

Massachusetts Adult Proficiency Tests

Technical Manual¹

Stephen G. Sireci, Peter Baldwin, Andrea Martone,
April Zenisky Laguilles, Ronald K. Hambleton, and Kyung T. Han

Center for Educational Assessment

October, 2006

¹ *Center for Educational Assessment Research Report No. 600*. Amherst, MA: School of Education, Center for Educational Assessment. Funding for this research was provided by the Massachusetts Department of Education under Project/Grant #S39000000700000. The opinions expressed in this report are those of the authors and do not necessarily represent official positions of the Massachusetts Department of Education.

Acknowledgements

This technical manual represents an extraordinary team effort including tireless work by staff at the Massachusetts Department of Education Office of Adult and Community Learning Services (especially Jane Schwerdtfeger, Anne Serino, Donna Cornellier, Bob Bickerton, and Robert Foreman), graduate students in the Research and Evaluation Methods Program at UMASS (particularly Karen Addesso, Jill Delton, Ana Karantonis, Rob Keller, Christine Lewis, Tie Liang, and Polly Parker), additional faculty and staff at UMASS (Craig Wells, Pat Paugh, Mercedes Valle, and Lisa Keller), staff at the UMASS Center for Educational Software Development (Stephen Battisti, David Hart, Gordon Anderson, Cindy Stein, and Gail Parsloe), and the Office Manager for the Center for Educational Assessment—Kelly Smiaroski. Most of all, we are particularly grateful for the hard work of the hundreds of adult educators and ABE learners throughout Massachusetts who contributed to the development of the MAPT in many important ways.

Massachusetts Adult Proficiency Tests

Technical Manual

Table of Contents

I. INTRODUCTION.....	5
PURPOSES OF THIS TECHNICAL MANUAL.....	6
II. DESCRIPTION OF MAPT	7
1. THE PURPOSE OF MAPT	7
2. DESCRIPTION OF THE TEST SPECIFICATION DEVELOPMENT PROCESS	8
3. DESCRIPTION OF THE MAPT FOR MATHEMATICS AND NUMERACY	11
4. DESCRIPTION OF THE MAPT FOR READING	17
5. DESCRIPTION OF MULTISTAGE TESTING	21
III. DESCRIPTION OF TEST CONSTRUCTION PROCEDURES.....	23
1. ITEM REVIEW AND REVISION	23
2. DEVELOPMENT OF COMPUTERIZED TEST SUPPORTING MATERIAL	23
3. PILOT TESTING	25
4. ITEM ANALYSIS	32
5. SENSITIVITY REVIEW	42
6. DEVELOPMENT OF MST MODULES AND PANELS.....	44
IV. DEVELOPING THE MAPT SCORE SCALE	47
1. ITEM RESPONSE THEORY CALIBRATION	47
2. ESTABLISHING THE MAPT SCORE SCALE.....	48
V. STANDARD SETTING: ESTABLISHING THE NRS EDUCATIONAL FUNCTIONING LEVELS	51
1. STANDARD SETTING METHODOLOGY	51
2. VALIDITY EVIDENCE FOR STANDARD SETTING.....	55
3. SETTING NRS CUT-SCORES ON THE MAPT FOR MATHEMATICS AND NUMERACY	57
4. SETTING NRS CUT-SCORES ON THE MAPT FOR READING	61
VI. MEASUREMENT PRECISION	65
1. TEST INFORMATION	65
2. CONDITIONAL STANDARD ERROR OF ESTIMATION	68
3. DECISION ACCURACY	70
VII. VALIDITY	73
1. UNDERSTANDING VALIDITY AND VALIDATION	73
2. VALIDITY EVIDENCE BASED ON TEST CONTENT	76

3. VALIDITY EVIDENCE BASED ON INTERNAL STRUCTURE	89
4. VALIDITY EVIDENCE BASED ON RELATIONS TO OTHER VARIABLES	95
5. VALIDITY EVIDENCE BASED ON CONSEQUENCES OF TESTING	97
6. VALIDITY EVIDENCE SUPPORTING EVALUATION OF EDUCATIONAL GAIN	100
7. OTHER VALIDITY EVIDENCE	107
8. TEST SECURITY	109
VIII. INTERPRETING MAPT TEST SCORES	111
IX. LOOKING FORWARD: NEXT STEPS FOR THE MAPT	113
REFERENCES	114
APPENDICES	120
APPENDIX A	120
LISTS OF MAPT COMMITTEE MEMBERS	120
APPENDIX B	130
LINKS TO ADDITIONAL RESOURCES	130
APPENDIX C	131
MAPT FOR READING STANDARDS BY LEVEL	131
APPENDIX D	138
SENSITIVITY REVIEW GUIDELINES	138
APPENDIX E	139
MATH STANDARD SETTING MEETING – AGENDA, DESCRIPTION, RATING FORM, SURVEY	139
APPENDIX F	146
READING STANDARD SETTING MEETING – AGENDA, DESCRIPTION, RATING FORM, SURVEY ..	147
APPENDIX G	153
MATH CONTENT VALIDITY MEETING – AGENDA, DESCRIPTION, RATING FORM, SURVEY	153
APPENDIX H	162
READING CONTENT VALIDITY MEETING – AGENDA, DESCRIPTION, RATING FORM, SURVEY ..	162

I. Introduction

Since January 2003, the Center for Educational Assessment at the University of Massachusetts Amherst (UMASS) has worked closely with the Massachusetts Department of Education's Office of Adult and Community Learning Services (ACLS) to develop achievement tests in math and reading that are appropriate for adult learners in Massachusetts. Our collaborative vision in creating these tests involved ensuring the tests are (a) aligned with the curriculum frameworks established by ACLS and the adult basic education (ABE) community in Massachusetts, (b) consistent with the instruction in ABE classrooms as directed by these frameworks, and (c) developed with comprehensive input from teachers and administrators from the ABE community in Massachusetts.

Early on in this collaboration we realized that measuring ABE learners' gains required a highly specialized test design. Given that ABE learners span a wide range of proficiency and typically are unable to take classes full-time, the testing system would need to measure progress along a wide spectrum of achievement and would need to be sensitive to small, but meaningful, educational gains. These two needs are best met by *adaptive* test designs that adjust the difficulty of an exam according to how well learners perform while taking the exam (Hambleton, Zaal, & Pieters, 1991; Sands, Waters, & McBride, 1997; Sireci, 2004; Wainer, 2000). The decision to use an adaptive test administration design is consistent with the recommendations from a national panel of experts convened by the National Academy of Sciences to address issues in ABE assessment (Mislevy & Knowles, 2002).

The end results of our collaborative endeavor are the *Massachusetts Adult Proficiency Test for Mathematics and Numeracy* and the *Massachusetts Adult Proficiency Test for Reading*. To develop these tests, we convened groups of ABE teachers and administrators who are experts in ABE mathematics or reading instruction to help us create test specifications congruent with Massachusetts ABE curriculum frameworks. In addition, we trained over 200 Massachusetts ABE teachers and staff to write test items in accordance with accepted principles of item writing (e.g., Downing & Haladyna, 1997; Haladyna, 1994; Haladyna & Downing, 1989). Expert staff and teachers also helped us review items and set the cut-scores on the MAPT that correspond to the U.S. Department of Education's Educational Functioning Levels (as dictated in the National Reporting System for Adult Basic Education). We also designed a sophisticated multistage adaptive testing system to tailor the administration of the MAPT to each learner and to maximize the precision of our estimates of their achievement levels and educational gains. To accomplish that feat, we collaborated with the Center for Educational Software Development at UMASS.

We are delighted that ABE learners in Massachusetts are currently taking computerized math and reading tests over the World Wide Web that are tied to Massachusetts ABE curriculum frameworks, developed in part by their teachers, and targeted to their proficiency levels. The remaining sections in this manual describe the complex nature of the MAPT and the procedures used to develop and evaluate these tests. This manual represents the first compendium of technical information on the MAPT and will be updated as changes are implemented. Additional reports pertaining to the psychometric characteristics of the MAPT are available from the Center for Educational Assessment (<http://www.UMASS.edu/rempe>).

Purposes of this Technical Manual

Proper test score use and interpretation requires a firm understanding of the capabilities and limitations of a testing program. As described in the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education, 1999), “The provision of supporting documents for tests is the primary means by which test developers, publishers, and distributors communicate with test users” (p. 67). The purpose of this Technical Manual is to inform all current and potential users of the MAPT in math and reading of (a) the purpose of the tests and their intended uses; (b) the content of the tests; (c) the processes used to develop, validate, and set standards on the tests, (d) the processes used to ensure assessment equity across diverse learners; (e) the technical characteristics of the test such as measurement precision, score scale characteristics, and multistage test administration design; and (f) how to properly interpret the test scores. Additional information regarding test administration and other policies related to assessing the proficiencies of adult learners in Massachusetts can be found on the ACLS web site (<http://www.doe.mass.edu/acls/assessment>). That web site contains links to practice test and other information regarding the MAPT, as well as the *Assessment Policies and Procedures Manual*.

This Technical Manual is intended to be informative to several audiences; however, some sections may require familiarity with statistics and other aspects of psychometrics. The types of professionals who may be interested in the material in this manual include ABE teachers, administrators, and staff; representatives from the US Department of Education and other organizations interested in evaluating the technical merits of the MAPT; representatives from ABE programs in other states who want to understand the strengths and limitations of MAPT; and members of the psychometric and educational research communities who are interested in test evaluation and computer-based testing.

The MAPT represents an evolution in testing and so it is likely to be of interest to many researchers studying educational assessment. The multistage adaptive nature of the test, the linking of the score scale to the NRS standards (EFLs), the alignment of the tests to the state frameworks, and the explicit recruitment and inclusion of teachers in the test development and evaluation processes are relatively unique features of an educational assessment program. For this reason, we have tried to be as comprehensive as possible in documenting all processes used in the development and validation of these tests, as well as the rationales behind our decisions.

Although this manual represents a comprehensive description of the technical qualities and features of the MAPT, these tests are not intended to remain static, and the process of test validation is ongoing. We expect the MAPT to evolve and improve as we continue to monitor its functioning, and as we continue to conduct validation research. The final chapter in this manual (Chapter IX, Looking Forward: Next Steps for MAPT) discusses some forward-looking issues we hope to explore as well as some potential directions for the MAPT program.

II. Description of MAPT

1. The Purpose of MAPT

There are many assessment needs in ABE. These needs include placement of learners into appropriate instructional programs, measuring learners' educational gains over time, and diagnosing their strengths and weaknesses in specific subject areas. Another important need is for accountability. The Massachusetts Department of Education, which funds ABE programs in the Commonwealth, needs to evaluate programs with respect to the degree to which they are helping learners meet their educational goals according to the Massachusetts ABE Curriculum Frameworks. In addition to accountability at the state level, there is also accountability at the Federal level, because Massachusetts receives federal funding for its ABE programs. The US Department of Education (DOE)'s Office of Vocational and Adult Education (OVAE) established the National Reporting System (NRS) for Adult Basic Education, which requires states to measure ABE learners' educational gains as a core outcome measure of program effectiveness (see <http://www.nrsweb.org/>). All states receiving funds from OVAE must comply with the NRS requirements. These federal and state accountability demands were the primary factors motivating development of the MAPT.

The purposes of the MAPT are to measure ABE learners' knowledge and skill in math and reading so that their progress in meeting educational goals can be evaluated. The MAPT is designed to measure learners' educational gains for the purposes of state monitoring and accountability under the NRS. Learners' MAPT scores and score gains can be aggregated to provide meaningful summative measures of program effectiveness. The specific knowledge and skills measured by MAPT come from the Massachusetts ABE curriculum frameworks and are described in detail in Chapter III of this manual.

At this juncture, the MAPT assessments are not designed for diagnosing learners' specific strengths and weaknesses within each subject area. Diagnostic evaluation should be based on more than MAPT scores, however MAPT scores may be useful as one piece of information in a more comprehensive assessment of learners' strengths and weaknesses that should include classroom assessments, homework assignments, classroom observation, and performance on the MAPT practice items.

MAPT assessments are also not designed for placing students into instructional programs. As described in the *Assessment Policies and Procedures Manual* (<http://www.doe.mass.edu/acls/assessment/>), students should be in instructional programs for at least two weeks before they are pre-tested with the MAPT.

2. Description of the Test Specification Development Process

A primary factor in deciding to develop assessments specifically for ABE learners in Massachusetts was a desire to link assessments with the Massachusetts ABE curriculum frameworks developed by the Department of Education. Prior to commissioning test development, the Massachusetts Department of Education assembled a Performance Accountability Working Group that comprehensively studied available assessments for ABE learners. This Group evaluated the overlap of several national tests (ABLE, AMES, CASAS, and TABE) with the Massachusetts frameworks and concluded

The commercially available assessment that came closest to alignment with the ELA and Math Curriculum Framework standards was the TABE...Unfortunately, “closest” still only represents a 56% correlation of the TABE’s test items with learning standards in the math framework and even less for alignment with the English Language Arts standards...the TABE is also limited in the degree to which it meets the PAWG’s third overriding goal as stated earlier - assessments must provide useful information to teachers and students. These are three of the reasons why the TABE is not recommended as a long-term solution for assessing ABE Reading, Writing and Math. (Massachusetts Adult Basic Education Performance Accountability Working Group, 2002, p. 6)

To ensure the MAPT would be aligned with the curriculum frameworks and to instruction in ABE classrooms, UMASS and ACLS (with input from the System for Adult Basic Education Support, SABES) worked together to assemble test development specification committees. Separate committees were assembled for the MAPT for Mathematics and Numeracy and for the MAPT for Reading. The MAPT for Mathematics and Numeracy Test Specifications Development Committee consisted of five leaders in ABE mathematics instruction in Massachusetts, a senior staff person from ACLS, and four representatives from UMASS (see Appendix A for committee membership; see Appendix B for Internet links to additional assessment resources related to the MAPT). The MAPT for Reading Test Specifications Development Committee consisted of four leaders in ABE reading instruction in Massachusetts, one expert in educational psychology with a concentration in reading, a senior staff person from ACLS, and four representatives from UMASS (see Appendix A for committee membership).

Each committee was sent materials in advance of the meeting describing the purpose of the Committee (to develop test specifications aligned with the curriculum frameworks) and two sets of potential test specifications (see Baldwin & Sireci, 2003 for math and Valle & Sireci, 2003 for reading). Each committee met for a full day and had several follow-up conference calls and multiple rounds of correspondence via e-mail. At the full-day meeting, each committee was oriented to the purposes of the forthcoming tests and asked to respond to a set of specific questions designed to facilitate the test specification development process. Both committees met during December 2003. Each committee produced a final report (see Sireci, Baldwin et al., 2004 for the math report and Sireci, Li et al., 2004, for the reading report) and these reports contained the test specifications that governed the item and test development processes.

Deriving test specifications from the ABE curriculum frameworks was a first step in fostering test-curriculum alignment. However, an important aspect of such alignment is the degree to which the actual test items reflect tasks on which students work in the classroom. To maximize the degree to which the test items would reflect instructional practice, an explicit goal of test development was to train teachers to write items for the new tests. This strategy encouraged teachers to be part of the test development process so that they could take some “ownership” of the new tests and not feel like these tests were being “forced” on them from the outside.

Asking teachers to write items for the new test is not straightforward for at least two reasons. First, teachers are not trained in writing the types of items appropriate for large-scale assessment, such as multiple-choice items (American Federation of Teachers, National Council on Measurement in Education, & National Education Association, 1990), and second, it takes time to write quality test items—time most teachers do not have. To address these obstacles, we held a series of item writing workshops across the state from March 2004 through June 2005. These workshops briefly informed teachers of the characteristics of quality educational assessments (e.g., content validity) and of the appropriate uses of test scores. The bulk of the workshops focused on (a) reviewing guidelines for writing multiple-choice items (e.g., Haladyna & Downing, 1989) (b) giving teachers opportunities to practice item writing, and (c) providing feedback to teachers on the items they wrote. We also offered a two-week, full-time test construction graduate course at the University of Massachusetts in August 2004 in which 14 ABE teachers/administrators participated.

Over 200 ABE teachers and administrators were trained to write items targeted to the test specifications (and to specific objectives within the curriculum frameworks). These activities led to the production of over 2,000 test items that made up the initial pool of potential MAPT items.

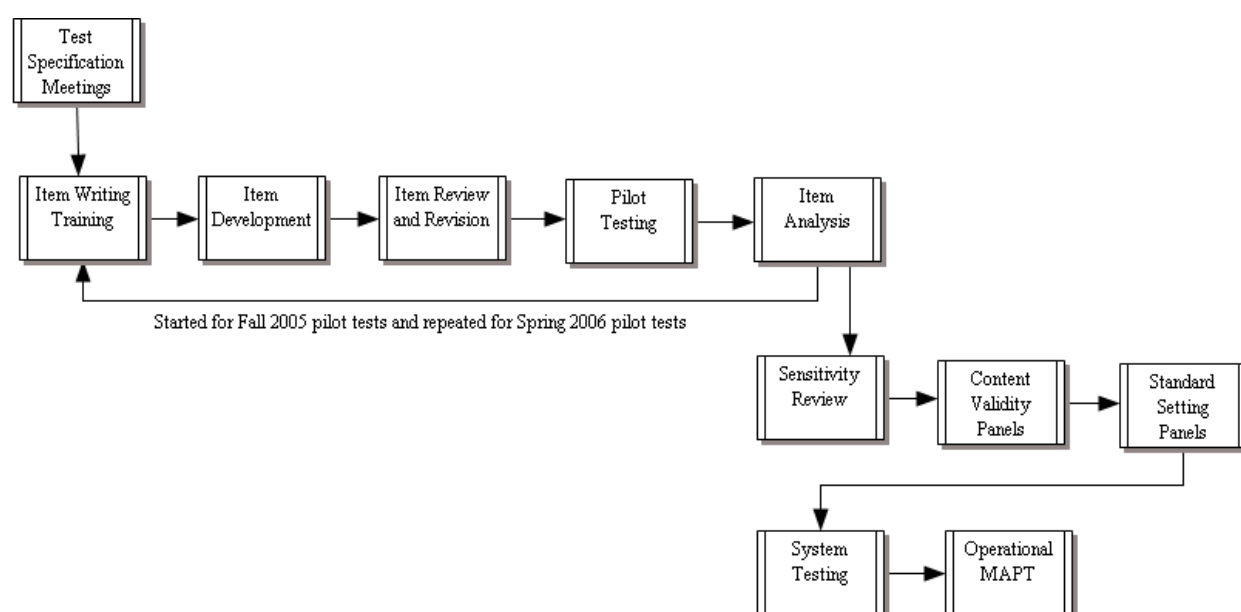
As the item pool for each test was created, item review committees were assembled to review items for content quality (degree of match to intended objective, technical accuracy, appropriateness for adult learners, clarity, etc.). The review committees also rated items to determine how well the item measured the objective to which it was matched. The reviewers used a scale of 1-6, with 6 being exemplary, and only items rated a 4 or above were considered for pilot testing. Items were also reviewed for psychometric quality by UMASS staff. Items were edited and then subsets of items were selected for pilot testing. Further details regarding item development, review, revision, and pilot testing are described in Chapter III. Following pilot testing, the statistical functioning of the items was evaluated and the items considered to be most appropriate for operational use were selected. These items then underwent sensitivity (item bias) review and content validity review (see Chapters III and VII for information on the sensitivity and content validity reviews, respectively).

The end results of the item development, revision, and review processes were pools of items considered most appropriate for assembling the operational multistage tests. After the operational tests were developed, standard setting studies were conducted to determine the most appropriate cut-scores on the MAPT to signify the Educational Functioning Levels established by the NRS (see chapter V).

The process of developing the MAPT, from the development of test specifications to development of the operational tests and standard setting, spanned more than three years. From across the Commonwealth, literally hundreds of people wrote, edited, or reviewed items, or provided valuable input in some other way. Figure I.1 presents an illustration of the test development process. Each of the steps in Figure I.1 is explained in one or more sections of this manual. It should also be noted that refinements to the MAPT testing program are ongoing and so Figure 1 illustrates the development process, but not the continuation and improvement of MAPT, which is dynamic.

Figure I.1

Illustration of MAPT Test Development Process



Incorporating Universal Test Design Principles

The MAPT assessments are designed for all types of ABE learners, regardless of any disabling conditions they may have. For this reason, the principles of *Universal Test Design* (UTD) were used throughout the test development process. UTD involves developing tests as flexibly as possible to minimize the need to provide accommodations to the test administration, and also flexible enough so that accommodations can be provided for any learners who may need them (see Thompson & Thurlow, 2002, or Thompson, Thurlow, & Malouf, 2004 for more information regarding UTD). Examples of how UTD principles governed MAPT development include the use of clear and simple language in developing items, clear definitions of the construct measured, legible presentation of items, extremely generous time limits for all learners and the opportunity to restart the test at a subsequent date (if necessary), simple instructions for taking the test, and amenability to accommodations. A discussion of the MAPT accommodations suggested for learners with disabilities is presented in Chapter VII.

3. Description of the MAPT for Mathematics and Numeracy

The “construct” of Mathematics and Numeracy was operationally defined for the MAPT using two dimensions. The first dimension is a *content strand* dimension comprising the four content strands contained in the *Massachusetts Adult Basic Education Curriculum Framework for Mathematics and Numeracy* (ACLS, 2005b). These strands are Number Sense; Patterns, Functions, and Algebra; Statistics and Probability; and Geometry and Measurement. The second dimension is a cognitive (reasoning) dimension containing three levels derived from Bloom’s (1956) taxonomy. The original Bloom’s taxonomy contained six levels, but for the MAPT for Mathematics and Numeracy, the two lowest levels were combined into one category and the three highest levels were combined into another. This collapsing of categories is consistent with contemporary theories regarding how cognitive skill dimensions should be used in test specifications (Linn & Gronlund, 2000; Popham, 1992) because it is hard to distinguish among the original six levels when classifying test items. The three cognitive skill levels in the MAPT for Mathematics and Numeracy are Knowledge and Comprehension; Application; and Analysis, Synthesis, and Evaluation.

In addition to the two-dimensional framework for defining the MAPT for Mathematics and Numeracy, the test development framework needed to address the different levels of ABE learners. The *Massachusetts Adult Basic Education Curriculum Framework for Mathematics and Numeracy* (ACLS, 2005b) describes six *Learning Levels* that distinguish the different performance levels present in the ABE student population. These levels are similar to the NRS Education Functioning Levels for math and range from 1 (Beginning Adult Numeracy) to Level 6 (Bridge to College). The test specifications that governed development of this test stipulated specific percentages of test items in each of the intersections of content strands and cognitive levels, but these percentages were set within each Learning Level.

MAPT for Mathematics and Numeracy Test Specifications

As described in Sireci, Baldwin et al. (2004), the Test Specifications Development Committee used the content strand-by-cognitive level matrix to suggest the relative emphases of these strands and levels across the different Learner Levels. It was decided *not* to develop a test for the lowest Learner Level, Beginning Adult Numeracy, because learners at that level are working on literacy skills that are necessary for taking a text-based test. The Committee also expressed concern that students at the highest level (level 6) would not be motivated to take the MAPT because they were focused on the GED. Therefore, it was decided that test specifications for the current (first edition) of the MAPT for Mathematics and Numeracy be designed for ABE learners at Levels 2 through 5.

The test specifications that governed item development and the assembly of pilot test forms for the MAPT for Mathematics and Numeracy are presented in Table III.1. As is evident from this table, all content strands are measured at all Learner Levels, but there are some differences in emphasis across levels (e.g., Patterns, Functions, and Algebra increases across learner levels). It can also be seen that the emphasis of the Analysis, Synthesis, and Evaluation cognitive level increases as Learner Level increases (from 18% at Level 2 to 41% at Level 5).

Table III.1

Initial Test Specifications for ACLS Mathematics and Numeracy Proficiency Tests

Test Level 2 (GLE 2.0-3.9)				
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Total
Number Sense	15	15	5	35
Patterns, Functions, Algebra	6	6	3	15
Statistics and Probability	10	10	5	25
Geometry and Measurement	10	10	5	25
	41%	41%	18%	100%
Test Level 3 (GLE 4.0-5.9)				
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Total
Number Sense	10	15	5	30
Patterns, Functions, Algebra	10	5	5	20
Statistics and Probability	5	15	5	25
Geometry and Measurement	10	10	5	25
	35%	45%	20%	100%
Test Level 4 (GLE 6.0-8.9)				
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Total
Number Sense	10	10	5	25
Patterns, Functions, Algebra	5	15	5	25
Statistics and Probability	5	10	10	25
Geometry and Measurement	5	15	5	25
	25%	50%	25%	100%
Test Level 5 (GLE 9.0-10.9)				
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Total
Number Sense	3	6	6	15
Patterns, Functions, Algebra	5	15	10	30
Statistics and Probability	5	10	15	30
Geometry and Measurement	5	10	10	25
	18%	41%	41%	100%

Note: Cell entries represent percentages of test items devoted to each content strand and cognitive level.

It should be noted that these Learner Levels are linked to the NRS Education Functioning Levels (described further in Chapter V) and that these levels are the entry points for learners taking the MAPT. That is, a learner classified by her/his teacher as Level 3 (Intermediate ABE Mathematics or Low Intermediate) will begin the MAPT at a point where the multistage test algorithm targets the test specifications for Learning Level 3. However, the learner's performance may cross levels from one stage of the test to another, which will affect the specific distribution of items across content strands. Thus, a learner who begins the test at Level 3, but ends up at Level 4, will have taken a test with test specifications somewhere between those stipulated for Level 3 and Level 4 (more details regarding content representation are provided in this chapter and in chapter VII). In the following sections, detailed descriptions of the areas within the content strand dimension and the levels within the cognitive dimension are provided.

Content Strands

As mentioned above, the four content strands in the MAPT for Mathematics and Numeracy were taken directly from the *Massachusetts Adult Basic Education Curriculum Framework for Mathematics and Numeracy* (ACLS, 2005b). The definitions that follow are taken directly from these frameworks, since they were used to guide development of the items for the MAPT for Mathematics and Numeracy.

Number sense

Number Sense is the foundation of numeracy. Sound number sense enables us to interpret and represent the world in which we live. It is evident in all we do, whether in complex examples such as the Gross National Product, basic issues such as the family budget, or as personal as a blood pressure reading. Mathematical intuition grows with a strong basic understanding of numbers and, with that, our ability to do mathematical problem solving.

To be efficient workers or consumers in today's world, adults must have a strongly developed conceptual understanding of arithmetic operations, as well as the procedural knowledge of computation and number facts. They must be able to perceive the idea of place value and be able to read, write, and represent numbers and numerical relationships in a wide variety of ways. Simple paper-and-pencil computation skills are not enough. Adults must be able to make decisions regarding the best method of computation (mental math, paper-and-pencil, or calculator/computer) to use for a particular situation. Knowledge of numbers, operations and computation must include both a well-developed number sense and the ability to use basic mathematics-related technologies.

Number sense promotes accuracy in estimation and flexibility and efficiency in mental math. While calculators and computers are used to do most of the complex computations in today's world, the ability to estimate is critical for lifelong learners. Adults use informal measurements in life skill activities such as cooking, shopping, buying clothes, or estimating the time required for daily tasks. Estimation is a valuable skill for checking the reasonableness of computation or accuracy in problem solving, and is an aid in timed-test situations such as the GED. It builds on adult experience and knowledge. Good estimators use a variety of strategies and techniques for computational estimation that can be explored and shared by learners.

Patterns, functions, and algebra

Mathematics has been defined as the study of patterns. Learning to recognize, analyze, describe, and represent patterns and number relationships connects math to the world and helps us to appreciate fully the intrinsic value of such pleasures as poetry, art, music, and science. Math concepts formerly taught only in basic algebra courses are increasingly part of the culture and vocabulary of modern life. Headlines and news reports speak of exponential growth of the national debt, a variable rate mortgage, or a balanced budget, while medical literature uses terms like “HIV-positive,” or “RH-negative.”

Being able to see and use patterns has been identified as a fundamental skill needed for developing mathematical understanding. The Patterns, Functions, and Algebra strand is positioned after the Number Sense strand because of the importance of building pre-number skills such as patterning which, in turn, enable adult learners to learn multiplication tables and number relationships necessary for efficient and fluent computation skills. The strand also encompasses skills that are necessary for developing concepts in the Data and Geometry and Measurement strands.

