Massachusetts Department of
ELEMENTARY & SECONDARY
EDUCATION

Massachusetts English Proficiency Assessment (MEPA)

2011–12 Technical Report

February 2013

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

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

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

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 ELL 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 a former English language learner (former ELL) who no longer needs ELL services. The MEPA is based on *English Language Proficiency Benchmarks and Outcomes for English Language Learners* developed by Massachusetts 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 the following 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 provided 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. Each level provides separate single-session tests for reading and writing.

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 combined 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 a former ELL is not required for a student at a particular MEPA performance level, the Massachusetts 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; Massachusetts Comprehensive Assessment System (MCAS) scores; and the observations and recommendations of educators.

1.2 DESCRIPTION OF THIS REPORT

The purpose of this report is to document the technical quality and characteristics of the 2011 and 2012 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, administration, scoring, analysis, and reporting of student results.

Chapter two describes updates to the MEPA program that occurred in 2011 and 2012. Each remaining chapter is dedicated to a specific aspect of the program; together, they provide detailed descriptions of each step of the testing process and detailed analyses of the 2011–2012 MEPA results. Numerous appendices are referenced throughout the report and are included at the end of the document.

CHAPTER 2. UPDATES FOR 2011–2012

2.1 COMPUTER-BASED TESTING

Computer-based testing continued for the second and third operational MEPA-R/W administrations in the spring of 2011 and the spring of 2012. The spring testing window was extended to allow time for schools participating in computer-based testing to rotate students through computer labs. The 2011 fall test continued to be strictly paper-based.

For the 2011 and 2012 administrations, emphasis was placed on creating test items in such a manner that identical wording could be used in both the computer-based tests (CBT) and the paper-based tests (PBT). As a result, no adaptations had to be made to item wording as were made in 2010 to make PBT items fit the online model.

In the 2010 online tests, passages were displayed at the top of the screen with the corresponding items appearing below. In order to allow enough room on the screen for the item to be fully visible, a considerable amount of scrolling was required when reading the passages. In 2011 and 2012 the display was changed so that each passage appeared on a screen by itself for the students to read with minimal scrolling. On the following screens, the passage appeared on the left-hand side of the screen while the corresponding items appeared on the right-hand side. Changing to the side-by-side format allowed the students to see more of the passage at one time.

2.2 COMPUTER-BASED TEST AND PAPER-BASED TEST COMPARABILITY STUDY

As in 2010, the Department elected to have additional studies conducted in 2011 and 2012 to determine if CBT and PBT results were comparable. The studies conducted in these years were similar to the one conducted in 2010, but a secondary analysis was introduced to match students on additional covariates of gender, economic disadvantage, and primary language. The details and findings of these studies are discussed in the MEPA Comparability Studies in Appendix A.

2.3 PARENT/GUARDIAN REPORTS

Following the reporting of results for the spring 2010 testing, the Department held focus group meetings to gather feedback on the effectiveness of the information displays on the parent reports. These focus groups involved parents of ELL students, as well as ELL instructors and principals. The layout of the reports was redesigned for 2011, and many of the focus group suggestions were incorporated.

The Department also introduced information on the parent reports to indicate whether the students were making progress relative to their performance in previous administrations. The purpose of this information was to help parents know whether their student's English language skills were improving at a rate that would enable the student to no longer require ELL services within five to six years of attending schools in the United States.

CHAPTER 3. TEST DEVELOPMENT AND DESIGN

3.1 ITEM TYPES

Custom test items, based on the Department's *English Language Proficiency Benchmarks and Outcomes for English Language Learners*, are written for the MEPA-R/W tests.

- Multiple-choice items appear in both reading and writing. Each multiple-choice item requires students to select a single best answer from three or four response options. These items are machine scored, and students receive a score of 1 for a correct response or 0 for an incorrect or blank response. In multiple-choice items in K–2 tests, there are three response choices. In the other grade spans there are four choices.
- 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 using item-specific rubrics. K–2 responses are scored on a scale of 0 to 1. Short-answer responses at the other grade spans are scored on a scale of 0 to 2.
- Open-response items in reading require students to generate a response of typically one or more paragraphs. Open-response items are scored on a scale of 0 to 4 by trained readers using item-specific rubrics.
- Short-answer items in writing require students to respond to a graphic or short prompt. Trained readers assign scores of 0 to 2 based on item-specific rubrics (except those K–2 short-answer items described below).
- A short-answer item type that appears only in the K–2 writing tests requires students to review a storyboard and develop the next event in the story. Trained readers assign scores of 0 to 3 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 up to two pages in length based on a prompt. Trained readers assign scores of 0 to 4 based on item-specific rubrics.

3.2 OPERATIONAL TEST DEVELOPMENT PROCESS

The MEPA-R/W development process is typical of large-scale assessments that utilize the common/matrix test design.

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 issues, and content accuracy (see Appendix B for committee membership). These selections provide the bases for the passage-based items on the tests.

3.2.1 Item and Scoring Rubric 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 rubrics are used for all item types except multiple-choice, to indicate how to evaluate student responses and assign appropriate scores. Development of these rubrics begins in conjunction with the items' creation. The 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 based on the *Massachusetts English Language Arts Curriculum Framework* of June 2001 and aligned to *English Language Proficiency Benchmarks and Outcomes for English Language Learners* (see highlights in Appendix C or view the entire document at www.doe.mass.edu/ell/benchmark.pdf), which was created by the Department in June 2003. 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 prewriting, writing, revising, editing, and media.

3.2.3 Internal Item Review

The Department's lead test developers and its testing contractor, Measured Progress, 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 of 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

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 accuracy, content appropriateness, and age appropriateness.

3.2.5 Bias and Sensitivity Review

Following review by the assessment development committees, a separate committee of educators reviews 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 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 MCAS English Language Arts test 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. 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 (TCCs) 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 TCCs are reviewed. If TCCs are similar to those from the prior year, and if test information functions (TIFs) are acceptable, 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 nonitem 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. Several rounds of edits are made, and new hard-copy pages are 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 file of the standard test is printed in a larger format with no other modifications from the original.

3.3 GUIDELINES FOR TEST DESIGNS AND BLUEPRINTS

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

3.3.1 Selection Guidelines

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

- Match the framework;
- Fulfill blueprint specifications;
- Contain accurate and clear content;
- Are age and content appropriate;
- Achieve a critical threshold for item statistics;
- Are free from bias;
- Are sensitive to students with special needs

In addition, TCCs, TIFs, 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 and Blueprints

Once 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. Tables 3-1 through 3-12 show the test blueprints for the 2011 and 2012 MEPA tests.

Table 3-1. 2011–12 MEPA: Reading Test Blueprint, Grade Span K–2
Level A: 16 points

Outcomes Assessed	*2012*	*2011*
	MC	*MC*
Vocabulary & Syntax in Print	4	8
Beginning to Read in English	6	2
Comprehension	4	6
Literary Elements & Techniques	1	0
Informational/Expository Text	1	0
Total Items by Type	16	16
Total Points	16	16

MC = multiple-choice, 1 point

Table 3-2. 2011–12 MEPA: Writing Test Blueprint, Grade Span K–2
Level A: 11 points in 2012, 12 points in 2011

Outcomes Assessed	*2012*			*2011*	
	MC	*SA1*	*SA2*	*SA1*	*SA2*
Writing	0	0	3	0	4
Editing	2	3	0	4	0
Total Items by Type	2	3	3	4	4
Total Points	2	3	6	4	8

MC = multiple-choice, 1 point; SA1 = short-answer, 1 point; SA2 = short-answer, 2 points

Table 3-3. 2011–12 MEPA: Reading Test Blueprint, Grade Span K–2
Level B: 14 points

Outcomes Assessed	*2012*	*2011*
	MC	*MC*
Vocabulary & Syntax in Print	5	4
Beginning to Read in English	3	4
Comprehension	5	6
Literary Elements & Techniques	1	0
Total Items by Type	14	14
Total Points	14	14

MC = multiple-choice, 1 point

Table 3-4. 2011–12 MEPA: Writing Test Blueprint, Grade Span K–2
Level B: 20 points

Outcomes Assessed	*2012*				*2011*			
	MC	*SA1*	*SA2*	*STW*	*MC*	*SA1*	*SA2/SW*	*STW*
Writing	0	0	4	2	0	0	4	2
Editing	2	4	0	0	2	4	0	0
Total Items by Type	2	4	4	2	2	4	4	2
Total Points	2	4	8	6	2	4	8	6

MC = multiple-choice, 1 point; SA1 = short-answer, 1 point; SA2 = short-answer, 2 points; SW = sentence-writing, 2 points; STW = story-writing, 3 points

