPA-R/W readers met or exceeded the minimum standard on a qualifying set and began scoring, they were constantly monitored throughout the entire scoring process to be sure they scored student responses as accurately and consistently as possible. Readers were required to meet or exceed the minimum standard of 70 percent exact and 90 percent exact and adjacent agreement on the following:

- Recalibration assessments
- Embedded committee reviewed responses (CRRs)
- Read behinds
- Double-blind readings
- Compilation reports, end-of-shift reports combining recalibration assessments and read-behind readings

If a reader fell below standard on any of these quality control tools, leadership initiated a reader intervention ranging from counseling to retraining to dismissal. If a reader fell below standard on any three daily compilation reports for a particular item, he or she was automatically dismissed from scoring that item. If a reader was dismissed from scoring two MEPA items within a grade and content area, the reader was not allowed to score any additional items within that grade and content area. If a reader was dismissed from two different grade and content areas within one scoring session, the reader was dismissed from the project, and he/she was not allowed to score any additional items from that test administration.

Recalibration assessments, given to readers at the very beginning of a scoring shift, consisted of a set of five responses representing the entire range of possible scores. If readers had an exact score match on four of the five responses and were at least adjacent on the fifth response, they were allowed to begin scoring operational responses. Readers who had discrepant scores, or only two or three exact score matches, were retrained and, if approved by the senior reader, given extra monitoring assignments such as additional read behinds and allowed to begin scoring. Readers who had zero or one out of the five exact were not retrained and not allowed to score that item on that day.

CRRs were chief reader-approved responses loaded into iScore for blind distribution to readers at random points during the scoring of their first 200 operational responses. While the number of CRRs ranged from 5 to 30 depending on the item, for most items MEPA-R/W readers received 10 of these previously scored responses during the first day of scoring that particular item. Readers who fell below the accuracy standard were counseled and, if approved by the senior reader, given extra monitoring assignments such as additional read behinds and allowed to resume scoring.

For read behinds, responses were first read and scored by a reader, then read and scored by a senior reader. The senior reader would, at various points during the scoring shift, command iScore to forward the next one, two, or three responses to be scored by a particular reader to his or her own computer. Without knowing the score given by the reader, the senior reader would first give his or her own score to the response and then be allowed to compare that score to the reader’s score. Each full-time day shift reader was read behind at least 10 times, and each evening shift and half-day reader at least five times. Readers who fell below the 70 percent exact and 90 percent exact and adjacent score match standard were counseled, given extra monitoring assignments such as additional read behinds, and allowed to resume scoring.

Double-blind readings involved responses scored independently by two different readers. Readers knew 10 percent or more of their responses were to be scored by others, but they had no way of knowing whether a particular response had already been scored or was scheduled to be scored by another. Over the course of a scoring shift, readers who fell below the score match standard were, if necessary, counseled, given extra monitoring assignments such as additional read behinds, and allowed to resume scoring. Responses given discrepant scores by two independent readers were scored by a senior reader.

Compilation reports combined a reader’s percentage of exact, adjacent, and discrepant scores on the recalibration assessment with that reader’s percentage of exact, adjacent, and discrepant scores on read behinds. Once the senior reader completed the minimum number of required read behinds on a reader—5 for a half shift and 10 for a full shift—the reader’s overall percentages on the compilation reports were automatically calculated. Readers who were below standard were counseled and, if approved by the senior reader, given extra monitoring assignments such as additional read behinds and allowed to resume scoring.

A final compilation report for the scoring group was run at the end of each scoring shift. If there were individuals who were still below the 70 percent exact/90 percent exact and adjacent level, their scores for that day were voided and the responses they scored were returned to the scoring queue for other readers to score.

5.2 MELA-O

Scoring (or rating) of students on the MELA-O took place in each student’s school, and scores were subsequently provided to the scoring contractor for inclusion in the student’s overall MEPA performance level.

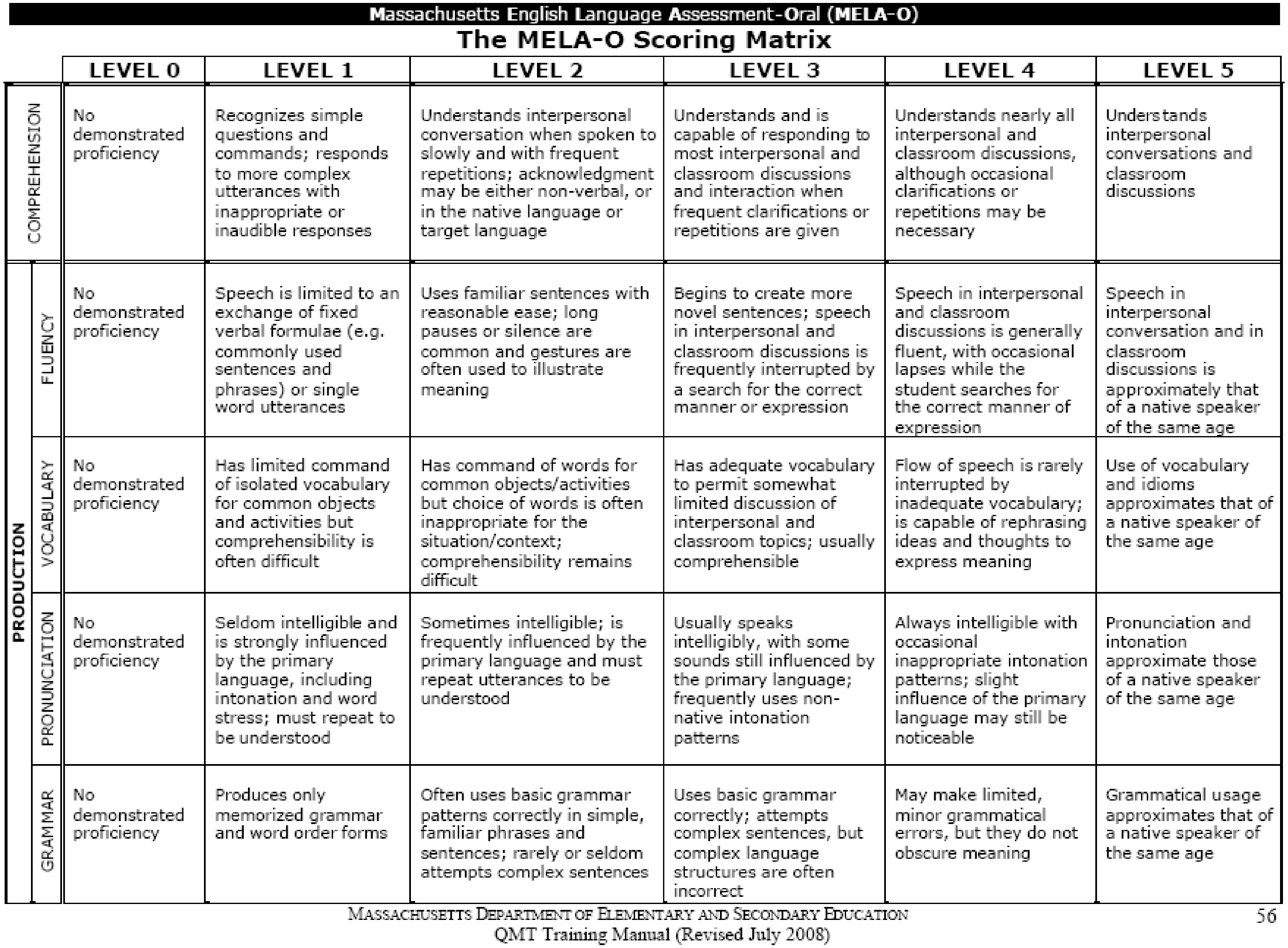
5.2.1 Scoring Matrix

Administrators used a scoring matrix, presented as Figure 5-1, to assign each student separate scores for each of the following areas:

- Listening (comprehension) and
- Speaking (production), which was broken down into the subdomains of
    - fluency,
    - vocabulary,
    - pronunciation, and
    - grammar.

The scores ranged from 0 to 5 points in each of the five areas, with a score of 0 indicating no demonstrated proficiency and a score of 5 approximating the proficiency of a native speaker.

Figure 5-1. 2009 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)

5.2.2 Collection of MELA-O Scores

Once scores were assigned, schools recorded them by filling in the appropriate bubbles on the students' MEPA-R/W answer booklets.

5.2.3 Weight of MELA-O Scores in Student Performance Level

The MELA-O scores were incorporated into the students' overall score along with the MEPA-R/W scores. A natural weighting was used in combining MELA-O scores with MEPA-R/W; the weighting consisted simply of totaling the possible points for each component (5 for listening and 20 for speaking).