Algebra serves as a bridge between arithmetic and more broadly generalized mathematical situations. These generalizations can be expressed in words, tables and charts, the notation of formulas, and graphs. Life experience has afforded adult basic education learners with a broad base of real-world ties that can be readily linked to the concepts of equation, function, variable, and graph. From baby formulas to chemical formulas, algebra offers a succinct way to define real-world situations that can aid adults in the home and in the workplace.

Algebra impacts the competency of workers, parents and citizens, and algebraic thinking skills are crucial if adults are to compete in the global economy. Workplace skills requiring competencies in “information,” “systems,” and “technology” stress the need for organizing, interpreting and communicating information and employing computers as a tool for those tasks, as well as the ability to “discover a rule or principle underlying the relationship between two or more objects and apply it in solving a problem.” Identifying and expressing pattern, relation and function are the algebraic skills imbedded within these competencies.

Statistics and probability

The Statistics and Probability strand links numeracy and literacy learning. Numbers, logical reasoning, and texts interweave to describe phenomena visually, numerically and verbally in what we term *data*, which is the heart of this strand.

Data are wide-ranging and touch on many areas of academic study and inform us about our world. For instance, we learn about preferences, predilections and group characteristics when we read and interpret data. We learn about the power of evidence as we develop the skills to make statements and evaluate arguments based on data. We learn the power of the question and the framer of the question when we collect and represent data, and we learn that sometimes true, sometimes false, pictures are created when we compress data into statistics. Data is a

powerful descriptive tool.

So powerful are data that agencies of authority often use them to generate, promote and, sometimes, evaluate decisions. Citizens, therefore, must understand the ways of data to exercise their collective and individual intelligence by responding to the expanding presence of this dense expression of information.

The learning standards in the Statistics and Probability strand provide adult learners with the tools for dealing with data.

Geometry and measurement

Geometry and measurement help us represent in an orderly fashion what we see in our world. Whether we are cooking or cartooning, shopping or shipping, painting a canvas or a wall, designing an addition for a house or a play yard for preschool, we continually bump up against these mathematical organizers. Lifelong learners should know and understand these interconnected and symbiotic mathematical domains.

Adult learners who attend basic mathematics classes at any level share a wealth of pragmatic experience surrounding geometric and spatial concepts. They have probably built a bookcase, laid out a garden, applied wallpaper or tiled a floor, all the while discovering informally the rules which formally govern the study of geometry itself.

Geometry and measurement often spark a renewed interest in mathematics for those students who have been turned off for some reason or heretofore have felt unsuccessful with mathematics learning. Investigating problems that involve geometry and measurement broadens all students' mathematical understanding and engages them as they explore mathematical ideas.

Cognitive Levels

As mentioned earlier, MAPT for Mathematics and Numeracy items were written to and classified according to three cognitive levels—Knowledge and Comprehension; Application; and Analysis, Synthesis, and Evaluation. This dimension was included in specifying the content of the test to ensure that items were written to measure skills necessary for the different purposes for which math is used. As described above, these three levels represent a simplification of Bloom's (1956) six levels of cognitive complexity. The Test Specifications Development Committee believed it was important to separate the lower-level knowledge and comprehension areas from the higher cognitive levels, and they also believed it was important to distinguish application of math knowledge and comprehension from the areas of analysis, synthesis, and evaluation. Brief descriptions of the three cognitive levels used on the MAPT follow.

Knowledge and comprehension

Knowledge refers to remembering material that was previously learned, such as specific

facts, events, or principles. Examples include remembering mathematical terms and definitions or identifying terms, symbols, equations, etc. Comprehension refers to the ability to comprehend the meaning of material. Comprehension goes beyond mere memorization and is signified by an understanding of a concept. Being able to explain, translate, or summarize material are ways in which comprehension is demonstrated. The knowledge/comprehension level is characterized by both recognition of a concept and understanding the concept.

Application

Application refers to the ability to use learned material in new, but specific (concrete) situations. This cognitive level is invoked in applying rules, methods, concepts, principles, and other learned material in a realistic context. Examples include working with data in realistic context (calculating change, sales tax, discounts, etc.) and using a formula to solve a problem.

Analysis, synthesis, and evaluation

Analysis refers to the ability to break material down into its component parts to understand its underlying structure and the relationships between the components. Synthesis refers to the ability to combine parts to form a new whole, which may involve creation of a product from a variety of elements. Evaluation refers to the ability to make a judgment about a situation for a given purpose or to accomplish a goal. The analysis/synthesis/evaluation cognitive level is characterized by maneuvering through complex stimuli to fully understand a problem, and using several aspects of the situation to solve the problem. This cognitive level is distinguished from application in that it typically involves situations with more than one variable, or multiple pieces of information that must be understood, deciphered, and brought back together in a logical way.

The Dynamic Interplay Between Frameworks and Test Specifications

Early in the item writing process UMASS staff noticed that the Mathematics and Numeracy Curriculum Frameworks needed to be revised. While the strands from the original curriculum frameworks were used to guide the test specification development, specific objectives within the different strands and at different levels were either redundant or missing. UMASS worked with three ABE Math Curriculum Experts to review and revise the frameworks. Details of the changes to the curriculum frameworks can be found in *Refinements to ABE Mathematics Standards for Assessment Purposes* (Martone, Goodridge, et al., 2004). The result of this revision was a cleaner progression of objectives across levels and within strands. The final document was reviewed by ACLS staff and then released to ABE practitioners to reflect these changes. After this process, items could be more closely aligned to specific objectives.

4. Description of the MAPT for Reading

The “construct” of Reading was operationally defined for the MAPT for Reading using two dimensions. One dimension pertains to *reading skill* areas and the other dimension pertains to different *purposes for reading*. The skill dimension consists of three areas: Word Recognition, Vocabulary and Meaning, and Comprehension. The Purposes for Reading dimension consists of four areas. The first area is reading to “Acquire Basic Word Analysis Skills,” which is most relevant to the lower learner levels (GLE 0-3.9), as it is considered a pre-requisite to reading. The remaining areas in the Purposes dimension are Reading for Literary Experience, Reading to Gain Information, and Reading to Perform a Task. The Purposes for Reading dimension was seen as capturing the spirit of the reading skills articulated in the Massachusetts ABE Curriculum Frameworks and *Equipped for the Future* (Curry, et al., 1996; Stein, 2000).

In addition to the two-dimensional framework for defining the MAPT for Reading, the test development framework needed to address the different levels of ABE learners. The *Massachusetts Adult Basic Education Curriculum Framework for the English Language Arts* (ACLS, 2005a) describes five *Learning Levels* that distinguish the different performance levels present in the ABE student population. These levels are similar to the NRS Education Functioning Levels for reading and range from 1 (Initial) to Level 5 (Adept). However, the Test Specifications Development Committee concluded it was not appropriate to test essentially preliterate students with a text-based test and so it was decided to begin development of the test at Level 2. The Committee also noted that a test developed for level five would be appropriate for adults working to earn a high school diploma or gain access to postsecondary education (see Sireci, Li, et al., 2004). However, they thought these students will probably take the Tests of General Educational Development (GED) and would not be likely to sit for an ACLS proficiency test. Therefore, test specifications for the MAPT for Reading targeted Levels 2 through 4 (i.e., Beginning Basic through Low Adult Secondary). These specifications are described next.

MAPT for Reading test specifications

The test specifications that governed item development and the assembly of pilot test forms for the MAPT for Reading are presented in Table III.² A comparison of the proportions of test content allocated to the different area-by-purpose cells across test levels indicates that as test level increases, the proportion of the test measuring the purpose “Acquisition of Basic Word Analysis Skills” decreases. At test level 2, 50% of test content is associated with this purpose. At test level 3, this percentage drops to 20%, and at test level 4 it drops to 0%. The same pattern appears for the “Word Recognition” skill area. The purposes “Reading for Literary Experience” and “Reading to Gain Information” increase across test levels, as does the skill area of Comprehension. The changes in test specifications across test level clearly indicate that more advanced skills and more challenging reading purposes are being measured as test level

² As was mentioned in describing the math test specifications, these levels are the entry points for learners taking the MAPT and a learner’s performance may cross levels from one stage of the test to another, which will affect the specific distribution of items across content strands and purposes for reading. Thus, a learner who begins the test at Level 3, but ends up at Level 4, will have taken a test with test specifications somewhere between those stipulated for Level 3 and Level 4.

increases. The Committee agreed that these specifications reflect the changing emphasis in these areas across instructional levels in ABE programs. The specific standards measured on the MAPT for Reading are presented in Appendix C. A description of the Skill Areas and Purposes for Reading is presented in the next section.

Table III.2

Initial Test Specifications for ACLS Reading Proficiency Tests

Test Level 2 (GLE 2-3.9)					
Skill Areas	Purposes for Reading				
	Acquire Basic Skills& Word Analysis	Literary Experience	Gain Information	Perform Task	Total
Word Recognition	50	--	--	--	50
Vocabulary & Meaning	--	10	10	10	30
Comprehension	--	8	7	5	20
	50%	18%	17%	15%	100%
Test Level 3 (GLE 4-5.9)					
Skill Areas	Purposes for Reading				
	Acquire Basic Skills& Word Analysis	Literary Experience	Gain Information	Perform Task	Total
Word Recognition	20	--	--	--	20
Vocabulary & Meaning	--	10	10	10	30
Comprehension	--	20	20	10	50
	20%	30%	30%	20%	100%
Test Level 4 (GLE 6-10.9)					
Skill Areas	Purposes for Reading				
	Acquire Basic Skills& Word Analysis	Literary Experience	Gain Information	Perform Task	Total
Word Recognition	--	--	--	--	--
Vocabulary & Meaning	--	12	12	6	30
Comprehension	--	30	30	10	70
	--	42%	42%	16%	100%

Note: Cell entries represent percentages of test items devoted to each skill/purpose.

Skill Areas

As mentioned above, the three skill areas in the MAPT for Reading were developed by the Reading Test Specification Committee (Sireci, Li, et al., 2004). The definitions that follow are based on the guidance provided in that report and were used to guide development of the items on this assessment. A detailed listing of the standards within each skill area is presented in Appendix C.

Word recognition

This skill area involves recognizing controlled words, sight words, and visual cues. Word recognition skills are not comprehension skills. They simply allow the learner to read the word off the page. Examples of MAPT for Reading objectives that fall under this skill area include (a) Identify short vowels with single consonants (one syllable words); (b) Identify compound words; and (c) Apply the rules of syllabication to unfamiliar words from the content areas (social studies, science, and literature) and be able to know what syllable should be accented.

Vocabulary and meaning

This skill area involves the ability to understand word meanings and the ability to infer word meanings from context. Examples of MAPT for Reading objectives that fall under this skill area include (a) Recognize synonyms, antonyms, homonyms, and homophones (e.g., dear-deer) for identified vocabulary words presented in isolation or within a group of words; (b) Use context clues to determine meaning of unfamiliar words, including increasingly sophisticated multi-syllabic words originating from Latin and Greek prefixes, suffixes, and roots; (c) Identify idioms and their use in passages; and (d) Recognize and comprehend the meaning of moderately complex vocabulary using word, sentence, and paragraph clues to determine meaning.

Comprehension

Comprehension refers to the ability to extract meaning from connected discourse that is read. Examples of MAPT for Reading objectives that fall under this skill area include (a) Locate pertinent information in simple familiar materials (e.g., want ads, job listings, schedules, signs, food packages) including who, what, which, when, where; (b) Distinguish between fact and opinion in a short paragraph; (c) Determine author's purpose from a variety of written pieces; and (d) Draw conclusions and make inferences based on passages read.

Purposes for Reading

As mentioned earlier, MAPT for Reading items were written to and classified according to four purposes for reading: Reading to Acquire Basic Word Analysis Skills, Reading for Literary Experience, Reading to Gain Information, and Reading to Perform a Task. This dimension was included in specifying the content of the test to ensure that items measured skills relevant to adult learners as articulated in Massachusetts ABE Curriculum Frameworks and in *Equipped for the Future* (Curry, et al., 1996; Stein, 2000). The Test Specifications Development Committee

believed it was important to separate the different purposes for reading to ensure a balanced representation of different types of reading purposes. Brief descriptions of the four purposes for reading used on the MAPT follow.

Reading to acquire basic word analysis skills

This purpose for reading highlights learners' ability to develop an understanding of how phonemes, or speech sounds, are connected to print, as well as the ability to decode unfamiliar words.

Reading for literary experience

To read for literary experience is to listen to voices not one's own and to gain access to new experiences and perspectives. The encounter with literature is likely to engage the heart and the senses, as well as the mind. This encounter may be challenging but also offers the possibility of great pleasure.

Reading to gain information

Reading is one of the primary means of gathering the information that is required to make appropriate choices and to analyze and appreciate the complexity of life. A wide variety of texts offer information, from government documents to medical advisories, from newspaper articles to notes sent home from a child's school.

Reading to perform a task

On a daily basis, people need to be able to do things: for example, operate a machine, open a bank account, cook a meal, vote in an election. The challenge to the reader is to understand the purposes of the task, the vocabulary specific to the task, and the activity required to successfully complete the task.

5. Description of Multistage Testing

The National Academy of Sciences commissioned a Committee to investigate assessment needs and issues in adult education. In December 2001, this group, called the Committee for the Workshop on Alternatives for Assessing Adult Education and Literacy Programs (Mislevy & Knowles, 2002), held a national workshop where experts in adult education and psychometrics discussed assessment issues for adult learners. One of the concluding themes of the workshop was that, given the wide diversity of proficiency among adult education students and their limited time in the classroom, some form of *computerized adaptive testing* would be beneficial. Computerized adaptive testing (CAT) is a test administration model that uses the computer to select and deliver test items to examinees. The model is called *adaptive* because the computer selects the items to be administered to a specific examinee based, in part, on the proficiency of the examinee. Unlike many traditional tests where all examinees take a single form of an exam, the computer tailors the exam to each examinee. This tailoring is done by keeping track of an examinee's performance on each test question and then using this information to select the next item to be administered. Computerized adaptive testing reduces testing time by shortening the length of the test without losing information about examinee proficiency (Wainer, 2000).

The statistical model underlying computerized-adaptive testing is item response theory (IRT). IRT places test items and examinees on a common scale (i.e., the scale that indicates the difficulty of an item is the same scale that is used to assign scores to examinees). There are several attractive features of IRT, including the ability to provide scores on a common scale for examinees who take different items. This feature allows us to place ABE learners of different levels, who take different tests, on the same (wide) proficiency continuum. By using IRT in a CAT, learners' proficiency estimates are updated as they progress through the test, and a new item (or set of items) is selected based on the updated estimate. Items that are targeted at a learner's proficiency provide more information about her/his proficiency than items that are much too easy or much too difficult. This reduction explains why accurate measurement of ABE learners can be accomplished using fewer items than is typically required on a paper-based test such as the TABE.

There are several computerized-adaptive testing models (see Hambleton, et al., 1991; Sands, Waters, & McBride, 1997, and Wainer, 2000, for comprehensive discussions of these models), but most can be described as falling into one of two categories—adaptive at the item level or adaptive at a testlet level. A testlet (sometimes referred to as a *module* in multistage testing) is a group of items that function together as a “mini-test.” Item-level adaptive models select a new item for each learner after the learner answers an item. If the learner answered the item correctly, a slightly more difficult item is selected. If the learner answered the item incorrectly, a slightly easier item is selected. Adaptivity at the item level is the best way to tailor a test to the performance of the learner, however, item-level adaptive models require very large item pools and large sample sizes for piloting these items. In addition, item-level adaptive testing allows far less control over test content, since an almost infinite number of test “forms” could be administered. Testlet-adaptive models have emerged to address these drawbacks. In a testlet-adaptive model, such as the multistage testing model designed for the MAPT, learners answer a set of several items before the computer scores their performance and selects the next set of items (testlet) to be administered. Like item-level models, if the learner is doing well, s/he will be routed to a testlet at or above the difficulty level of the last testlet administered. If a

learner does not do well on the set of items, s/he will be routed to an easier testlet.

Multistage testing (MST) is often viewed as an intermediary step between a test that has a fixed number of items (such as a paper-and-pencil test) and item-level computerized-adaptive testing (CAT). MST combines the adaptive features of CAT with the opportunity to pre-assemble portions of the tests prior to test administration, which allows for content-balancing and general form integrity to be evaluated and controlled prior to operational administration. Other potential advantages of MST include simplification of the delivery mechanism and tempering the impact of response aberrancy (Jodoin, Zenisky, & Hambleton, 2006; Luecht, 2005; Sireci, 2004).

The specific MST architecture underlying the MAPT is complex and was specifically designed to meet the needs of adult assessment in Massachusetts. The details of this architecture, including how learners are routed as they progress through the assessment, are provided in Chapter III.

III. Description of Test Construction Procedures

The procedure for developing items for the MAPT (i.e., training ABE teachers and staff to write items aligned with the curriculum frameworks) was described in Chapter II. In this chapter, we provide descriptions of how the operational version of the MAPT was created from these items. The chapter begins with a discussion of the initial item reviews and ends with a description of how the multistage test modules and panels were assembled. A description of the comprehensive pilot tests conducted is also provided, as is a description of the item analysis statistics used to evaluate the statistical characteristics of the items, and the sensitivity reviews conducted to promote a more equitable assessment for learners from a variety of backgrounds.

1. Item Review and Revision

All potential MAPT test items went through a multistage review process. Initially, for both the math and reading areas, a content area specialist at UMASS reviewed each item. Items were approved, slightly revised, rewritten or deleted based on the item content and the standard to which it was matched. Prior to pilot testing, ABE specialists in each content area also reviewed each item. The math ABE Item Review Committee comprised two current math ABE teachers and one person with extensive experience in ABE math standards (see Appendix A). The reading Item Review Committee comprised three current or former ABE teachers and a literacy specialist (see Appendix A). These item reviewers examined how well each item measured the objective to which it was matched. They rated each item on scale of 1-6 with 6 being an exemplary item-objective match. The reviewers also suggested changes to the content of the item and to the cognitive area/purpose for reading areas as needed. Only items rated a 4 or above were considered for pilot testing. The UMASS content area specialists reviewed all comments and suggestions from the ABE reviewers. Based on these comments, some items were deleted, matched to a different standard, or revised. At this stage items were ready to be considered for selection into the pilot test. The content validity studies that were conducted on the items that performed well in the pilot tests are described in Chapter VII. The sensitivity reviews conducted on these items are reported later in this chapter.

2. Development of Computerized Test Supporting Material

We considered learners' unfamiliarity with computers to be a potential source of "construct-irrelevant variance," which refers to a factor that unnecessarily impedes learners from fully demonstrating their knowledge and skills. Thus, we considered it crucial to provide all learners with an opportunity to become familiar with the basic computer functions needed to answer test questions. A Computer Basics tutorial was developed in collaboration with the Massachusetts Department of Education and several ABE educators. This tutorial is accessible at <http://owl.cs.umass.edu/Departments/CtrEdAssess/Tutorial/index.html>.

The Computer Basics tutorial is comprised of two portions. The first section is an overview of basic computer functionality including using a mouse, clicking, and scrolling. Each page in this section demonstrates these skills with pictures and words in simple language. Upon completion of that section, students then move on to several web pages in which the information needed to navigate the actual testing system is provided. This section included a review of the login process, initial name and site confirmation screens, directions on the Student Home page,

and a summary of some important things to know about the system functionality.

Within the MAPT system there are two additional ways learners can prepare for the online test and teachers can gain familiarity with the MAPT. Prior to taking the MAPT for Math or MAPT for Reading all students are required (by their teachers) to complete the Sample Questions in the appropriate subject area. For each area there are four sample questions that provide an overview of the specific skills or knowledge required for that test. The Sample Questions are not meant to test content knowledge and they are specifically designed for a low reading level. In addition, the correct answers for three out of the four items are presented, with the fourth item designed for the student to practice on their own. The Math Sample Questions include clicking radio buttons, calculator usage, and the formula page link. The Reading Sample Questions include clicking on radio buttons, exploring how multiple items can relate to the same reading passage, and reviewing how reading graphics might appear. A raw score is given at the end of the four questions. The students can retake the Sample Questions as often as they like and they can be accessed outside of a testing session to allow for more practice. Teachers and test administrators can also access the Sample Questions so they will also be familiar with these features of the MAPT.

Practice Tests are also available to students, teachers, and test administrators in each subject area for each testing level. The practice tests are a fixed form of 20 questions for each test and the students receive a raw number correct score. The practice tests mirror the test specifications used to guide the operational test development. They help both students and teachers to see the type of content that is tested at each level and serve as another way to determine what level would be most appropriate. Students can access multiple levels and can retake the Practice Tests as often as necessary.

3. Pilot Testing

As described in Chapter II, pilot tests are a critical aspect of test development. Through the pilot test process, items are administered to learners in the target population so that statistical information about the items can be computed and evaluated. The development of the MAPT involved two separate and comprehensive pilot tests. The first pilot test occurred between March 2005 and June 2005 and is referred to as Pilot 1. The second pilot test occurred between October 2005 and January 2006 and is referred to as Pilot 2. In both pilots, items developed for the MAPT for Reading and MAPT for Mathematics and Numeracy were administered over the Internet to adult learners throughout Massachusetts. Trained test administrators managed this process and administered the pilot tests to students during regular class hours. The specific goals of the pilot tests were to:

- (a) evaluate the statistical characteristics of items developed for the MAPT
- (b) evaluate the feasibility of delivering the tests over the Internet
- (c) evaluate ABE learners' ability to take tests on computer
- (d) (Pilot 1 only) select specific, high quality items for setting the scale for Pilot 2
- (e) (Pilot 2 only) select items for operational use

In this section, we describe the procedures followed by the Center for Educational Assessment (CEA) at UMASS Amherst in collaboration with ACLS and UMASS' Center for Educational Software Development (CESD) during pilot testing. Included are descriptions of participation rates and a summary of the results, and a description of how the results were used to develop the operational tests.

Pilot 1 Design

Pilot 1 took place from March through June 2005. To appropriately target the pilot items to the intended learners, multiple fixed-length pilot test forms were created. These forms followed the test specifications stipulated in the reports from the Test Specification Development Committees (Sireci, Baldwin, et al., 2004; Sireci, Li, et al., 2004) and included 5 forms at levels 2, 3, and 4 for both reading and math and two additional forms at math level 5. The booklet design for Pilot 1 is shown in Tables III.3 and III.4. Each form had 30 items: 20 items associated with its given level and 10 items from an adjacent level (or levels). All items from adjacent levels also appeared on forms within their respective level. This design resulted in a total of approximately 100 piloted items at each level (20 items $\times$ 5 forms) for each subject (reading and math).

The Pilot 1 design included some linking items (i.e., items that appear on more than one form), which allowed item statistics for within-subject items to be placed on a common scale for comparison. However, the primary goal of Pilot 1 was to evaluate items relative to quality criteria—not relative to one another—and for this reason, links were weak in some cases. Using less than high-quality links across forms and levels allowed for more items to be piloted and was permissible given that the item statistics were not going to be used operationally.

Table III.3: Spring 2005 (Pilot 1) Test Booklet Design for Reading

Booklet		Level 2 Items				Level 3 Items				Level 4 Items				Number of Items in Booklet		
		20-item blocks	10-item blocks	5-item blocks		10-item blocks				20-item blocks	10-item blocks	5-item blocks				
Level 2 Booklets	L2.1	20												30		
	L2.2		20						10					30		
	L2.3			20							10			30		
	L2.4				10	10			10					30		
	L2.5							5	5	5	5			30		
Level 3 Booklets	L3.1						10			10			10	30		
	L3.2				5			10			10			30		
	L3.3					5				10		5		30		
	L3.4							10					5	30		
	L3.5				10					10				30		
Level 4 Booklets	L4.1										10			30		
	L4.2								10					30		
	L4.3									10				30		
	L4.4										10	10		30		
	L4.5												5	5	5	5
			= block of unique items					= block of common (anchor) items								



= block of unique items



= block of common (anchor) items

Table III.4: Spring 2005 (Pilot 1) Test Booklet Design for Mathematics and Numeracy

Booklet	Level 2 Items			Level 3 Items		Level 4 Items			Level 5 Items		Number of Items in Booklet
	20-item blocks	10-item blocks	5-item blocks	10-item blocks		20-item blocks	10-item blocks	5-item blocks	20-item blocks	10-item blocks	
Level 2 Booklets	L2.1	20									30
	L2.2		20								30
	L2.3			20							30
	L2.4				10	10					30
	L2.5										30
Level 3 Booklets	L3.1				10						30
	L3.2					10					30
	L3.3						10				30
	L3.4							10			30
	L3.5								10		30
Level 4 Booklets	L4.1										30
	L4.2										30
	L4.3										30
	L4.4										30
	L4.5										30
Level 5 Booklets	L5.1										30
	L5.2										30

 = block of unique items = block of common (anchor) items

Pilot 2 Design

A key objective of Pilot 1 was to identify linking items for Pilot 2. In addition to evaluating the statistical characteristics of the pilot items, a key objective of Pilot 2 was to place all items onto a common scale for the operational administration of the MAPT. Pilot 2 occurred between October 2005 through mid-January 2006 and included 4 forms at levels 2, 3, and 4 for both reading and math and 2 additional forms at math level 5. These forms were also developed according to the Test Specification Development Committee recommendations. Each form had 40 items with a minimum of 15 linking items per form. As discussed in the previous paragraph, *linking items* are items that appear on multiple forms, creating a link between forms. Such links allow item parameter estimates from multiple forms to be placed on a common scale through a statistical adjustment called *scaling*, which is discussed in greater detail in Chapter IV. For Pilot 2, it was critical that all items within each subject area (math and reading) were placed on a common scale, so that the item statistics from the pilot 2 could be used in the operational administration of the MAPT. Thus, the booklet linking design was more extensive for Pilot 2 than for Pilot 1. Tables III.5 and III.6 show the Pilot 2 design for Reading and Mathematics and Numeracy, respectively.

Participation

To estimate item parameters using the IRT model selected for the MAPT (i.e., the one-parameter logistic model or 1-PL), it was decided that a minimum of 100 learners needed to respond to each item to properly gauge its difficulty, but our goal was to achieve double that sample size for as many items as possible. Figures III.1 and III.2 show the mean sample sizes for each subject and form level, excluding within-level linking items for Pilot 1 and 2, respectively. It can be seen that the goal of $n \geq 200$ was satisfied for all items except for Pilot 1 math level 4 items, which had an average exposure of $n=166$, Pilot 1 math level 5 items, which had a mean exposure of $n=138$, and Pilot 2 math level 2, which had a mean exposure of $n=192$. For Pilot 1, linking items typically had the same within-level exposure as non-linking items because in general they did not appear on more than one form in a given level. For Pilot 2, on the other hand, linking items typically had twice the within-level exposure as non-linking items because, excepting Mathematics and Numeracy Level 5, all linking items appear on two within-level forms. Figure III.3 shows the within-level exposure of the Pilot 2 linking items.