Table 3-5. 2011–12 MEPA: Reading Test Blueprint, Grade Span 3–4

	Outcomes Assessed	Session 1: 15 Points 0 Reading Passages		Session 2: 12 Points 2 Reading Passages		Session 3: 14 Points 2 Reading Passages		
		MC	SA2	MC	SA2	MC	SA2	OR
2012	Vocabulary & Syntax in Print	5	0	0	0	0	0	0
	Beginning to Read in English	4	0	0	0	0	0	0
	Comprehension	0	3	6	1	4	0	1
	Literary Elements & Techniques	0	0	0	1	1	1	0
	Informational/Expository Text	0	0	2	0	3	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4
2011	Vocabulary & Syntax in Print	5	0	2	0	3	0	0
	Beginning to Read in English	4	0	0	0	0	0	0
	Comprehension	0	3	4	0	3	1	0
	Literary Elements & Techniques	0	0	1	1	1	0	1
	Informational/Expository Text	0	0	1	1	1	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4

MC = multiple-choice, 1 point; SA2 = short-answer, 2 points; OR = open-response, 4 points
Total points possible for sessions 1 & 2: 27
Total points possible for sessions 2 & 3: 26

Table 3-6. 2011–12 MEPA: Writing Test Blueprint, Grade Span 3–4

	Outcomes Assessed	Session 1: 15 Points 0 Reading Passages			Session 2: 12 Points 2 Reading Passages			Session 3: 14 Points 2 Reading Passages		
		SA1	SA2	SW	MC	SW	WP	MC	SW	WP
2012	Prewriting		1	0	0	0	0	0	0	0
	Writing		1	5	0	1	1	0	2	2
	Editing		0	0	4	0	0	4	0	0
	Total Items by Type		2	5	4	1	1	4	2	2
	Total Points		4	10	4	2	4	4	4	8
2011	Prewriting	4		0	0	0	0	0	0	0
	Writing	0		5	0	1	1	0	2	2
	Editing	0		0	4	0	0	4	0	0
	Total Items by Type	4		5	4	1	1	4	2	2
	Total Points	4		10	4	2	4	4	4	8

MC = multiple-choice, 1 point; SA1 = short-answer, 1 point; SA2 = short-answer, 2 points;
SW = sentence-writing, 2 points; WP = writing-prompt, 4 points
Total points possible for sessions 1 & 2: 24
Total points possible for sessions 2 & 3: 26

Table 3-7. 2011–12 MEPA: Reading Test Blueprint, Grade Span 5–6

	Outcomes Assessed	*Session 1: 15 Points 0 Reading Passages*		*Session 2: 12 Points 2 Reading Passages*		*Session 3: 14 Points 2 Reading Passages*		
		MC	*SA2*	*MC*	*SA2*	*MC*	*SA2*	*OR*
2012	Vocabulary & Syntax in Print	6	1	1	0	3	0	0
	Beginning to Read in English	3	0	0	0	0	0	0
	Comprehension	0	2	3	1	4	0	0
	Literary Elements & Techniques	0	0	2	1	1	0	1
	Informational/Expository Text	0	0	2	0	0	1	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4
2011	Vocabulary & Syntax in Print	4	0	2	0	3	0	0
	Beginning to Read in English	5	0	0	0	0	0	0
	Comprehension	0	3	4	0	4	1	1
	Literary Elements & Techniques	0	0	1	1	0	0	0
	Informational/Expository Text	0	0	1	1	1	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4

MC = multiple-choice, 1 point; SA2 = short-answer, 2 points; OR = open-response, 4 points
Total points possible for sessions 1 & 2: 27
Total points possible for sessions 2 & 3: 26

Table 3-8. 2011–12 MEPA: Writing Test Blueprint, Grade Span 5–6

	Outcomes Assessed	*Session 1: 15 Points 0 Reading Passages*			*Session 2: 12 Points 2 Reading Passages*			*Session 3: 14 Points 2 Reading Passages*		
		SA1	*SA2*	*SW*	*MC*	*SW*	*WP*	*MC*	*SW*	*WP*
2012	Prewriting		2	0	0	0	0	0	0	0
	Writing		0	5	0	1	1	0	2	2
	Editing		0	0	4	0	0	4	0	0
	Total Items by Type		2	5	4	1	1	4	2	2
	Total Points		4	10	4	2	4	4	4	8
2011	Prewriting	4		0	0	0	0	0	0	0
	Writing	0		5	0	1	1	0	2	2
	Editing	0		0	4	0	0	4	0	0
	Total Items by Type	4		5	4	1	1	4	2	2
	Total Points	4		10	4	2	4	4	4	8

MC = multiple-choice, 1 point; SA1 = short-answer, 1 point; SA2 = short-answer, 2 points;
SW = sentence-writing, 2 points; WP = writing-prompt, 4 points
Total points possible for sessions 1 & 2: 24
Total points possible for sessions 2 & 3: 26

Table 3-9. 2011–12 MEPA: Reading Test Blueprint, Grade Span 7–8

	Outcomes Assessed	*Session 1: 15 Points 0 Reading Passages*		*Session 2: 12 Points 2 Reading Passages*		*Session 3: 14 Points 2 Reading Passages*		
		MC	*SA2*	*MC*	*SA2*	*MC*	*SA2*	*OR*
2012	Vocabulary & Syntax in Print	7	1	2	0	1	0	0
	Beginning to Read in English	2	0	0	0	0	0	0
	Comprehension	0	2	6	2	2	0	1
	Literary Elements & Techniques	0	0	0	0	2	1	0
	Informational/Expository Text	0	0	0	0	3	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4
2011	Vocabulary & Syntax in Print	5	0	3	0	4	0	0
	Beginning to Read in English	4	0	0	0	0	0	0
	Comprehension	0	3	3	1	4	1	1
	Literary Elements & Techniques	0	0	0	0	0	0	0
	Informational/Expository Text	0	0	2	1	0	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4

MC = multiple-choice, 1 point; SA2 = short-answer, 2 points; OR = open-response, 4 points
Total points possible for sessions 1 & 2: 27
Total points possible for sessions 2 & 3: 26

Table 3-10. 2011–12 MEPA: Writing Test Blueprint, Grade Span 7–8

	Outcomes Assessed	*Session 1: 15 Points 0 Reading Passages*			*Session 2: 12 Points 2 Reading Passages*			*Session 3: 14 Points 2 Reading Passages*		
		SA1	*SA2*	*SW*	*MC*	*SW*	*WP*	*MC*	*SW*	*WP*
2012	Prewriting		2	0	0	0	0	0	0	0
	Writing		0	5	0	1	1	0	2	2
	Editing		0	0	4	0	0	4	0	0
	Total Items by Type		2	5	4	1	1	4	2	2
	Total Points		4	10	4	2	4	4	4	8
2011	Prewriting	3		0	0	0	0	0	0	0
	Writing	1		5	0	1	1	0	2	2
	Editing	0		0	4	0	0	4	0	0
	Total Items by Type	4		5	4	1	1	4	2	2
	Total Points	4		10	4	2	4	4	4	8

MC = multiple-choice, 1 point; SA1 = short-answer, 1 point; SA2 = short-answer, 2 points;
SW = sentence-writing, 2 points; WP = writing-prompt, 4 points
Total points possible for sessions 1 & 2: 24
Total points possible for sessions 2 & 3: 26

Table 3-11. 2011–12 MEPA: Reading Test Blueprint, Grade Span 9–12

	Outcomes Assessed	*Session 1: 15 Points 0 Reading Passages*		*Session 2: 12 Points 2 Reading Passages*		*Session 3: 14 Points 2 Reading Passages*		
		MC	*SA2*	*MC*	*SA2*	*MC*	*SA2*	*OR*
2012	Vocabulary & Syntax in Print	7	0	1	0	2	0	0
	Beginning to Read in English	2	0	0	0	0	0	0
	Comprehension	0	3	5	2	3	1	0
	Literary Elements & Techniques	0	0	0	0	1	0	1
	Informational/Expository Text	0	0	2	0	2	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4
2011	Vocabulary & Syntax in Print	5	0	3	0	2	0	0
	Beginning to Read in English	4	0	0	0	0	0	0
	Comprehension	0	3	4	2	2	1	1
	Literary Elements & Techniques	0	0	0	0	3	0	0
	Informational/Expository Text	0	0	1	0	1	0	0
	Total Items by Type	9	3	8	2	8	1	1
	Total Points	9	6	8	4	8	2	4

MC = multiple-choice, 1 point; SA2 = short-answer, 2 points; OR = open-response, 4 points
Total points possible for sessions 1 & 2: 27
Total points possible for sessions 2 & 3: 26