In addition, the MELA-O scores were treated as items (one item for listening and four for speaking) and included with MEPA-R/W items in the item calibrations. For a more detailed explanation of the item calibrations, refer to section 7.2.

Chapter 6. CLASSICAL ITEM ANALYSIS

6.1 Classical Difficulty and Discrimination Indices

Both *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999) and *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1988) include standards for identifying quality items. Items should assess only the 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 advantage or 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 met these standards. Previous chapters of this report have delineated various qualitative checks. This chapter presents two categories of quantitative statistical evaluations: difficulty indices and item-test correlations. Item response theory analyses are discussed in the next chapter.

The results presented here are based on the spring 2009 and fall 2009 MEPA administrations. The item-level classical statistics, including difficulty and discrimination indices, are presented in more detail in Appendix D.

6.1.1 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 responded to the item. Multiple-choice items were scored dichotomously (correct versus incorrect), so for these items the difficulty index was simply the proportion of students who answered the item correctly. Open-response items were scored on a scale of either 0–2, 0–3 (grade span K–2 only), 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, open-response, and MELA-O indicators were placed on the same 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 correctly answered by almost all students provide little information about differences in student 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 student 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 open-response items) to 0.90. Items with indices outside this range signify items that may be either too difficult or too easy for the target population. Items outside this range are used only if content specialists agree that they are essential to the construct tested.

6.1.2 Discrimination Indices

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.

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

6.1.3 Summary of Item Analysis Results

Summary statistics of the difficulty and discrimination indices for each item type are provided in Tables 6-1 and 6-2. In general, the item difficulty and discrimination indices are within 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 2009 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 open-response items is inappropriate because multiple-choice items can be answered correctly by guessing. Thus, it is not surprising that, in many cases, the difficulty indices for multiple-choice items are higher (indicating easier items) than the difficulty indices for open-response items. The partial credit allowed for open-response items is advantageous in the computation of item-test correlations; therefore, the discrimination indices for these items tend to be larger than the discrimination indices of other item types.

Table 6-1. 2009 MEPA: Average Difficulty and Discrimination and (Standard Deviation) of Different Item Types for Composite Score—Spring 2009 Administration

Grade span	*Session*	*Statistic*	*Item type*		
			All	*Multiple-choice*	*Open-response*
K–2	A	Difficulty	0.60 (0.18)	0.69 (0.16)	0.50 (0.14)
		Discrimination	0.45 (0.16)	0.36 (0.08)	0.55 (0.16)
		N	30	16	14
	B	Difficulty	0.71 (0.14)	0.74 (0.15)	0.67 (0.13)
		Discrimination	0.45 (0.12)	0.39 (0.08)	0.52 (0.12)
		N	32	16	16
3–4	1:2	Difficulty	0.66 (0.15)	0.71 (0.11)	0.62 (0.18)
		Discrimination	0.55 (0.13)	0.46 (0.08)	0.65 (0.10)
		N	43	22	21
	2:3	Difficulty	0.76 (0.16)	0.81 (0.13)	0.67 (0.19)
		Discrimination	0.39 (0.08)	0.34 (0.05)	0.47 (0.05)
		N	41	26	15
5–6	1:2	Difficulty	0.60 (0.17)	0.58 (0.16)	0.61 (0.18)
		Discrimination	0.54 (0.18)	0.41 (0.12)	0.68 (0.10)
		N	43	22	21
	2:3	Difficulty	0.70 (0.14)	0.70 (0.12)	0.69 (0.18)
		Discrimination	0.37 (0.10)	0.32 (0.08)	0.47 (0.05)
		N	41	26	15
7–8	1:2	Difficulty	0.54 (0.17)	0.56 (0.18)	0.53 (0.17)
		Discrimination	0.47 (0.18)	0.34 (0.11)	0.61 (0.13)
		N	43	22	21
	2:3	Difficulty	0.68 (0.15)	0.71 (0.13)	0.63 (0.17)
		Discrimination	0.39 (0.09)	0.34 (0.06)	0.48 (0.07)
		N	41	26	15
9–12	1:2	Difficulty	0.55 (0.16)	0.54 (0.15)	0.55 (0.18)
		Discrimination	0.49 (0.13)	0.40 (0.09)	0.59 (0.07)
		N	43	22	21
	2:3	Difficulty	0.65 (0.16)	0.65 (0.17)	0.65 (0.15)
		Discrimination	0.40 (0.10)	0.34 (0.07)	0.49 (0.06)
		N	41	26	15

Table 6-2. 2009 MEPA: Average Difficulty and Discrimination and (Standard Deviation) of Different Item Types for Composite Score—Fall 2009 Administration

Grade span	*Session*	*Statistic*	*Item type*		
			All	*Multiple-choice*	*Open-response*
K–2	A	Difficulty	0.55 (0.19)	0.67 (0.15)	0.40 (0.11)
		Discrimination	0.52 (0.18)	0.41 (0.08)	0.65 (0.17)
		N	30	16	14
	B	Difficulty	0.63 (0.14)	0.67 (0.14)	0.60 (0.14)
		Discrimination	0.52 (0.16)	0.42 (0.08)	0.62 (0.16)
		N	32	16	16
3–4	1:2	Difficulty	0.42 (0.15)	0.49 (0.13)	0.36 (0.14)
		Discrimination	0.60 (0.16)	0.48 (0.09)	0.73 (0.10)
		N	43	22	21
	2:3	Difficulty	0.72 (0.16)	0.77 (0.13)	0.62 (0.19)
		Discrimination	0.43 (0.10)	0.38 (0.08)	0.52 (0.05)
		N	41	26	15
5–6	1:2	Difficulty	0.43 (0.15)	0.45 (0.14)	0.41 (0.16)
		Discrimination	0.55 (0.19)	0.41 (0.12)	0.70 (0.11)
		N	43	22	21
	2:3	Difficulty	0.68 (0.15)	0.71 (0.14)	0.65 (0.16)
		Discrimination	0.40 (0.14)	0.32 (0.09)	0.53 (0.10)
		N	41	26	15
7–8	1:2	Difficulty	0.42 (0.16)	0.47 (0.17)	0.36 (0.14)
		Discrimination	0.51 (0.20)	0.36 (0.13)	0.68 (0.12)
		N	43	22	21
	2:3	Difficulty	0.67 (0.13)	0.70 (0.11)	0.61 (0.15)
		Discrimination	0.47 (0.17)	0.37 (0.10)	0.64 (0.12)
		N	41	26	15
9–12	1:2	Difficulty	0.49 (0.16)	0.51 (0.13)	0.48 (0.18)
		Discrimination	0.49 (0.16)	0.36 (0.09)	0.63 (0.11)
		N	43	22	21
	2:3	Difficulty	0.67 (0.16)	0.68 (0.16)	0.65 (0.16)
		Discrimination	0.42 (0.11)	0.36 (0.08)	0.52 (0.07)
		N	41	26	15

6.2 Differential Item Functioning

Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 1988) explicitly states that subgroup differences in performance should be examined when sample sizes permit, and action should be taken to ensure that differences in performance are due to construct-relevant, rather than irrelevant, factors. *Standards for Educational and Psychological Testing* (AERA et al., 1999) includes similar guidelines.

As part of the effort to identify such problems, 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.

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 & Kulick, 1986) was used to evaluate differences among three MEPA subgroups: male versus female, white versus black, and white versus Hispanic or Latino. This procedure calculates the average item performance for each subgroup at every total score. An overall average is then calculated, weighting the total score distribution so it is the same for the reference and focal groups (e.g., male and female). The index ranges from -1 to 1 for multiple-choice items; the index is adjusted to the same scale for open-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. The authors 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). DIF analyses were performed for grade spans K–2, 3–4, 5–6, 7–8, and 9–12 for the spring 2009 and fall 2009 administrations. The results of these analyses are reported in Appendix E. For K–2, separate analyses were conducted on the two total test forms associated with that grade span.

- Form A: Reading Level A, Writing Level A, and MELA-O
- Form B: Reading Level B, Writing Level B, and MELA-O

For each of the other grade spans, there were four total test forms. A student could take either Sessions 1 and 2 of reading or Sessions 2 and 3 of reading. Independently, the same student could take either Sessions 1 and 2 of writing or Sessions 2 and 3 of writing. Crossing these two choices with each other gives four possible combinations of reading and writing. In addition, all students also took the MELA-O assessment. Of the four possible total test forms for each of the 3–4, 5–6, 7–8, and 9–12 grade spans, two were by far more frequently administered, and so analyses were limited to those two forms.

- Form 1-2: Reading Sessions 1 and 2, Writing Sessions 1 and 2, and MELA-O
- Form 2-3: Reading Sessions 2 and 3, Writing Sessions 2 and 3, and MELA-O

The numbers of students taking the other two possible combinations were insufficient to report DIF results.