Table III.5: Fall 2005 (Pilot 2) Test Booklet Design for Reading

Booklet		Level 2 Items			Level 3 Items			Level 4 Items			Number of Items in Booklet
		25-item blocks		5-item blocks	20-item blocks		10- item blocks	25-item blocks		5-item blocks	
Level 2 Booklets	L2.1	25		5			10				40
	L2.2		25				10				40
	L2.3			25				10			40
	L2.4				25	5			10		40
Level 3 Booklets	L3.1			5	20		10			5	40
	L3.2					20	10				40
	L3.3						20	10			40
	L3.4			5			20	10		5	40
Level 4 Booklets	L4.1						10	25		5	40
	L4.2						10		25		40
	L4.3							10		5	40
	L4.4							10		25	40

= block of unique items
 = block of common (anchor) items

Table III.6: Fall 2005 (Pilot 2) Test Booklet Design for Mathematics and Numeracy

Booklet		Level 2 Items		Level 3 Items		Level 4 Items			Level 5 Items		Number of Items in Booklet
		25-item blocks	5-item blocks	20-item blocks	10-item blocks	20-item blocks	10-item block	5-item blocks	25-item blocks	5-item block	
Level 2 Booklets	L2.1	25		5			10				40
	L2.2		25		5		10				40
	L2.3			25	5			10			40
	L2.4				25	5		10			40
Level 3 Booklets	L3.1			5		20		10			40
	L3.2				5		20		10		40
	L3.3					5		20		10	40
	L3.4				5			20		10	40
Level 4 Booklets	L4.1					10		20		5	40
	L4.2					10			20		40
	L4.3						10		20		40
	L4.4						10			10	40
Level 5 Booklets	L5.1								5	5	40
	L5.2								5	5	40

□ = block of unique items ■ = block of common (anchor) items

Figure III.1

Pilot 1: Mean Item Exposure by Subject and Test Form Level (excluding within-level linking items)

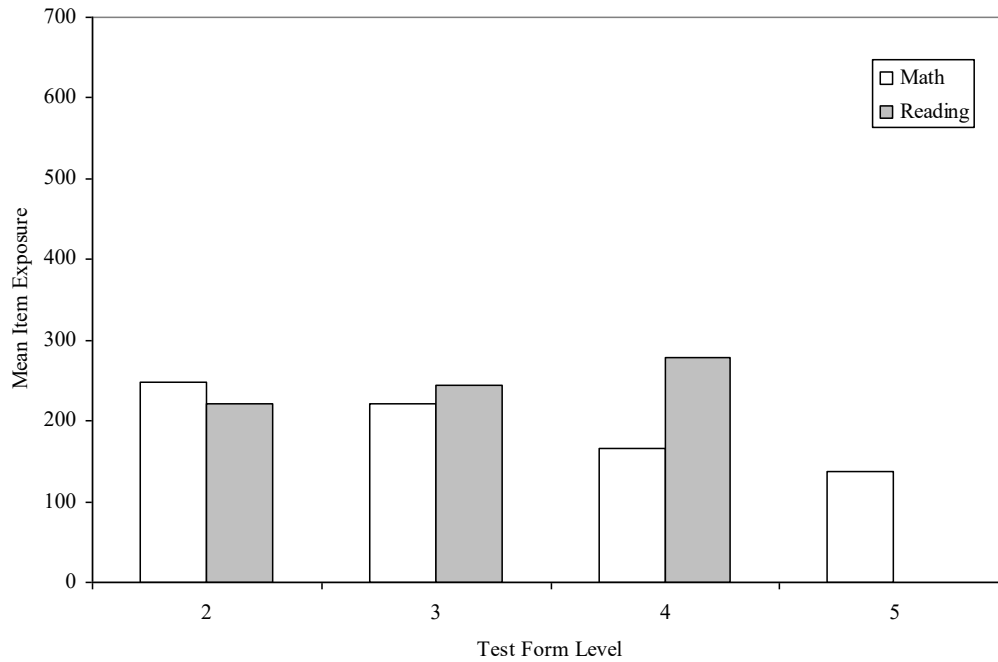


Figure III.2

Pilot 2: Mean Item Exposure by Subject and Test Form Level (excluding within-level linking items)

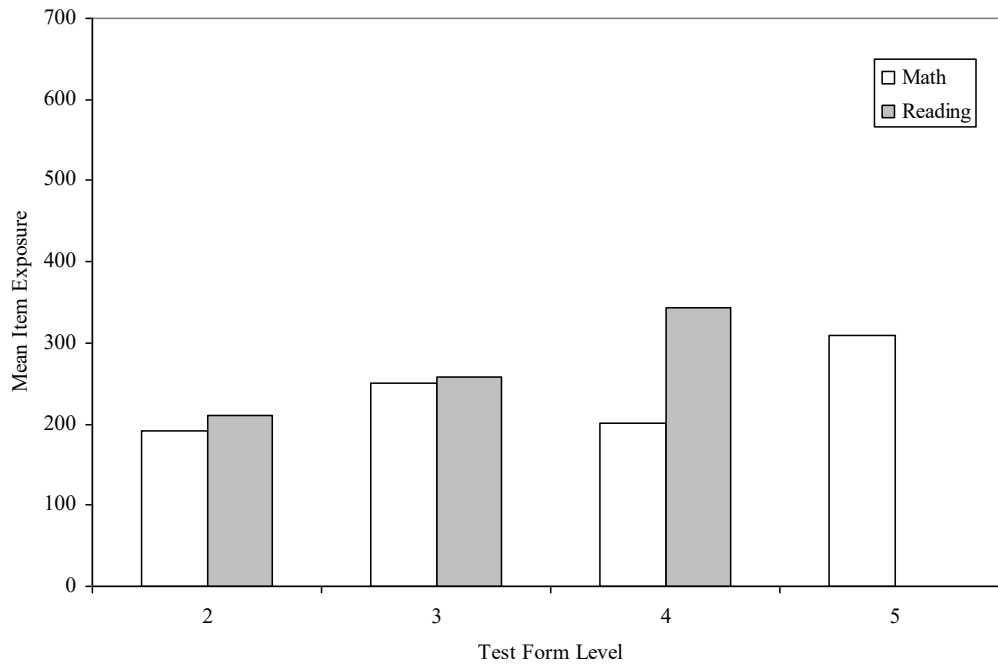
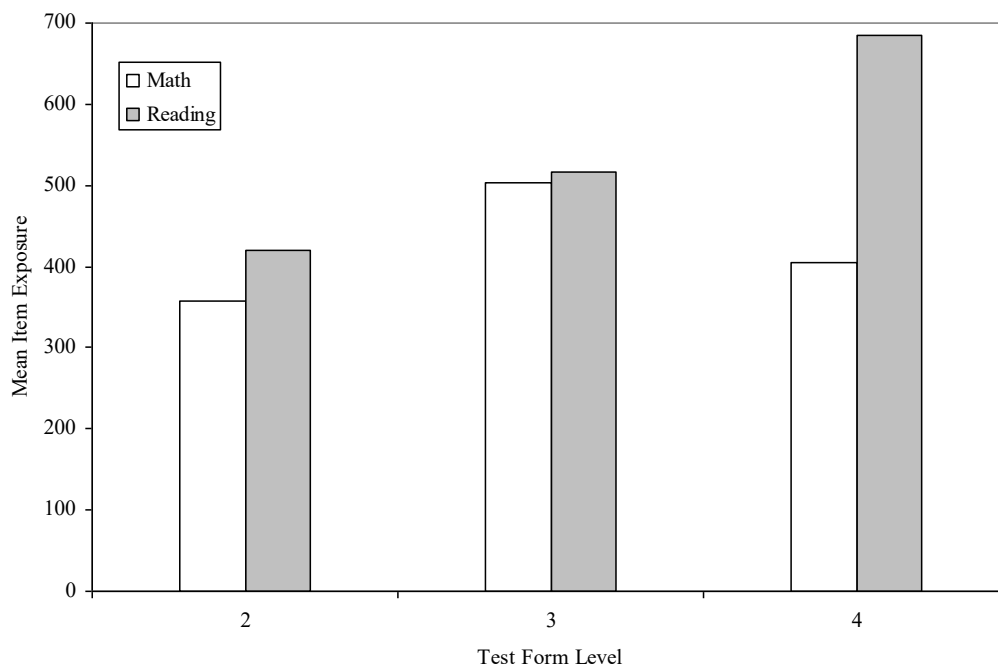


Figure III.3

Pilot 2: Mean Item Exposure by Subject and Test Form Level (within-level linking items only)



4. Item Analysis

A total of 1,101 unique items were evaluated during Pilots 1 and 2, including 239 items that appeared on both Pilot 1 and Pilot 2. For each item, classical item difficulty and discrimination statistics were computed³. Figure III.4 shows an example of the item analysis statistics computed for each Pilot 1 item (similar information was gathered for Pilot 2). Only one item is illustrated in the figure, but a complete item analysis report for all items is available.

The top of the item analysis report provides the name of the item writer, the subject area, and test level. Next, the number of learners who responded to the item is shown, in the “item exposure” field. Many of the items appeared on multiple test forms each targeted to a different level and therefore the frequencies for each level are reported when applicable. In the “response summary” field, the proportions of learners (at each level) who correctly answered the item are reported along with the average number of seconds it took learners to respond to the item. These proportion correct statistics are referred to as “p-values” and are an index of item difficulty. The next section reports the numbers and proportions of learners who selected each of the four

³ For the first pilot tests, the Winsteps IRT software was used for item analysis and the item infit and outfit statistics were also analyzed. For pilot test 2, and for the operational calibration, Bilog software was used and so both item fit (residual plots, chi-square values) and classical item statistics were used to statistically evaluate the items. Only the classical results are described in this manual, but we do address item fit in Chapter VII. For pilot 1, we found that essentially the same items were flagged whether we used corrected point-biserials, infit/outfit, or residual analysis. Analysis of differential item functioning across groups defined by sex, ethnicity, and native language (English vs. other) were also conducted, where sample sizes permitted, and are summarized in Han, Baldwin, and Sireci (2006).

response options for the item. The correct answer is denoted with an asterisk. The correlation between the selection of each response and total test score is also reported. Items that “perform” as expected will have large and positive correlations (i.e., $\geq .25$) for the correct answer and negative correlations for the incorrect response options.

The “point-biserial” statistics for the item are presented next. The point-biserial is a measure of “item discrimination.” It represents the correlation between item scores (0 if answered incorrectly and 1 if answered correctly) and total test scores. Since this statistic is a correlation, it ranges from -1.0 (perfect negative correlation) to $+1.0$ (perfect positive correlation), with values at or above $.25$ reflecting an acceptable relationship between learners’ performance on an item and their total test score. For Pilot 1, total test score was computed by summing all the items on a test form correctly answered by each learner, minus the item of interest. Positive point-biserials indicate that the item is performing as expected—i.e., higher-scoring examinees are answering the item correctly more often than lower-scoring examinees. Thus, the closer the point-biserial is to 1.0 , the better. The point-biserial statistic can also be found in the distractor analysis for the row representing the correct answer. For convenience, it is repeated separately so that the statistic can be viewed easily when reviewing lots of items. The “infit” and “outfit” statistics indicate how well the item fits the 1PL IRT model. These statistics provide similar information to the point biserial, but are on different scales. The last section of the item analysis report is a histogram that illustrates the distributions of item difficulties for all items within a subject area and for all items within that test level. The difficulty location of the item of interest is represented by a small diamond ($\diamond$) on the x-axis of the figure. These item difficulties are not the same as the p-values described earlier. Rather, they are IRT difficulty parameters reported on a scale ranging from about -4.0 to 4.0 , with a mean of 0 and a standard deviation of 1.0 for level 3 items. On this scale, a larger number indicates a more difficult item.

As can be seen in Figure III.4, the item analysis report provides a great deal of information for each item (e.g., item difficulty, item discrimination, item response time, and how well the distractors are functioning). In selecting the “best” items to be used on Pilot 2, the point-biserial statistics were of particular interest because they are good indicators of how well an item will perform when calibrated using item response theory (IRT). We excluded most Pilot 1 items with point biserials less than $.25$ from the item calibration process. The specific criteria for inclusion in item calibration were:

1. Minimum of 150 people responded to item (minimum of 120 for math Level 5)
2. Corrected point-biserial greater than $.25$ (on at least one form, if item appears on multiple forms) *or* proportion correct (p-value) less than $.15$ or greater than $.85$

The minimum N criterion helped ensure that the item characteristics could be reasonably estimated. It can be seen that when only the minimum number of examinees ($N = 150$) responded to an item and when p-values are extreme (i.e., very easy or very hard items), discrimination indices must be based on performance differences of an extremely small number of examinees ($150 \times .15 \approx 23$). Under such conditions, corrected point-biserials may be difficult to estimate. Thus, the minimum corrected point-biserial criterion of $.25$ was relaxed for items that had extreme p-values ($<.15$ or $>.85$). This allowed potentially usable items to be included in the second round of pilot testing before a final determination of their suitability for operational use is made.

Figure III.4

Sample Item Analysis Report for a Level 2 Math Item

author: ckotori
attributed to: Glenn

Math Item 1008
Level 2

Item Exposure		
L2F2	L3F4	Total
240	225	465

Response Summary

	level 2	level 3	level 3-2
p-value	.65	.81	.15
mean response time (sec.)	45	40	-5

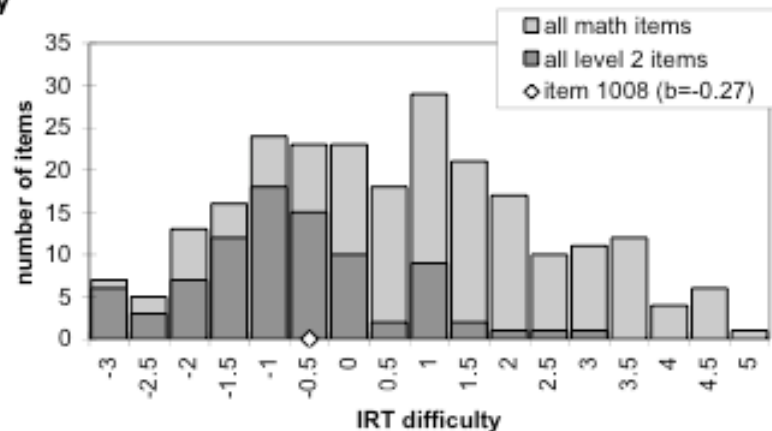
Distractor Analysis

	% (N)			correlation with total score	
	level 2	level 3	all levels	level 2	level 3
A	3.3(8)	3.1(7)	3.2(15)	-.05	-.09
B*	65.4(157)	80.9(182)	72.9(339)	.40	.32
C	24.2(58)	12.4(28)	18.5(86)	-.26	-.25
D	5.4(13)	3.1(7)	4.3(20)	-.25	-.13
omit	1.7(4)	0.4(1)	1.1(5)	-	-

Model Fit

	L2F2	L3F4	
point biserial	.40	.32	> .4: good
in fit	.94	Productive	
out fit	.84	Productive	

IRT difficulty



For the item shown in Figure III.4, we can see it has a p-value of .65 for level 2 learners and .81 for level 3 learners. This is consistent with our expectations—lower-level learners had less success with the item than higher-level learners. This item also performed well with respect to the point-biserial criterion: .40 and .32 for level 2 and 3 learners, respectively. It is also interesting to note that the time it took for learners to respond to the item decreased as level increased, which is also consistent with expectation. Thus, this is an item that appears to function as intended and would probably be selected for Pilot 2.

Unfortunately, not all piloted items exhibited the same pattern of results. One example is presented in Figure III.5. For this item, level 4 learners performed slightly better than level 5 learners, although the most striking finding is that the item was very difficult for learners at both levels, and learners took a substantial amount of time to answer the item. This particular item was removed from the pool of potential items and was not selected for operational use.

A summary of the item difficulty and discrimination statistics for all Pilot 1 items is presented in Table III.7. This table provides the mean p-value and point-biserials for (a) all Pilot 1 items, and (b) the subset of Pilot 1 items that passed the statistical screening criteria.

Table III.7

Mean P-Values and Point-Biserials for Pilot 1 (Spring 2005) Items

	Mean (standard deviation) p-value	Mean (standard deviation) corrected point-biserial
All piloted items	.63 (.22)	.27 (.12)
Subset of items that passed the statistical screening criteria	.67 (.22)	.31 (.11)

Figures III.6 and III.7 show the numbers of Pilot 1 items that passed the statistical screening criteria within each level for Reading and Math, respectively. Overall, about 2/3 (69%) of the items piloted passed our statistical screening criteria. This success rate allowed 448 of the 650 items to be eligible for Pilot 2.

Figure III.5

Sample Item Analysis Report for a Level 4 Math Item

author:

Math Item 444

attributed to:

Level 4

Item Exposure		
L4F3	L5F5	Total
155	66	221

Response Summary

	level 4	level 5	level 5-4
p-value	.21	.17	-.04
mean response time (sec.)	87	88	2

Distractor Analysis

	% (N)			correlation with total score	
	level 4	level 5	all levels	level 4	level 5
A	23.9(37)	40.9(27)	29.0(64)	.03	.04
B	7.1(11)	7.6(5)	7.2(16)	-.02	-.19
C	47.1(73)	34.8(23)	43.4(96)	.03	-.01
D*	20.6(32)	16.7(11)	19.5(43)	-.11	.08
omit	1.3(2)	0.0(0)	0.9(2)	-	-

Model Fit

	L4F3	L5F5	
point biserial	-.11	.08	malfunctioning item
in fit	-	excluded from calibration	
out fit	-	excluded from calibration	

IRT difficulty

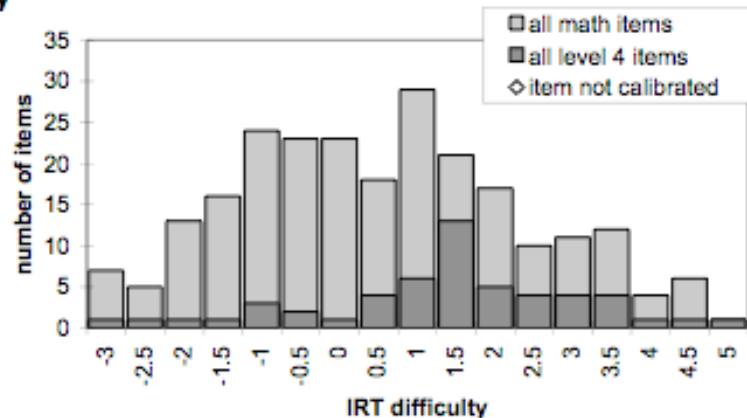


Figure III.6

Approximate Frequencies of Pilot 1 Reading Items Included/Excluded from IRT Calibration

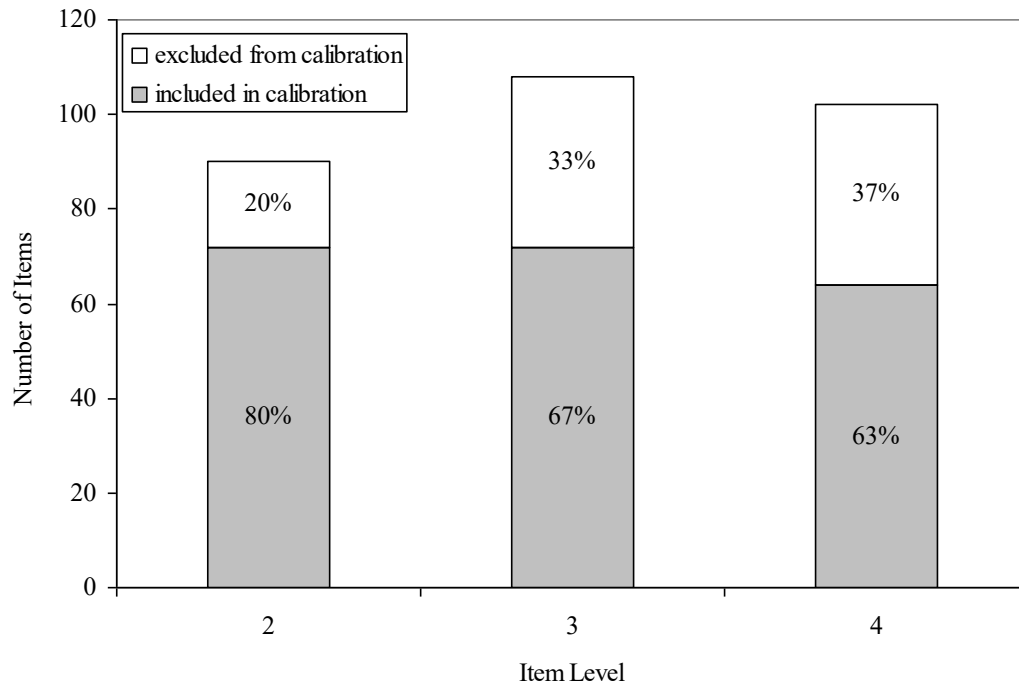
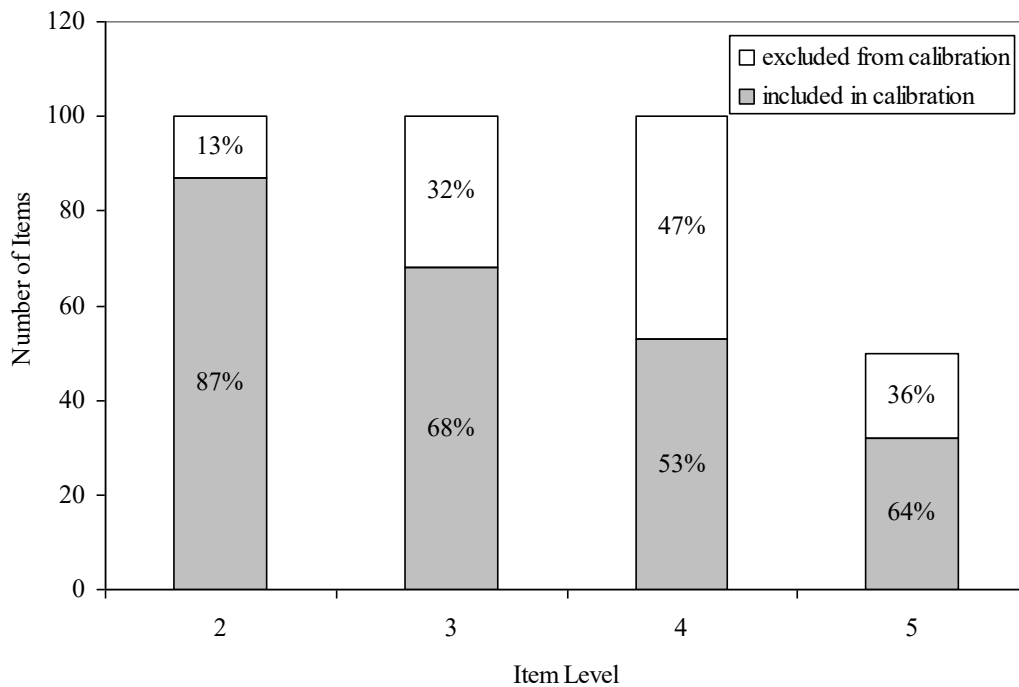


Figure III.7

Approximate Frequencies of Pilot 1 Math Items Included/Excluded from IRT Calibration



For Pilot 2, all items with point-biserials greater than .10 were calibrated. For content reasons, it was necessary to include a small number (<6%) of items with relatively low point-biserials (between .13 and .25) when assembling the operational forms. Table III.8 shows the mean p-value and point-biserials for (a) all Pilot 2 items, and (b) the subset of Pilot 2 items that comprise the operation forms.

Table III.8

Mean P-Values and Point-Biserials for Pilot 2 (Fall 2005) Items

	Mean (standard deviation) p-value	Mean (standard deviation) point-biserial
All piloted items	.61(.20)	.37 (.11)
Subset of items that are being used operationally	.60 (.17)	.40 (.10)

Figures III.8 through III.11 show the distribution of point-biserials and IRT difficulty (*b*-parameters) for the subset of Pilot 2 items that were selected for operational use. The *b*-parameter figures (III.10 and III.11) are also of interest because they represent the relationship between intended test level and item difficulty. One advantage of the IRT difficulty scale is that items across levels can be placed on the same scale, which allows relationships across levels (and forms) to be examined (unlike p-values, which are level and form dependent). The expectation is that as level increases, so does item difficulty.

For math, the items increase in difficulty as intended level increases: each distribution of level-specific item difficulties moves to the right from one test level to the next. The results for reading are less consistent with considerably more overlap across levels. In particular, the distributions of level 2 and 3 reading item difficulties exhibited more overlap than expected. Thus, learners who start the test at either of these entry points may see items of similar difficulty, particularly at stage 1. However, the routing rules and standard setting procedures help ensure that learners at all levels are administered items of appropriate difficulty, given their current proficiency estimate.

Figure III. 8

Distribution of Point-Biserials by Level (Operational MAPT for Reading)

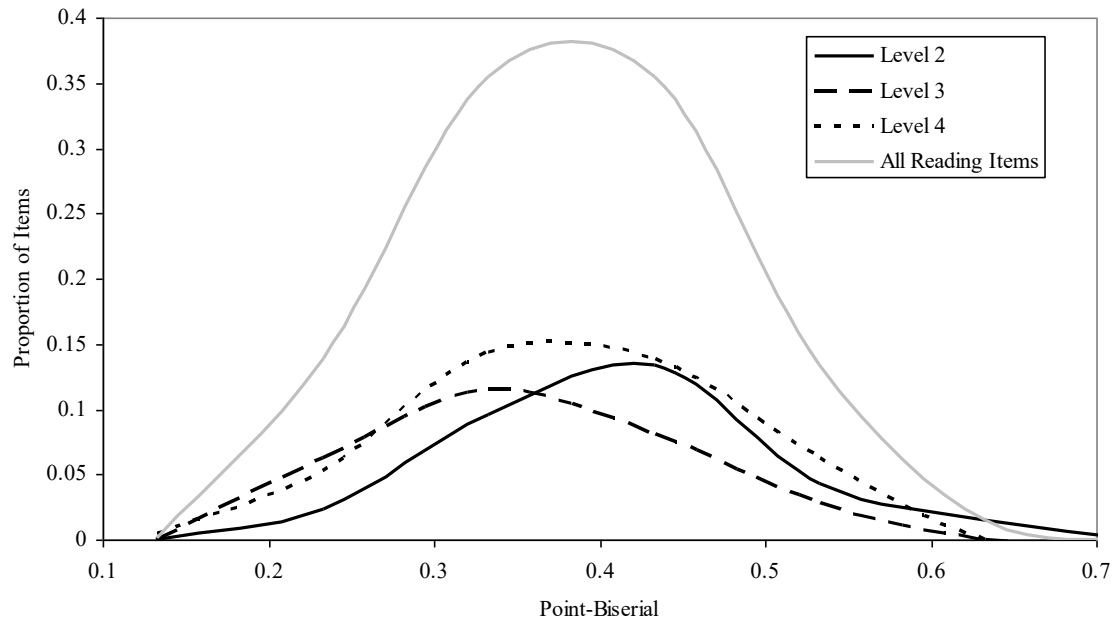


Figure III.9

Distribution of Point-Biserials by Level (Operational MAPT for Math)

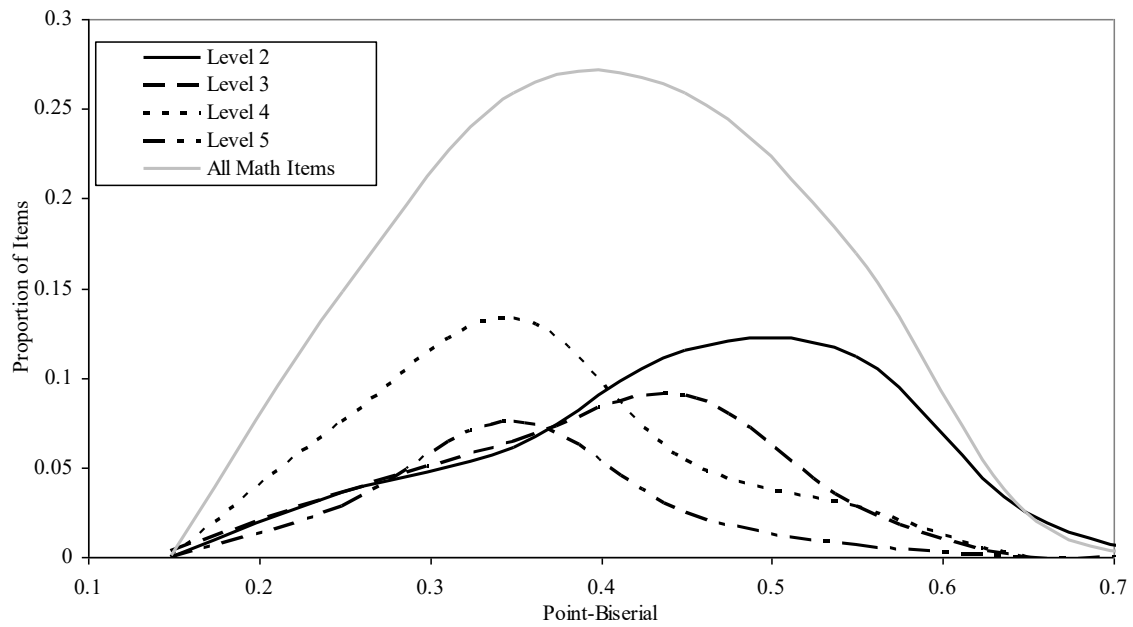


Figure III.10

Distribution of IRT Difficulty by Level (Operational MAPT for Reading)