Table 3-12. 2011–12 MEPA: Writing Test Blueprint, Grade Span 9–12

	Outcomes Assessed	*Session 1: 15 Points 0 Reading Passages*			*Session 2: 12 Points 2 Reading Passages*			*Session 3: 14 Points 2 Reading Passages*		
		SA1	*SA2*	*SW*	*MC*	*SW*	*WP*	*MC*	*SW*	*WP*
2012	Prewriting		0	0	0	0	0	0	0	0
	Writing		2	5	0	1	1	0	2	2
	Editing		0	0	4	0	0	4	0	0
	Total Items by Type		2	5	4	1	1	4	2	2
	Total Points		4	10	4	2	4	4	4	8
2011	Prewriting	4		0	0	0	0	0	0	0
	Writing	0		5	0	1	1	0	2	2
	Editing	0		0	4	0	0	4	0	0
	Total Items by Type	4		5	4	1	1	4	2	2
	Total Points	4		10	4	2	4	4	4	8

MC = multiple-choice, 1 point; SA1 = short-answer, 1 point; SA2 = short-answer, 2 points;
SW = sentence-writing, 2 points; WP = writing-prompt, 4 points
Total points possible for sessions 1 & 2: 24
Total points possible for sessions 2 & 3: 26

3.3.3 Field-Test Design

New field-test items were embedded in the spring 2011 matrix forms to generate common items for the 2012 administration. Eight matrix forms were tested in each of the K–2 levels, A and B. Twelve matrix forms were tested in each of the higher grade spans. 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 included items that were read aloud to the students,

the spring matrix-sampled forms were distributed so that each school administered only one of the forms. The inclusion of items that were read aloud made it impossible to spiral the forms at the individual student level, so spiraling had to be done at the school level. The number of students who took the MEPA tests in each school varied greatly. In order to ensure the greatest consistency in sample size across forms, a stratified random assignment procedure was used, with school size as the stratification variable. Further small school assignment adjustments were made based on the number of students targeted per form and the average 2010 MEPA score for the students assigned to each form.

The fall 2011 test used the same common items as the spring test, in a single form. It did not include any field-test items.

Only one common form was administered in each grade span and/or level in spring 2012; since spring 2012 was the final administration of MEPA, no field-test items were included in this administration.

3.3.4 Equating Design

An assessment program such as the MEPA, which has multiple forms and levels (i.e., sessions 1 and 2 versus sessions 2 and 3), 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 2011 within-year equating, all test items for grade spans 3–4 and above were concurrently calibrated to an IRT scale, independently for each grade span. For these grade spans, the nonmatrix items in session 2 were common to all students in the grade span. Details regarding the IRT models and the results of that analysis can be found in chapter 7.

3.3.4.2 Across-Year Equating

The operational scale for MEPA was originally established in 2009. In subsequent years, therefore, across-year equating was necessary to maintain that scale. For the 2011 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. Both the 2011 and the 2012 test forms were mostly constructed from scaled field-test items. That is, the majority of items in levels A and B for grade span K–2 and in sessions 1, 2, and 3 for other grade spans were pre-equated from the previous administrations. A small number of items—including all field-test items for the 2011 administration—were brought onto scale using the within-year equating procedure.

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 full page-width items are used, 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 through 12 use separate test booklets and answer booklets.

3.4 TEST SESSIONS

In grades 3 through 12, three sessions are available for reading and for writing. Students take two adjacent sessions in each content area, 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. Estimating how a student might perform, teachers select the assessment that is most appropriate for that student. In addition, a locator test is available to help teachers determine the appropriate sessions to administer to each student. Other relevant factors in determining which sessions to assign to a student are his or her performance on local assessments, MCAS tests, previous MEPA tests, and in the classroom. The locator tests for grades 3 through 12 include three passages and 13 to 14 multiple-choice items for reading, as well as 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 test level 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 ADMINISTRATION OF THE MEPA-R/W

The reading and writing portions of the assessment were administered in most schools as traditional paper-and-pencil tests. Some schools administered online versions of the same tests. This section contains details of test administration.

4.1.1 Responsibility for Administration

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

- Enforcing test security
- Ensuring participation of all ELL students at the appropriate grade spans
- Providing accurate student information
- Coordinating the testing schedule
- Ensuring proper test administration
- Ensuring the 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 Test Administration Window

The spring 2011 administration period for the MEPA-R/W was March 7–14. Schools that tested online were given an extended testing window through March 18. The fall 2011 tests were administered October 24–31. In the spring of 2012, the testing window was March 5–16, but schools that tested online were allowed to begin on February 27.

4.1.3 Administration Procedures

Administration procedures were explained in manuals provided to the principals and test administrators. Principals ordered test materials for their schools, designated and trained the test administrators, assigned spaces in which to administer the tests, and accounted for and returned the test materials.

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

The reading test for each grade span 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*s. In the spring 2011 administration, multiple forms were produced for each grade span. In order to accommodate the items that were read aloud by the test administrator, all students within a school took the same form of the test for their grade span. The fall 2011 and spring 2012 tests had only one form per grade span, containing only common items.

4.1.4 Participation Requirements and Documentation

All enrolled ELL students were required to participate in the spring MEPA-R/W administrations. Participation in the fall 2011 MEPA-R/W was only required for ELL students enrolled in grades 1 through 12 who did not participate in the spring 2011 test. Exceptions were made to the participation requirements for all administrations for students with medically documented absences, students who required accommodations that were not available (such as Braille), students who were deaf or hard of hearing, and students who required alternate assessments due to significant disabilities. Appendix D provides participation information for all students and for students in various demographic categories.

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 needed in order to participate. These decisions were based on the needs of the individual students consistent with the accommodations that they regularly used in the classroom. Examples of standard accommodations included scheduling 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 by which the student responded to test questions. IEP and 504 teams could also allow nonstandard accommodations if students met the specific criteria for each accommodation. Examples of nonstandard accommodations included having an administrator read the reading test aloud to the student or scribe 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. Schools that administered the tests online could enter accommodation information in the Principal's Administration System (PAS). Appendix E provides information about the numbers of students tested with and without accommodations, as well as about the numbers of students by grade span who received each type of accommodation.

4.1.6 Administrator Training

Prior to each spring administration, training workshops were held throughout the state to prepare principals or their designees to administer the MEPA tests. Principals received manuals specifically written for the overall MEPA test administration. Principals then held local meetings with test administrators to review testing security, schedules, logistics, and materials. Administrators were provided with manuals that included general testing policies and tasks, as well as specific instructions and scripts for each test session within the appropriate grade span.

4.1.7 Test Security

Test security was addressed extensively in the 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.

For schools participating in online testing, the Department provided a document containing numerous suggestions to address security; this document was posted on the Department's website and was also available online during testing. Schools were able to request security carrels from the testing contractor for use during the online administration.

All test materials were to be inventoried upon receipt, and any discrepancies reported immediately. Materials were then to be stored in secure locations until the time designated 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 manuals for principals and test administrators provided specific examples of security violations and contact information to be used in the event of any testing irregularity. Penalties for testing irregularities and security violations were listed in the manuals and included delays in reporting of test results, invalidation of test results, ineligibility of school personnel to participate in future test administrations, and possible employment and/or licensure consequences.

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 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. The service center also assisted in contacting schools to solicit participation in online testing. Following testing, the service center contacted schools to resolve discrepancies between materials shipped and returned.

During the time leading up to and during the online test administration, the MEPA technical service center was also available. The technical service center representatives were trained on the online system and were able to provide technical assistance and guidance for schools.

4.2 ADMINISTRATION OF THE MELA-O

The 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 were evaluated: fluency, grammar, pronunciation, and vocabulary.

4.2.1 Responsibility for Administration

Students were assessed by trained and qualified MELA-O trainers (QMTs) or administrators (QMAs) in the schools. Each district was responsible for ensuring that staff within the district had been trained to administer this assessment.

4.2.2 Test Administration Window

The spring 2011 administration of the MELA-O took place February 14–March 14. The fall 2011 MELA-O was administered October 3–31. In the spring of 2012, the MELA-O was administered February 13–March 16.

4.2.3 Administration Procedures

Students were observed in a classroom setting by a QMT or QMA as they engaged in normal academic interactions with the teacher and with other students. In order to attain an adequate sample of the students' language skills, students may have been observed on multiple occasions during the administration window. Students were rated (i.e., scored) on a 0–5 MELA-O scoring matrix in each category and subdomain. Scores were reported on the student's answer booklets or in the PAS for those schools testing online.

4.2.4 Participation Requirements and Documentation

ELL students in kindergarten through grade 12 were required to participate in the spring administrations of the MELA-O listening and speaking test. The fall 2011 assessment was required for ELL students in grades 1 through 12 who did not participate in spring 2011. A very small number of students were not required to participate because they had extended medically documented absences from school during the testing window or they had IEPs identifying them as deaf or hard of hearing.

4.2.5 Administrator Training

To qualify as MELA-O administrators or MELA-O trainers (who may administer the test and also train new administrators), individuals were required to participate in a training session and pass a qualifying test. The qualifying test consisted of a video showing clips of students in a classroom setting. The training participants used 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 were checked. A score assigned by a trainee that was more than two score points from the correct score was deemed discrepant. The minimum scores needed to qualify were:

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

The Department conducted two training sessions for new MELA-O trainers in 2011. Because spring 2012 was the last administration of the MEPA program, no additional QMT trainings were held after November 2011.