Tables E-1 and E-2 in Appendix E show the number of items classified into each category separately by item type (multiple-choice versus open-response) for the following subgroup comparisons: male versus female, white versus black, and white versus Hispanic or Latino. (Blank cells indicate comparisons for which there were insufficient numbers of students to compute reliable results.) Tables E-3 and E-4 of Appendix E give the number of items, by item type, that favor males or females in each of the three DIF categories. As can be seen in Tables E-1 through E-4, most MEPA items fell within the negligible range; very few items were classified as exhibiting high DIF.

6.3 Dimensionality Analyses

The DIF analyses described previously were performed to identify items that showed evidence of differences in performance between pairs of subgroups beyond what would be expected based on the primary construct that underlies total test score (also known as the primary dimension; for example, general achievement in English language arts). When items are flagged for DIF, statistical evidence points to their measuring dimensions in addition to the primary dimension.

Because tests are constructed with multiple content area subcategories and their associated knowledge and skills, the potential exists for a large number of dimensions being invoked beyond the common primary dimension. Generally, the subcategories are highly correlated with each other; therefore, the primary dimension they share typically explains an overwhelming majority of variance in test scores. In fact, the presence of just such a dominant primary dimension is the psychometric assumption that provides the foundation for the unidimensional item response theory (IRT) models that are used for calibrating, linking, scaling, and equating the MEPA test forms. As noted in the previous section, a statistically significant DIF result does not automatically imply that an item is measuring an *irrelevant* construct or dimension. An item could be flagged for DIF because it measures one of the construct-*relevant* dimensions of a subcategory's knowledge and skills.

The purpose of dimensionality analysis is to investigate whether violation of the assumption of test unidimensionality is statistically detectable and, if so, (1) the degree to which unidimensionality is violated and (2) the nature of the multidimensionality. Dimensionality analyses were performed on MEPA common items for grade spans K–2, 3–4, 5–6, 7–8, and 9–12 for the spring 2009 administration. The results of these analyses are reported below. As with the DIF analyses, for K–2 separate analyses were conducted on the two total test forms associated with that grade span.

- Session A: Reading Level A, Writing Level A, and MELA-O
- Session B: Reading Level B, Writing Level B, and MELA-O

Similarly, for grade spans 3–4, 5–6, 7–8, and 9–12, analyses were performed on the following:

- Session 1-2: Reading Sessions 1 and 2, Writing Sessions 1 and 2, and MELA-O
- Session 2-3: Reading Sessions 2 and 3, Writing Sessions 2 and 3, and MELA-O

In addition to analyzing these 10 test forms (two per grade span across five grade spans), psychometricians also analyzed the same 10 forms without including the MELA-O items. (Note: Only common items were analyzed, since they are used for score reporting.)

The dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both use as their basic statistical building block the estimated average conditional covariances for item pairs. A conditional covariance is the covariance between two items conditioned on total score for the rest of the test, and the average conditional covariance is obtained by averaging over all possible conditioning scores. When a test is strictly unidimensional, all conditional covariances are expected to take on values within random noise of zero, indicating statistically independent item responses for examinees with equal expected scores. Non-zero conditional covariances are essentially violations of the principle of local independence, and such local *dependence* implies multidimensionality. Thus, non-random patterns of positive and negative conditional covariances are indicative of multidimensionality.

DIMTEST is a hypothesis-testing procedure for detecting violations of local independence. The data are first randomly divided into a training sample and a cross-validation sample. An exploratory analysis of the conditional covariances is conducted on the training sample data to find the cluster of items that displays the greatest evidence of local dependence. The cross-validation sample is then used to test whether the conditional covariances of the selected cluster of items display local dependence, conditioning on total score on the non-clustered items. The DIMTEST statistic follows a standard normal distribution under the null hypothesis of unidimensionality.

DETECT is an effect-size measure of multidimensionality. As with DIMTEST, the data are first randomly divided into a training sample and a cross-validation sample (these samples are drawn independent of those used with DIMTEST). The training sample is used to find a set of mutually exclusive and collectively exhaustive clusters of items that best fit a systematic pattern of positive conditional covariances for pairs of items from the same cluster and negative conditional covariances from different clusters. Next, the clusters from the training sample are used with the cross-validation sample data to average the conditional covariances. The within-cluster conditional covariances are summed, and from this sum the between-cluster conditional covariances are subtracted. The resulting difference is divided by the total number of item pairs, and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. DETECT values less than 0.2 indicate very weak multidimensionality (or near unidimensionality); values of 0.2 to 0.4, weak to moderate multidimensionality; values of 0.4 to 1.0, moderate to strong multidimensionality; and values greater than 1.0, very strong multidimensionality.

DIMTEST and DETECT were applied to the spring 2009 MEPA tests for grade spans K–2, 3–4, 5–6, 7–8, and 9–12. The data for each test form analyzed were split into a training sample and a cross-validation sample. There was a large amount of variability in sample size across the different test forms, as shown in Table 6-3. The smallest sample size was about 1,800, which occurred for Session 1-2 in grade span 7–8, resulting in about 900 examinees for the training and cross-validation samples. All the other grade spans had sample sizes of at least 2,000. DIMTEST simulation studies have indicated 99 percent power rates for total sample sizes (training sample combined with cross-validation sample) as small as 750 (Stout, Froelich, & Gao, 2001), while also adhering well to nominal Type 1 error rates. After randomly splitting the data into training and cross-validation samples, DIMTEST was applied to each data set to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each data set for which the DIMTEST null hypothesis was rejected in order to estimate the effect size of the multidimensionality.

6.3.1 Analysis of Full Test Sessions (Reading, Writing, & MELA-O)

The results of the DIMTEST analyses on the test data that included all three subtests (reading, writing, and MELA-O) indicated that the null hypothesis was rejected for every data set at level 0.05. Indeed, the rejections were very strong, with the DIMTEST hypothesis-testing *p*-value less than 0.00005 in every case. Because strict unidimensionality is an idealization that almost never holds exactly for a given data set, the strong statistical rejections in the DIMTEST results were not necessarily surprising. A strong rejection of the null hypothesis is possible for even weak violations of unidimensionality if the sample size is large enough. Thus, it was important to follow up the DIMTEST analyses with DETECT analyses to estimate the size of the multidimensionality.

DETECT was used to estimate the effect size for the violations of local independence in the full test forms, as indicated by the DIMTEST results. Table 6-3 displays the DETECT effect-size estimates for the full test forms in the column labeled “R, W, MELA-O.”

Table 6-3. 2009 MEPA: Multidimensionality Effect Sizes by Grade Span: Spring 2009 Administration

			Multidimensionality effect size	
Grade span	*Session*	*Sample size (rounded)*	*R, W, MELA-O*	*R & W only*
K–2	A	10,500	1.15	0.36
	B	10,600	0.62	0.33
3–4	1:2	3500	0.47	0.23
	2:3	7400	0.32	0.15
5–6	1:2	2000	0.48	0.19
	2:3	5100	0.39	0.18
7–8	1:2	1800	0.39	0.27
	2:3	4000	0.45	0.17
9–12	1:2	3300	0.66	0.32
	2:3	6100	0.48	0.26

All the DETECT values indicated moderate to strong multidimensionality for every test form. Indeed, the results for Session A in grade span K–2 indicated very strong multidimensionality. These results are not surprising, given the inclusion of both the MEPA-R/W and MELA-O in the analyses. The MELA-O is administered *because* it is presumed to provide information beyond that provided by the MEPA-R/W; the results of the DETECT analyses simply confirm that hypothesis. Closer investigation of both the DIMTEST and DETECT results indicates that the multidimensionality is predominantly caused by MELA-O measuring a construct different from reading or writing, while reading and writing displayed much less difference between each other. This can be seen by comparing the multidimensionality effect sizes in the "R, W, MELA-O" column in Table 6-3 to those in the "R & W only" column: the values based on reading and writing only are dramatically lower than those for reading, writing and MELA-O.

6.3.2 Analysis of Only Reading and Writing

Because the analysis of the total test forms indicated that MELA-O measures a construct that is more different from reading and writing than reading and writing are from each other, a follow-up analysis focusing on only the reading and writing subtests was conducted to provide a more accurate picture of that dimensionality structure. DIMTEST again rejected the null hypothesis of unidimensionality for every data set at level 0.05, but the *p*-values were usually not as low as when MELA-O was present, indicating weaker multidimensionality.

DETECT was used to estimate the effect size for the violations of local independence in the combined reading and writing subtests, as indicated by the DIMTEST results. Table 6-3 displays the DETECT effect-size estimates for these analyses in the column labeled "R, W only." These results clearly confirm that the constructs measured by the reading and writing subtests are much more similar to each other than to the construct measured by the MELA-O subtest. Moreover, except for K–2 Session A, K–2 Session B, and 9–12 Session 1:2, the multidimensionality was either very weak (less than 0.2) or weak (less than 0.3). And only K–2 Session A had a DETECT effect size that was closer to 0.4 than to 0.3.