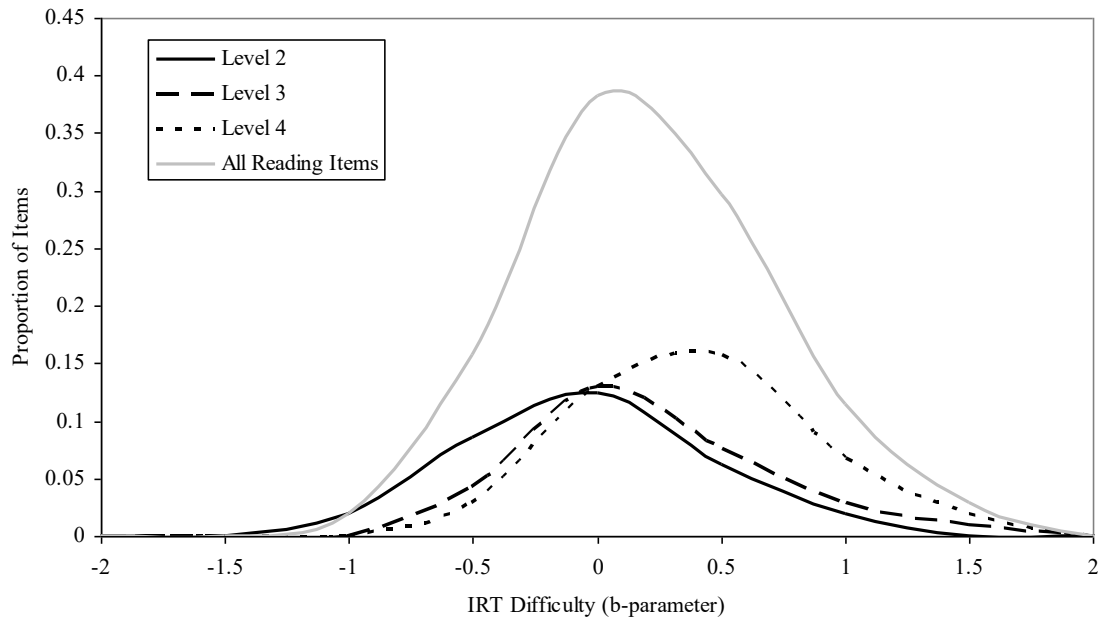
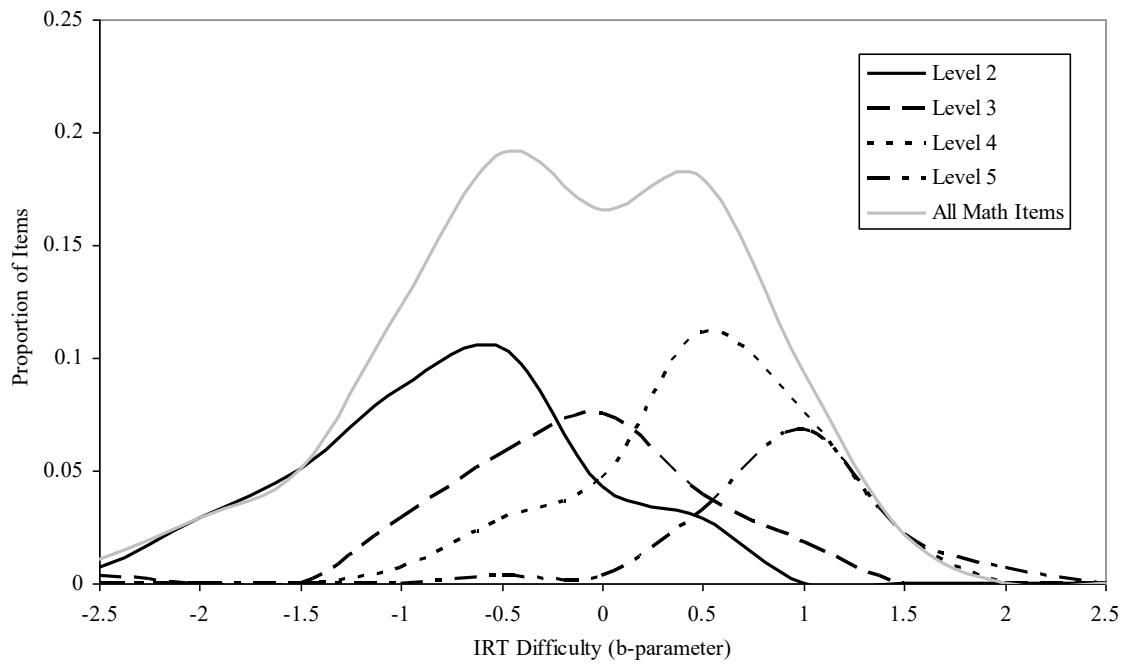


Figure III.11

Distribution of IRT Difficulty by Level (Operational MAPT for Math)



Item response time

The amounts of time it took learners to answer each item and to complete the test were also calculated. These analyses were conducted to ensure learners would have sufficient time for completing the exam and to further evaluate item functioning. Table III.9 presents a summary of the data for Pilot 1, in which 2,817 learners took a 30-item Mathematics pilot test, and 2,758 learners took a 30-item Reading pilot test. The results across the two tests were similar. On average, learners spent about 53 seconds per item on the math items and 55 seconds on the reading items, which resulted in a total testing time of less than 30 minutes.

Table III.9

Descriptive Statistics for Item and Test Completion Times for Pilot 1 (30 Items)

Test		Mean	SD
MAPT for Mathematics and Numeracy	Item Response Time (seconds)	53.2	18.9
	Time to Complete Test (minutes)	26.6	9.5
MAPT for Reading	Item Response Time (seconds)	54.7	20.8
	Time to Complete Test (minutes)	27.3	10.4

The results for Pilot 2 (n=3,284 for Math and n=3,254 for Reading) are presented in Table III.10. The average time to respond to an item went up about 6 seconds for Math, but went down about 4 seconds for reading. The mean time to complete the test was about 39 minutes for math and 34 minutes for reading. The vast majority of learners (93%) completed the test within one hour. The MAPT is not intended to be speeded and so the time limit set for the exam is due to practical factors such as scheduling learners to take the test. The current time limit is two hours. The results from the pilot test suggest all learners should have no trouble completing the test within the two-hour time limit and that the vast majority will finish in half that time. However, for learners who need additional time, additional time will be provided.

Table III.10

Descriptive Statistics for Item and Test Completion Times for Pilot 2 (40 Items)

Test		Mean	SD
MAPT for Mathematics and Numeracy	Item Response Time (seconds)	58.8	51.3
	Time to Complete Test (minutes)	39.2	16.3
MAPT for Reading	Item Response Time (seconds)	51.2	44.0
	Time to Complete Test (minutes)	34.1	12.9

5. Sensitivity Review

Sensitivity review refers to the process of having a diverse group of professionals review tests to flag material that may unintentionally interact with demographic characteristics of some test takers. As the *Standards for Educational and Psychological Testing* (AERA, et al., 1999) describes

Test developers should strive to identify and eliminate language, symbols, words, phrases, and content that are generally regarded as offensive by members of racial, ethnic, gender, or other groups, except when judged to be necessary for adequate representation of the domain. (AERA, et al., 1999, p. 82)

Although item writers and item reviewers were asked to consider issues of diversity and sensitivity when constructing and reviewing items, we believed it was important to add a separate, independent review that would concentrate solely on sensitivity issues. The sensitivity review process for the MAPT sought to do more than identify and eliminate test content that may be construed as offensive (Ramsey, 1993; Sireci & Mullane, 1994). Rather, the MAPT sensitivity review also sought to identify test material that was irrelevant to the construct measured, but may provide an advantage or disadvantage to learners from different backgrounds. That is, our sensitivity reviews attempted to uncover unintended bias in the MAPT. Specifically, the MAPT sensitivity reviewers were asked to review pools of MAPT items for material that may

- (a) be construed as offensive to particular groups of learners,
- (b) portray groups of learners unfavorably, or in a stereotypical fashion,
- (c) be advantageous to one group, and/or disadvantageous to another, or
- (d) be unfamiliar to certain groups of learners.

Sensitivity review is often seen as a final, qualitative check on potential bias in a test⁴. Unlike a technical content review, the reviewers need not necessarily be experts in the subject domain tested, although such expertise is desirable. Instead, the reviewers are selected for their knowledge of specific cultural groups and how test takers from such groups may interpret test material. Sensitivity reviewers are typically a diverse group consisting predominantly of minority group members. By reviewing tests for potential bias, sensitivity review can improve the content validity of a test, but it should not be confused with a content review. As Sireci and Mullane described

“[test developers] strive to ensure that the content of their tests is representative of and relevant to the domain tested, and that trivial or superfluous material is not present. Therefore, it can be argued that sensitivity review is built into the test development process under the rubric of content validation. After all, if the entire content of a test has been demonstrated to be relevant to the content domain, there should be no material contained therein that is offensive or unfamiliar to examinees who are knowledgeable of the domain. This argument is flawed, however, because evaluations of test content focus on how well the items that comprise a test represent the content domain, rather than on evaluating the context within which the items are framed. (pp. 22-23)

Eight sensitivity reviewers were hired to review MAPT items. Five of the reviewers were men and three were women. The reviewers were selected to represent a variety of backgrounds including African/African-American (4), Hispanic/Latino (2), Asian-American (1), and Euro-American (1). The reviewers taught or worked with adult learners from a variety of backgrounds, particularly new immigrants, learners with limited English proficiency, and traditionally under-represented minority groups such as Hispanic/Latino, Haitian, Asian, and African American, and other ethnic groups within the adult education community in Massachusetts such as Polish, Russian, Cape Verdean, and Vietnamese.

The reviewers were trained in the sensitivity review process during a four-hour meeting held in Worcester, Massachusetts. The principles of sensitivity review were discussed and instructions were given regarding the types of problems to look for in items and their associated text and stimuli (e.g., reading passages, graphs, tables, etc.). Following these discussions, panelists reviewed sample items that had been flagged for reasons of sensitivity in different testing programs (see Ramsey, 1993; Sireci, 2004; and Sireci & Mullane, 1994 for some examples of these types of items) and the issues associated with these items were discussed as a group.

After training, the sensitivity reviewers were given a booklet containing 120 MAPT items, with about half coming from each subject area. The item booklets were partitioned into sets of about 30 items to facilitate the review process. The booklets were assembled so that all MAPT items being considered for operational use were reviewed by at least two (and up to four) sensitivity reviewers. The reviewers were also given sets of Sensitivity Review Rating Sheets

4 Analyses of differential item functioning are also conducted on the MAPT, where sample sizes permit, as a quantitative check on potential item bias. These analyses were conducted on pilot 2 data and included analysis across groups based on sex, ethnicity, and native language (English vs. other). See Han, Baldwin, and Sireci (2006) for a summary of these analyses.

(see Appendix D). They were asked to review all items and provide any comments regarding items with potential sensitivity issues on a review sheet. When they were unsure whether an item could be problematic, they were told to err on the side of flagging the item and to provide a description of why the item might have a sensitivity issue. The rating sheet also asked whether the packet of test items was appropriate for 12 specific groups of learners (see Appendix D). The reviewers spent two hours reviewing items at the meeting and were asked to complete their remaining reviews within a one-week period. All reviewers completed this task within the requested time period.

The sensitivity reviewers provided numerous comments on items and reading passages. Approximately 40 items were flagged for various reasons. Examples of why items were flagged included a lack of diversity in the names that were associated with the items (particularly in math problems), concerns over terms related to money, negative stereotypes for certain body types associated with reading passages, and concerns over items that described violence during World War II.

All comments received from the sensitivity reviewers were carefully considered by the test development team at UMASS, which included test development staff and a Professor of Psychometrics with 17 years of experience in test development and sensitivity review. Changes were made to approximately 20 items in response to the reviewers' comments. In some cases, an offensive sentence was removed from a reading passage, after concluding the sentence was not needed to correctly answer any items associated with the passage. In other cases, more diverse names, such as "Julio," "Djata," and "Zoe," were used to replace names such as "John" and "Sue." Some items were left unchanged because it was concluded the issue raised by the sensitivity reviewer was content-relevant. For example, one sensitivity reviewer thought that some learners would not understand the term "dime," on a math item measuring "carry out basic calculations with money." This issue was discussed with members of the Math Content Validity Panel who concluded the term was necessary for appropriately measuring the objective.

To summarize the MAPT sensitivity review, a carefully selected panel of reviewers who were familiar with learners from a variety of backgrounds, were trained to review items for sensitivity issues. These reviewers flagged many items and explained why they should be discussed before being considered for operational use. UMASS staff reviewed and considered all comments received from the sensitivity reviewers and made several changes to many items.

6. Development of MST Modules and Panels

To address the challenge of testing a population with a heterogeneous distribution of proficiency, the MAPT uses a multistage adaptive test design. The basic concepts underlying multistage testing (MST) were described in Chapter II. In this section, we describe how the MAPT modules and panels were developed. A *module* is a group of items administered to learners at a particular stage of the test. A *panel* is a collection of modules that spans the learner levels and stages.

In the MAPT MST design, tests are divided into three stages. After completing a set of items (referred to as a *module*) at stage 1, learners are routed to a stage 2 module targeted to their proficiency as estimated by the number of items they correctly answered in stage 1. After

completing the stage 2 module, they are routed to a stage 3 module based on their performance on the module at stage 2.

For the purpose of routing, *performance* refers to the number of correct responses in a given module—not IRT proficiency estimates. Nevertheless, even though routing decisions were based on a number correct, routing rules were determined such that IRT information would be maximized on average. (IRT information, which is discussed in greater detail in section VI.1, is somewhat analogous to the concept of reliability.) This was accomplished by determining the various IRT proficiency estimates—i.e., thetas (θ) or theta ranges—that produced the expected (number correct) scores of 0, 1, ... n , where n is the number of items in the module being exited. Each such expected score was then matched to whichever subsequent module yielded the maximum information at its corresponding θ , with the exception that routing was limited to adjacent modules only. These matches were then taken as the routing rules. For the learner, the presentation of items is seamless—all routing decisions are made within the test delivery system. Upon completion of the entire test (i.e., all three stages), examinee IRT proficiency is estimated based on *all* of their responses.

Figures III.12 and III.13 illustrate the multistage, multi-panel adaptive test structure used for the MAPT. Note that there are two sets of stages and modules for each subject (one light grey and one dark grey). These sets are called *panels* and in no instance does routing occur across panels. Furthermore, each panel comprises a unique set of items. First-time learners are randomly assigned to one of two parallel panels. In subsequent administrations, a learner is randomly assigned a panel from those panels he or she has *not* yet seen. If this is not possible (because the examinee has seen all panels), the learner is assigned the panel he or she has seen least recently. Even though considerable time (at least 2 months) may elapse between test administrations, this design minimizes that chance that a learner will have an unfair advantage as a result of being previously exposed to an item.

It should be noted that the test administration architecture for the MAPT is controlled by the Online Web-based Learning (OWL) system developed by the Center for Educational Software Development (CESD) at UMASS. The test selection and item routing rules were developed by the Center for Educational Assessment and were implemented in collaboration with CESD. The OWL system keeps track of a learner's MAPT history so that appropriate panel assignments can be made.

Figure III.12: Multistage, Multi-Panel Test Structure (Reading).

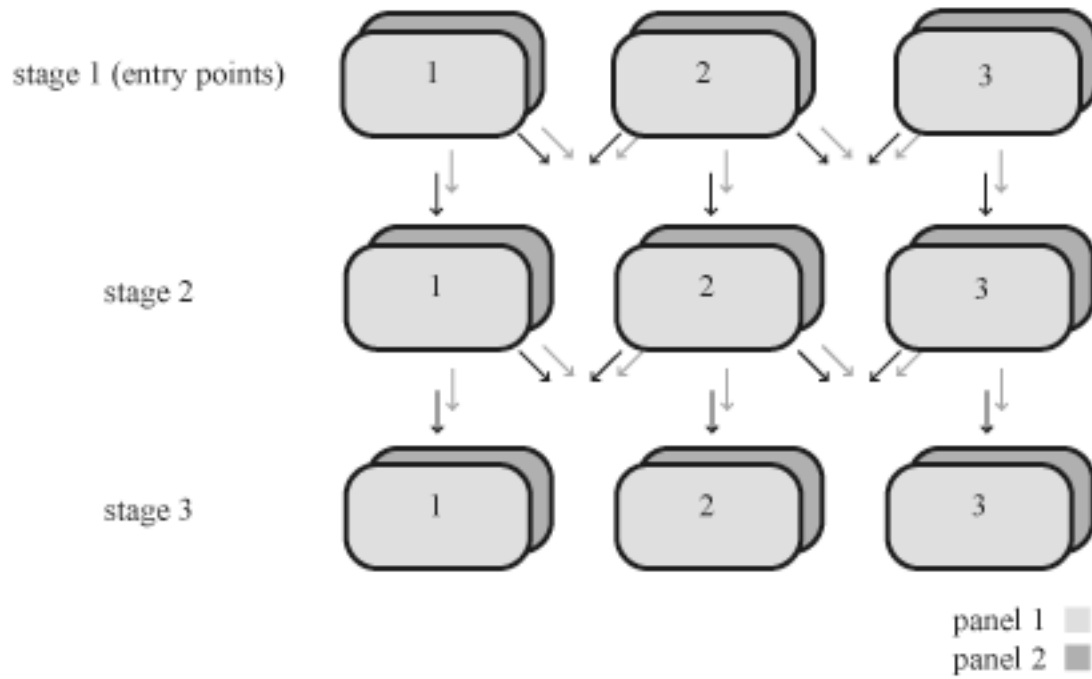
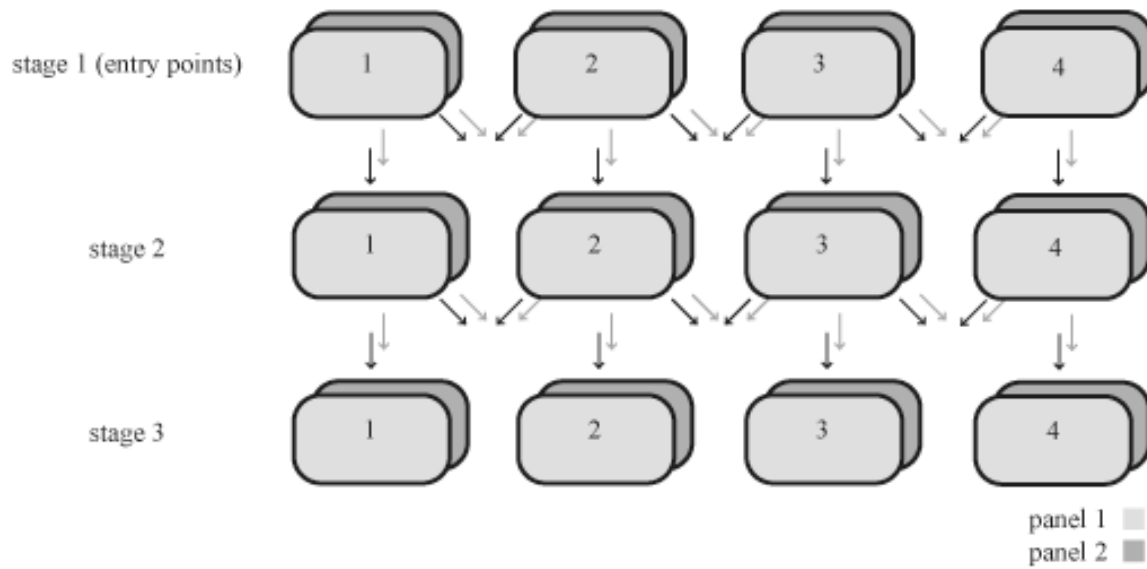


Figure III.13: Multistage, Multi-Panel Test Structure (Math).



IV. Developing the MAPT Score Scale

1. Item Response Theory Calibration

Item difficulty and proficiency scores for examinees were estimated using the 1-parameter logistic (1PL) item response theory (IRT) model, which is defined as (Hambleton & Swaminathan, 1985):

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}}$$

where:

P_i is probability of correct answer on item i ;

θ (theta) is examinee proficiency;

b_i is item difficulty parameter;

D is a scaling factor equal to 1.7;

e is base of the natural logarithm.

Two advantages of IRT models are that the model parameters (item difficulty and examinee proficiency) are on a common scale and that these parameters are invariant—up to a linear transformation—for a given population of examinees. The qualifier *up to a linear transformation* is necessary because transforming the scale by adding a constant to b_i and θ does not affect the probability of a correct response. That is,

$$P_i(\theta) = \frac{e^{D(\theta - b_i)}}{1 + e^{D(\theta - b_i)}} = \frac{e^{D((\theta + k) - (b_i + k))}}{1 + e^{D((\theta + k) - (b_i + k))}}$$

where k is some constant value. The implication of this scale indeterminacy is that separate estimations (calibrations) of the model parameters cannot be assumed to be on a common scale. When multiple calibrations are needed to estimate model parameters, a statistical adjustment must be made to place these estimates on a common scale. This was the case when item difficulties were estimated using examinee response data from the Fall 2005 pilot test.

The Fall 2005 pilot test (Pilot 2) included 26 unique test forms (12 Reading and 14 Mathematics and Numeracy). (See Chapter III, Section 3 for the Pilot 2 booklet design.) For Pilot 2, a test form, or booklet, is a set of 40 items targeted to a specific *Educational Functioning Level* (EFL) (DAEL, 2006). For each subject, items were placed on a common scale by first concurrently calibrating all within-level test forms and then transforming the resulting parameter estimates to the level 3 metric using the mean equating method in an anchor test design. The level 3 scale was defined by transforming the difficulty estimates for items appearing on level 3 forms such that their mean was zero.

The Pilot 2 linking design was chosen as a reasonable compromise between the competing goals of maximizing the number of unique items piloted and maximizing the number of linking items, which ensures that items can be confidently placed on a common scale. In theory, within-level links were not necessary because examinees were randomly assigned forms within a given level and therefore the randomly equivalent examinee samples can serve as the links across forms rather than linking items. Nevertheless, approximately 30 linking items were distributed across the four forms per level (with a maximum anchor length of 10 items) so that the assumption of equivalency could be verified. Whether linking by items or examinee sample, item parameter estimates were highly similar suggesting that the assumption of equivalent groups was met.

At each level, the same 30 within-level linking items were also used, *en masse*, to link to the adjacent level (e.g., level 2 to level 3). Additionally, parameter estimates for linking items were further adjusted by averaging each estimate from the destination scale (generally level 3) with the newly transformed estimate. For example, for some linking item i appearing on both level 2 and level 3 forms:

$$b_{i_{final}} = \frac{b_{i_2}^* + b_{i_3}}{2}$$

where $b_{i_{final}}$ is the operational b-parameter (difficulty) estimate for item i , $b_{i_2}^*$ is the b-parameter estimate for item i from the level 2 calibration transformed to the level 3 scale using mean equating, and b_{i_3} is the b-parameter estimate for item i from the level 3 calibration.

To the extent possible, the content on each form as well as each anchor test was selected to match the test specifications. All item calibrations were done with BILOG-MG (2002) using the RASCH scaling keyword and the default settings.

2. Establishing the MAPT Score Scale

As discussed above, the indeterminacy of the IRT metric was resolved (i.e., the model was identified) by setting the distribution of item difficulty for level 3 forms to have a mean of zero. However, scores reported on this metric can be difficult for users to interpret. Consequently, proficiency estimates are transformed to the MAPT score scale, which goes from 200 to 700, for reporting purposes. Transformations were determined such that the cut-point between the NRS Educational Functioning Level (EFL) 1 and 2 corresponds to a MAPT score of 200; the cut-point between EFL levels 2 and 3 corresponds to a MAPT score of 300, and so on. (The processes for establishing the EFL cut-points are described in Chapter IV.) This transformation strategy was chosen entirely to aid in score interpretation; for example, a learner scoring in the 200's would be classified as performing at EFL 2.

Figures III.13 and III.14 show the transformation functions for MAPT Reading and MAPT Math, respectively. Note that the transformation is highly—but not perfectly—linear within the range of reported scores. To the extent that the transformation is *not* linear, the theoretical equal interval property of the theta metric will not exist for the MAPT score scale except conditionally in the ranges 201-399, 400-499, and 500-699. Because equal interval

measurement is often desirable, an additional word about the decision to compromise it is warranted.

Equal interval scales have attractive properties—most notably the property implied by their name: the comparability of equal size intervals throughout the scale range. A calendar is an everyday example of an equal interval scale: 10 days in January refer to the same quantity of time as 10 days in October. For test score scales, equal interval measurement means that a 10 point gain at, say, level 2 refers to the same increase in amount of proficiency as 10 points gained at level 4. This property allows certain arithmetic operations and certain comparisons among values to be meaningfully made. However, when single scores are of interest (e.g., an examinee wishes to know how many points he or she needs to move to the next EFL), the advantages of equal interval measurement are not needed. Equal interval measurement is also not necessary for measuring gain under the NRS because in this context gain is defined by advancement from one level to another *not* the magnitude of gain. Indeed, the advantages of equal interval measurement are only present when summarizing or comparing scores.

Because a non-linear score scale transformation is made, the arithmetic mean and standard deviation cannot meaningfully be computed for MAPT scores except conditionally in the ranges 201-399, 400-499, and 500-699. (Of course, these statistics can still be computed for MAPT scores but they are arguably not meaningful or at least not as meaningful.) For this reason, it is recommended that θ scores be used for internal analyses that require computation of these statistics.

In addition to permitting certain arithmetic operations, the equal interval property allows certain comparisons to be made. For example, it may be informative to compare the effectiveness of two learning centers in raising proficiency (as measured by test score gains) even when one program teaches high-proficiency learners while the other teaches low-proficiency learners. Again, θ scores are the appropriate scores to use in such comparisons. Still, while equal interval scales support such comparisons theoretically, it may be inappropriate to make such comparisons in practice. For example, it may not be meaningful to compare reading gains for low-literacy learners to gains made by high-literacy learners, even when these gains are reported on a common equal interval metric. In this case, even though the construct of interest is *reading* for both low and high literacy learners, low literacy reading and high literacy reading may refer to very different skill sets and cognitive processes that cannot be meaningfully compared.

In summary, for reporting individual scores or for computing gain under the NRS, MAPT scale scores are convenient and comprehensible. When summarizing multiple scores or comparing scores gains spanning multiple educational functioning levels, it is advisable to use IRT proficiency scores rather than MAPT scale scores. Moreover, caution should be exercised when comparing scores or score gains that differ greatly because differences in the construct being measured at widely divergent points along the scale may cast doubt on the meaningfulness and appropriateness of such comparisons.

Figure IV.1. MAPT Scale Score Transformation Function (Reading).