4.2.6 Service Center

As was the case for 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 SCORING OF THE MEPA-R/W

Once received after testing by the testing contractor, Measured Progress, each 2011 and 2012 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. Digitized student responses to open-response, short-answer, sentence-writing, and writing-prompt test items were sorted, by grade span, into item-specific groups.

5.1.1 Machine-Scored Items

Multiple-choice items were used in all sessions of the reading and writing tests. Student responses to these items were machine scored by applying a scoring key to the captured responses. Correct answers were assigned a score of 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 by readers, 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 are all based in Dover, New Hampshire, MEPA-R/W responses were also scored in three other Measured Progress scoring locations. Table 5-1 shows the distribution of grade-span testing across all four scoring locations.

Table 5-1. 2011–12 MEPA: Scoring Center Locations

Grade Span	*Subject*	*Spring 2012*	*Fall 2011*	*Spring 2011*
K–2	Writing	Dover	Dover	Dover
3–4	Reading	Menands	Louisville	Louisville
	Writing	Menands	Louisville	Louisville
5–6	Reading	Menands	Louisville	Louisville
	Writing	Menands	Louisville	Louisville
7–8	Reading	Menands	Louisville	Louisville
	Writing	Menands	Louisville	Louisville
9–12	Reading	Menands	Louisville	Louisville
	Writing	Menands	Louisville	Louisville

The following staff members were involved with scoring the 2011 and 2012 MEPA-R/W responses:

- The MEPA-R/W scoring manager, located in Dover, oversaw communication and coordination of scoring among the scoring sites.
- The iScore operations manager, located in Dover, coordinated technical communication among the scoring sites.
- The 2011 scoring site manager in Louisville, Kentucky, and the 2012 scoring site manager in Menands, New York, provided all on-site logistical coordination.
- A chief reader in writing and a chief reader in reading at each scoring location ensured consistency of benchmarking and scoring across all grade spans. Chief readers monitored and read behind both on-site and off-site quality assurance coordinators (QACs).
- Several QACs, selected from a pool of experienced senior readers, participated in benchmarking, training, scoring, and cleanup activities for specified content areas and grade spans. QACs 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 4 to 11 readers.

5.1.2.2 2011 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 in spring 2011. There were no field-tested items and therefore no benchmarking meetings in 2012. All spring 2011 benchmarking meeting results were recorded and considered final upon Department signoff.

The primary goals of the 2011 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 a particular item are 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 responses 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

2011 and 2012 MEPA-R/W readers were obtained primarily 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 reader was eliminated from the applicant pool. Teachers, tutors, and administrators (principals, guidance counselors, etc.) currently under contract with or employed by or in Massachusetts schools, and people under 18 years of age, were ineligible to score MEPA-R/W responses.

Table 5-2 is a summary of MEPA-R/W reader background across all 2012 scoring shifts at all locations.

Table 5-2. 2012 MEPA-R/W: Summary of Reader Background Across Scoring Shifts and Scoring Locations

	Background	*Number*	*Percent*
Education	Less than 48 college credits	0	0.0
	Associate degree/More than 48 college credits	5	2.6
	Bachelor's degree	81	42.2
	Master's/Doctorate	106	55.2
Teaching Experience	No teaching certificate or experience	76	39.6
	Teaching certificate or experience	89	46.4
	College Instructor	27	14.1
Scoring Experience	No previous experience as reader	44	22.9
	1–3 years experience	82	42.7
	3+ years experience	66	34.4

Table 5-3 is a summary of MEPA-R/W reader background across all 2011 scoring shifts at all locations.

Table 5-3. 2011 MEPA-R/W: Summary of Reader Background Across Scoring Shifts and Scoring Locations

	Background	*Number*	*Percent*
Education	Less than 48 college credits	0	0.0
	Associate degree/More than 48 college credits	50	14.7
	Bachelor's degree	184	54.1
	Master's/Doctorate	106	31.2
Teaching Experience	No teaching certificate or experience	174	51.2
	Teaching certificate or experience	152	44.7
	College Instructor	14	4.1
Scoring Experience	No previous experience as reader	77	22.7
	1–3 years experience	166	48.8
	3+ years experience	97	28.5

5.1.2.4 Methodology for Scoring Polytomous Items

The 2011 and 2012 MEPA-R/W contained polytomous items (including some short-answer items), for which scores of 0–2 were assigned, that required students to generate a brief response. They also contained open-response items, for which scores of 0–3 or 0–4 were assigned, that required longer or more complex responses.

In addition to assigning a score point between 0 and 4, depending on the item, readers could optionally 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 5-4 presents a K–2 level B writing open-response scoring guide, one of the many different MEPA-R/W scoring guides used in 2011 and 2012. 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 5-4. 2011–12 MEPA R/W: 3-Point Open-Response Item Scoring Guide—Writing, Grade Span 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 have been flagged as a "crisis" if it indicated

- 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; or
- 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 case each response was scored only once, or double-blind scored, in which case each response was independently read and scored by two separate readers. For each 2011 and 2012 MEPA-R/W item, at least 10% of the responses were randomly double-blind scored; neither reader knew the response 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 QAC.

Above and beyond the 10% 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 5-5 and 5-6 provide examples of how the resolution rules were applied when the two read-behind or two double-blind scores were not identical (i.e., were adjacent or discrepant).

Table 5-5. 2011–12 MEPA R/W: Example 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 5-6. 2011–12 MEPA R/W: Example 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 discussion of 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 all 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 QACs 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 training was delivered from a remote location; that is, the chief reader communicated 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 through on-site 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 with no 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.

Some of the MEPA-R/W examinees took the computer-based test, and these examinees typed their responses into the computer rather than writing them by hand in a test booklet or answer document. In addition to participating in discussions on similarities and differences between handwritten and

typed responses, readers who scored the typed responses were given several typed sample responses along with the same anchor, training, ranking, and qualifying sets given to readers of handwritten responses.

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% exact and 90% exact and adjacent agreement (at least 7 out of the 10 scores were exact matches and at least 9 of 10 were either 0 or 1 point discrepant) on either of two potential qualifying sets. Scoring leadership (QACs and senior readers) had to meet or surpass the higher qualification standard of at least 80% exact and 90% 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% exact and 90% 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 that could range from counseling to retraining to dismissal. If a reader did not qualify for or was dismissed from scoring any two MEPA items within a grade span, the reader was not allowed to score any additional items within that grade span. If a reader was dismissed from two different grade spans within one scoring session, the reader was dismissed from the project, and he or 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 (such as additional read-behinds) and allowed to begin scoring. Readers who had zero or one out of the five exact scores were not retrained and were not allowed to score that item on that day.

Embedded CRRs were responses approved by the chief reader and loaded into iScore for blind distribution to readers at random points during the scoring of their first 200 operational responses. While the number of embedded 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 (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% exact and 90% exact and adjacent score match standard were counseled, given extra monitoring such as additional read-behinds, and allowed to resume scoring.

Double-blind readings involved responses scored independently by two different readers. Readers knew 10% 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 such as additional read-behinds, and were 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—five 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 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% exact and 90% 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 SCORING OF THE 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 scores to students in 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 indicating the approximate proficiency of a native speaker.

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. Schools testing online were able to provide MELA-O scores in the PAS.

5.2.3 Weight of MELA-O Scores in Student Performance Level

Each student's MELA-O score was incorporated into the student's overall score along with his or her MEPA-R/W score. A natural weighting was used in combining MELA-O scores with MEPA-R/W by totaling the possible points for each component (5 for listening and 20 for speaking) and combining this with the score point total of the reading and writing tests.

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 sections 7.1 and 7.2.

Figure 5-1. 2011–12 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
QMT Training Manual (Revised July 2008)
56

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, 2004) 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 for the combined spring and fall 2011 MEPA administrations (which used the same test form) as well as for the spring 2012 administration. The item-level classical statistics, including difficulty and discrimination indices, are presented in more detail in Appendix F.

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 and only a handful answered at near-perfect rates. Similarly, the positive discrimination indices indicate that most items were assessing closely related constructs, and students who performed well on individual items tended to perform well overall. There were a small number of items with low discrimination indices, but none was near-zero or 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 the 2011 and 2012 MEPA administrations.