A more detailed analysis of the DETECT results was also conducted to investigate the degree to which differences between reading and writing contributed to whatever multidimensionality is evident in the DETECT results. Only for K–2 Session A, 7–8 Session 1:2, and 9–12 Session 2:3 was there strong evidence of separation of the reading and writing subtests being the predominant cause of the multidimensionality in the respective data sets. Significant differentiation of the reading and writing subtests also occurred for K–2 Session B, 3–4 Session 2:3, and 9–12 Session 1:2, but the DETECT results indicated that such separation was not the primary cause of the data sets' multidimensionality in these cases. This suggests that there was some multidimensionality within either the reading or writing subtest (or both).

6.3.3 Summary

Overall, the results indicate that significant levels of multidimensionality exist in the 2009 MEPA tests, and that MELA-O is the primary cause of that multidimensionality in most cases. Moreover, the multidimensionality tends to be larger for the K–2 grade span, especially for Session A. The remaining multidimensionality beyond that caused by MELA-O is much weaker, but again tends to be stronger for grade span K–2. There is some evidence of reading and writing measuring consistently different constructs, but the difference is not consistent across all data sets and what differences there are appear to be negligibly weak. The results clearly indicate that the special procedures implemented for MELA-O items (see Chapter 7) were fully justified and necessary in order to control for the dimensionality differences due to those items. Similarly, the use of the more conservative one-parameter logistic and partial credit IRT models with the K–2 grade span are also supported by the dimensionality analysis results. Indeed, dimensionality analyses that were conducted on field test data collected in fall 2008 had already given an accurate preview of the differing multidimensionality associated with K–2 and had guided modeling decisions. The present results serve to act as a confirmation of those earlier results and the subsequent modeling decisions.

In summary, the results of the dimensionality analyses support the continued use of the current test construction, modeling, and calibration procedures developed for the MEPA program. In particular, MEPA should continue to employ the more conservative psychometric models and special calibration procedures for the K–2 grade span, the special procedures for including the MELA-O items in the total test calibrations, and the test construction procedures that focus on reading and writing but do not overemphasize separate analyses.

Chapter 7. ITEM RESPONSE THEORY SCALING AND EQUATING

7.1 Item Response Theory

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).

Several IRT models can be used to specify the relationship between θ and p (Hambleton & van der Linden, 1997; Hambleton & Swaminathan, 1985). After careful consideration to choose the best-fitting IRT model, psychometricians decided to use a different model for grade span K–2 than for the other grade spans. The partial credit model (PCM) was employed for MELA-O indicators and polytomous MEPA-R/W items for grade span K–2. The other grade spans used the graded response model (GRM) for polytomous items, including MELA-O indicators. Additionally, a Rasch model was used for K–2 dichotomous items, whereas a two-parameter logistic (2PL) model was used for dichotomous items in all other grade spans.

The generalized form of the PCM can be defined as

Error! Objects cannot be created from editing field codes.

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.

For grade span K–2, the a_j term in the above equation is equal to 1.0 for all polytomous 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:

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

For the remaining grade spans, the 2PL model and the GRM were used for dichotomous and polytomous items, respectively. The 2PL model for dichotomous items can be defined as follows:

$$P_i\left(1\middle|\theta_{j,}\xi_i\right) = \frac{\exp\left[Da_i\left(\theta_j - b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j - b_i\right)\right]}$$

where

i indexes the items,
j indexes students,
a represents item discrimination,
b represents item difficulty,
ξ_i represents the set of item parameters (*a* and *b*), and
D is a normalizing constant equal to 1.701.

In the GRM, an item is scored in *k* + 1 graded categories, which can be viewed as a set of *k* dichotomies. At each point of dichotomization (i.e., at each threshold), a 2PL model can be used. This implies that a polytomous item with *k* + 1 categories can be characterized by *k* item category threshold curves (ICTCs) of the 2PL form:

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

where
ξ_i represents the set of item parameters for item *I*,
i indexes the items,
j indexes students,
k indexes threshold,
a represents item discrimination,
b represents item difficulty,
d represents threshold, and
D is a normalizing constant equal to 1.701.

After computing *k* ICTCs in the GRM, *k* + 1 item category characteristic curves (ICCCs) are derived by subtracting adjacent ICTCs:

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

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

The GRM is also commonly expressed as follows:

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

where
all components are as defined above.

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 that 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, refer to Lord and Novick (1968), Hambleton and Swaminathan (1985), or Baker and Kim (2004).

7.2 Model Selection

As stated previously, for grade span K–2 the 1PL model and PCM were used, and for all other grade spans the 2PL model and GRM were used. For each grade span, data from the fall 2008 field test were used to evaluate model fit. The PARSCALE program was used to estimate all item parameters for the fall 2008 field test and for all subsequent analyses. These results were presented to the MCAS Technical Advisory Committee along with recommendations as to which models should be used. Several factors contributed to model selection: model fit (comparing observed data to estimated parameters), standard errors associated with estimated parameters, number of Newton cycles to converge, sample size and future anticipated sample sizes, and classical test theory statistics. The IRT models selected for each grade span optimized model fit with an eye toward future data collection activities.

Although IRT model selection was principally performed using the fall 2008 field test data, all models were again evaluated using the spring 2009 operational data (i.e., the data used in standard setting). Spring 2009 was the first time MELA-O (speaking and listening) data were available for IRT calibration. Several different calibration approaches were evaluated in an attempt to find a set of procedures that produced a scale that was dominated by reading and writing, and where test characteristic curve (TCC) charts from Sessions 1 and 2 and Sessions 2 and 3 reflected the intentions of the test developers. Test developers worked to align the tests with the goals of the program—in other words, to ensure that Sessions 1 and 2 were indeed easier than Sessions 2 and 3. For each grade span a simple calibration approach—a concurrent calibration across all reading, writing, speaking, and listening items—yielded optimal results. For the MELA-O observations it was necessary to fix the discrimination parameters to a value of 1.00 (for the *a* parameters K–2 was automatically set to 1.0 because of the use of the PCM). Due to factors such as high inter-item correlations among MELA-O observations, discrimination parameters were fixed so as to not inflate overall test level information (see the discussion on dimensionality in section 6.3). This ultimately resulted in information functions (see Appendix F) that could accurately assess student performance across the performance continuum while also ensuring consistent (i.e., parallel) test forms from one year to the next.

During the spring 2009 operational calibration activities, staff members at the Research and Evaluation Methods Program at the University of Massachusetts at Amherst evaluated the procedures and results. All analysis input, output, and decisions were thoroughly reviewed to ensure that a proper scale had been established for each grade span. Particular emphasis was placed on discussing information functions to ensure that estimated parameters adequately reflected student-item interaction while providing adequate measurement across the entire performance continuum.

7.3 Item Response Results

The tables in Appendix G give the IRT item parameter calibration results of all items in the spring 2009 administration by grade span.

The TCCs in Appendix F display the expected (average) raw score associated with each θ value between -4.0 and 4.0. Mathematically, the TCC is computed by summing the ICCs of all items that contribute to the raw score. Using the notation introduced in Section 7.1, the expected raw score at a given value of θ_j is as follows:

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

where
i indexes the items (and n is the number of items contributing to the raw score),
j indexes students (here, θ_j runs from -4 to 4), and
$E(X \mid \theta_j)$ is the expected raw score for a student of ability θ_j.

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

Graphics that compare the TCCs between the two session combinations (reading 1 and 2 and writing 1 and 2, reading 2 and 3 and writing 2 and 3) are presented in Appendix F along with the graphics of the test information functions (TIFs). The TIFs display the amount of statistical information associated with each θ value. TIFs essentially depict test precision across the entire latent trait continuum. For detailed information about TIFs, refer to the references listed in section 7.1.

7.4 Achievement Standards (Development of and Match to Content Standards)

Following the spring 2009 operational administration, it was necessary to conduct a standard-setting meeting in order to determine where along the performance continuum the various cut scores resided. A full, detailed standard-setting report (including methods and results) is available at http://www.mcasservicecenter.com/McasDefault.asp?ProgramID=14. All item parameters, TCCs, and information functions used in the standard-setting process were the same as those used operationally to build the raw score to scaled score lookup tables once the results of standard setting were finalized. A summary of the standard-setting process follows.

7.4.1 Standard-setting Process

MEPA assessments are intended to measure English language proficiency across an extensive range of language ability in four subdomains, for students from extremely diverse backgrounds and more than five dozen language groups. Scores from each subdomain are combined and reported using a common scale that is reliable and valid for all students, and allows comparisons across students, schools, and districts, regardless of whether the students took Sessions 1 and 2, or Sessions 2 and 3.