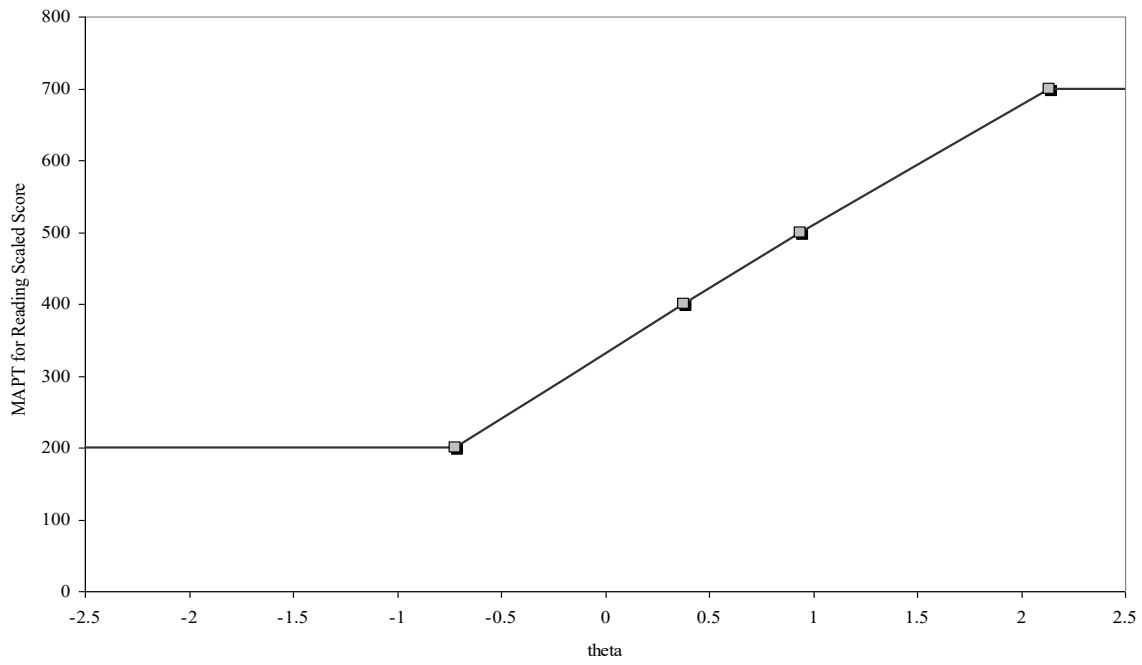
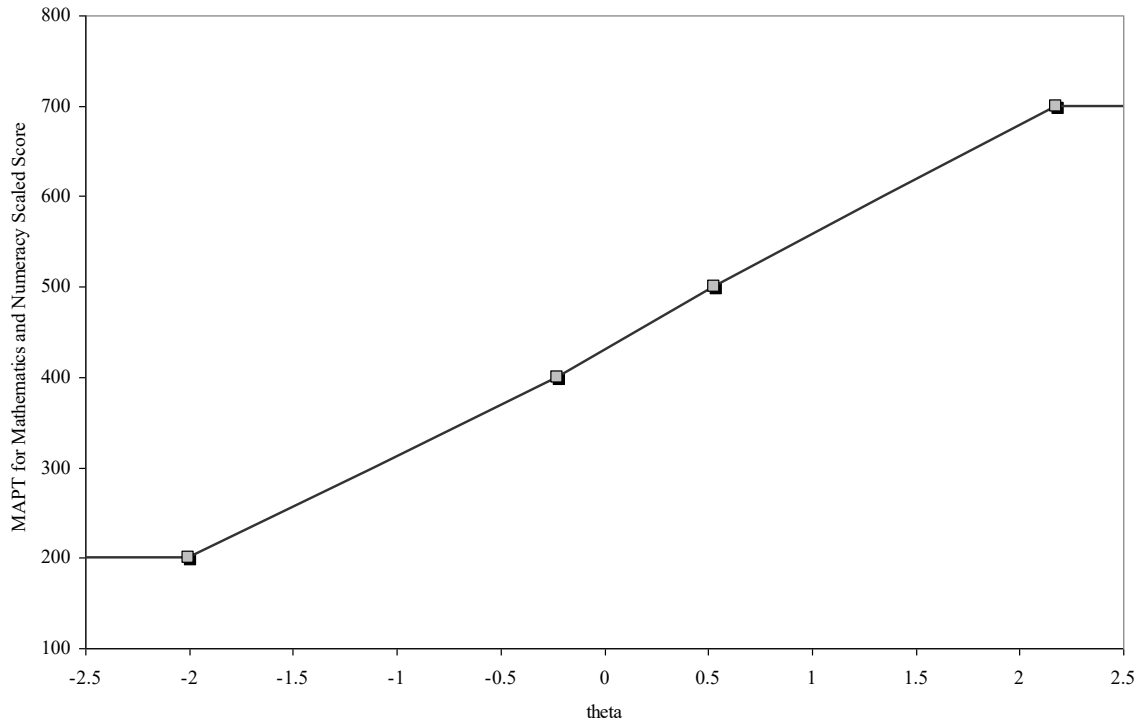


Figure IV.2. MAPT Scale Score Transformation Function (Mathematics and Numeracy).



V. Standard Setting: Establishing the NRS Educational Functioning Levels

As mentioned in Chapter I, a primary purpose of the MAPT is to measure learners' educational gains for both evaluation of individual learners' progress as well as for program accountability. In the National Reporting System (NRS), educational gain is a core outcome measure that is defined by a set of Educational Functioning Levels (EFLs). There are six EFLs defined in the NRS that relate to Mathematics and Reading—Beginning ABE Literacy, Beginning Basic Education, Low Intermediate Basic Education, High Intermediate Basic Education, Low Adult Secondary Education, and High Adult Secondary Education (Division of Adult Education and Literacy (DAEL), 2006). Under the NRS, learners' educational gains are measured by comparing their initial EFL at intake with their EFL at some later point in time after instruction. The following excerpt from the NRS illustrates this framework for measuring gain.

The NRS approach to measuring educational gain is to define a set of *educational functioning levels* at which students are initially placed based on their abilities to perform literacy-related tasks in specific content areas. After a set time period or number of instructional hours set by the State, students are again assessed to determine their skill levels. If their skills have improved sufficiently to be placed one or more levels higher, an “advance” is recorded for that student. (DAEL, 2006, p. 13)

Descriptions of the NRS EFLs that pertain to the MAPT are provided in Exhibit V.1. These descriptions were the focal point on which the MAPT standard setting process centered.

For the MAPT to be useful for measuring educational gain within the NRS, cut-scores on the MAPT score scale are needed to demarcate the NRS EFLs. To establish these cut-scores, standard setting studies were conducted using standard setting panels consisting of experts in math and reading instruction for adult learners. Separate studies were conducted for the MAPT for Mathematics and Numeracy and the MAPT for Reading; however, the methodology used was the same across studies. To describe this process, we first present the methodology. Then, we present the results for each test.

1. Standard Setting Methodology

Standard setting is the process of dividing a continuous variable, such as a test score scale, into a discrete variable with two or more categories (sometimes referred to as performance or achievement levels). The demarcations between these categories are characterized by cut-scores, which are points along the continuum that divide one category from another. Many different methods exist for setting cut scores (or standards) on educational tests (see Cizek, 1996, 2001; or Hambleton, 2001 for descriptions of a variety of these methods). However, all methods are inherently subjective because there is no “true” standard to discover—that is, the optimal cut score is not simply a parameter to be estimated. Hence, setting standards on educational tests is essentially the establishment of a policy, albeit one that is informed by data. These data are typically in the form of judgments from subject matter experts (standard setting panelists) regarding the knowledge and skills possessed by students at specific proficiency levels and how they are likely to perform on specific items.

Exhibit V.1

NRS Educational Functioning Level Descriptions

Literacy Level	Basic Reading	Numeracy Skills
Beginning Basic Education <i>(grade level 2–3.9):</i>	Individual can read simple material on familiar subjects; Comprehend simple and compound sentences in single or linked paragraphs containing a familiar vocabulary.	Individual can count, add, and subtract three digit numbers Can perform multiplication through 12 Can identify simple fractions Perform other simple arithmetic operations.
Low Intermediate Basic Education	Individual can read text on familiar subjects that have a simple and clear underlying structure (e.g., clear main idea, chronological order); Can use context to determine meaning; Can interpret actions required in specific written directions.	Individual can perform with high accuracy all four basic math operations using whole numbers up to three digits Can identify and use all basic mathematical symbols.
High Intermediate Basic Education	Individual is able to read simple descriptions and narratives on familiar subjects or from which new vocabulary can be determined by context; Can make some minimal inferences about familiar texts; Compare and contrast information from such texts but not consistently.	Individual can perform all four basic math operations with whole numbers and fractions Can determine correct math operations for solving narrative math problems Can convert fractions to decimals and decimals to fractions Can perform basic operations on fractions.
Low Adult Secondary Education	Individual can comprehend expository writing and identify spelling, punctuation, and grammatical errors; Can comprehend a variety of materials such as periodicals and nontechnical journals on common topics; Can comprehend library reference materials; Can identify the main idea in reading selections; Use a variety of context issues to determine meaning.	Individual can perform all basic math functions with whole numbers, decimals, and fractions Can interpret and solve simple algebraic equations, tables, and graphs Can develop own tables and graphs Can use math in business transactions.
High Adult Secondary Education	Individual can comprehend, explain, and analyze information from a variety of literacy works, including primary source materials and professional journals; Can use context cues; Higher order processes to interpret meaning of written material.	Individual can make mathematical estimates of time and space Can apply principles of geometry to measure angles, lines, and surfaces Can also apply trigonometric functions.

Adapted from exhibit 2.1 of the DAEL (2006) *Implementation guidelines*.

For the MAPT, the goal of the standard setting process was to identify cut-scores that corresponded to these NRS EFLs. Given that Federal descriptions of these levels existed, we used a standard setting method that focused on the knowledge and skills possessed by students within a functioning level, rather than a method that focuses on the performance of students at the borderline of proficiency categories. The method we used is based on the *Item Descriptor Matching Procedure* developed by Ferrara and his colleagues (Ferrara, Perie, & Johnson, 2002, in press). Across the two subject areas, 21 panelists participated in the MAPT standard setting.

The Item Descriptor Matching Method

Like most standard setting methods, the Item Descriptor Matching (IDM) procedure relies on the judgments of subject matter experts who serve as standard setting panelists. The panel reviews descriptions of the achievement level categories of interest (in our case, the EFL descriptions) and discusses the knowledge and skills possessed by students in each category. Next, a booklet of test items is distributed to the panelists for review. The items within the booklet are ordered by their difficulty, which is estimated using item response theory (IRT) (Hambleton, Swaminathan, & Rogers, 1991). That is, the first item appearing in the booklet had the lowest IRT difficulty parameter estimate, and so on, until the last item in the booklet, which had the highest difficulty parameter estimate. This feature of the method is similar to the bookmark standard setting procedure (Karantonis & Sireci, 2006), but the task required of panelists is different, as we explain below. In addition to the item booklet, the content standards and the correct answer for the item are also provided to the panelists. The task for the panelists is to review each item, consider the knowledge and skills required to correctly answer the item, and then match the item to a proficiency category (EFL). Panelists are trained to complete this task, and after they rate the items independently, they discuss their ratings as a group. Through multiple rounds of discussion, they are invited to change their ratings based on the discussions, if they so desire.

The first step in our application of the IDM method was discussion of the knowledge and skills possessed by learners in each of the EFLs (see Appendices E and F for standard setting materials). After explanation of the matching task and discussion, the panelists were asked to review 10 practice items (selected to span the EFLs), answer the items, and check the answer key (to get an appreciation of the item difficulties). Next, they were asked to match each item to an EFL. The specific question presented to the panelists for matching items to EFLs was

To which NRS Educational Functioning Level description are the knowledge, skills, and cognitive processes required to correctly answer this item most closely matched? (I.e., match each item to an NRS Educational Functioning Level.)

Following discussion of the practice items, panelists reviewed and rated items within an item booklet, where the items were ordered by their difficulty (IRT *b*-parameter) from easiest to hardest. Two 60-item booklets were created, each booklet was developed to represent the content specifications of the tests and to provide an appropriate range of difficulty designed to span the EFLs. Hence, the booklets were parallel with respect to item pool representation and difficulty. We created two booklets to balance our desire to use as many items as possible in deriving the cut-scores and our fear that panelists might not be able to complete their ratings in

the amount of time for which we hired them⁵. This strategy would allow for deriving cut-scores based on at least 60 representative items for each panelist, in case we ran out of time. As described later, all panelists completed both booklets (i.e., all panelists rated 120 items).

Ferrara, et al., (in press) describe the advantages of the IDM procedure relative to other standard setting methods. One advantage is that the cognitive task does not require probability judgments, but rather relies on content expertise that is more familiar to teachers and others who serve on standard setting panels. Another advantage is that the panelists do not have to make difficult, probabilistic judgments regarding the likelihood that certain types of learners will answer an item correctly at some specific probability level (as is required in the Bookmark and Angoff standard setting methods, Karantonis & Sireci, 2006). In addition, the IDM is ideally suited to the present situation where cut-scores must be set to preexisting achievement level definitions. Using a procedure like the Angoff method would involve consideration of the EFLs and four different types of “borderline” learners, which would be difficult for panelists to distinguish and articulate.

Given that IRT provides the statistical framework for the item calibration, scaling, and scoring architecture for the MAPT (see chapter IV), the IRT item difficulty parameters are on the same scale as learners’ scores⁶. Therefore, the difficulty parameters for items classified by the standard setting panelists within each EFL were used to determine the cut-scores signifying the boundaries of each EFL. Specifically, logistic regression was used to calculate the cut-score boundaries between each EFL. Separate regressions were conducted for each cut-score using the equation

$$p = \frac{1}{1 + e^{-(a+bx)}}$$

where p =the probability of a panelists’ rating being in a specific EFL or higher for an item with difficulty= x , and a and b are the intercept and slope, respectively, of the logit of p . In this regression, a dichotomous (criterion) variable was created for each panelist’s EFL classification indicating whether it was in a particular EFL or higher (e.g., Beginning Basic ABE, or higher) and the IRT difficulty parameter for each item was used as the predictor. Using these data, the slope and intercept in the logistic regression equation were calculated, and the cut-scores were computed by setting $p=.50$ (indicating a 50/50 chance of a point on the scale being in either of two adjacent EFLs) and solving for x . The selection of $p=.50$ means that learners with a proficiency estimate at the cut-score resulting from the calculation (x) have a 50% chance of being in either EFL, and so it is the point that minimizes false positive and false negative errors (Livingston & Zieky, 1989). Logistic regression has been used in this fashion to derive cut-scores in several contexts including those based on contrasting groups (Livingston & Zieky, 1989) and cluster analysis (Sireci, Robin, & Patelis, 1999).

⁵ Panelists were paid \$500 for their participation in the study, which included reviewing the EFLs and MAPT test specifications before the meeting, participating in an 8.5-hour standard setting meeting, and completing unfinished ratings after the meeting and returning them to the researchers.

⁶ In item response theory items and examinees are placed on the same score scale metric, but for reporting purposes, such as the MAPT, the scores reported for examinees are usually transformed to a more interpretable scale (which for the MAPT is currently 200-700).

Although the process we followed in gathering panelists' standard setting data closely mirrored the IDM procedure described by Ferrara, et al., (2002, in press), our use of logistic regression departs from their method for calculating the cut-scores. Ferrara, et al., conducted several additional rounds of panelist review and discussion of preliminary results from the IDM and used the panelists to help determine the final cut-score regions from which the standards would be set. We used the logistic regression approach because in our context more than two rounds of discussion were not possible, due to the time and budget available for the study. For example, we did not have six days to set the standards across the two tests as would be required if we exactly followed the Ferrara, et al., procedure. In addition to the time needed to complete such a study, it was not possible to recruit the types of panelists we needed (i.e., experienced ABE educators) for a three-day meeting, or pay them for that amount of time.

As described below, the panelists discussed their classifications for over half of the items and so training and discussion were integral parts of our application of the IDM method. To make up for the lack of discussion on the other items and to address specific item ratings that appeared to be way off base relative to the item statistics and majority of panelists' judgments, we used a data cleaning process to eliminate widely discrepant ratings before applying the logistic regression model. In some standard setting studies extremely discrepant panelists or items are eliminated before calculating cut-scores (Livingston & Zieky, 1982). Our data cleaning did *not* eliminate any panelists or items⁷. Instead, ratings for particular panelists on specific items were omitted. The omit rule was to eliminate the rating from a panelist where inclusion of the rating would place the item in a EFL category in which its IRT difficulty estimate was more than two standard deviations from the mean difficulty of the remaining items in that ELF. This data-cleaning rule was applied for two iterations and led to retention of 90% or more of the rating data.

2. Validity Evidence for Standard Setting

Kane (1994, 2001) suggested three types of validity evidence for evaluating standard setting studies—procedural evidence, internal evidence, and external evidence. Procedural evidence for evaluating standard setting “focuses on the appropriateness of the procedures used and the quality of the implementation of these procedures” (Kane, 1994, p. 437). This category of evidence includes the selection of qualified standard setting panelists, appropriate training of panelists, clarity in defining the tasks and goals of the procedure, appropriate data collection procedures, and proper implementation of the method. The preceding and subsequent sections report procedural validity evidence for the MAPT standard setting. In reporting the results for the MAPT standard setting, we also include data related to *internal* validity evidence, which refers to the expected consistency of results, if the study were replicated. One internal validity evaluation criterion is the standard error of the cut-score, although calculation of this standard error is not straightforward due to dependence among panelists' ratings (i.e., due to discussion) and practical factors (e.g., time and expense in conducting independent replications). Other evidence for

⁷ Several different approaches were considered in calculating the cut-scores, including conducting the logistic regressions separately for each panelist and then averaging their cut-scores and calculating the cut-scores without any data cleaning. However, the other methods we explored were rejected because the approach on which we ultimately decided used most of the data available (all but those eliminated through the cleaning process) and provided the most sufficient estimates of the cut-scores, as well as the estimates most resistant to outliers.

internal validity includes evaluation of the variability across panelists within a single study, and the degree to which this variability decreases across subsequent rounds of the study⁸.

External validity evidence for standard setting refers to the degree to which the classifications of examinees are consistent with other performance data. Such data are not available at this time, but we plan on gathering external validity data during the first year these tests are operational (see Chapter VII). External validity evidence for standard setting is hard to gather because valid, external criteria rarely exist (hence the need for tests and standards in the first place) and construction of such measures involves considerable time, personnel, and money. Even when these data are gathered and analyzed the results may be hard to interpret because the validity of the external data would need to be established. This problem of the validity of the criterion has been discussed for well over 60 years (e.g., Guilford, 1946; Jenkins, 1946; Toops, 1944). As illustrated below, we believe the procedural and internal validity evidence for the MAPT standard setting process provides sufficient justification for the defensibility and use of the cut-scores.

Kane (1994) described the validation of standard setting as building an argument that the process was reasonable and leads to defensible decisions. As he put it,

The best that we can do in supporting the choice of a performance standard and an associated [cut] score is to show that the [cut] score is consistent with the proposed performance standard and that this standard of performance represents a reasonable choice, given the overall goals of the assessment program. In practice, however, we seldom, if ever, achieve even this goal. A more modest, but realistic goal in most cases is to assemble evidence showing that the passing score and its associated performance standard are not unreasonable. (p. 437)

Given the strong link between the NRS EFLs and the procedure we used to set standards on the MAPT, we believe our results are far better than the relatively low standard of being “not unreasonable.” They represent Massachusetts ABE educators’ consensus impressions of how performance on the MAPT relates to the NRS EFLs. As described below, evidence in support of these standards includes procedural and internal validity evidence.

⁸ However, Kane (2001) pointed out that interpretations of the variability of panelists’ ratings are not always clear: “A high level of consistency across participants is not to be expected and is not necessarily desirable; participants may have different opinions about performance standards. However, large discrepancies can undermine the process by generating unacceptably large standard errors in the cutscores and may indicate problems in the training of participants.” (p. 73)

3. Setting NRS Cut-scores on the MAPT for Mathematics and Numeracy

The previous section described the overall IDM procedure for setting the NRS standards on the MAPT. In this section, we give the specific details in setting standards on the MAPT for Mathematics and Numeracy.

Participants

Ten subject matter experts served as panelists for setting the NRS standards on the MAPT for Mathematics and Numeracy (see Appendix A for a list of these panelists). Half of the panelists were classroom teachers and the others were administrators or teacher/administrators. Eight of the ten panelists were female. The panelists were recruited through nominations from ACLS. Of the 10 panelists, one taught for just under 5 years, three taught for 5-10 years, three taught for 10-20 years, and three taught for more than 20 years. One panelist was a non-native English speaker, the rest were native speakers of English. Panelists were selected to ensure a representation of expertise across the continuum of ABE students' learning levels. Two of the 10 panelists felt very comfortable working with students at all of the learning levels and the remaining eight participants were evenly divided between a preference for the lower levels and a preference for the higher levels. Through the training and discussions all the panelists shared their expertise about learners at the different levels.

Special care was also taken to invite panelists from different regions of the state and who taught diverse learner populations. Four of the 10 panelists were from the Eastern region, three were from the Western region, two were from the Central region, and one panelist was from Rhode Island, but worked for years in Massachusetts and consults with ACLS on curriculum issues. The panelists work with learners from a variety of international backgrounds with many non-native English speakers in their classes. The predominant ethnicities of their students include Hispanic/Latino, Haitian, Asian, and African American.

Procedure

In advance of the standard setting meeting, panelists were sent a demographic survey, the MAPT for Mathematics and Numeracy test specifications, and the NRS EFL descriptors for numeracy skills. They were asked to review the EFL descriptors and think about the different types of learners to whom they referred, and to add annotations regarding the types of knowledge and skills possessed by learners at these different levels.

The standard setting meeting was conducted in one day from 8:30 a.m. to 5:00 p.m. The agenda for the meeting is reproduced in Appendix E. After a welcome, introductions, and overview of the MAPT, a facilitated review and discussion of the EFLs was conducted. Panelists were asked to add specific knowledge and skills to the EFL descriptors that characterized learners at that level. Large flip charts were used to post the embellished (more detailed) descriptors agreed upon by the Panel. The review and elaboration of the EFLs took about three hours. At the end of this discussion, flip charts for each EFL were hung on the walls of the room and were referred to by panelists throughout the remainder of the meeting.

Next, the panelists were introduced to the item-matching task. They were asked to

review and rate 10 specific items, which were subsequently discussed as a group. This discussion ensured that all panelists were on-task, were using appropriate criteria for making their ratings, and could understand the perceptions and opinions of their fellow panelists. After this step, panelists took a brief lunch break, and then rated the remaining 50 items in Booklet 1 on their own. Items with less than 50% agreement across the panelists regarding the EFL classification were targeted for discussion and all these items, and many others, were discussed. Panelists were invited to revise their ratings for any items based on the group discussions as they regarded necessary. Panelists recorded both the original and revised ratings on the rating forms.

About half of the panelists completed their ratings for Booklet 2 before the meeting ended. The other panelists took the unfinished pages from Booklet 2 with them and returned them with their ratings within four days of the meeting. All panelists completed their ratings on all 120 items, yielding a data set of 1,200 item/EFL matches (10 panelists X 120 items).

As mentioned earlier, the panelists' rating data went through a data cleaning procedure where highly aberrant ratings were eliminated. Specifically, we eliminated a panelist's rating for an item if it resulted in placing the item into an EFL with a mean difficulty more than 2 standard deviations from the item's difficulty. Using this process, 102 of the 1,200 ratings (8.5%) were omitted before implementing the logistic regression analyses.

Results

Using the procedure described above, the cut-scores for the MAPT for Mathematics and Numeracy were calculated for each of the boundaries between the NRS EFLs, with the exception of the lowest boundary (Beginning Literacy versus Beginning Basic), since the MAPT is not designed for learners at the Beginning Literacy Level. The cut-scores are presented in Table V.1. It should be noted that the initial cut-scores, presented in the 2nd column in the table, are reported on the IRT metric before transformation to the 100-700 MAPT score scale. The transformations from the IRT scale to the MAPT scale are described in Chapter III.

Table V.1

Cut-Scores for the MAPT for Mathematics and Numeracy

NRS EFL Boundary	Score Scale	
	IRT Scale	MAPT Scale
Beginning Basic/Low Intermediate	-1.1154	300
Low Intermediate/High Intermediate	-.2278	400
High Intermediate/Low Adult Secondary	.5258	500
Low Adult Secondary /High Adult Secondary	1.3494	600

Variability among panelists

After completing Booklet 1, there was 60% (6 of 10 panelists) or greater agreement (majority agreement) on 39 of the 60 items with respect to its EFL classification. After

discussion, 55 of the 60 items had 60% or greater agreement. The mean item agreement for Booklet 1 was 64% before discussion (s.d.=16.7) and 72% after discussion (s.d.=15.6). These data indicate the discussions helped foster consensus among the panelists regarding the way in which they rated the items. For Booklet 2, 43 of the 60 items had 60% or greater agreement, and the mean agreement rate was 68% (s.d.=17.9). Across all 120 items, only 7 items had less than 50% agreement, and 15 other items had less than 60% agreement. The mean agreement across all 120 items was 70%. Eleven items (9%) exhibited unanimous agreement across all ten panelists. Unanimous agreement on all items is not expected, or even desired, but some degree of consensus should exist to show that most panelists generally agreed regarding the appropriate placement of most items. The degree of agreement observed here suggests the panelists generally concurred on the EFL classifications for most items, but the ratings were far from unanimous, which reflects the diversity of teaching and learning styles prevalent in ABE classes.

Two estimates of the standard errors of the cut-scores were calculated. The first estimate was computed by calculating the cut-scores separately for each panelist, and then taking the standard error of the mean of these cut-scores. This estimate has two drawbacks. First, the standard error of the mean assumes independence among observations, which is violated in standard setting studies when discussion takes place among the panelists. Second, because the cut-scores were computed using logistic regression, the regression equation using the ratings for all panelists (about 1,100 ratings) will be more stable than the equations based on individual panelists (about 120 ratings). Thus, these two caveats must be kept in mind when interpreting that estimate of error.

The second estimate of standard error was computed by calculating separate cut-scores based on Booklet 1 and Booklet 2 and dividing the difference between these two separate cut-scores by the square root of 2, as suggested by Brennan (2002). This estimate may overestimate the true standard error for two reasons. First, computing the cut-scores based on 60 items is sure to be less stable than computing it based on 120 items. Second, most of the items in Booklet 2 were not discussed by the panelists, and so there is likely to be less agreement across items not discussed than across random subsets of items.

The standard error estimates for Mathematics and Numeracy are presented in Table V.2. The estimates based on Method 2 are smaller than those for Method 1 for 3 of the 4 cut-scores. It should be noted that the scale of the cut-scores is based on an IRT metric with a mean of .38 and standard deviation of .70 (based on the subset of 120 items used in the standard setting study). Thus, the standard error estimates based on the standard error of the mean (Method 1) for the Beginning Basic/Low Intermediate cut-score and for the Low Adult Secondary /High Adult Secondary cut-score, represent about one-third of a standard deviation. The estimates based on either method for the other two cut-scores are much lower. All standard errors estimated using either method are small relative to the widths of the score intervals for each EFL.

Table V.2

Estimates of Standard Errors for Mathematics and Numeracy Cut Scores (n=10)

NRS EFL Boundary	Cut-Score (IRT Scale)	Standard Error: Method 1	Standard Error: Method 2
Beginning Basic/Low Intermediate	-1.1154	0.249	0.021
Low Intermediate/High Intermediate	-.2278	0.054	0.033
High Intermediate/Low Adult Secondary	.5258	0.046	0.120
Low Adult Secondary /High Adult Secondary	1.3494	0.247	0.067

Notes: Method 1 refers to the standard error of the mean across panelists. Method 2 refers to the difference across cut-scores derived separately for Booklet 1 and Booklet 2.

Math standard setting post-meeting survey results

The post-meeting survey administered to the panelists inquired about the clarity of the goals of the meeting, adequacy of time spent on specific activities, the degree to which they understood their tasks, and the confidence they had in their ratings (see Appendix E for a copy of the survey). Nine of the 10 panelists indicated the goals of the meeting were “very clear,” with the remaining panelist marking “somewhat clear.” With respect to the NRS EFLs, 7 of the 10 panelists reported they were “very clear,” while the other 3 reported they were “somewhat clear.” When asked how well they understood the rating task, 5 panelists responded “Understood Completely” and 5 panelists responded “Mostly Understood.” Eight of the 10 panelists responded that the time spent discussing the EFLs was “about right,” while the other 2 panelists thought more time was needed. With respect to the amount of time needed to complete their item descriptor matching ratings, 8 of the 10 panelists thought the amount of time was “about right,” and 2 thought more time was needed.

Panelists were also asked “How confident are you that your item ratings will be useful for helping us set cut-points on the MAPT score scale to demarcate the NRS Educational Functioning Levels?” Four panelists responded “Very Confident,” 4 responded “Confident,” and 2 responded “Not Sure.” All panelists indicated they felt comfortable expressing their opinions during the group discussions. When asked about the criteria they used in making their ratings, virtually all panelists used the expected criteria (e.g., EFL descriptions, knowledge and skills measured by the items, group discussions, experience working with students), and 3 panelists listed additional criteria—knowledge of current curriculum, a more holistic view of how an item matches to a level, and reading requirements within an item. The overwhelming conclusions from the panelists survey data are that they understood the tasks, had sufficient time to provide their ratings, had confidence in the standard setting data, and enjoyed their experience.

4. Setting NRS Cut-scores on the MAPT for Reading

In this section, we give the specific details regarding the standard setting for the MAPT for Reading, including the specifics regarding the participants and procedures, as well as the results.

Participants

Eleven subject matter experts served as panelists for setting the NRS standards on the MAPT for Reading (see Appendix A for a list of these panelists). Three panelists were administrators, three were teachers, and four were both teachers and administrators. Eight of the eleven panelists were female. The panelists were recruited through nominations from ACLS. Of the 11 panelists, one taught for less than 5 years, three taught for 5-10 years, five taught for 10-20 years, and two taught for more than 20 years. Five of the 10 panelists have additional professional and curriculum development experience. Given the range of ABE learners, panelists were selected to ensure a representation of expertise across the continuum of learning levels. One of the 11 panelists felt very comfortable working with students at all of the learning levels and the remaining ten participants were evenly divided between a preference for the lower levels and a preference for the higher levels. Through the training and discussions all the panelists shared their expertise about learners at the different levels.