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 somewhat 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. 2011–12 MEPA: Summary of Item Difficulty and Discrimination Statistics by Grade Span and Session—Spring 2012

Grade Span	*Session*	*Item Type*	*Number of Items*	*P-Value*		*Discrimination*	
				Mean	*Standard Deviation*	*Mean*	*Standard Deviation*
K–2	A	ALL	29	0.62	0.17	0.44	0.18
		MC	18	0.64	0.17	0.33	0.10
		OR	11	0.59	0.18	0.62	0.13
	B	ALL	31	0.75	0.10	0.45	0.12
		MC	20	0.76	0.11	0.38	0.07
		OR	11	0.74	0.07	0.58	0.07
3–4	1&2	ALL	40	0.73	0.12	0.53	0.15
		MC	21	0.77	0.12	0.44	0.10
		OR	19	0.70	0.11	0.63	0.12
	2&3	ALL	39	0.73	0.16	0.43	0.18
		MC	24	0.77	0.16	0.35	0.11
		OR	15	0.67	0.16	0.56	0.19
5–6	1&2	ALL	40	0.67	0.13	0.49	0.19
		MC	21	0.70	0.12	0.35	0.09
		OR	19	0.65	0.14	0.64	0.14
	2&3	ALL	39	0.69	0.15	0.40	0.20
		MC	24	0.70	0.10	0.31	0.09
		OR	15	0.66	0.20	0.55	0.23
7–8	1&2	ALL	40	0.65	0.13	0.51	0.17
		MC	21	0.68	0.11	0.39	0.09
		OR	19	0.62	0.16	0.64	0.15
	2&3	ALL	39	0.64	0.14	0.41	0.21
		MC	24	0.64	0.13	0.30	0.11
		OR	15	0.63	0.16	0.59	0.21
9–12	1&2	ALL	40	0.60	0.12	0.48	0.18
		MC	21	0.61	0.12	0.36	0.08
		OR	19	0.59	0.11	0.61	0.16
	2&3	ALL	39	0.63	0.11	0.41	0.20
		MC	24	0.64	0.10	0.31	0.11
		OR	15	0.62	0.12	0.56	0.22

Table 6-2. 2011–12 MEPA: Summary of Item Difficulty and Discrimination Statistics by Grade Span and Session—Spring & Fall 2011

Grade Span	Session	Item Type	Number of Items	P-Value		Discrimination	
				Mean	Standard Deviation	Mean	Standard Deviation
K–2	A	ALL	29	0.60	0.16	0.47	0.15
		MC	16	0.64	0.15	0.37	0.06
		OR	13	0.56	0.16	0.59	0.15
	B	ALL	31	0.76	0.14	0.44	0.12
		MC	16	0.79	0.18	0.36	0.05
		OR	15	0.73	0.08	0.52	0.12
3–4	1&2	ALL	42	0.72	0.11	0.53	0.16
		MC	21	0.75	0.10	0.42	0.09
		OR	21	0.69	0.12	0.64	0.12
	2&3	ALL	39	0.69	0.15	0.43	0.18
		MC	24	0.73	0.12	0.36	0.09
		OR	15	0.62	0.18	0.55	0.22
5–6	1&2	ALL	42	0.72	0.12	0.52	0.16
		MC	21	0.73	0.11	0.43	0.08
		OR	21	0.71	0.13	0.61	0.18
	2&3	ALL	39	0.70	0.13	0.45	0.19
		MC	24	0.73	0.09	0.37	0.10
		OR	15	0.65	0.16	0.58	0.22
7–8	1&2	ALL	42	0.65	0.12	0.52	0.16
		MC	21	0.65	0.11	0.41	0.07
		OR	21	0.64	0.12	0.63	0.16
	2&3	ALL	39	0.65	0.13	0.43	0.20
		MC	24	0.65	0.13	0.33	0.10
		OR	15	0.66	0.13	0.60	0.22
9–12	1&2	ALL	42	0.60	0.12	0.45	0.18
		MC	21	0.63	0.12	0.33	0.08
		OR	21	0.57	0.11	0.58	0.17
	2&3	ALL	39	0.62	0.11	0.41	0.20
		MC	24	0.64	0.09	0.29	0.08
		OR	15	0.59	0.14	0.59	0.20

6.2 DIFFERENTIAL ITEM FUNCTIONING

Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 2004) 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-R/W items and MELA-O indicators 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 six MEPA subgroups: male versus female, white versus African American/black, white versus Hispanic or Latino, white versus Asian, not low income versus low income, and no disability versus disability. A minimum of 200 matched students were required for these calculations. 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 students in focal groups. 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-R/W item and MELA-O indicator was categorized according to the guidelines adopted 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 combined spring 2011 and fall 2011 administrations as well as for the spring 2012 administration.

Appendix G presents the number of items classified as either "low" or "high" DIF, overall and by focal group. Results are shown separately for multiple-choice (MC) versus open-response (OR) items. Overall, relatively few items exhibited either low or high DIF and the numbers were fairly consistent with results obtained for previous administrations of the test.

6.3 DIMENSIONALITY ANALYSIS RESULTS FOR SPRING 2011 AND SPRING 2012

The DIF analyses of the previous section were performed to identify items that showed evidence of differences in performance between pairs of subgroups beyond that which 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 an additional dimension(s) 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, (a) the degree to which unidimensionality is

violated and (b) the nature of the multidimensionality. Dimensionality analyses were performed on the MEPA common items and indicators for grade-spans K–2, 3–4, 5–6, 7–8, and 9–12 for the spring 2011 and spring 2012 administrations. (Note: Only common items were analyzed since they are used for score reporting.) The results for these analyses are reported below. For K–2, separate analyses were conducted on the two total test forms associated with that grade span, which are delineated as follows:

Form A: The combination of reading session A, writing session A, and MELA-O.

Form B: The combination of reading session B, writing session 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 test. 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 analyses were limited to those two forms, which are delineated as follows:

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.

In addition to analysis of these 10 test forms (two per grade span across five grade spans) for each administration, the same 10 forms were also analyzed without including the MELA-O items.

The dimensionality analyses were conducted using the nonparametric IRT-based methods DIMTEST (Stout, 1987; Stout, Froelich, & Gao, 2001) and DETECT (Zhang & Stout, 1999). Both of these methods 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, nonrandom 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. Then 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 nonclustered 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 independently 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: within-cluster conditional covariances are summed; from this sum the between-cluster conditional covariances are subtracted; this difference is divided by the total number of item pairs; and this average is multiplied by 100 to yield an index of the average violation of local independence for an item pair. 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 MEPA 2011–12 spring assessments for grade spans K–2, 3–4, 5–6, 7–8, and 9–12. The data for each test form analyzed (see above) 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 Tables 6.3 and 6.4 below. The smallest sample size was about 1500, which occurred for the spring 2011 administration of Form 1–2 for grade span 7–8, resulting in about 750 examinees for the training and cross-validation samples. All the other grade spans had sample sizes of at least 2000. DIMTEST simulation studies have indicated 99% 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 dataset to see if the null hypothesis of unidimensionality would be rejected. DETECT was then applied to each dataset 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 Forms (Reading, Writing, and 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 dataset at level 0.01. Because strict unidimensionality is an idealization that almost never holds exactly for a given dataset, the statistical rejections in the DIMTEST results were not necessarily surprising. A 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.

Next, 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. Tables 6.3 and 6.4 below display the DETECT effect size estimates for spring 2011 and spring 2012, respectively. The results for the full test forms are in the columns labeled "R, W, MELA-O."

Seven of the 10 DETECT values for spring 2011 and nine of 10 for spring 2012 indicated moderate to strong multidimensionality. Indeed, the results for Form A for grade span K–2 indicated strong multidimensionality for both spring 2011 and spring 2012. A closer investigation of both the DIMTEST and DETECT results indicated that the multidimensionality in every case was predominantly caused by MELA-O measuring a construct that is different from reading or writing, while reading and writing displayed much less difference between each other. For comparison purposes, Tables 6.3 and 6.4 also include the corresponding DETECT values for the preceding year (thus, the 2011 results are presented in both tables). The results appear to be consistent across all three years presented in the tables, 2010 to 2012. In particular, all the results indicated the same pattern of the MELA-O items being the predominant cause of the multidimensionality.

Table 6-3. 2011–12 MEPA: Multidimensionality Effect Sizes by Grade Span—Spring 2011

Grade Span	*Form*	*Sample Size (Rounded)*	*Multidimensionality Effect Size*			
			R, W, MELA-O		*R & W Only*	
			2011	*2010*	*2011*	*2010*
K–2	A	14,900	0.96	0.96	0.35	0.47
	B	9,900	0.55	0.58	0.23	0.29
3–4	1-2	3,300	0.41	0.37	0.18	0.18
	2-3	7,500	0.27	0.28	0.17	0.17
5–6	1-2	2,000	0.39	0.34	0.10	0.15
	2-3	6,200	0.36	0.34	0.18	0.14
7–8	1-2	1,500	0.22	0.27	0.16	0.16
	2-3	4,200	0.42	0.40	0.14	0.22
9–12	1-2	3,000	0.24	0.34	0.17	0.18
	2-3	6,500	0.48	0.42	0.23	0.18

Table 6-4. 2011–12 MEPA: Multidimensionality Effect Sizes by Grade Span—Spring 2012

Grade Span	*Form*	*Sample Size (Rounded)*	*Multidimensionality Effect Size*			
			R, W, MELA-O		*R & W Only*	
			2012	*2011*	*2012*	*2011*
K–2	A	15,800	0.91	0.96	0.33	0.35
	B	9,600	0.50	0.55	0.23	0.23
3–4	1-2	3,200	0.41	0.41	0.21	0.18
	2-3	8,100	0.24	0.27	0.13	0.17
5–6	1-2	2,000	0.31	0.39	0.13	0.10
	2-3	6,400	0.30	0.36	0.17	0.18
7–8	1-2	1,600	0.30	0.22	0.14	0.16
	2-3	4,300	0.45	0.42	0.13	0.14
9–12	1-2	2,900	0.37	0.24	0.19	0.17
	2-3	7,000	0.46	0.48	0.21	0.23

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 different from each other, a follow-up analysis focusing on only the reading and writing subtests was conducted to provide a more accurate picture of the reading and writing dimensionality structure. DIMTEST again rejected the null hypothesis of unidimensionality for every dataset at level 0.01.