The Department began the standard setting process in summer 2009 when it selected 81 panelists to serve on five grade span committees and make recommendations on MEPA cut scores. Panelists were selected for their diversity as well as their common interest and involvement in English language learner (ELL) education. Approximately 50 percent worked directly with ELLs and approximately 40 percent were grade-level teachers with expertise in English language arts and other content areas, with much overlap. About 10 percent were either school administrators or higher education faculty.

Because MEPA is composed of multiple components (reading, writing, listening, and speaking), decisions were required as to whether cut scores would have to be determined for each. Reading-writing and listening-speaking are assessed using different instruments with different item types and psychometric characteristics. The reading-writing component uses more traditional question formats (e.g., multiple-choice, short-answer, and open-response) to measure achievement. The MELA-O is an observational assessment during which educators rate students on five separate 6-point scales (0–5) for listening and speaking, and then report the scores to the Department. Because of these distinctions, and based on discussions with the Technical Advisory Committee (TAC), two separate standard-setting procedures were used during the MEPA standard setting: one for reading-writing (R/W), and one for listening-speaking (L/S).

MEPA standard setting was conducted in four steps: (1) establish revised performance standards for the R/W components, (2) establish revised performance standards for the L/S components, (3) combine the R/W performance standards with the L/S performance standards, and (4) make recommendations for cross-grade comparability. A modified bookmark procedure was used to establish revised performance standards for the R/W components, with two rounds of ratings. A modified body of work procedure with student profiles was used to establish revised performance standards for the L/S components, with two rounds of ratings. The purpose of the third step was to combine the results of the R/W standard setting with the results of the L/S standard setting to create a single set of cut points for each grade span. Finally, in the fourth step, the combined results from step 3 were reviewed in order to assess the comparability of the cut points across grade spans and to validate, in particular, the *Level 4/5* cut point with empirical evidence, since this level would serve as a threshold for recommending the reclassification of a student as non- (or formerly) LEP.

7.4.1.1 Reading and Writing

During the standard setting for reading and writing, test items were rank ordered by difficulty and panelists were oriented and trained on the standard-setting process, in which they were asked to identify the point in an ordered set of test items at which the students at the borderline of two performance levels no longer had a two-thirds chance of answering the item correctly. Constructed-response questions were represented multiple times in the rank-ordered list—once for each possible score point. Each panelist reviewed the ordered item booklet item by item, considering the knowledge, skills, and abilities students needed to respond to each item. Panelists conducted this task individually, then discussed the ordered item list together, after which they reviewed the MEPA performance level definitions for reading and writing. This was intended to ensure that panelists thoroughly understood the knowledge, skills, and abilities needed for students to be classified into *Levels 1–5*. Panelists developed a consensus definition of borderline students for each performance level; that is, students who are “just able enough” to be categorized into each performance level. Each panelist then participated in two rounds of bookmark placements.

7.4.1.2 Listening and Speaking

Panelists were asked to classify each student profile into a single performance level by considering the holistic pattern of student scores across subdomains, using the body of work standard-setting method, which was developed specifically for use with assessments that allow for a range of student responses, such as portfolio- and performance-based assessments. A modified version of the method has been in use for a number of years that substantially reduces the logistical burden of the procedure and has been found to yield reasonable and defensible cut points. Because the MELA-O does not result in work samples per se, student profiles were used instead, defining typical patterns of subdomain scores for a given total score. Panelists were familiarized with the MELA-O and shown several videotaped student samples that were used to train MELA-O administrators. Panelists then convened in grade span groups to discuss and clarify the performance level definitions for listening and speaking. This was intended to ensure that panelists fully understood the listening and speaking knowledge, skills, and abilities needed for students to be classified into *Levels 1–5*.

The panelists reviewed forty profiles ranging from 0–25 points, rating each profile based on the performance level definitions. While the profiles were presented in order of total score, panelists were encouraged to regard scores holistically and to review the *pattern* of subdomain scores, plus the knowledge, skills, and abilities required to obtain those scores, rather than making a judgment solely on a total raw score. After sharing the average cut-point locations with panelists and conducting a group discussion, panelists performed a second round of ratings.

7.4.1.3 Combining Ratings from Four Subdomains

In order to combine the cut scores, the Department used a weighted average approach in which the cut scores for both R/W and L/S were placed on the underlying theta metric established during the IRT calibration. These theta cut scores were weighted according to the number of points each component (R/W and L/S) contributed to the total MEPA score, and then averaged. This weighted average represents, on the theta metric, the combined cut scores from the modified bookmark and modified body of work processes. This formula was also used on previous versions of MEPA and was regarded as giving an appropriate weight to each subdomain of reading, writing, listening, and speaking.

7.4.1.4 Cross-grade Comparability and Final Standard-setting Decisions

Careful examination of the Round 2 R/W results led to consideration of an important policy decision and its implications regarding the number of students identified in *Level 5* across grade spans. Standard-setting panels determined differing points at each grade span at which they considered a student's results to be *Level 5*. After much discussion, a decision was reached to make the levels at which a student's score became *Level 5* consistent across all grade spans, and parallel with those attaining a score of *Proficient* on MCAS ELA tests (in cases where the same student took both tests), and the relatively fixed percentage of students each year being reclassified as non-LEP. It might have been misinterpreted if a student was placed in *Level 5*, with a recommendation to reclassify, if educators and parents did not believe the student was ready, and different *Level 5* cuts at each grade span would have been equally confusing. This issue, which had been discussed with the MCAS TAC prior to standard setting, was further explored in meetings with standard-setting facilitators immediately following the MEPA standard setting.

The Department decided to further evaluate the MCAS performance of those students who also participated in MEPA. The evaluation conducted by the Department revealed that, with the exception of the 5–6 grade span, performance expectations at the *Level 4/5* cut score were inconsistent with one another (ranging from 5 to 25 percent), and were not in line with historical MEPA reclassification rates which range between 20 to 23 percent in grades 3–12 and 13 percent in K–2. They also showed no consistent relationship with the MCAS ELA performance of ELL students. It was difficult to justify having only five percent of the ELL students reaching *Level 5* on

MEPA when 20 percent of those same students were identified as *Proficient* on the MCAS ELA test in the same year.

Consequently, the Department decided to adjust all of the *Level 4/5* cuts to bring them in line with each other and with the *Proficient* standard on the grade 3–10 ELA tests. The rationale for readjusting the *Level 4/5* cuts final results are depicted in Table 7-1, and the final adjusted MEPA performance level distribution is presented in Figure 7-1.

Table 7-1. 2009 MEPA: Empirical Evidence for Adjusting *Level 4/5* Cuts

Grade Span	Percent of Students at Level 5, determined by Standard Setting Committees	Percent of LEP Students Scoring Proficient on MCAS ELA tests	Percent of LEP students reclassified annually as non-LEP (State)	Percent of Students at Level 5, Readjusted
K–2	9	NA	10–13	13
3–4	11	18–23	18–24	20
5–6	23	18	21–28	22
7–8	5	16–24	18–19	23
9–12	5	20	14–20	20