Special care was also taken to invite panelists from different regions of the state and who taught diverse learner populations. Five of the 11 panelists were from the Northeast, five were from the Eastern region, and one was from the Western region. The panelists work with learners from a variety of international backgrounds with many non-native English speakers in their classes and from ethnic backgrounds such as Haitian, Caribbean, Hispanic/Latino, Asian, and African American.

Procedure

In advance of the standard setting meeting, panelists were sent a demographic survey, the MAPT for Reading test specifications, and the NRS EFL descriptors for reading and writing. They were asked to review the EFL descriptors and think about the different types of learners to whom they referred, and to add annotations regarding the types of knowledge and skills possessed by learners at these different levels.

The standard setting meeting was conducted in one day from 8:30 a.m. to 5:00 p.m. The agenda for the meeting is reproduced in Appendix F. Similar to the math standard setting meeting, after a welcome, introductions, and overview of the MAPT, a facilitated review and discussion of the EFLs was conducted. Panelists were asked to add specific knowledge and skills to the EFL descriptors that characterized learners at that level. Large flip charts were used to post the embellished (more detailed) descriptors agreed upon by the Panel. The review and elaboration of the EFLs took just over three hours. At the end of this discussion, flip charts for each EFL were hung on the walls of the room and were referred to by panelists throughout the remainder of the meeting.

Next, the panelists were introduced to the item-matching task. They were asked to

review and rate 10 specific items, which were subsequently discussed as a group. This discussion ensured that all panelists were on-task, were using appropriate criteria for making their ratings, and could understand the perceptions and opinions of their fellow panelists. After this step, panelists took a brief lunch break, and then rated the remaining 50 items in Booklet 1 on their own. Items with less than majority agreement across the panelists regarding the EFL classification were targeted for discussion and all these items were discussed. Many other items were also discussed based on panelists' requests. Following these discussions, panelists were invited to revise their ratings for any items based on the group discussions as they regarded necessary. Panelists recorded both the original and revised ratings on the rating forms.

About half of the panelists completed their ratings for Booklet 2 before the meeting ended. The other panelists took the unfinished pages from Booklet 2 with them and returned them with their ratings within four days of the meeting. All panelists completed their ratings on all 120 items, yielding a data set of 1,320 item/EFL matches (11 panelists X 120 items).

As mentioned earlier, the panelists' rating data went through a data cleaning procedure where highly aberrant ratings (i.e., those that placed an item within an EFL with a mean difficulty more than 2 standard deviations from the item's difficulty) were eliminated. The complete data set consisted of 1,320 ratings (11 panelists X 120 items). After data cleaning, 137 of the 1,320 ratings (10%) were omitted before implementing the logistic regression analyses, which was based on the remaining 1,183 ratings. This data-cleaning step was necessary to reduce the influence of extreme ratings on the final cut-scores. Given that panelists made 120 ratings, we expected that some of them may be off the mark due to fatigue and other factors.

Results

Using the procedure described above, the cut-scores for the MAPT for Reading were calculated for each of the boundaries between the NRS EFLs, with the exception of the lowest (Beginning Literacy versus Beginning Basic), since the MAPT is not designed for learners at the Beginning Literacy Level. The cut-scores are presented in Table V.3. It should be noted that the initial cut-scores, presented in the 2nd column in the table, are reported on the IRT metric before transformation to the 100-700 MAPT score scale.

Table V.3
Cut-Scores for the MAPT for Reading

NRS EFL Boundary	Score Scale	
	IRT Scale	MAPT Scale
Beginning Basic/Low Intermediate	-.1700	300
Low Intermediate/High Intermediate	.3777	400
High Intermediate/Low Adult Secondary	.9380	500
Low Adult Secondary /High Adult Secondary	1.5376	600

Variability among panelists

After completing Booklet 1, there was majority agreement (6 of 11 panelists or 55%) on 47 of the 60 items with respect to its EFL classification. After discussion, 56 of the 60

items had majority agreement. The mean item agreement for Booklet 1 was 62% before discussion (s.d.=13.7) and 76% after discussion (s.d.=13.2). These data indicate the discussions helped foster consensus among the panelists regarding the way in which they rated the items. For Booklet 2, 56 of the 60 items had 55% or greater agreement, and the mean agreement rate was 65% (s.d.=13.2). Across all 120 items, only 17 items (14%) had less than 55% agreement. The mean agreement across all 120 items was 70%, which was the same observed for the math test. The degree of agreement observed here suggests there was strong, but not perfect, consensus among the panelists regarding matching the items to EFLs.

As described in reporting the results for the Math test, two estimates of the standard errors of the cut-scores were calculated. These estimates are presented in Table V.4. The estimates based on the standard error of the mean (Method 1) were larger for 3 of the 4 cut-scores, and all standard error estimates were relatively small with the exception of the Low Adult Secondary/High Adult Secondary cut-score. The scale of the cut-scores is based on an IRT metric with a mean of .54 and standard deviation of .49 (based on the items used in the standard setting study). Therefore, the estimate based on Method 1 for the Low Adult Secondary/High Adult Secondary cut-score is about eight-tenths of a standard deviation, and over 1.5 standard deviations based on Method 2. These findings indicate the cut-score for the Low Adult Secondary/High Adult Secondary is likely to exhibit the least stability, if the study were replicated. This relatively larger standard error for this cut-score is not surprising, since the MAPT for Reading was originally designed to stop at the Low Adult Secondary boundary and so there are fewer items written at that level. Given this result, we also investigated a different means for setting the Low Adult Secondary/High Adult Secondary cut-score by subtracting one standard deviation from a perfect score on the MAPT (i.e., 700-1SEM). The conditional error at that point was just under 100 points and so the result was very close to the cut-score resulting from the standard setting study. At this juncture, we recommend using the cut-scores listed in Table V.3 for the MAPT for Reading, but we recommend adding more difficult items to the Reading item pool so that the Low Adult Secondary/High Adult Secondary cut-score can be revisited after one year of operational testing.

Table V.4
Estimates of Standard Errors for Reading Cut Scores (n=11)

NRS EFL Boundary	Cut-Score (IRT Scale)	Standard Error: Method 1	Standard Error: Method 2
Beginning Basic/Low Intermediate	-.1700	0.079	.067
Low Intermediate/High Intermediate	.3777	0.055	.011
High Intermediate/Low Adult Secondary	.9380	0.057	.034
Low Adult Secondary /High Adult Secondary	1.5376	0.394	.775

Notes: Method 1 refers to the standard error of the mean across panelists. Method 2 refers to the difference across cut-scores derived separately for Booklet 1 and Booklet 2.

Reading post-meeting survey results

The post-meeting survey administered to the panelists inquired about the clarity of

the goals of the meeting, adequacy of time spent on specific activities, the degree to which they understood their tasks, and the confidence they had in their ratings (see Appendix G for a copy of the survey). Eight of the 11 panelists indicated the goals of the meeting were “very clear,” and 3 panelists indicated the goals were “clear.” With respect to the NRS EFLs, 9 of the 11 panelists reported they were “very clear,” while the other 2 reported they were “somewhat clear.” When asked how well they understood the rating task, 4 panelists responded “Understood Completely” and 7 panelists responded “Mostly Understood.” All 11 panelists responded that the time spent discussing the EFLs, and the amount of time needed to complete their item descriptor matching ratings, was “about right.”

Panelists were also asked “How confident are you that your item ratings will be useful for helping us set cut-points on the MAPT score scale to demarcate the NRS Educational Functioning Levels?” Two panelists responded “Very Confident,” 8 responded “Confident,” and 1 responded “Not Sure.” All panelists indicated they felt comfortable expressing their opinions during the group discussions. When asked about the criteria they used in making their ratings, virtually all panelists used the expected criteria (e.g., EFL descriptions, knowledge and skills measured by the items, group discussions, experience working with students). Two panelists listed other criteria that influenced their ratings. These criteria were (a) specific knowledge of some content that was appearing more challenging (e.g. vowel sounds), (b) past experience reviewing reading assessments. The overwhelming conclusions from the panelists survey data are that they understood the tasks, had sufficient time to provide their ratings, had confidence in the standard setting data, and enjoyed their experience.

Summary of MAPT Standard Setting

The method for setting cut-scores on the MAPT score scale was specifically designed to ensure the cut-scores corresponded to the EFL descriptions stipulated in the NRS. By using these descriptions as a starting point, and systematically gathering judgments from carefully selected ABE teachers and administrators, who were familiar with the Massachusetts ABE curriculum frameworks and learners at various educational levels, the cut-points resulting from the study result in achievement levels that differ from one another in a way that is congruent with the differences among the EFLs described in the NRS. All standard setting panelists reviewed 120 items and a logistic regression procedure was applied to their data to derive the cut-scores. The panelists’ survey data generally indicated comfort with the process as well as confidence that their ratings would be useful for establishing cut-scores for distinguishing EFLs. The internal validity evidence indicated good consensus among the SMEs and adequate anticipated replication of the results if the study were repeated, for all cut-scores except the Low Adult Secondary/High Adult Secondary boundary for the MAPT for Reading. Additional data should be used to confirm that students classified as High Adult Secondary by the MAPT for Reading, have the knowledge and skills associated with that EFL. Additional items designed for that EFL should be added to the MAPT for Reading item bank so that more items can be used to set that cut-score. Aside from this cut-score, the preponderance of procedural and internal validity evidence suggests the cut-scores derived from these standard setting studies are appropriate for demarcating the NRS EFLs on the MAPT score scales.

VI. Measurement Precision

The *reliability* of a test score is a measure of the consistency or stability of the score. If a given test yields widely discrepant scores for the same individual on separate testing occasions, and the individual did not change on the proficiency measured, then the test scores are not reliable. As described by Anastasi (1988):

Reliability refers to the consistency of scores obtained by the same persons when reexamined with the same test on different occasions or with different sets of equivalent items, or under other variable examining conditions. This concept of reliability underlies the computation of the *error of measurement* of a single score, whereby we can predict the range of fluctuation likely to occur in a single individual's score as a result of irrelevant, chance factors. (p. 109)

Reliability is inversely related to the amount of measurement error associated with test scores, and so it is a crucial index of test quality. Clearly, if a person's test score on different occasions changes dramatically according to unintended variations in the testing process, little faith can be put into a particular score obtained on a particular day. For this reason, reliability has been described as a "...*necessary*, but not *sufficient* condition for validity" (Nunnally, 1978, p. 192).

A great deal of statistical theory has been developed to provide indices of the reliability of test scores as well as measures of measurement error throughout the test score scale. Classical test theory defines reliability as the squared correlation between observed test scores and their unbiased values ("true scores"). Reliability indices typically range from zero to one, with values of .80 or higher signifying test scores that are likely to be consistent from one test administration to the next.

Reliability indices are based on "classical" theories of testing, which are somewhat re-conceptualized in modern measurement theory, such as IRT, that characterizes measurement precision in terms of test information and conditional standard error. In this chapter, we describe measurement error within an IRT context and present estimates of the precision of MAPT scores in terms of (a) test information, (b) conditional standard error, and (c) decision accuracy.

1. Test Information

In IRT, the classical concept of *reliability* is extended and replaced by the concept of *test information*. Test information takes into account the lack of uniformity in reliability across the proficiency scale and can be computed as

$$I(\theta) = \sum_{i=1}^n \frac{D^2}{e^{D(\theta-b_i)} [1 + e^{-D(\theta-b_i)}]^2} \quad [\text{VI.1}]$$

where $I(\theta)$ is the information function, θ is examinee proficiency, b_i is the IRT difficulty for item i , n is the test length, D is a scaling constant equal to 1.7, and e is the base of the natural logarithm (Hambleton, Swaminathan, & Rogers, 1991). It is evident from equation VI.1 that

given a fixed length test, information is maximized when item difficulty (b_i) equals proficiency (θ). It is this very property that makes adaptive testing so attractive: *by targeting items to learners' proficiencies, measurement information is maximized.*

Figures VI.1 and VI.2 show the expected test information function, $I(\theta)$, for the MAPT for Reading and MAPT for Math, respectively. Expected information was computed by taking the weighted average of the test information associated with each unique path through the adaptive multistage MAPT. Information was weighted based on the probability that a given path would be taken by an examinee with a given θ . For example, every possible path, p , and the probability that it would be taken, $G_p(\theta = .5)$, was determined for $\theta = .5$. Total information, $I_p(\theta = .5)$, for each path was also computed. Expected information, $E(\theta = .5)$, was then computed in the following manner:

$$E(\theta = .5) = \sum_{p=1}^k P_p(\theta = .5) I_p(\theta = .5) \quad [\text{VI.2}]$$

Note that in Figures VI.1 and VI.2 the proficiency metric has been transformed from the θ metric to the reporting scale metric to aid in interpretation. Typically, information functions are relatively smooth. However, in this case there are a few abrupt changes in the information functions, particularly for math. These changes are due to the nonlinear proficiency scale transformation.

It can be seen that test information is at or above 20 throughout the score ranges the test was designed to measure (200-600 for Reading; 200-700 for Mathematics and Numeracy). Given the characteristics of the MAPT score scale from Pilot test 2 (i.e., standard deviation of the proficiency distribution = .63), the estimated internal consistency reliability when the test information = 20 is about .88⁹. The information dips below 20 at the upper end of the Reading score scale (at about 550) and at the extremes of the Math score scale. When information = 15, the estimated internal consistency reliability is about .84. These results indicate adequate reliability for scores on the MAPT scale.

As information increases, measurement precision increases. Figure VI.3 illustrates this relationship by plotting test information against the standard error of estimation (of learner proficiency). It can be seen that the standard error of estimation stabilizes when information is in the range of about 20-25. Furthermore, it can be seen from Figures VI.1 and VI.2 that for most of the score range 200-600, information exceeds 20.

⁹ To compute this estimate we first related test information to standard error using equation VI.2 (below) and then use the formula for the standard error of measurement ($SEM = \sigma \sqrt{1 - r_{tt}}$) and solve for the reliability estimate (r_{tt}).

Figure VI.1

Expected Information (Reading)

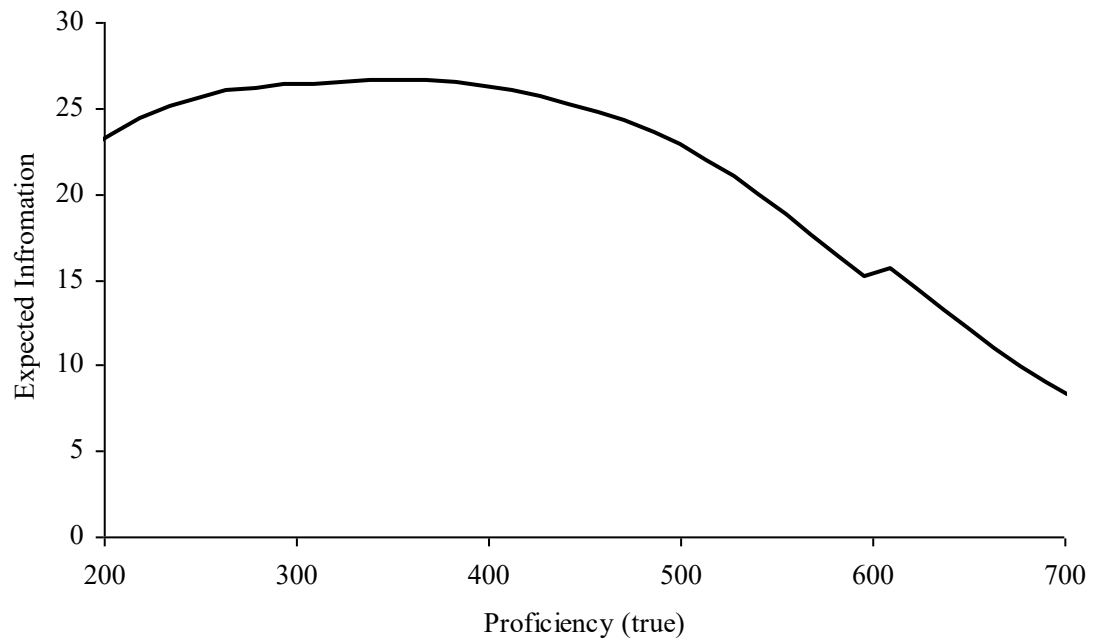


Figure VI.2

Expected Information (Mathematics and Numeracy)

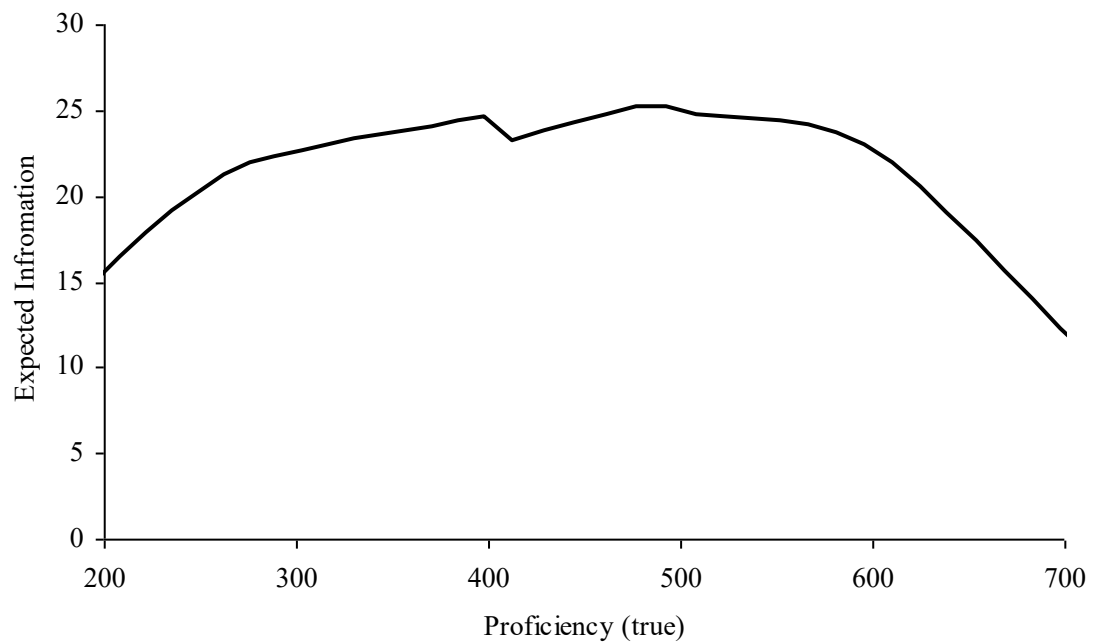
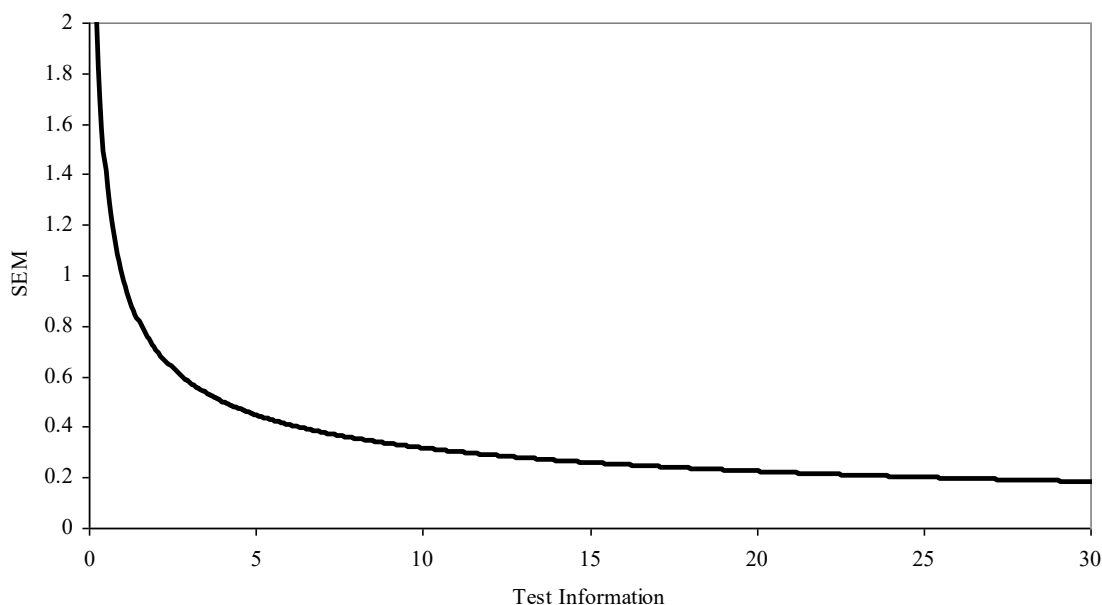


Figure VI.3

Test Information vs. Standard Error of Measurement for Proficiency



2. Conditional Standard Error of Estimation

Conditional standard error refers to estimates of measurement error at specific points along the score scale. Test information and conditional standard errors of estimation ($SE(\hat{\theta})$) are inversely related as

$$I(\theta) = \frac{1}{[SE(\hat{\theta})]^2} \quad [\text{VI.3}]$$

where $SE(\hat{\theta})$ is standard error of estimation for a given examinee proficiency estimate $\hat{\theta}$. Figures VI.4 and VI.5 show the expected conditional standard errors of estimation, $SE(\hat{\theta})$, transformed to the MAPT score scale, for MAPT for Reading and MAPT for Mathematics and Numeracy, respectively.

Expected conditional SE is the weighted average standard deviation of the maximum likelihood estimates of examinee proficiency reported on the MAPT score scale (weighting was accomplished in the same manner as described above for test information). For example, if some very large number of examinees all with a true proficiency of 250 took the MAPT for Reading, the expected standard deviation in their MAPT scores would be about 36 points.

Figure VI.4

Expected Conditional Standard Error of Estimation vs. Proficiency (Reading)

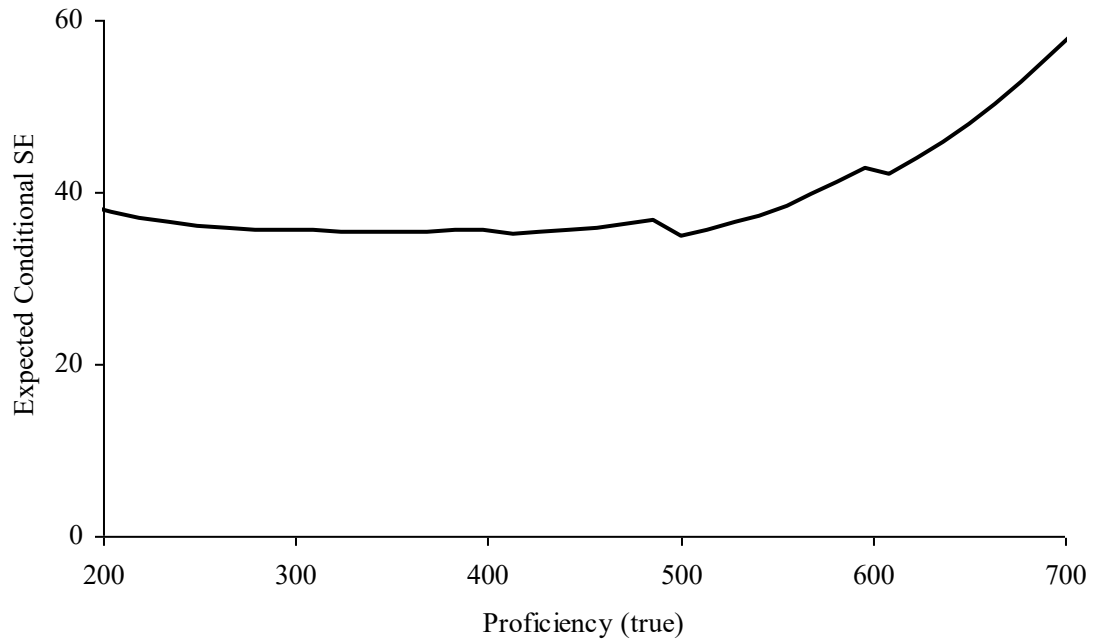
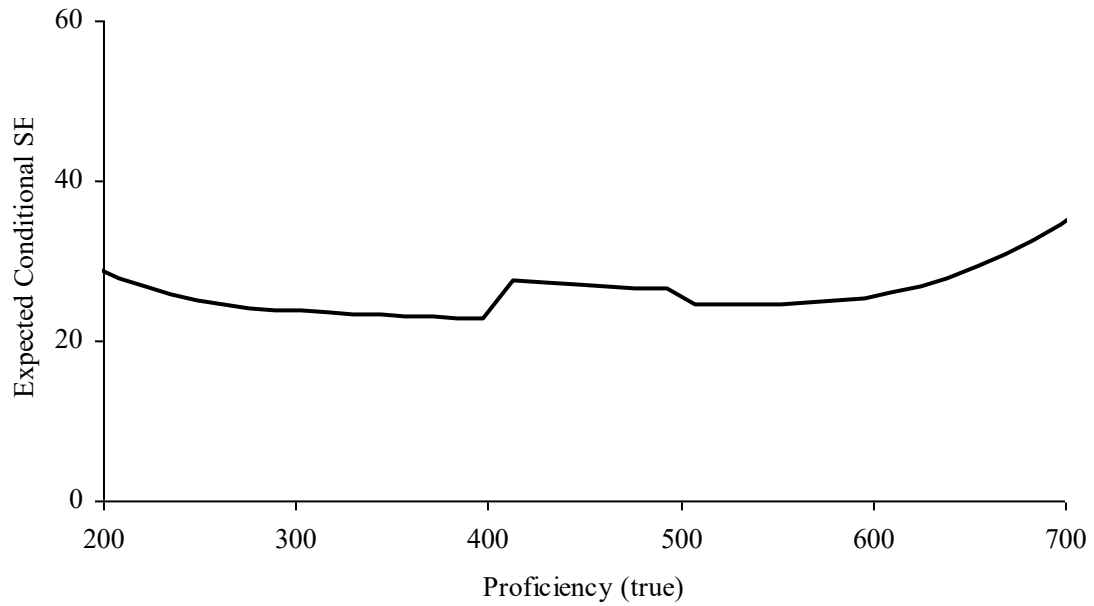


Figure VI.5

Expected Conditional Standard Error of Estimation vs. Proficiency
(Mathematics and Numeracy)



3. Decision Accuracy

In addition to providing a score on the 200-700 score scale, the MAPT classifies learners into one of five Educational Functioning Levels. As discussed elsewhere in this report (e.g., Chapter V), under the NRS, learners' educational gains are measured by comparing their initial EFL with their EFL at some later point in time. For this reason, decision accuracy is of particular interest—perhaps even more so than conditional standard errors of measurement, although the two are highly related.

Decision accuracy can be thought of as the extent to which the observed classifications of learners agree with their *true* classification. Estimates of decision accuracy compare the classifications into which learners are placed based on their test score with an estimate of their “true” classification. However, because learners' true proficiencies are never known, simulation studies or some type of split-half estimate (Livingston & Lewis, 1995) are typically used to estimate decision accuracy. To provide estimates of the decision accuracy for the MAPT, a simulation study was conducted.

Data simulation

Twenty-five-thousand simulees for each subject (Reading and Mathematics and Numeracy) were assigned a true proficiency (θ) drawn at random from a uniform distribution spanning the range of θ associated with MAPT scores between 100 and 700. Each simulee was then classified into a true EFL based on their true θ . Two response vectors were then generated for each simulee (one for each panel) based on their true θ and the operational item parameter estimates. Simulees were assigned to the correct initial module based on their true θ , except in three cases. The first exception was for Level 1 simulees, who were assigned to an initial module associated with Level 2. The second exception was for MAPT for Reading Levels 5 and 6, where simulees were assigned to an initial module associated with Reading Level 4. The third exception was for MAPT for Mathematics and Numeracy Level 6, where simulees were assigned to an initial module associated with Level 5. These reassignments were done because not every EFL has a corresponding initial module. Simulees were then scored and classified into an EFL. These ‘observed’ classifications were then compared to their corresponding true classifications. Given the 2 subject areas and 2 panels, 100,000 response vectors were scored for this simulation.