Next, 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. Tables 6.3 and 6.4 display the DETECT effect size estimates for these analyses under the columns labeled "R & W Only." These results clearly confirm that the constructs measured by the reading and writing tests are much more similar to each other than to the construct measured by the MELA-O test. Moreover, except for K–2 Form A, the multidimensionality was either very weak (less than 0.2) or weak (less than 0.3). Comparing across the three years of results presented in the tables, the same pattern of results is seen to have occurred over all three years; that is, the multidimensionality was much weaker for reading and writing alone in comparison to the analyses that included MELA-O.

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 "R & W Only" DETECT results. The investigation of the DETECT clusters revealed that for all 20 datasets there existed item clusters that were dominated by either reading or writing items. In some cases, such dominant clusters were the only clusters found by DETECT, while in other cases there were other clusters that had a nearly even mix of reading and writing items. These results correspond closely with results reported in previous years.

6.3.3 Summary

Overall, the results indicate that significant levels of multidimensionality exist in the MEPA tests, and that MELA-O is the primary cause of that multidimensionality. Moreover, the multi-dimensionality tends to be larger for the K–2 grade span, especially for Form 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 strong evidence of reading and writing measuring different constructs, but what differences there are appear to be negligibly weak, as reflected in the DETECT effect size values. The results clearly indicate that the special procedures that were implemented with regard to the inclusion of the MELA-O items in the IRT calibrations were fully justified and were necessary in order to control for the dimensionality differences due to those items. Similarly, the use of the more conservative 1PL and PCM IRT models with the K–2 grade span (see section 7.1) are also supported by the dimensionality analysis results. Indeed, dimensionality analyses that were conducted on field-test data collected in the fall of 2008 had already given a preview of the differing multi-dimensionality associated with K–2, and this information was used in guiding the modeling decisions for the current MEPA program. The dimensionality analysis results for spring 2011 and spring 2012, like those for spring 2009 and spring 2010, continue to confirm those earlier results and the resulting modeling decisions. Current and prior MEPA results and technical reports are available on the Department's Web site at www.doe.mass.edu/mcas/mepa/results.html

In summary, the results of the dimensionality analyses support the continued use of the current test construction, modeling, and calibration procedures that have been developed for the MEPA program. In particular, there should be continued use of 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 answering a dichotomous item correctly 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). Model selection originally occurred for the 2009 administration. At that time, 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. (See the *MEPA 2009 Technical Report* for a complete description of the processes followed to determine the best models to use for MEPA.)

The generalized form of the PCM can be defined as

$$P_{ijk}\left(k\middle|\theta_i,\zeta_j\right)=\frac{\exp\sum_{v=0}^{k}\left[Da_j\left(\theta_i-b_j+d_v\right)\right]}{\sum_{c=1}^{m}\exp\sum_{v=1}^{c}\left[Da_j\left(\theta_i-b_j+d_v\right)\right]} \quad (7.1)$$

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:

$$P_j\left(1\middle|\theta_i,b_j\right)=\frac{\exp\left[D\left(\theta_i-b_j\right)\right]}{1+\exp\left[D\left(\theta_i-b_j\right)\right]} \quad (7.2)$$

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:

$$P_i\left(1\middle|\theta_j,\xi_i\right)=\frac{\exp\left[Da_i\left(\theta_j-b_i\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i\right)\right]} \tag{7.3}$$

where
i indexes the items,
j indexes students,
α 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, as defined below, 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:

$$P_{ik}^{*}\left(1\middle|\theta_j,\xi_i\right)=\frac{\exp\left[Da_i\left(\theta_j-b_i+d_{ik}\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_{ik}\right)\right]} \tag{7.4}$$

where
ξ_i represents the set of item parameters for item I,
i indexes the items,
j indexes students,
k indexes threshold,
α 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:

$$P_{ik}\left(1\middle|\theta_j\right)=P_{i(k-1)}^{*}\left(1\middle|\theta_j\right)-P_{ik}^{*}\left(1\middle|\theta_j\right) \tag{7.5}$$

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:

$$P_{ik}\left(k\middle|\theta_j,\xi_i\right)=\frac{\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_k\right)\right]}-\frac{\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]}{1+\exp\left[Da_i\left(\theta_j-b_i+d_{k+1}\right)\right]} \tag{7.6}$$

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 nonlinear, monotonically increasing relationship between θ and p. Once the item parameters are known, $\hat{\theta}$ 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 & Novick (1968), Hambleton & Swaminathan (1985), or Baker & Kim (2004).

7.2 ITEM RESPONSE THEORY RESULTS

The number of Newton cycles required for convergence for each grade span during the IRT analysis can be found in Table 7-1. The number of cycles required fell within acceptable ranges.

Table 7-1. 2011–12 MEPA: Number of Newton Cycles Required for Convergence

Grade Span	*Cycles*	
	Spring 2012	*Spring & Fall 2011*
K–2, A	21	18
K–2, B	46	46
3–4	69	89
5–6	142	97
7–8	135	32
9–12	121	109

The tables in Appendix H give the IRT item parameter calibration results of all items in the spring 2011 administration by grade span.

The Test Characteristic Curves (TCCs) in Appendix I 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:

$$E\left(X\middle|\theta_j\right) = \sum_{i=1}^{n} P_i\left(1\middle|\theta_j\right) \tag{7.7}$$

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
$E\left(X\middle|\theta_j\right)$ 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 years (2010/2011 and 2011/2012) are presented in Appendix I for each session (A versus B for grade span K–2; reading sessions 1 and 2 and writing sessions 1 and 2 versus reading sessions 2 and 3 and writing sessions 2 and 3 for the remaining grade spans). Also shown are graphics of the corresponding 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. Note that, because of the use of the one-parameter model, TIFs are not provided for grade span K–2. For detailed information about TIFs, see the references provided in section 7.1.

7.3 EQUATING

The purpose of equating is to ensure that scores obtained from different forms of a test are equivalent to each other. Equating may be used to ensure equivalency of multiple test forms administered in the same year, as well as to equate one year's forms to those used in the previous year. Equating ensures that students are not given an unfair advantage or disadvantage because the test form they took is easier or harder than forms taken by other students.

The 2011 and 2012 administrations of the MEPA used a raw-score-to-theta equating procedure in which test forms were equated to the theta scale established on the reference form (i.e., the form used in the most recent standard setting). This is accomplished through the chained linking design, in which every new form is equated back to the theta scale of the previous year's test form. It can therefore be assumed that the theta scale of every new test form is the same as the theta scale of the reference form, since this is where the chain originated.

The groups of students who took the equating items on the 2011 and 2012 MEPA tests are not equivalent to the groups who took them in the reference year. IRT is particularly useful for equating scenarios that involve nonequivalent groups (Allen & Yen, 1979). Equating for the MEPA uses the anchor-test-nonequivalent-groups design described by Petersen, Kolen, & Hoover (1989). In this equating design, no assumption is made about the equivalence of the examinee groups taking different test forms (that is, naturally occurring groups are assumed). Comparability is instead evaluated by utilizing a set of anchor items (also called equating items). However, the equating items are designed to mirror the common test in terms of item types and distribution of emphasis.

MEPA tests are pre-equated. Item parameter estimates for 2011 and 2012 were placed on the same scale as previous years by using the Fixed Common Item Parameter method (FCIP2) (Kim, 2006), which is based on the IRT principle of item parameter invariance. According to this principle, the equating items for the 2011 and 2012 MEPA tests should have the same item parameters as when these items were used in previous administrations. After the item parameters for each test were estimated using PARSCALE (Muraki & Bock, 2003) to check for parameter drift of the equating items, the FCIP2 method was employed to place the non-equating items onto the operational scale. This method is performed by fixing the parameters of the equating items to their previously obtained on-scale values, and then calibrating using PARSCALE to place the remaining items on scale.

7.4 EQUATING RESULTS

Prior to equating the 2011 and 2012 tests, a variety of evaluations of the equating items were conducted. Items that were flagged as a result of these evaluations are listed in Tables 7-2 and 7-3. These items were scrutinized and a decision was made as to whether to include the item as an equating item or to discard it. The procedures used to evaluate the equating items are described below.