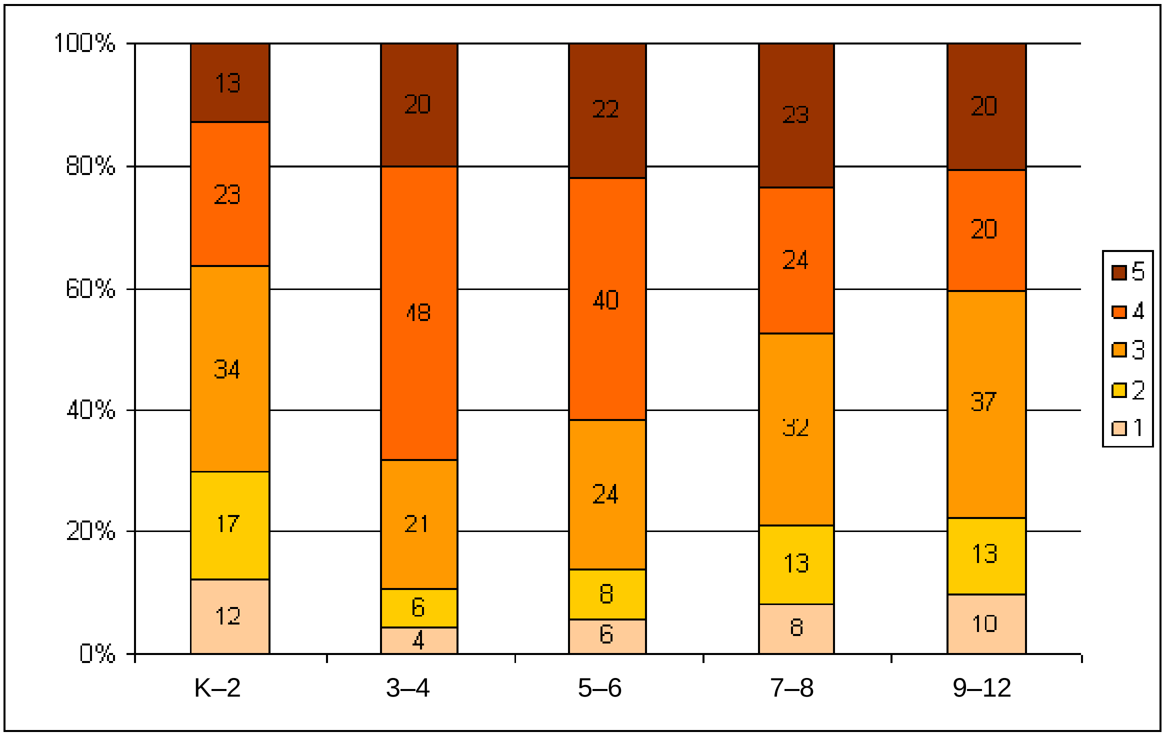


Figure 7-1. 2009 MEPA: Performance Level Distribution

7.5 Reported Scaled Scores

7.5.1 Description of Scale

Overall scaled scores for MEPA range from 400 to 550. This 150-point scale was selected because it minimized problems of scale compression (the number of raw score points collapsing to a single scaled score) and scale expansion (the number of unused scaled score values within the 150-point range). The scaled score cutpoint of 500 for the *Level 4*/*Level 5* cut was fixed across grade spans.

The *Level 1*/*Level 2*, *Level 2*/*Level 3*, and *Level 3*/*Level 4* cutpoints varied across grade spans depending on the location of the theta (θ) cut score established during MEPA standard setting in 2009. Scaled score cutpoints are presented in Table 7-2. A policy decision was made that students taking Sessions 1 and 2 in either reading or writing could not achieve a scaled score greater than 499 (scale truncation).

Table 7-2. 2009 MEPA: Scaled Score Cutpoints by Performance Level

Grade span	*Level 1/Level 2*	*Level 2/Level 3*	*Level 3/Level 4*	*Level 4/Level 5*
K–2	453	466	485	500
3–4	432	452	474	500
5–6	436	456	479	500
7–8	443	464	486	500
9–12	450	464	489	500

7.5.2 Calculations

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

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

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:

$$m = \frac{500 - 400}{\theta_4 - (-4.0)}$$

$$b = SS_1 - m\theta_1$$

where

500 and 400 are the scaled score cuts for the *Level 4/Level 5* and minimum scaled score, respectively, and

θ_4 is the theta cut corresponding to the scaled score cut of 500.

The transformation constants (slope and intercept) for each grade span are presented in Table 7-3.

Table 7-3. 2009 MEPA: Transformation Constants for Composite MEPA Scores

Grade	*Transformation constants*	
span	*Slope*	*Intercept*
K–2	19.93	479.73
3–4	20.06	480.24
5–6	20.16	480.65
7–8	20.08	480.32
9–12	20.12	480.48

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 TCC. While the rubric and procedure for assigning MELA-O scores were the same for all students, students took different combinations of

MEPA reading and writing sessions. Therefore, the IRT parameters for the MELA-O indicators and the MEPA-R/W items were used together to calculate four TCCs (and four TIFs) for each administration of the MEPA—one for each possible combination of reading and writing sessions in each grade span except K–2:

- 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

Because most students took the same combination of sessions in reading and writing (i.e., the first or fourth combination listed above), these two combinations are the focus of the scaled score calculation report. Grade span K–2 consisted of two levels (A or B) instead of three sessions; these two levels are included in the report.

Appendix H provides tables showing each raw score with its corresponding theta and overall scaled score. Tables are included for both the spring and fall 2009 tests.

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 by session, reading and writing raw scores were translated to a scale that ranged from 0 to 30. This was necessary because the reading and writing components were designed to exhibit different levels of difficulty depending on the session. Putting these components onto a 0–30 metric enables comparisons when evaluating student performance.

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:

$$m = \frac{SS_{\max} - SS_{\min}}{\theta_{\max} - \theta_{\min}} = \frac{30 - 0}{4.0 - (-4.0)} = \frac{30}{8} = 3.75$$

$$b = SS_{\min} - m\theta_{\min} = 0 - 3.75(-4.0) = 15$$

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 for reading. The conversion tables from raw score to scaled score for both reading and writing are also provided in Appendix H.

7.5.3 Distributions

The composite scaled score distributions for each grade span for the spring 2009 and fall 2009 administrations are displayed in Appendix I.

Chapter 8. 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: some students will receive scores that underestimate their true level of knowledge, and other students will receive scores that overestimate their true level of knowledge. Items that function well together produce assessments that have less measurement error (i.e., errors 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 is minimal.

In the determination of assessment reliability for MEPA, MELA-O indicators were treated in the same manner as MEPA-R/W items. MELA-O indicators have been included with open-response data.

8.1 Reliability and Standard Errors of Measurement

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,
Error! Objects cannot be created from editing field codes. represents individual item variance, and
Error! Objects cannot be created from editing field codes. represents the total test variance.

Tables J-1 and J-2 in Appendix J present descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement (SEMs) for each MEPA grade span.

As described previously, the SEM 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 2.4.2 for details).

8.2 Subgroup Reliability

The reliability coefficients described in the previous section were based on the overall population of students who took the 2009 MEPA tests. Tables J-3 and J-4 of Appendix J present reliabilities for various subgroups of interest (by gender, ethnicity, LEP status, income level, and special education status), as required for AYP reporting. Cronbach's α coefficients for each subgroup were calculated using the formula defined above including only the members of the subgroup in question in the computations.

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

8.3 Reporting Categories Reliability

In previous sections, the reliability coefficients were calculated based on form and item type. Item type represents just one way of breaking an overall test into subtests. Of even more interest are reliabilities for the reporting subcategories within MEPA content areas. Cronbach's α coefficients for subcategories were calculated via the same formula defined previously using just the items of a given subcategory in the computations. Results are presented in Tables J-5 and J-6 of Appendix J. As expected, because they were based on a subset of items rather than the full test, computed subcategory reliabilities were lower (sometimes substantially so) than overall test reliabilities, and interpretations should take this information into account.

8.4 Interrater Reliability

Interrater reliability is the degree of agreement among raters or scorers. It gives a value of how much homogeneity, or consensus, there is in the scorings given by different scorers. A number of statistics can be used to determine interrater reliability. Different statistics are appropriate for different types of measurement and different purposes of interpretation. Some options are listed here.

The joint probability of agreement is probably the most simple and least robust measure. It is the number of times each score point (e.g., 1, 2, . . . 5) is assigned by each scorer divided by the total number of scorings. However, it does not take into account that agreement may happen solely based on chance.

Cohen's kappa does consider the amount of agreement that could be expected to occur through chance. Tables K-1 through K-4 in Appendix K provide the kappa coefficients and their standard errors. Also shown are the proportions of exact agreement, adjacent agreement, and exact plus adjacent agreement. All statistics are provided for both the double-blind and read-behind scoring procedures (see section 5.1.2.6 for complete details of scoring quality monitoring procedures).

8.5 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 in the next section 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.

8.5.1 Accuracy and Consistency

Accuracy refers to the extent to which decisions based on test scores match decisions that would have been made if the scores did not contain any measurement error. Accuracy must be estimated because errorless test scores do not exist.

Consistency measures the extent to which classification decisions based on test scores match the decisions based on scores from a second, parallel form of the same test. Consistency can be evaluated directly from actual responses to test items if two complete, 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 it works with both open-response and multiple-choice items.

8.5.2 Calculating Accuracy

Following Livingston and Lewis (1995), 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 (1995) method, the estimated true scores are used to classify each student into his or her true performance category, which is labeled “true status.” After various technical adjustments (described in Livingston & Lewis, 1995), a four-by-four contingency table of accuracy is created for each grade span. 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).

8.5.3 Calculating Consistency

To estimate consistency, the true scores are used to estimate the distribution of classifications on an independent, parallel test form. Following statistical adjustments per Livingston and Lewis (1995), a four-by-four consistency contingency table is created for each grade span to show the proportions of students who are classified into each performance category by the actual test and by a (hypothetical) parallel test form.

8.5.4 Calculating 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. In this case, the second form 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.

> Note The lowest performance level of the spring administration and the highest performance level of the fall administration do not have enough students to provide meaningful results. Thus, the two lowest levels (*Level 1* and *Level 2*) for the spring administration and the two highest levels (*Level 4* and *Level 5*) for the fall administration are merged in the calculations.