Simulation Results

Tables VI.1 and VI.2 show the true and observed proportion of examinees in each EFL for MAPT for Reading and MAPT for Mathematics and Numeracy, respectively. Diagonal cells represent proportions of simulees correctly classified, whereas off-diagonal cells represent proportions misclassified. The margins of the table represent the total proportions of simulees classified into each EFL by either observed proficiency (horizontal) or true proficiency (vertical). Of special interest are comparisons between diagonal and off-diagonal values. For example, cell [2,2] of Table VI.1 (i.e., the cell associated with row 2 and column 2) indicates that 14% of the simulees had both a true EFL and an observed EFL of 2. However, the marginal [2,Total] shows that in total, 16% of simulees had a true EFL of 2. It can be seen from [2,3] the 2% discrepancy is due to 2% of the simulees having a true EFL of 2, but being wrongly classified as EFL 3.

Table VI.1
Proportions of MAPT for Reading Simulees Classified into EFLs:
True Scores vs. Observed Scores

True NRS Level	Observed NRS Level						Total
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	
1	0.00	0.16	0.00	0.00	0.00	0.00	0.16
2	0.00	0.14	0.02	0.00	0.00	0.00	0.16
3	0.00	0.03	0.12	0.02	0.00	0.00	0.16
4	0.00	0.00	0.03	0.12	0.02	0.00	0.17
5	0.00	0.00	0.00	0.04	0.12	0.02	0.18
6	0.00	0.00	0.00	0.00	0.05	0.12	0.18
Total	0.00	0.33	0.16	0.18	0.19	0.14	1.00

Table VI.2
Proportions of MAPT for Mathematics and Numeracy Simulees Classified into EFLs: True
Scores vs. Observed Scores

True NRS Level	Observed NRS Level						Total
	<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>	<i>5</i>	<i>6</i>	
1	0.00	0.18	0.00	0.00	0.00	0.00	0.18
2	0.00	0.15	0.02	0.00	0.00	0.00	0.17
3	0.00	0.02	0.14	0.02	0.00	0.00	0.18
4	0.00	0.00	0.02	0.12	0.01	0.00	0.15
5	0.00	0.00	0.00	0.02	0.13	0.01	0.16
6	0.00	0.00	0.00	0.00	0.02	0.14	0.16
Total	0.00	0.35	0.18	0.15	0.17	0.15	1.00

Figures VI.6 and VI.7 show the proportions of simulees correctly classified at each EFL for MAPT for Reading and MAPT for Mathematics and Numeracy, respectively. These values can also be computed from Tables VI.1 and VI.2 by dividing each diagonal value by its respective (vertical) marginal value. The proportion of accurate classifications is higher for Math than for Reading, which can be explained by the relatively closer cut-scores for Reading. Overall, it can be seen that 77% of Reading simulees with a true proficiency between 200 and 500 and 83% of Mathematics and Numeracy simulees with a true proficiency between 200 and 600 were correctly classified. The decision accuracy estimates for both Reading and Math are at or above those observed in many K-12 statewide testing programs. For example, the decision accuracy for the Massachusetts Comprehensive Assessment System (MCAS) 2004 Grade 10 English Language Arts Exam is .76, and for the 2004 Grade 10 Mathematics Exam it is .80 (Massachusetts Department of Education, 2005).

Figure VI.6

Decision Accuracy Rates for MAPT for Reading

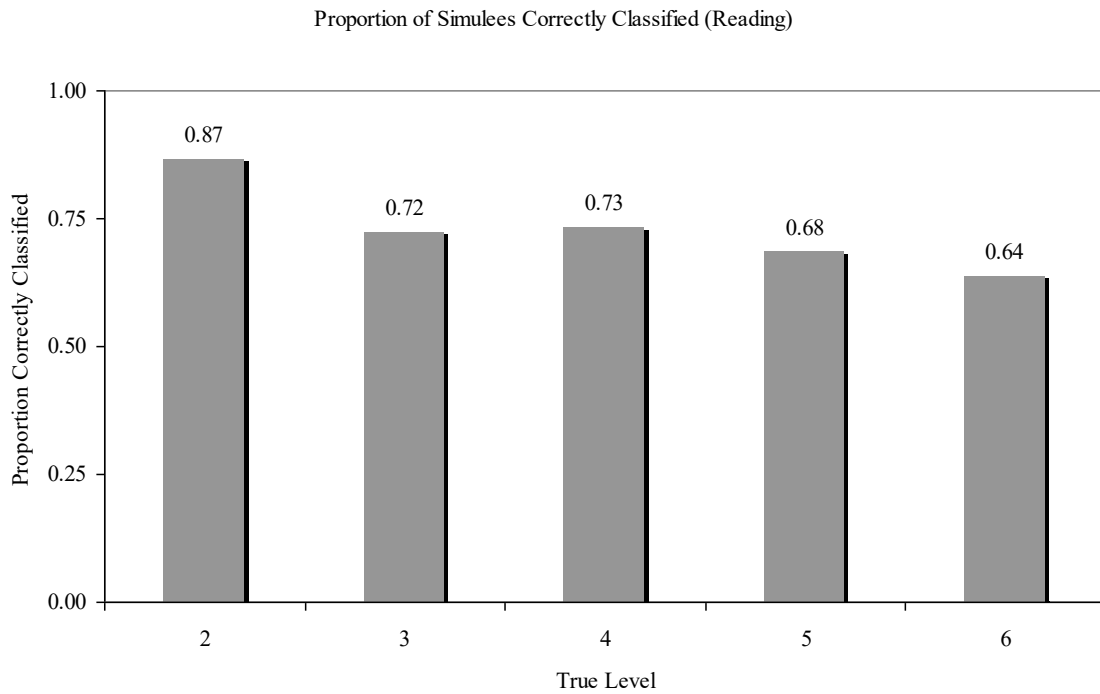
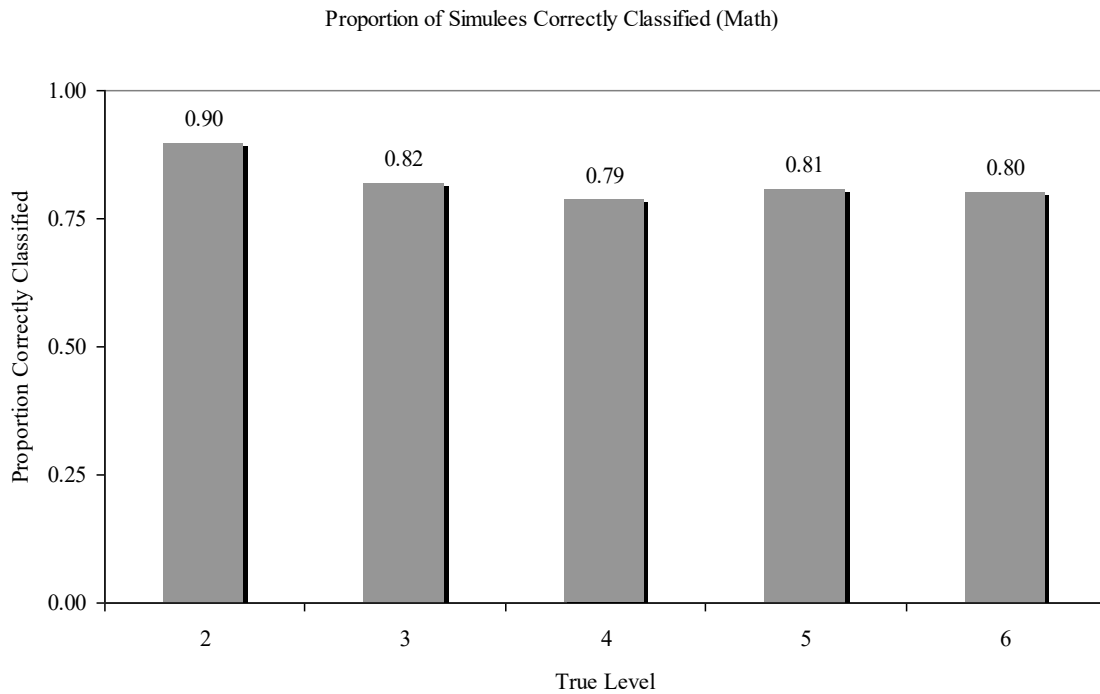


Figure VI.7

Decision Accuracy Rates for MAPT for Mathematics and Numeracy



VII. Validity

According to the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), “validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (p. 9). In other words, *validity* is the extent to which the inferences (interpretations) derived from test scores are justifiable from both scientific and equity perspectives. For decisions based on test scores to be valid, the use of a test for a particular purpose must be supported by theory and empirical evidence, and biases in the measurement process must be ruled out.

Much of the information contained in earlier chapters of this manual pertains to the validity of inferences derived from MAPT scores. Validity starts with careful and comprehensive definition of the domain to be tested (Chapter II), continues with sound test construction procedures that include several quality control checks and concerns for fairness (Chapter III), and includes adherence to technical quality in item calibration and scaling (Chapter IV), and standard setting (Chapter V). With respect to standard setting, explicit information supporting the validity of the MAPT cut-scores for the NRS EFLs was presented in Chapter V. In this chapter, we present additional information to support the use of the MAPT for the purposes for which these exams were intended. Before presenting specific validity evidence, we first provide some background on validity theory and test validation.

1. Understanding Validity and Validation

From the outset, it is important to bear in mind that validity is not an intrinsic property of a test. As many psychometricians have pointed out (e.g., Cronbach, 1971; Messick, 1989; Shepard, 1993), in judging the worth of a test, it is the inferences derived from the test scores that must be validated, not the test itself. Therefore, the specific purpose(s) for which test scores are being used must be considered when evaluating validity.

Contemporary definitions of validity borrow largely from Messick (1989) who stated “validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment” (p. 13). From this definition, it is clear that validity is not something that can be established by a single study and that tests cannot be labeled ‘valid’ or ‘invalid.’ Given that (a) validity is the most important consideration in evaluating the use of a test for a particular purpose, and (b) such utility can never be unequivocally established, establishing that a test is appropriate for a particular purpose is an arduous task. Thus, the following facts about validity should be clear: (a) tests must be evaluated with respect to a particular purpose, (b) what needs to be validated are the inferences derived from test scores, not the test itself, (c) evaluating inferences made from test scores involves several different types of qualitative and quantitative evidence, and (d) evaluating the validity of inferences derived from test scores is not a one-time event; it is a continuous process. In addition, it should be noted that although test developers must provide evidence to support the validity of the interpretations that are likely to be made from test scores, ultimately, it is the responsibility of the *users* of a test to evaluate this evidence to ensure the test is appropriate for the purpose(s) for which it is being used.

Like all testing programs, evaluating the validity of MAPT scores must be evaluated with respect to the intended purposes of the testing program, which were described in Chapter I. Specifically, interpretations based on MAPT scores should be valid for (a) evaluating adult learners' proficiencies with respect to the knowledge and skills specified in the relevant Massachusetts ABE Curriculum Frameworks, (b) evaluating ABE learners' educational gains over time, and (c) monitoring instructional quality for state and federal accountability. Factors that might threaten the validity of the use of MAPT scores for these purposes include poor congruence between MAPT items and the standards and objectives they are designed to measure, measurement of knowledge or skills that are extraneous to the subject matter tested, and imprecision in the scale scores and classifications (i.e., EFLs) assigned to learners. Information regarding the precision of MAPT scale scores and EFL classifications was provided in Chapter VI. In this chapter, we present further information in support of the utility and appropriateness of MAPT scores for their intended purposes. Before presenting these descriptions and summaries, some basic terms in test validity and validation are described.

Test Validation

To make the task of validating inferences derived from test scores both scientifically sound and manageable, Kane (1992) proposed an "argument-based approach to validity." In this approach, the validator builds an argument based on empirical evidence to support the use of a test for a particular purpose. Although this validation framework acknowledges that validity can never be established absolutely, it requires evidence that (a) the test measures what it claims to measure, (b) the test scores display adequate precision, and (c) test scores display relationships with other variables in a manner congruent with their predicted properties. Kane's practical perspective is congruent with the *Standards for Educational and Psychological Testing* (AERA, et al., 1999), which provide detailed guidance regarding the types of evidence that should be brought forward to support the use of a test for a particular purpose. For example, the *Standards* state

A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses...Ultimately, the validity of an intended interpretation...relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction; adequate score reliability; appropriate test administration and scoring; accurate score scaling, equating, and standard setting; and careful attention to fairness for all examinees... (p. 17)

The information in this chapter, taken together with the information presented in previous chapters, represents the current validity argument for the MAPT.

Gathering validity evidence

To build a validity argument for a test, there are several types of evidence that can be brought forward. Traditionally, the major forms of validity evidence are *content validity*,

criterion-related validity, and *construct validity*. Content validity evidence involves gathering data from subject matter experts regarding the degree to which the behaviors sampled on the test represent the behaviors the test is designed to measure. Criterion-related validity evidence involves evaluating correlations among test scores and other variables related to the construct measured. Construct validity involves gathering data that show test scores are indicative of the construct measured. Many test theorists (e.g., AERA, et al., 1999; Messick, 1989) consider content and criterion validity to be subcomponents of construct validity because such evidence assists in evaluating test-construct congruence. The current version of the *Standards* revised these three traditional forms of validity evidence as five sources for validity evidence. These five sources are evidence based on (a) test content, (b) response processes, (c) internal structure, (d) relations to other variables, and (e) testing consequences.

Evidence based on test content refers to traditional forms of content validity evidence such as subject matter expert review and rating of test items and test specifications (Sireci, 1998a, 1998b). Evidence based on response processes refers to “evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees” (AERA et al., p. 12). Such evidence can include interviewing test takers about their responses to test questions, systematic observations of test response behavior, and evaluation of the criteria used by judges when scoring performance tasks. Evidence based on internal structure refers to statistical analysis of item and sub-score data to investigate the dimensions measured by an assessment. Such evidence includes factor analysis (both exploratory and confirmatory) and analysis of the fit of IRT models to the observed item response data (Hambleton, 1989). Evidence based on relations to other variables refers to traditional forms of criterion-related evidence for validity such as correlations with external criteria relevant to the attributes measured (e.g., other test scores, grades, supervisor ratings). Finally, evidence based on consequences of testing refers to evaluation of the intended and unintended consequences associated with a testing program.

For MAPT scores, evidence of content validity is of paramount importance because these tests are designed to measure instruction as defined by the Massachusetts ABE Curriculum Frameworks and test specifications. In fact, it was the lack of congruence between commercially available tests and the Massachusetts ABE Frameworks that motivated development of the MAPT in the first place. For this reason, emphasis on quality test content and congruence among the frameworks, test specifications, and tests have been a priority in conceptualizing and developing these tests. Thus, the first source of evidence we present in the validity argument for the MAPT is evidence based on test content.

2. Validity Evidence Based on Test Content

MAPT Content Validity Studies

In addition to the content reviews that took place for all items (described in Chapter III), two separate content validity studies were conducted—one for the MAPT for Mathematics and Numeracy and one for the MAPT for Reading. In each study, carefully selected subject matter experts were recruited and trained to evaluate test items with respect to their congruence to the test specifications. The study involved reviewing material in advance of the meeting and attending a full-day meeting.

Participants

Thirteen subject matter experts (SMEs) participated in these studies—6 for the MAPT for Math and 7 for the MAPT for Reading. In addition, three facilitators from UMASS participated in both meetings. Lists of the members of the Content Validity Committees are presented in Appendix A.

The SMEs came from diverse backgrounds with a range of experiences. Both groups described the learners in their ABE classes as coming from very diverse backgrounds including native English speakers and students for whom English is a second language, and cultural heritages such as Euro-American, Hispanic/Latino, African-American, Asian and Cape Verdean. For the math study, five of the SMEs were women. Four of the math SMEs were from eastern Massachusetts, the other two were from the western part of the state. Five of the math SMEs had over 15 years of experience in teaching adult learners; the other SME had 5 years of experience. For the reading study, five of the SMEs were women, and all were from the East and Northeast parts of Massachusetts. Six of the participants had over 10 years of teaching experience; the other panelist had three years of teaching experience.

Materials

For each subject area, two item review booklets were created for each SME. Each page in the booklet contained a screen shot of a MAPT item along with a unique item ID number. The booklets were designed so that all items selected for the study would be reviewed by at least 3 SMEs if there were only time to complete one booklet, and all selected items would be reviewed by all 6 SMEs if there were time to complete both booklets. As described below, all SMEs in both subject areas completed both booklets. Almost all items that were being considered for operational use were included in the study. For Math, there were 352 items contained in the two booklets. For Reading, there were 240 items across the two booklets.

In addition to the item booklets, SMEs were given handouts that described the content strands and cognitive levels (Math test) or the skill areas and purposes for reading (Reading test). Item rating forms were also created for the SMEs to mark their item ratings. The rating forms included the unique item IDs and columns for the SMEs to mark their ratings regarding the content, cognitive, skill, and purpose areas being measured by each item. Math content validity materials are presented in Appendix G and Reading Content Validity materials are presented in Appendix H. An evaluation form was also created for each group of SMEs.

Procedure

For each subject area, the meeting began with a description of the test, including its purpose, and a description of the content and cognitive (Math) or skill and purpose (Reading) dimensions of the test specifications. Next, the purpose of the meeting was described and the SMEs were given instructions on how to review and rate the test items. Specifically, they were asked to read each item and consider the knowledge and skills being measured. The Math SMEs were asked to indicate the content strand and cognitive level they thought each item measured. They were instructed to not mark any content strand or cognitive level if they did not think the item matched one of the content strands or cognitive levels in the test specifications or if they could not determine the content strand or cognitive level being measured. For Reading, the SMEs were asked to indicate the skill area and purpose for reading being measured by each item. For both subject areas, the SMEs were *not* informed of the content specifications for the items and so their judgments were independent appraisals of what they thought the items were measuring.

After describing the rating task, the SMEs were asked to rate 10 items that were selected to span the content specifications for each test. These ratings were discussed as a group to make sure that the SMEs understood their task and were using appropriate criteria in making their ratings. After discussing these practice ratings, they were asked to rate the remaining items in their booklets. As questions arose, they were discussed as a group. After the SMEs rated both booklets, they were given an evaluation form that inquired about their impressions of the meeting and their confidence in the ratings.

Data analysis

Popham (1992) noted there are no absolute criteria for determining when a test item is adequately “matched” to its intended specifications based on SME rating data. To solve this problem, he recommended use of a 70% criterion for indicating adequate content validity at the item and area levels. That is, if 70% of SMEs match an item to its intended area, it should be considered to be content valid. O’Neil, Sireci, and Huff (2004) and Sireci (1998a) supported use of this criterion for evaluating the content validity of educational assessments. For the MAPT, an item was considered correctly matched to its intended areas (i.e., content strand, cognitive level, skill area, or purpose for reading) if *approximately* 70% of the SMEs placed the item in area the item was designed to measure. For math, this was operationalized as 4 of the 6 SMEs (67%). For reading, this criterion was operationalized as 5 of 7 SMEs (71%).

This (approximately) 70% criterion was used to classify each item as congruent or incongruent with the areas it was designed to measure. In addition, several summary statistics were calculated to summarize the overall content validity of the MAPT item pools and of the different aspects of the test specifications. First, we calculated the percentage of items within each subject area that were unanimously classified in the intended area (i.e., content strand, cognitive level, skill area, or purpose for reading). Second, we calculated the percentage of items in each area that met or surpassed the (approximately 70%) congruence criterion. The percentages of items that did not meet this criterion were also calculated. These summary statistics were broken down by the areas within each dimension of the test specifications to

evaluate the content representation of each dimension.

In addition to providing summative data regarding content representation, the SMEs' ratings were also used to reevaluate items with respect to their specifications. In some cases, there was consensus among the SMEs that the item was measuring an area different from the one it was intended to measure. In such cases, the specifications for the item were discussed by the test development team, and if the team agreed with the consensus ratings of the SMEs, the item was reclassified.

Results

In this section, we present a summary of the results from the content validity studies for both the MAPT for Mathematics and Numeracy (MAPT for Math) and the MAPT for Reading. The results are summarized following the traditional means for reporting content validity data; however, virtually all studies of content validity have been conducted on static test forms, not on adaptive tests (Huff & Sireci, 2001). Thus, it should be kept in mind that the results here refer to a pool of items from which the operational versions of the MAPT were assembled. The MST algorithm that delivers sets of test items to adult learners who take the MAPT is designed to represent the content dimensions of each test (see Chapter III).

In addition to providing a summary of the content validity studies, the results reported here are also useful for targeting the types of items that will be added to the MAPT item bank in the near future. That is, new items will be constructed to address any noted "shallow" aspects of the item pool.

Math Content Validity Results

As mentioned earlier, there were 352 items that were reviewed by the 6 Math SMEs. With respect to the content strand dimension, 80% of these items met the congruence criterion. The association between the intended content strand and the strand into which at least 4 SMEs placed the item was statistically significant ($\chi^2_{(9)}=689, p < .001$). For the cognitive skill dimension, 61% of the items met this criterion, and the association between the intended and consensus SME skill level was also statistically significant ($\chi^2_{(4)}=232, p < .001$). The relatively lower congruence observed for the cognitive skill dimension is consistent with previous research in this area (e.g., O'Neil, et al., 2004, Sireci, Robin, Meara, Rogers, & Swaminathan, 2000), which has consistently found that SMEs have trouble distinguishing among cognitive levels because these levels overlap to some extent.

Although the overall results generally support the content validity of the MAPT for Mathematics and Numeracy, further breakdown of the results by content strand and cognitive level reveals the relative strengths and weaknesses of the assessment with respect to the congruence of the items to their intended specifications. A summary of the SMEs' content validity ratings is presented in Table VII.1, which reports the results by content strand. A visual display of these data is presented in Figure VII.1. The Statistics and Probability strand exhibited the highest level of congruence (94%), followed by the Geometry and Measurement Strand (87%) and the Number Sense strand (80%). The Patterns, Functions, and Algebra strand

exhibited the lowest level of congruence (53%). The relatively lower congruence of the “Patterns” content strand was primarily due to the fact that 16 of these items (21%) tended to be classified as Number Sense items by the SMEs. This finding should be revisited when reviewing the objectives for this strand contained in the Curriculum Frameworks and when selecting new items to expand the item pool. It should be noted that content strand sub-scores are not reported on the MAPT. If such scores ever were to be reported, improved congruence results for the Patterns, Functions, and Algebra strand would be needed to have confidence that the sub-score adequately represented that sub-domain.

Table VII.1

Summary of SMEs’ Content Validity Ratings by Content Strand: Mathematics and Numeracy

Content Strand	# Items	% Not Congruent	% Unanimous	% Congruent
Statistics and Probability	89	6%	55%	94%
Geometry and Measurement	85	13%	54%	87%
Number Sense	102	20%	50%	80%
Patterns, Functions, Algebra	76	47%	22%	53%
Total	352	20%	46%	80%

Notes: “Not Congruent” refers to items that were matched to the intended content strand by less than 4 of the 6 SMEs. “Unanimous” refers to items matched to the intended strand by all 6 SMEs. “Congruent” refers to items matched by 4-6 SMEs.

Table VII.2 presents a summary of the results for the cognitive dimension. These data are also visually summarized in Figure VII.2. The Knowledge/Comprehension level exhibited the highest congruence (81%) followed by the Application level (71%). The Analysis, Synthesis, Evaluation (ASE) level exhibited much lower congruence (18%). Further inspection of this result found that more than half of the ASE items (46 of 88 items) were classified as Application items by 4 of the 6 SMEs. These results suggest that the distinction between these cognitive levels needs to be clarified. It is interesting to note that only 3 ASE items had a consensus rating at the lowest cognitive level (Knowledge/Comprehension) and so the poor congruence of the two higher cognitive levels appears to stem from a lack of consensus regarding which of the higher cognitive levels an item belongs (i.e., whether the item is a Application or ASE), and is not due to ASE items measuring only basic knowledge or comprehension.

Table VII.2

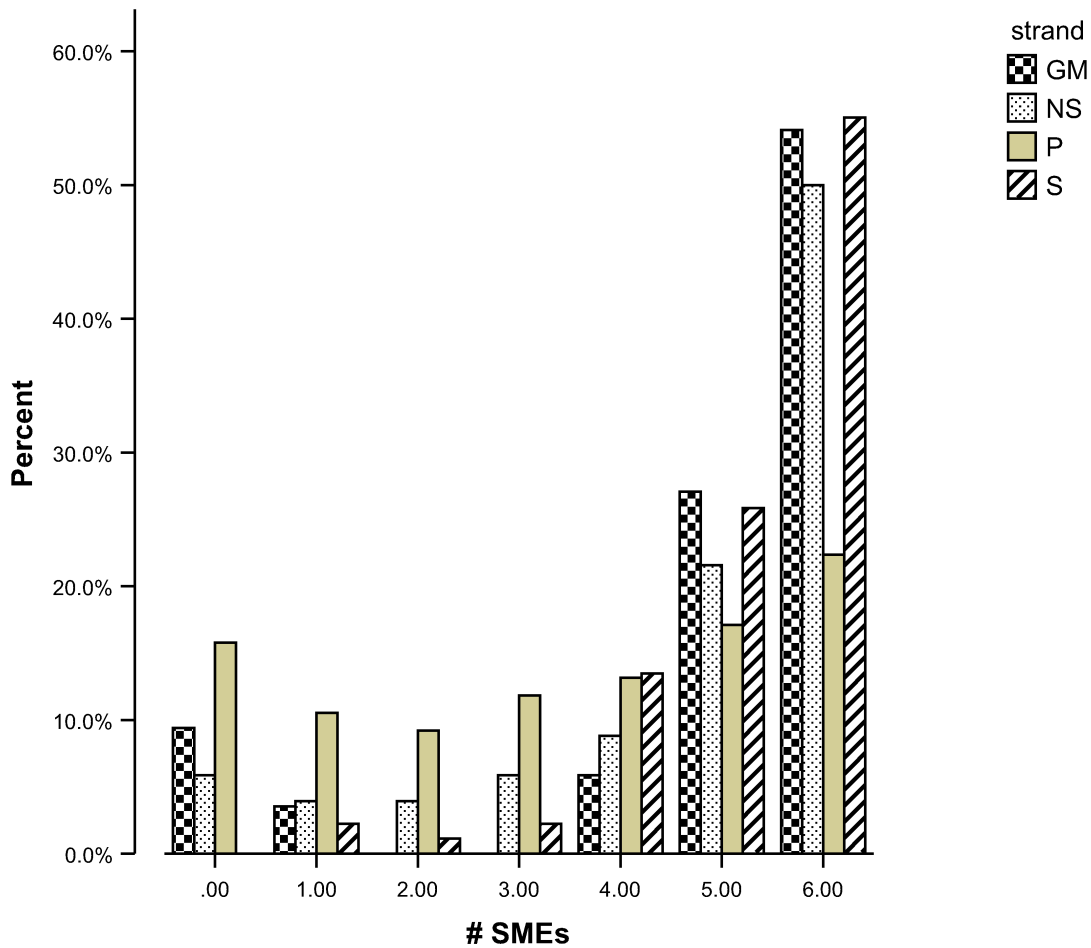
Summary of SMEs' Content Validity Ratings by Cognitive Level: Mathematics and Numeracy

Cognitive Level	# Items	% Not Congruent	% Unanimous	% Congruent
Knowledge/Comprehension	240	19%	51%	81%
Application	156	29%	22%	71%
Analysis, Synthesis, Evaluation	88	82%	3%	18%
Total	352	39%	26%	61%

Notes: "Not Congruent" refers to items that were matched to the intended cognitive level by less than 4 of the 6 SMEs, "Unanimous" refers to items that were matched to the intended level by all 6 SMEs, and "Congruent" refers to items matched by 4-6 SMEs.