Table 7-2. 2011–12 MEPA: Items That Required Intervention During IRT Calibration and Equating—Spring 2012

IREF	*Subject*	*Grade*	*Reason*	*Action*
185607W	ELA	K-2, A	IRT Plot Outlier	Removed from equating
185611W	ELA	K-2, B	Delta Analysis	Removed from equating
191438R	ELA	3-4	Delta Analysis	Removed from equating
191503R	ELA	3-4	Delta Analysis	Removed from equating
189319W	ELA	5-6	IRT Plot Outlier	Removed from equating
194103R	ELA	7-8	Delta Analysis	Removed from equating
194103R	ELA	7-8	IRT Plot Outlier	Removed from equating
194163W	ELA	7-8	IRT Plot Outlier	Removed from equating
10003V	ELA	7-8	EQ Removed Other	Removed from equating
10003V	ELA	7-8	a parameter	a set to initial
185076R	ELA	9-12	Delta Analysis	Removed from equating
192536W	ELA	9-12	IRT Plot Outlier	Removed from equating
192568W	ELA	9-12	Delta Analysis	Removed from equating

Table 7-3. 2011–12 MEPA: Items That Required Intervention During IRT Calibration and Equating—Spring & Fall 2011

IREF	*Subject*	*Grade*	*Reason*	*Action*
160469W	MEPA	K-2, A	Delta Analysis	Removed from equating
161084W	MEPA	K-2, B	Delta Analysis	Removed from equating
160641W	MEPA	3-4	a parameter	a set to initial
160723R	MEPA	3-4	a parameter	a set to initial
136231W	MEPA	5-6	Delta Analysis	Removed from equating
10002F	MEPA	7-8	a parameter	a set to initial

Appendix J presents the results from the delta analyses. The analysis was used to evaluate adequacy of equating items; the discard status presented in Appendix J indicates whether or not the item was flagged as potentially inappropriate for use in equating.

Also presented in Appendix J are the results from the rescore analyses. With this analysis, 200 random papers from 2010 were interspersed with papers from 2011 to evaluate scorer consistency from one year to the next, and the same procedure was conducted to evaluate consistency across 2011 and 2012. All effect sizes were well below the criterion value of 0.80 for excluding an item as an equating item.

Finally, *a*-plots and *b*-plots, which show IRT parameters for 2011 and 2012 plotted against the values for 2010 and 2011, respectively, are presented in Appendix K. Any items that appeared as outliers in the plots were evaluated in terms of suitability for use as equating items.

Once all flagged items had been evaluated and appropriate action taken, the FCIP2 method of equating was used to place the item parameters onto the previous year's scale, as described above.

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 (for complete details of the standard-setting meeting, see the standard-setting report available at www.mcasservicecenter.com/McasDefault.asp?ProgramID=14.). MEPA scaled score cutpoints are presented in Table 7-4. 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 7-4. 2011–12 MEPA: Scaled Score Cutpoints by Performance Level

Grade Span	Theta				Scaled Score					
	Cut 1	Cut 2	Cut 3	Cut 4	Minimum	Cut 1	Cut 2	Cut 3	Cut 4	Maximum
K–2	-1.3232	-0.6903	0.2859	1.0168	400	453	466	485	500	550
3–4	-2.4010	-1.4192	-0.3200	0.9850	400	432	452	474	500	550
5–6	-2.2342	-1.2255	-0.0935	0.9600	400	436	456	479	500	550
7–8	-1.8735	-0.8271	0.2607	0.9800	400	443	464	486	500	550
9–12	-1.5092	-0.8047	0.4144	0.9700	400	450	464	489	500	550

7.5.2 Calculations

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

$$SS = m \cdot \hat{\theta} + b \tag{7.8}$$

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:

$$m = \frac{500-400}{\theta_4 — 4.0} \tag{7.9}$$

$$b = SS_1 - m\theta_1 \tag{7.10}$$

where
500 and 400 are the scaled scores for the *Level 4/Level 5* cut 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 7-5.

Table 7-5. 2011–12 MEPA: Transformation Constants for Composite MEPA Scores—2011 & 2012

Grade Span	*Slope*	*Intercept*
K–2	19.93303	479.7321
3–4	20.06018	480.2407
5–6	20.16129	480.6452
7–8	20.08032	480.3213
9–12	20.12072	480.4829

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 L provides tables showing each raw score with its corresponding theta, overall scaled score, performance level, and conditional standard error of measurement (SEM). Results are shown for all session combinations for all grade spans.

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 (SSR) was calculated as follows:

$$SS_R = m\hat{\theta}_R + b \tag{7.11}$$

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

The slope and intercept were calculated as follows:

$$m = \frac{SS_{\max} - SS_{\min}}{\theta_{\max} - \theta_{\min}} = \frac{30-0}{4.0-(-4.0)} = \frac{30}{8} = 3.75 \quad (7.12)$$

$$b = SS_{\min} - m\theta_{\min} = 0 - 3.75(-4.0) = 15 \quad (7.13)$$

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 L.

7.5.3 Distributions

Graphs showing the composite scaled score distributions and tables showing performance-level distributions are displayed in Appendices M and N, respectively, for each grade span for the spring 2011, fall 2011, and spring 2012 MEPA administrations. For each administration, the results are compared with results from previous years.

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 limited 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:

$$\alpha \equiv \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} \sigma^2_{(Y_i)}}{\sigma^2_x}\right] \quad (8.1)$$

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

Table O-1 of Appendix O presents descriptive statistics, Cronbach's α coefficient, and raw score standard errors of measurement (SEMs) for each MEPA grade span for the spring 2012 administration as well as for the combined fall and spring 2011 administrations.

Error bars representing the SEM of each test were included on the MEPA *Parent/Guardian Report*. The SEMs were computed at the raw score level and converted to scaled scores (see section 7.5 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 2011 and 2012 MEPA tests. Tables O-2 (for the spring 2012 tests) and O-5 (for the fall and spring 2011 tests) of Appendix O present reliabilities for various subgroups of interest (by gender, race, income level, and disability status) required for accountability 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 O-3 and O-6 (corresponding to the 2012 and 2011 administrations, respectively) of Appendix O. 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

Chapter 5 of this report describes in detail the processes that were implemented to monitor the quality of the hand-scoring of student responses for open-response items. Results of these analyses were used during the scoring process to identify scorers that required retraining or other intervention. A summary of the interrater consistency results is presented in Tables P-1 and P-3 (corresponding to the 2012 and 2011 administrations, respectively) of Appendix P. Results in the table are collapsed across the hand-scored items by subject and grade span. The table shows the number of score categories, the number of included scores, the percent exact agreement, percent adjacent agreement, correlation between the first two sets of scores, and the percent of responses that required a third score. This same information is provided at the item level in Tables P-2 and P-4. 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

While related to reliability, the accuracy and consistency of classifying students into achievement categories are even more important statistics in a standards-based reporting framework (Livingston

& Lewis, 1995). After the achievement levels were specified and students were classified into those levels, decision accuracy and consistency (DAC) analyses were conducted to determine the statistical accuracy and consistency of the classifications. For the MEPA, students are classified into one of five achievement levels. This section of the report explains the methodologies used to assess the reliability of classification decisions, and results are given.

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 and parallel forms of the test are given to the same group of students. In operational test programs, however, such a design is usually impractical. Instead, techniques have been developed to estimate both the accuracy and consistency of classification decisions based on a single administration of a test. The Livingston and Lewis (1995) technique was used for the 2011 and 2012 MEPA administrations because it is easily adaptable to all types of testing formats, including mixed format tests.

The DAC results reported in Appendix Q make use of "true scores" in the classical test theory sense. A true score is the score that would be obtained if a test had no measurement error. Of course, true scores cannot be observed and so must be estimated.

In the Livingston and Lewis method, 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, 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). Note that the lowest performance level did not have enough students to provide meaningful results. Thus, the two lowest levels (*level 1* and *level 2*) are merged in the calculations.

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

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

$$\kappa = \frac{(\text{Observied agreement}) - (\text{Chance agreement})}{1 - (\text{Chance agreement})} = \frac{\sum_i C_{ii} - \sum_i C_{i.} C_{.i}}{1 - \sum_i C_{i.} C_{.i}} \quad (8.2)$$

where

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

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

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

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

8.5.2 Decision Accuracy and Consistency Results

The DAC analyses described above are provided in Tables Q-1 (spring 2012), Q-3 (fall 2011), and Q-5 (spring 2012) of Appendix Q. The tables include overall accuracy and consistency indices, including kappa. Accuracy and consistency values conditional upon achievement level are also given. For these calculations, the denominator is the proportion of students associated with a given achievement level. For example, the conditional accuracy value is 0.76 for *level 3* for grade span K–2 for the spring 2012 administration. This figure indicates that among the students whose true scores placed them in this classification, 76% would be expected to be in this classification when categorized according to their observed scores. Similarly, a consistency value of 0.69 indicates that 69% of students with observed scores in *level 3* would be expected to score in this classification again if a second, parallel test form were used.