8.5.5 Results of Accuracy, Consistency, and Kappa Analyses

Summaries of the MEPA accuracy and consistency analyses are provided in Tables 8-1 and 8-2. Detailed results can be seen in Appendix L.

The first part of Table 8-1 shows overall accuracy and consistency indices as well as kappa. The second part displays 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 0.74 for the *Level 3* category for grade span 3–4 for the spring 2009 administration. This indicates that, of the students whose actual scores placed them in the *Level 3* category, 74 percent would be expected to be in the *Level 3* category if they were categorized according to their true score. Similarly, the corresponding consistency value of 0.65 indicates that 65 percent of that same group of students would be expected to score in the *Level 3* category if a second, parallel test form were used.

Table 8-2 gives information at each of the cutpoints. 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 Table 8-2 indicate the accuracy and consistency of the dichotomous decisions either above or below the associated cutpoint. 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 8-1. 2009 MEPA: Summary of Decision Accuracy (and Consistency) Results—Overall and Conditional on Performance Level

				Conditional on Level			
	Grade span	*Overall*	*Kappa*	*Level 1 & Level 2*	*Level 3*	*Level 4*	*Level 5*
Spring 2009 Administration	K–2	0.76 (0.67)	0.55	0.85 (0.78)	0.73 (0.65)	0.67 (0.56)	0.85 (0.72)
	3–4	0.80 (0.72)	0.61	0.82 (0.73)	0.74 (0.65)	0.79 (0.72)	0.88 (0.80)
	5–6	0.78 (0.70)	0.59	0.83 (0.75)	0.74 (0.65)	0.72 (0.63)	0.89 (0.81)
	7–8	0.77 (0.69)	0.58	0.85 (0.79)	0.73 (0.64)	0.60 (0.48)	0.89 (0.80)
	9–12	0.78 (0.70)	0.59	0.85 (0.78)	0.79 (0.73)	0.55 (0.44)	0.88 (0.79)
				Conditional on Level			
	Grade span	Overall	Kappa	*Level 1*	*Level 2*	*Level 3*	*Level 4 & Level 5*
Fall 2009 Administration	K–2	0.79 (0.71)	0.6	0.88 (0.84)	0.63 (0.52)	0.74 (0.65)	0.88 (0.79)
	3–4	0.80 (0.73)	0.63	0.89 (0.86)	0.65 (0.54)	0.65 (0.54)	0.91 (0.83)
	5–6	0.79 (0.72)	0.62	0.89 (0.86)	0.64 (0.52)	0.65 (0.54)	0.89 (0.81)
	7–8	0.81 (0.74)	0.64	0.90 (0.88)	0.66 (0.55)	0.66 (0.55)	0.90 (0.83)
	9–12	0.78 (0.70)	0.59	0.85 (0.80)	0.59 (0.48)	0.76 (0.68)	0.89 (0.80)

Table 8-2. 2009 MEPA: Summary of Decision Accuracy (and Consistency) Results—Conditional on Cutpoint

		Level 1 & Level 2 / Level 3			*Level 3 / Level 4*			*Level 4 / Level 5*		
	Grade span	*Accuracy (consistency)*	*False positive*	*False negative*	*Accuracy (consistency)*	*False positive*	*False negative*	*Accuracy (consistency)*	*False positive*	*False negative*
Spring 2009 Administration	K–2	0.92 (0.88)	0.04	0.04	0.90 (0.87)	0.06	0.04	0.94 (0.92)	0.04	0.02
	3–4	0.96 (0.94)	0.02	0.02	0.92 (0.88)	0.04	0.04	0.93 (0.90)	0.05	0.03
	5–6	0.95 (0.93)	0.02	0.03	0.91 (0.88)	0.05	0.04	0.92 (0.89)	0.05	0.03
	7–8	0.93 (0.90)	0.03	0.03	0.91 (0.88)	0.05	0.04	0.93 (0.90)	0.05	0.03
	9–12	0.93 (0.91)	0.03	0.03	0.91 (0.88)	0.05	0.04	0.93 (0.90)	0.04	0.03
		Level 1 / Level 2			*Level 2 / Level 3*			*Level 3 / Level 4 & Level 5*		
	Grade span	*Accuracy (consistency)*	*False positive*	*False negative*	*Accuracy (consistency)*	*False positive*	*False negative*	*Accuracy (consistency)*	*False positive*	*False negative*
Fall 2009 Administration	K–2	0.92 (0.89)	0.04	0.04	0.92 (0.88)	0.05	0.04	0.94 (0.92)	0.03	0.02
	3–4	0.93 (0.90)	0.04	0.03	0.93 (0.90)	0.04	0.03	0.94 (0.92)	0.03	0.02
	5–6	0.93 (0.90)	0.04	0.03	0.93 (0.90)	0.04	0.03	0.94 (0.91)	0.04	0.02
	7–8	0.93 (0.90)	0.04	0.03	0.93 (0.90)	0.04	0.03	0.94 (0.92)	0.04	0.02
	9–12	0.92 (0.89)	0.04	0.04	0.91 (0.88)	0.05	0.04	0.94 (0.91)	0.04	0.02

Chapter 9. VALIDITY

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

Standards for Educational and Psychological Testing (AERA et al., 1999) provides a framework for describing sources of evidence that should be considered when evaluating validity. These sources include evidence on the following five general areas: test content, response processes, internal structure, consequences of testing, and relationship to other variables. Although each of these sources may speak to a different *aspect* of validity, they are not distinct *types* of validity. Instead, each contributes to a body of evidence about the comprehensive validity of score interpretations.

A measure of test content validity is to determine how well the test's tasks represent the curriculum and standards for each content area and grade level. This is informed by the item development process, including how test blueprints and test items align with the curriculum and standards. Validation through this content lens is extensively described in Chapter 3. In other words, the element's components discussed in the chapter—item alignment with content standards; item bias; sensitivity and content appropriateness review processes; adherence to the test blueprint; use of multiple item types; use of standardized administration procedures, with accommodated options for participation; and appropriate test administration training—are all components of content-based validity evidence. Finally, as described in Chapter 4, tests were administered according to mandated standardized procedures, with allowable accommodations, and all test coordinators and administrators were required to familiarize themselves with and adhere to all of the procedures outlined in the *MEPA Test Administrator's Manual.*

The scoring information in Chapter 5 describes the steps taken to train and monitor live scorers as well as the quality control procedures related to machine scanning and scoring. Additional studies might be helpful for evidence on student response processes. For example, think-aloud protocols could be used to investigate students' cognitive processes when confronting test items.

Evidence on internal structure is extensively detailed in the chapters on item analyses, scaling and equating, and reliability (Chapters 6 to 8). Technical characteristics of the internal structure of the tests are presented in terms of classical item statistics (*p*-values and discriminations), differential item functioning (DIF) analyses, several reliability coefficients, standard errors of measurement (SEMs), multidimensionality hypothesis-testing and effect-size estimation, and item response theory (IRT) analyses.

Evidence on the consequences of testing is addressed in the information on scaled scores and reporting in Chapters 7 and 10 and in the *Guide to Interpreting the MEPA Reports for Schools and Districts*, which is available at http://www.doe.mass.edu/mcas/mepa/2009/interpret.pdf. Each of these speaks to efforts undertaken to provide the public with accurate and clear test score information. Scaled scores simplify results reporting across content areas, grade levels, and successive years. Performance levels give reference points for mastery at each grade level—another useful and simple way to interpret scores. Evidence on the consequences of testing could be supplemented with broader research on MEPA's impact on student learning.

The remaining part of this chapter describes further studies of validity that could enhance the investigations that have already been performed. The proposed areas of validity to be examined fall into four categories: external validity, convergent and discriminant validity, structural validity, and procedural validity.

9.1 External Validity

In the future, investigations of external validity could involve targeted examination of the variables that one might expect to correlate with MEPA results, like classroom grades or classroom test scores in the same content areas as the MEPA test in question, for example.

9.2 Convergent and Discriminant Validity

The concepts of convergent and discriminant validity were defined by Campbell and Fiske (1959) as specific types of validity that fall under the umbrella of construct validity. *Convergent validity* is the notion that measures or variables that are intended to align should actually be aligned in practice. *Discriminant validity*, on the other hand, is the idea that measures or variables that are intended to differ should not be too highly correlated. Evidence for validity comes from examining whether the correlations among variables are as expected in direction and magnitude.

Campbell and Fiske (1959) introduced the study of different traits and methods as the means of assessing convergent and discriminant validity. *Traits* refer to the constructs that are being measured (e.g., mathematical ability), and methods are the instruments of measuring them (e.g., a mathematics test or grade). To utilize the framework of Campbell and Fiske, it is necessary that more than one trait and more than one method be examined. Analysis is performed through the multi-trait/multi-method matrix, which gives all possible correlations of the different combinations of traits and methods. For MEPA, convergent and discriminant validity could be examined by constructing a multi-trait/multi-method matrix in which the traits examined would be reading, writing, listening, and speaking, and the methods could include MEPA subscale scores and such variables as grades, teacher judgments, and scores on another standardized test.