For some testing situations, the greatest concern may be decisions around level thresholds. For example, for MEPA, the accuracy of the *level 4/level 5* threshold may be of greatest interest, since students scoring at *level 5* may be considered for transition out of LEP status. For the spring 2012, fall 2011, and spring 2011 MEPA administrations, Tables Q-2, Q-4, and Q-6, respectively, in Appendix Q provide accuracy and consistency estimates at each cutpoint, as well as false positive and false negative decision rates. (A false positive is the proportion of students whose observed scores were above the cut and whose true scores were below the cut. A false negative is the proportion of students whose observed scores were below the cut and whose true scores were above the cut.)

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

Note that, as with other methods of evaluating reliability, DAC statistics based on small groups can be expected to be lower than those based on larger groups. For this reason, the values presented in Appendix Q should be interpreted with caution.

CHAPTER 9. REPORTING OF RESULTS

9.1 UNIQUE REPORTING NOTES

Results for the spring 2011 and spring 2012 MEPA administrations 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*
- *Progress Report*
- *Parent/Guardian Report*

Each report is briefly described in this chapter; samples of the report shells are provided in Appendix R. MEPA tests are intended to measure students' performance and progress in acquiring proficiency in English. The results of the fall MEPA administration are used solely to determine baseline scores for new students and students who did not test the previous spring. Following the fall 2011 MEPA administration, only the *Roster of Student Results* report was generated and provided to schools and districts. Complete MEPA results for both the fall and spring 2011 administrations were reported following the spring 2011 administration.[1]

9.2 SCHOOL AND DISTRICT RESULTS REPORTS

9.2.1 Preliminary Reports of Participation and Performance

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

- *Preliminary Participation Report*
- *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/results.html.

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.

[1] For those students who participated in and had complete subcategory scores for both fall and spring 2011 MEPA testing, results were shown for both administrations in the spring reports. 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

9.2.1.1 Preliminary Participation Report

This report shows the following data for each grade span:

- The number of students enrolled
- The number of students who participated in testing
- The number of students who did not participate in testing in each category of nonparticipation (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)

9.2.1.2 Preliminary Results by Year of Enrollment in Massachusetts Schools

This report provides test results and administration data in each of the following categories:

- The number of students enrolled; 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 number and percentage of enrolled students who were tested
- 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.

9.2.2 Roster of Student Results

This report provides a school with the MEPA results for each ELL student at that school. A separate *Roster of Student Results* report for each grade span was generated following each MEPA administration. Each ELL student enrolled at the school in the grade span of the report is listed alphabetically by last name.[3] Each student's overall scaled score, performance level, scaled subscores in reading and writing, and raw scores in listening and speaking are shown. [4]

---

proficiency (LEP) enrollment data to determine actual participation rates.

[3] 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 one year, or if his or her records from previous administrations could not be matched based on SIMS, results from only that spring administration were reported.

[4] 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 of this report.

9.2.3 Progress Report

The MEPA *Progress Report* provides schools and districts with a summary of student progress toward English language proficiency. Progress is determined for all ELL students who fully participated in all portions of the MEPA test in the spring administration (2011 or 2012) and who had a baseline performance level, either from the previous spring or fall administration (2010 or 2011, respectively). Each of the lower four performance levels (levels 1–4) is divided into two performance level steps (low and high). All students at level 5 are assigned to a single step. For students whose baseline MEPA score is from the same grade span test as in the spring (2011 or 2012), the Department defines progress toward acquiring English language proficiency as advancing two or more performance level steps until they reach level 3 high. After a student reaches level 3 high, progress is defined as advancing one performance level step. For students whose baseline MEPA score is from an earlier grade span test than the test taken in spring 2010 or 2011 (with respect to spring 2011 or 2012, respectively), progress is defined as advancing at least one performance level step. Students with baseline MEPA scores in level 5 who remain level 5 in spring 2011 or 2012 (relative to the appropriate baseline) are considered to have made progress.

The report shows the following data:

- Number of students included; which consists of all students who tested in two consecutive administrations
- Number and percentage of students making progress
- Number and percentage of students in level 4 high and level 5 in the spring administration, by number of years in Massachusetts public schools
- Comparison of performance levels of students tested in consecutive administrations of the same grade span
- Comparison of performance levels of students tested in consecutive administrations of different grade spans
- Summary of average score change of students tested in consecutive administrations of the same grade span

9.3 PARENT/GUARDIAN 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 consecutive administrations at the same grade span, results were included for both administrations. If a student participated in consecutive administrations but was missing a score from one of the administrations 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 containing the missing score.

On the top half of the *Parent/Guardian Report* results page, the student's overall MEPA scaled score and performance level for the current year and prior year, if available, were shown. The score was 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 tools for comparing the student's scores to two other criteria: the performance of other students who have been enrolled in Massachusetts public schools for the same number of years, and the performance of a typical student performing at *level 5*.

9.4 Interpretive Materials and Workshops

Interpretive information to assist parents in understanding the displayed results was included in the *Parent/Guardian Report*s. A sample of the *Parent/Guardian Report* is available at www.doe.mass.edu/mcas/mepa/results.html.

9.5 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 test results is found in Appendix S.

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 *Preliminary Participation Report*.

9.6 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 with the MEPA results 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 and Reporting Services 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% agreement. Different exclusions assigned to students, which are used to determine whether a student receives a scaled score (or 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% 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, as explained below, though the sets may not be mutually exclusive.

The first set includes samples 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 samples that satisfy the following criteria:

- Private school serving students with disabilities at public expense
- Small school that receives no school reports
- Small district that receives no district reports
- District that receives a report, but all schools are too small to receive school reports
- 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.

CHAPTER 10. VALIDITY

Because the interpretations of test scores, and not the test itself, are evaluated for validity, 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 a determination of how well the test's tasks represent the curriculum and standards for each content area and grade level. This measure 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 chapters 3 and 4. 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—these 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. All principals and test administrators were required to familiarize themselves with and adhere to all of the procedures outlined in the MEPA administration manuals.

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 in gathering 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 through 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 9 and in the *Guide to Interpreting the MEPA Reports for Schools and Districts*, available at www.doe.mass.edu/mcas/mepa/results.html. The *Guide* provides the public with accurate and clear test score information. Scaled scores simplify reporting of results 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.

In 2010, the MEPA program embarked on a multi-year transition from a paper-based to a computer-based mode of administration. This transition continued in 2011 and in 2012. As part of this transition, a study was conducted each year to evaluate the comparability of the student test-taking experience between these two modes. Results of the studies for 2011 and 2012 continued to indicate that the effect sizes were small and that equating the two versions of the test was not necessary. Complete details of the comparability studies for 2011 and 2012 can be found in the two MEPA Comparability Study reports, which are included as Appendix A.

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.

10.1 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 used to measure 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.

10.2 Structural Validity

Though the previous types of validity examine the concurrence between different measures of the same content area, structural validity focuses on the relationships among 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.

10.3 PROCEDURAL VALIDITY

The MEPA *Principal's Administration Manual* and *Test Administrator's Manuals* delineated the procedures to which all MEPA test coordinators and test administrators were required to adhere. Procedural validity studies, if conducted, can document the procedures that were followed throughout the MEPA administration, thereby verifying 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 cheating among students or incorrect scanning of documents. 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.

REFERENCES

Allen, Mary J., & Wendy M. Yen. 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., & S. H. Kim. 2004. *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Dekker.

Campbell, D. T., & D. W. Fiske. 1959. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56:81–105.

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., & P. W. Holland. 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., & E. Kulick. 1986. Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement*, 23:355–368.

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

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

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

Joint Committee on Testing Practices. 2004. *Code of fair testing practices in education*. Washington, DC: National Council on Measurement in Education.

Kim, S. 2006. A comparative study of IRT fixed parameter calibration methods. *Journal of Educational Measurement*, 43(4):355–81.

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

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

Massachusetts Department of Education. 2001 and 2004. *English Language Arts Curriculum Framework.* Malden, MA.

—. 2003. *English Language Proficiency Benchmarks and Outcomes for English Language Learners*. Malden, MA.

—. 2009. *MEPA 2009 Technical Report*. Malden, MA.

—. 2011. *Guide to interpreting the 2011 MEPA reports for schools and districts*. Malden, MA.

—. 2012. *Guide to interpreting the 2012 MEPA reports for schools and districts*. Malden, MA.

Muraki, E., & R. D. Bock. 2003. PARSCALE 4.1. Lincolnwood, IL: Scientific Software International.

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

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

Stout, W. F., A. G. Froelich, & F. Gao. 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, NY: Springer-Verlag.

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

APPENDICES