9.3 Structural Validity

Though the previous types of validity examine the concurrence between different measures of the same content area, structural validity focuses on the relationship between strands *within* a content area, thus supporting content validity. *Structural validity* is the degree to which related elements of a test are correlated in the intended manner. For instance, it is desired that performance on different strands of a content area be positively correlated; however, as these strands are designed to measure distinct components of the content area, it is reasonable to expect that each strand would contribute a unique component to the test. Additionally, it is desired that the correlation between different item types (multiple-choice, short-answer, and open-response) of the same content area be positive.

As an example, an analysis of MEPA structural validity would investigate the correlation between performance in reading and writing and performance in MELA-O. The concordance between performance on multiple-choice items and open-response items would also be examined. Such a study would address the consistency of MEPA tests within each grade span. In particular, the dimensionality analyses of Chapter 6 could be expanded to include confirmatory analyses addressing these concerns.

9.4 Procedural Validity

As mentioned earlier, the *MEPA Test Coordinator* and *Test Administrator* manuals delineated the procedures to which all MEPA test coordinators and test administrators were required to adhere. A study of procedural validity would provide a comprehensive documentation of the procedures that were followed throughout the MEPA administration. The results of the documentation would then be compared to the manuals, and procedural validity would be confirmed to the extent that the two were in alignment. Evidence of procedural validity is important because it verifies that the actual administration practices were in accord with the intentions of the design.

Possible instances where discrepancies can exist between design and implementation include the following: cheating among students occurs; or answer documents are scanned incorrectly. These are examples of procedural error. A study of procedural validity involves capturing any such errors and presenting them within a cohesive document for review.

Chapter 10. Reporting of Results

10.1 Unique Reporting Notes

Results for the spring 2009 MEPA administration for students in kindergarten through grade 12 were provided in the following reports:

- *Preliminary Participation Report*
- *Preliminary Results by Year of Enrollment in Massachusetts Schools*
- *Roster of Student Results*
- *Parent/Guardian Report*

MEPA tests are intended to measure students' performance and progress in acquiring fluency in English. The fall MEPA administration is meant strictly to determine baseline scores for new students and students who did not test the previous spring. Following the fall MEPA administration, only the *Roster of Student Results* reports were generated and provided to schools and districts. Complete MEPA results for both administrations were reported following the spring administration.[1] Each report is briefly described in this chapter; copies of the report shells are provided in Appendix M.

10.2 School and District Results Reports

10.2.1 Preliminary Reports

For 2009, the following two reports were generated for each grade span in a school:

- *Preliminary Participation Report* (all years)
- *Preliminary Results by Year of Enrollment in Massachusetts Schools*

Each report is described here and in more detail in the *Guide to Interpreting the MEPA Reports for Schools and Districts*, available at www.doe.mass.edu/mcas/mepa/2009/interpret.pdf.

To ensure student confidentiality and to discourage generalizations about school and district 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 or district.

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.

10.2.1.1 Preliminary Participation Report

This report shows the following data for the grade span:

- 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

[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 Identifiers (SASIDs), 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' SASIDs; linked results were compared to the state's Student Information Management System (SIMS) limited English proficiency (LEP) enrollment data to determine actual participation rates.

- 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 (i.e., MEPA-R/W and MELA-O)

10.2.1.2 Preliminary Results by Year of Enrollment in Massachusetts Schools

This report gives student results 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 K–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)

Results were aggregated by the number of years students had been enrolled in Massachusetts schools: 1 year, 2 years, 3 years, 4 years, and 5 or more years.

10.2.2 Roster of Student Results

This report provides a school with the MEPA results for each LEP student at that school. A separate *Roster of Student Results* report for each grade span is 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. Each student's overall scaled score, performance level, and scaled subscores in reading, writing, listening, and speaking are shown.[3,4]

10.3 Student Report

The spring MEPA *Parent/Guardian Report* shows the student and his or her parents/guardians how the student performed in the MEPA administration(s) in which he or 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.

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

[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 7.5.2 of this report.

[4] If a student participated in more than one test administration, and his or her records from each administration could be matched based on student records from 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.

score is also depicted graphically on a 400–550 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 *Level 5* 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 Massachusetts. Each comparison is described in detail in the *Guide to the MEPA for Parents/Guardians.*

10.4 Interpretive Materials and Workshops

Interpretive information for the reports described in section 10.2 is provided in the Department's publication *Guide to Interpreting the MEPA Reports for Schools and Districts*.

Two conference calls were conducted to provide information to schools about the spring reports. These conference calls took place around the time that reports were released. Schools were invited to participate and members of the Department explained the reports that were available and the types of information on each. These presentations especially emphasized the new design of the tests, the new reporting scale, and the new performance levels.

A 2009 *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 displayed results.

10.5 Decision Rules

To ensure that reported results for the MEPA were accurate relative to collected data and other pertinent information, a document that delineated analysis and reporting rules was created. These decision rules were observed in the analyses of test data and in reporting the test results. Moreover, these rules were the main reference for quality assurance checks.

The decision rules document used for reporting results of the 2009 administrations of the MEPA is found in Appendix N.

The first set of rules pertains to general issues in reporting scores. Each issue was described, and pertinent variables were identified. The actual rules applied were described by the way they would impact analyses and aggregations and their specific impact on each of the reports. The general rules were further grouped into issues pertaining to test items, school type, student exclusions, and number of students for aggregations.

The second set of rules pertains to reporting student participation. These rules describe which students were counted and reported for each subgroup in the student participation report.

10.6 Quality Assurance

Quality assurance measures are embedded throughout the entire process of analysis and reporting. The data processor, data analyst, and psychometrician assigned to work on the MEPA implement quality control checks of their respective computer programs and intermediate products. Moreover, when data are handed off to different functions within the Psychometrics and Research and Data Services and Static Reporting Departments, the sending function verifies that the data are accurate before handoff. When a function receives a data set, the first step is to verify the data for accuracy.

Another type of quality assurance measure is parallel processing. Students' scaled scores for each content area are assigned by a psychometrician through a process of equating and scaling. The scaled scores are also computed by a data analyst to verify that scaled scores and corresponding performance levels are assigned accurately. Respective scaled scores and performance levels assigned are compared across all students for 100 percent agreement. Different exclusions assigned to students that determine whether each student receives scaled scores and/or is included in different levels of aggregation are also parallel-processed. Using the decision rules document, two data analysts independently write a computer program that assigns students' exclusions. For each grade and content area combination, the exclusions assigned by each data analyst are compared across all students. Only when 100 percent agreement is achieved can the rest of the data analysis be completed.

The third aspect of quality control involves the procedures implemented by the quality assurance group to check the veracity and accuracy of reported data. Using a sample of schools and districts, the quality assurance group verifies that reported information is correct. The step is conducted in two parts: (1) verify that the computed information was obtained correctly through appropriate application of different decision rules, and (2) verify that the correct data points populate each cell in the reports. The selection of sample schools and districts for this purpose is very specific and can affect the success of the quality control efforts. Two sets of samples are selected, though they may not be mutually exclusive.

The first set includes those that satisfy the following criteria:

- One-school district
- Two-school district
- Multi-school district

The second set of samples includes districts or schools that have unique reporting situations, as indicated by decision rules. This set is necessary to check that each rule is applied correctly. The second set includes the following criteria:

- Private school
- Small school that receives no school report
- Small district that receives no district report
- District that receives a report, but all schools are too small to receive a school report
- School with excluded (not tested) students

The quality assurance group uses a checklist to implement its procedures. After the checklist is completed, sample reports are circulated for psychometric checks and program management review. The appropriate sample reports are then presented to the Department for review and signoff.

REFERENCES

Allen, Mary J., & Yen, Wendy M. (1979). *Introduction to measurement theory.* Belmont, CA: Wadsworth.

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

Baker, F. B., & Kim, S.H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Dekker.

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

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, *23*, 355–368.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley and Sons, Inc.

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

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

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates.

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Massachusetts Department of Education. (2009). *Guide to interpreting the spring 2009 MEPA reports for schools and districts*. Malden, MA:

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

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52*, 589–617.

Stout, W. F., Froelich, A. G. & Gao, F. (2001). Using resampling methods to produce an improved DIMTEST procedure. In A. Boomsma, M. A. J. van Duign, & T. A. B. Snijders (Eds.) *Essays on item response theory*, (pp. 357–375). New York: Springer-Verlag.

Zhang, J., & Stout. W. F. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*, 213-249.

APPENDICES