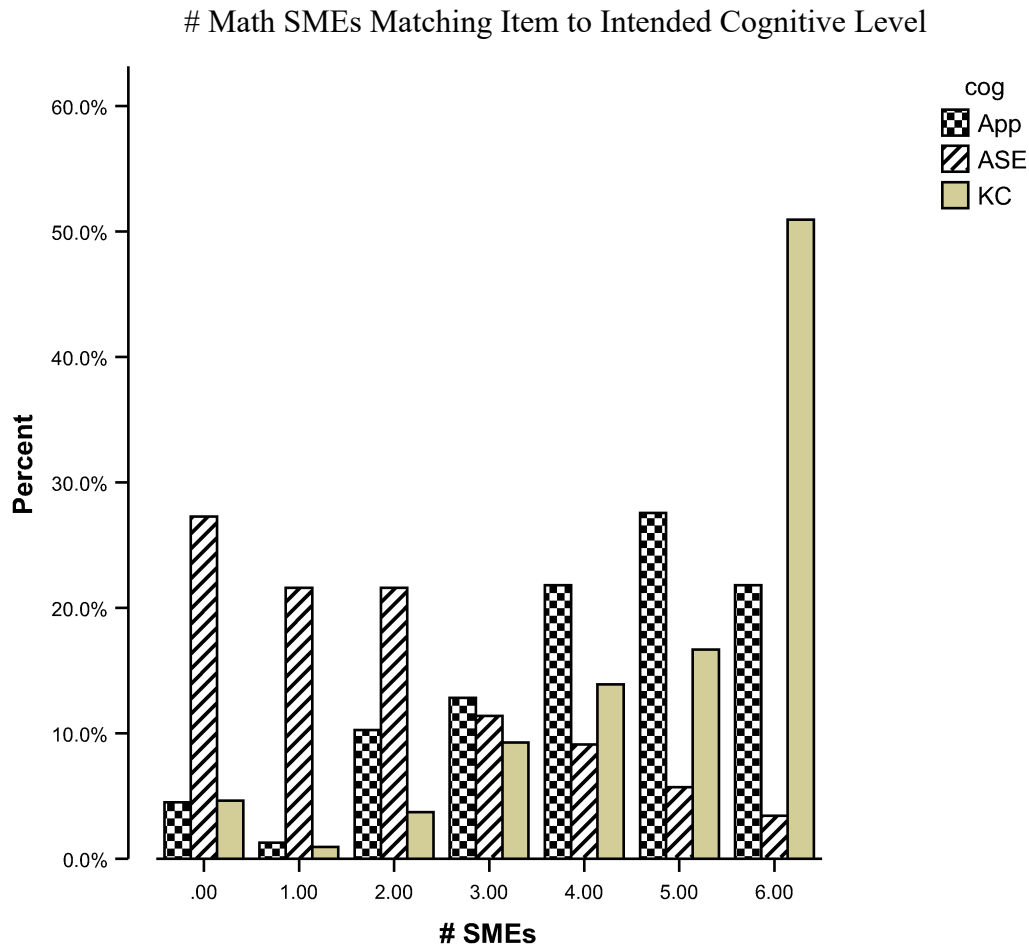
Figure VII.1

Math SMEs Matching Item to Intended Content Strand



Note: GM=Geometry & Measurement; NS=Number Sense; P=Patterns, Functions, & Algebra, and S=Statistics & Probability

Figure VII.2



Note: App=Application; ASE=Analysis/Synthesis/Evaluation;
KC=Knowledge/Comprehension

A final summary of the Math content validity results is presented in Table VII.3. In this table, we compare the SMEs' classifications of the items to the intended test specifications at each test level. Note that "test level" was used primarily for developing items and pilot test forms and does not pertain to how the MAPT is administered, since it is an MST. That is, although learners begin the test at an entry point corresponding to the specific test levels described in the test specification documents, they may be administered items at a higher or lower test level, depending on their performance as they progress through the test. Thus, there are not test "forms" specifically associated with test levels. Nevertheless, the comparison presented in Table VII.3 is useful for deciding the types of additional items that may be needed in future refreshments of the MAPT for Math item pool. For example, although all content strands are within 10% of the target percentages across all levels, Number Sense is consistently over-represented at test levels 2 through 4 and the Patterns, Functions, and Algebra content strand is under-represented at these test levels. These results suggest that more easy-to-moderate Patterns, Functions, and Algebra items should be added to the item pool in the future. It should be noted that the data presented in Table VII.3 pertain to the entire *pool* of MAPT for Math items

and that only those items meeting the consensus criterion for the content strand were used in assembling the operational modules.

Table VII.3
Comparison of Initial Test Specifications with SME Classifications: Mathematics and Numeracy

Test Level 2					
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Actual %	Target %
Number Sense	17	24	2	43%	35%
Patterns, Functions, Algebra	6	2	2	9%	15%
Statistics and Probability	8	13	5	26%	25%
Geometry and Measurement	14	6	2	22%	25%
Actual %	45%	45%	10%		
Target %	41%	41%	18%		
Test Level 3					
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Actual %	Target %
Number Sense	15	22	0	37%	30%
Patterns, Functions, Algebra	8	6	1	15%	20%
Statistics and Probability	5	15	4	25%	25%
Geometry and Measurement	13	11	0	24%	25%
Actual %	41%	54%	5%		
Target %	35%	45%	20%		
Test Level 4					
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Actual %	Target %
Number Sense	10	17	1	29%	25%
Patterns, Functions, Algebra	3	13	4	20%	25%
Statistics and Probability	5	16	8	30%	25%
Geometry and Measurement	7	14	0	21%	25%
Actual %	26%	61%	13%		
Target %	25%	50%	25%		
Test Level 5					
Content Strands	Knowledge & Comprehension	Application	Analysis, Synthesis, Evaluation	Actual %	Target %
Number Sense	2	10	0	12%	15%
Patterns, Functions, Algebra	12	19	6	37%	30%
Statistics and Probability	4	21	6	31%	30%
Geometry and Measurement	2	15	4	21%	25%
Actual %	19%	65%	15%		
Target %	18%	41%	41%		

Note: All cell entries represent percentages of items within a test level.

Survey of math SMEs' survey responses

The responses of the math SMEs to the post-meeting survey indicated they understood their tasks and the test specifications, and they believed their ratings would be useful for evaluating the degree to which the items they reviewed measured their intended specifications. In general, they indicated they thought the items generally represented the content strands and cognitive levels described in the test specifications. In terms of overall perception of the meeting, all of the SMEs reported they enjoyed the meeting and that it was easy for them to express their opinions. Five out of six panelists thought the group discussions were very helpful in the item rating process. In responding to the open-ended survey questions, all of the SMEs thought the content of the test is appropriate for measuring ABE learners' educational gains in Mathematics and Numeracy. For example, one panelist stated, "In general, I found the items were clearly written and the questions are in alignment with what I know to be in the frameworks." The SMEs were asked about the range of difficulty represented by the items they reviewed. In general, they thought the items they reviewed adequately represented an appropriate range of difficulty. Two panelists asked to see more ASE type questions and one panelist expressed a desire to see a few more straight computation questions.

The SMEs were also asked about the clarity of the Content Strand and Cognitive Area definitions. They all thought the Content Strand definitions were well written and helpful, but three recommended more extensive review of the frameworks in advance of the meeting to better understand how some topics are defined and categorized (for example, greater than/less than signs and exponents). This comment may explain why there may have been confusion between the Number Sense and Patterns, Functions, and Algebra content strands because some topics in Patterns, Functions, and Algebra intuitively seem like they deal with numbers. Regarding the cognitive levels, all of the SMEs thought the definitions were clear, but four out of the six noted that the ASE level was hard to identify in practice.

Summary

To summarize the content validity data for the MAPT for Math, all items were generally perceived as measuring the intended content strands and the intended cognitive levels. The results support the claim that the items represent the content domain defined by the tests specifications, but they also highlighted areas for future improvement. There appears to be a need to better distinguish between the Application and ASE cognitive levels, but it is clear the SMEs agreed that both lower and higher-level skills were being measured. Another area to be investigated is why many of the items from the Patterns, Functions, and Algebra strand were perceived as Number Sense items. This is most likely due to the specificity in the curriculum frameworks where topics such as inequality place value, number lines, and general operations expressed as equations are categorized as Patterns, Functions and Algebra, but intuitively seem to be dealing with aspects of number sense. This discrepancy would need to be resolved should content strand sub-scores become of interest.

Reading Content Validity Results

As mentioned earlier, there were 240 reading items reviewed in the content validity study, which represented nearly all of the items identified for potential use in the reading item pool. The content validity results for the reading skill dimension generally indicated a high level of congruence. Ninety percent of the items met or surpassed the congruence criterion (5 of 7 SMEs), and the association between the intended skill and the skill perceived to be measured by the SMEs was statistically significant ($\chi^2_{(4)}=366$, $p < .001$). With respect to the Purposes for Reading dimension, the overall congruence rating was 65%, and the association between intended purpose and the purpose perceived by the SMEs was also statistically significant ($\chi^2_{(9)}=356$, $p < .001$).

A summary of the results for the Skill Areas is presented in Table VII.4. The Comprehension and Vocabulary & Meaning areas exhibited strong congruence with more than 90% of the items achieving the 5 of 7 SME criterion. The congruence percentage for the Word Recognition area was lower, at 76%. The numbers of SMEs classifying the items in the intended Skill Area is displayed in Figure VII.3. In general, these results show a strong degree of congruence between the skill areas intended to be measured by these items and the SMEs' perceptions of the skill areas measured by each item.

Table VII.4

Skill Area	# Items	% Not Congruent	% Unanimous	% Congruent
Comprehension	109	6%	84%	94%
Vocabulary & Meaning	76	7%	55%	93%
Word Recognition	55	24%	27%	76%
Total	240	10%	62%	90%

Notes: "Not Congruent" refers to items that were matched to the intended content strand by less than 5 of the 7 SMEs. "Unanimous" refers to items matched to the intended strand by all 7 SMEs. "Congruent" refers to items matched by 5-7 SMEs.

With respect to the Purposes for Reading dimension, the congruence levels were generally lower, ranging from 85% (Acquire Basic Skills and Word Analysis) to 17% (Perform a Task). These results are summarized in Table VII.5. The very low congruence rating for the Perform a Task purpose is due to 10 of the 36 items being primarily rated as reading to Gain Information, and a split viewpoint on 19 other items between Gain Information and Perform a Task. Figure VII.4 displays the numbers of SMEs classifying the items in the intended Purpose for Reading Area.

Figure VII.3

Reading SMEs Matching Item to Intended Skill Area

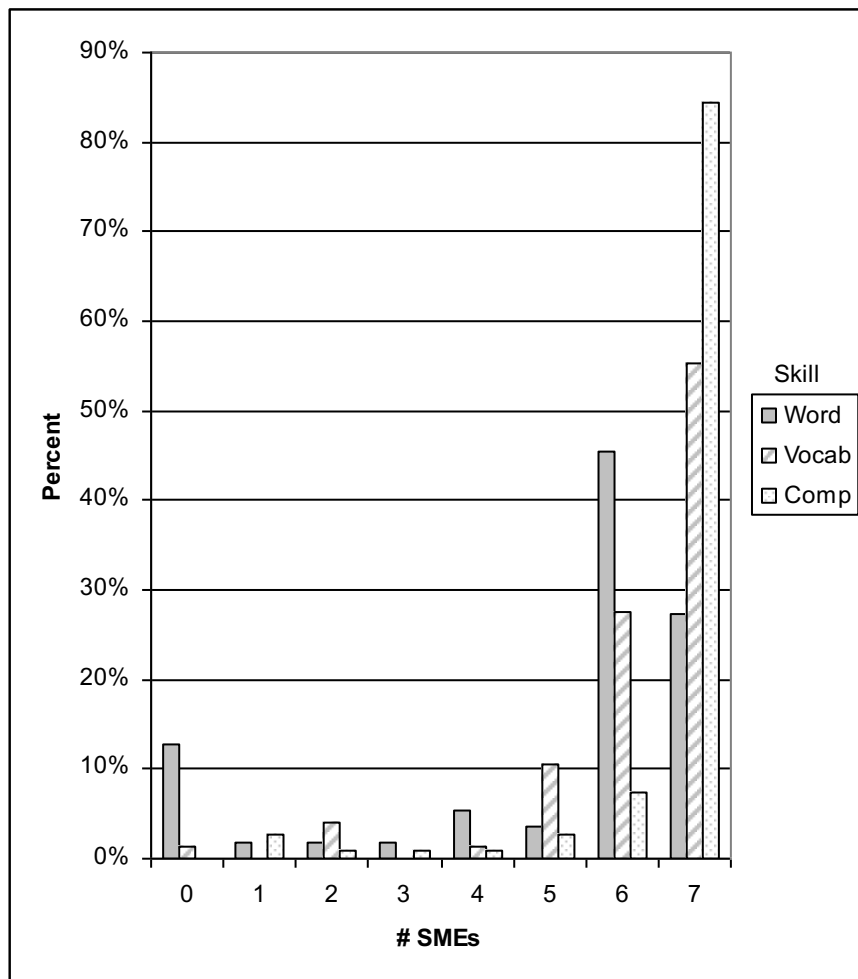


Table VII.5

Summary of Reading SMEs' Content Validity Ratings by Purpose for Reading

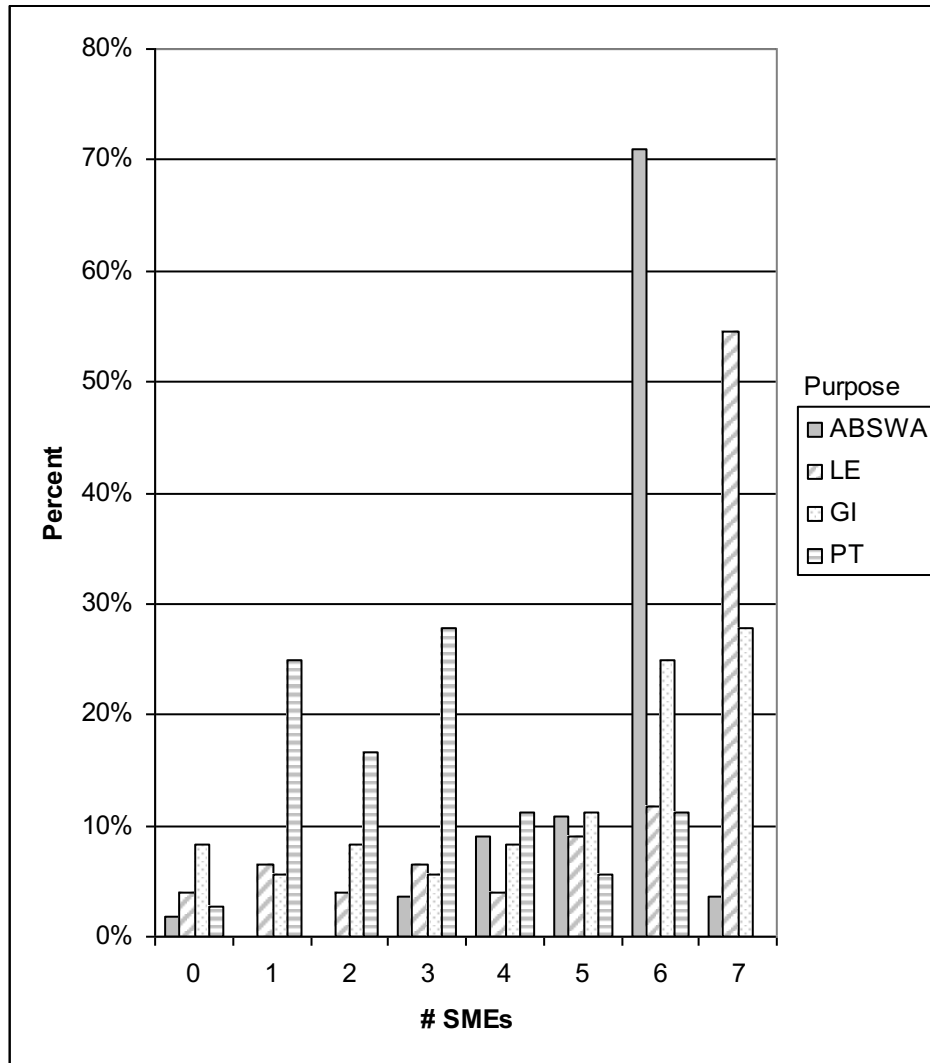
Purpose for Reading	# Items	% Not Congruent	% Unanimous	% Congruent
Acquire Basic Skills & Word Analysis	55	15%	4%	85%
Literary Experience	77	25%	55%	75%
Gain Information	72	36%	28%	64%
Perform a Task	36	83%	0%	17%
Total	240	35%	27%	65%

Notes: "Not Congruent" refers to items that were matched to the intended content strand by less than 5 SMEs.

Unanimous" refers to items matched to the intended strand by all 7 SMEs. "Congruent" refers to items matched by 5-7 SMEs.

Figure VII.4

Reading SMEs Matching Item to Intended Purpose for Reading



Note: ABSWA=Acquire Basic Skills & Word Analysis; LE=Literary Experience; GI=Gain Information; and PT=Perform a Task.

A final summary of the Reading content validity results is presented in Table VII.6. In this table, we compare the SMEs' classifications of the items to the intended test specifications at each test level. As mentioned earlier, "test level" does not accurately describe how the MAPT test is administered, since it is a multistage adaptive test. Thus, the "targets" in Table VII.6 do not reflect the content specifications for the MAPT per se. Nevertheless, by comparing the percentages of items in each Skill and Purpose area within these levels, we can see where additional items may be needed in future refreshments of the MAPT for Reading item pool.

The comparisons presented in Table VII.6 indicate very close Skill Area congruence

between the percentages of items classified by the SMEs and the percentages in the item pool for levels 2 and 4. At level 3, there are fewer Word Recognition items than expected and more Comprehension and Vocabulary and Meaning items. It should be noted that the data presented in Table VII.6 pertain to the entire *pool* of MAPT for Reading items and that only items meeting the consensus criterion for the skill area were used in assembling the modules that are used operationally. The congruence is lower for the Purposes for Reading dimension at all three levels, primarily due to the under-representation of the Perform a Task purpose at levels 3 and 4. Future item development efforts should target this purpose for reading area.

Table VII.6

Comparison of Initial Test Specifications with SME Classifications: Reading

Test Level 2						
Skill Areas	Purposes for Reading					
	Acquire Basic Skills & Word Analysis	Literary Experience	Gain Information	Perform Task	Actual %	Target %
Word Recognition	50	0	0	0	50%	50%
Vocabulary & Meaning	14	5	7	7	33%	30%
Comprehension	0	1	11	5	17%	20%
Actual %	64%	6%	18%	12%		
Target %	50%	18%	17%	15%		
Test Level 3						
Skill Areas	Purposes for Reading					
	Acquire Basic Skills & Word Analysis	Literary Experience	Gain Information	Perform Task	Actual %	Target %
Word Recognition	12	0	0	0	12%	20%
Vocabulary & Meaning	16	12	13	3	44%	30%
Comprehension	0	19	22	3	44%	50%
Actual %	28%	31%	35%	6%		
Target %	20%	30%	30%	20%		
Test Level 4						
Skill Areas	Purposes for Reading					
	Acquire Basic Skills & Word Analysis	Literary Experience	Gain Information	Perform Task	Actual %	Target %
Word Recognition	0	0	0	0	0%	0%
Vocabulary & Meaning	7	10	11	0	28%	30%
Comprehension	0	33	38	1	72%	70%
Actual %	7%	43%	49%	1%		
Target %	0%	42%	42%	16%		

Survey of reading SMEs

The responses of the reading SMEs to the post-meeting survey indicated they understood their tasks and the test specifications, although one SME indicated the directions for the rating task were confusing. Five of the 7 SMEs responded “definitely yes” when asked whether they thought their item ratings would be useful for evaluating the degree to which the Reading Skill areas were adequately represented by the MAPT items; the other 2 SMEs responded “probably yes.” With respect to the utility of their ratings for evaluating the Purposes for Reading, 3 SMEs responded “definitely yes,” 3 responded “probably yes,” and 1 responded “not sure.” Six of the 7 SMEs indicated they thought the MAPT items generally represented the Reading Skill areas. The other SME thought all Skill Areas were well represented with the exception of the Word Recognition items. Four SMEs thought the different Purposes for Reading were well represented; however, one noted that it was difficult to classify level 2 items with respect to purpose for reading. S/he commented that most of the level 2 items inquiring about word meaning and other facts did not seem to fit any of the listed purposes. The other 2 SMEs were not sure about how well the MAPT items represented the different Purposes for Reading.

In terms of overall perception of the meeting, all of the SMEs reported that they enjoyed the meeting and that it was easy for them to express their opinions. All SMEs indicated the group discussions were very helpful in the item rating process. On the open-ended questions, the SMEs provided several suggestions regarding future revisions of the MAPT for Reading, including (a) adding more “compare/contrast” items, (b) considering additional ways for testing word analysis skills for lower-level learners, and (c) adding items to better measure phonemic awareness and fluency. When asked to comment on the appropriateness of the range of difficulty of the items, 3 SMEs thought the range of difficulty was appropriate, but the other 4 thought more difficult items should be added for learners at level 4 (high intermediate) and above.

Summary

In summary, the content validity analysis of the MAPT for Reading items generally revealed strong congruence among the intended reading skill areas and the areas perceived to be measured by the SMEs. With respect to Purposes for Reading, the results indicate that many of the items measuring Perform a Task were not perceived to be associated with that purpose by the SMEs. The definition of that purpose, and the degree to which it differs from reading to Gain Information should be studied further. Adding more items that are more clearly associated with reading to perform a task to the item pool should be considered. The SMEs’ comments to the survey suggest that (a) more difficult items should be added to the pool, (b) different types of items for measuring word recognition for lower-level learners should be considered, and (c) a new purpose for reading area should be used to describe many of the word recognition items.

3. Validity Evidence Based on Internal Structure

Analysis of the internal structure of a test can provide validity information regarding the degree to which the dimensionality of the test data conforms to the expected dimensionality. The MAPT exams are designed to be unidimensional, that is, they are designed to measure a general continuum of reading or mathematics proficiency.

Methods for evaluating test structure include factor analysis and multidimensional scaling analysis. However, these procedures are typically applied to intact test forms, not adaptive tests where tests are uniquely constructed for each examinee. Thus, to evaluate the internal structure of the MAPT, a latent-modeling approach was needed. Fortunately, since the MAPT is scaled using IRT, such an approach is possible. Specifically, we examined the departure of the data from the hypothesized unidimensional IRT model, using residual analysis.

Examining residuals is one way to evaluate model fit in an IRT context. When the residuals are large, it could indicate multidimensionality, or under-parameterization of the data (i.e., a more general IRT model may be needed). Residuals are obtained by comparing the observed proportion correct, within a given range, to the proportion correct predicted by the IRT model. Residuals are typically standardized by dividing them by the standard error of the expected (model-predicted) proportion correct. Standardization is done to make residuals comparable.

The expected frequency distribution of standardized residuals is not known, although one might expect it to conform to a standard normal distribution (Hambleton, Swaminathan, & Rogers, 1991). In any event, the expected frequency need not be known: the observed distribution of standardized residuals can be compared to the frequency distribution using model-based (simulated) data. Under the null hypothesis that the model fits the data, the distribution of standardized residuals using simulated data can be thought of as an empirically generated sampling distribution against which the observed frequency distribution can be compared (Hambleton, Swaminathan, & Rogers, 1991).

Figures VII.5 and VII.6 show the frequency distribution of standardized residuals for MAPT for Reading and MAPT for Mathematics and Numeracy, respectively. In the case of the real data, residuals were computed using response data, proficiency estimates, and item parameter estimates from Pilot 2, although only those items that were eventually selected for operational use are included in the analysis. In the case of simulated data, the analysis was identical to the real data analysis except that the observed examinee responses were replaced with simulated responses based on the proficiency estimates, and item parameter estimates from Pilot 2. Additionally, as a point of reference, a standard normal probability density function is also shown on each figure.

It can be seen from these figures that the frequency distribution of standardized residuals for the observed data are highly similar to the distribution for the simulated data. This suggests that the IRT model is fitting the response data quite well, which provides evidence that the data for each exam are unidimensional, as they are intended to be.

Figures VII.7 through VII.13 present the frequency distribution of standardized residuals for each EFL within each subject area. This was accomplished by analyzing each subset of level-

specific Pilot 2 forms separately. Note that these forms do *not* contain items from their respective levels exclusively; some number of linking items from adjacent levels also appear on every form. In general, the patterns of results indicate adequate fit of the data to the unidimensional model.

Figure VII.5

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Operational MAPT for Reading Items

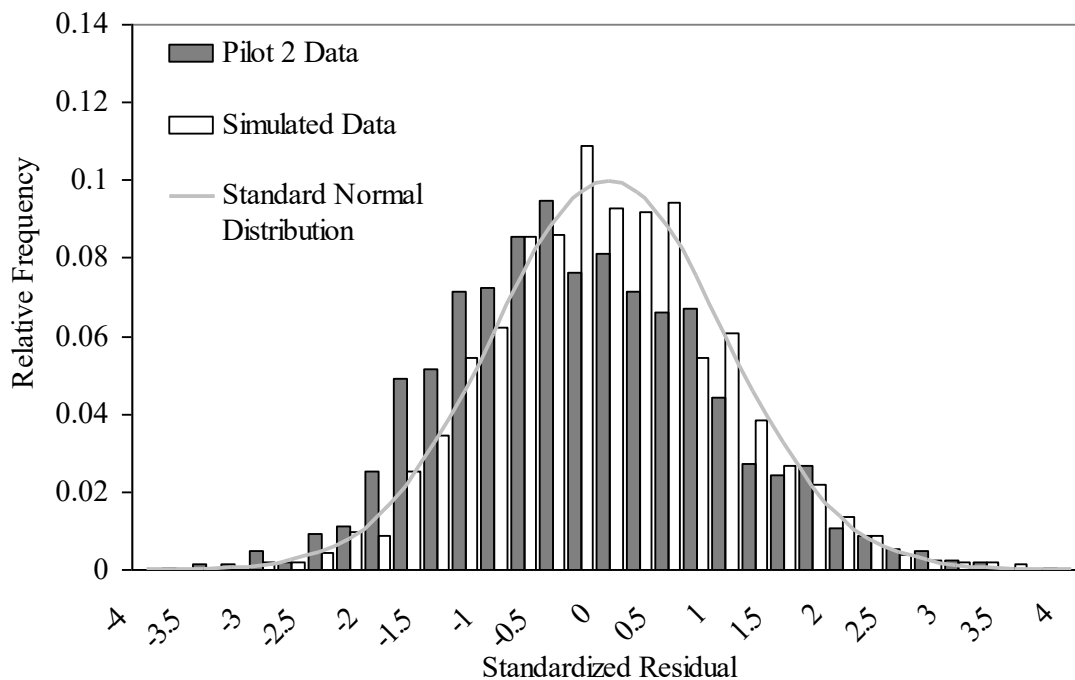


Figure VII.6

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Operational MAPT for Math and Numeracy Items

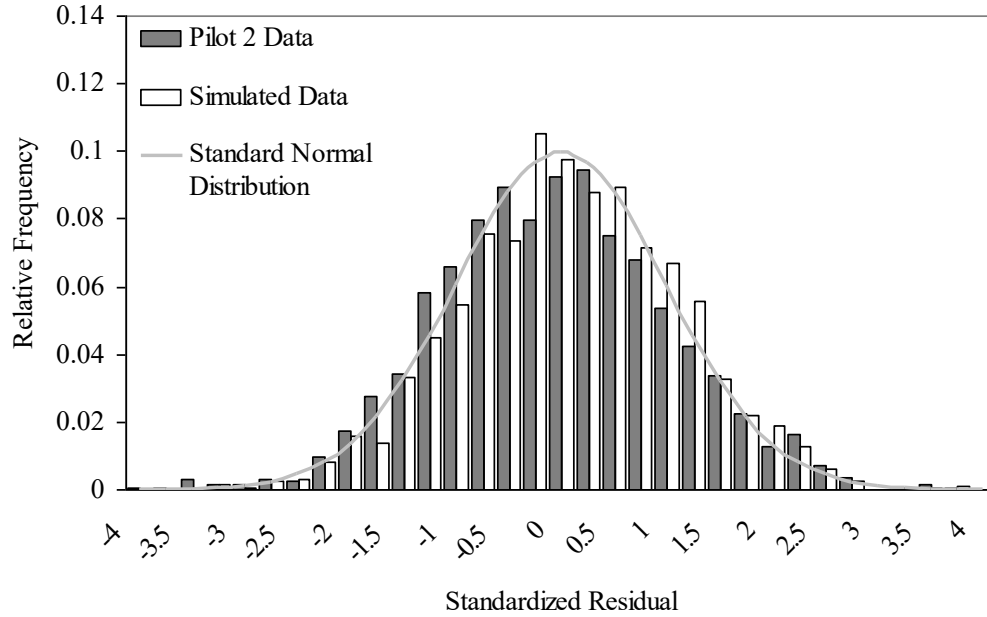


Figure VII.7

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 2 Operational MAPT for Reading Items

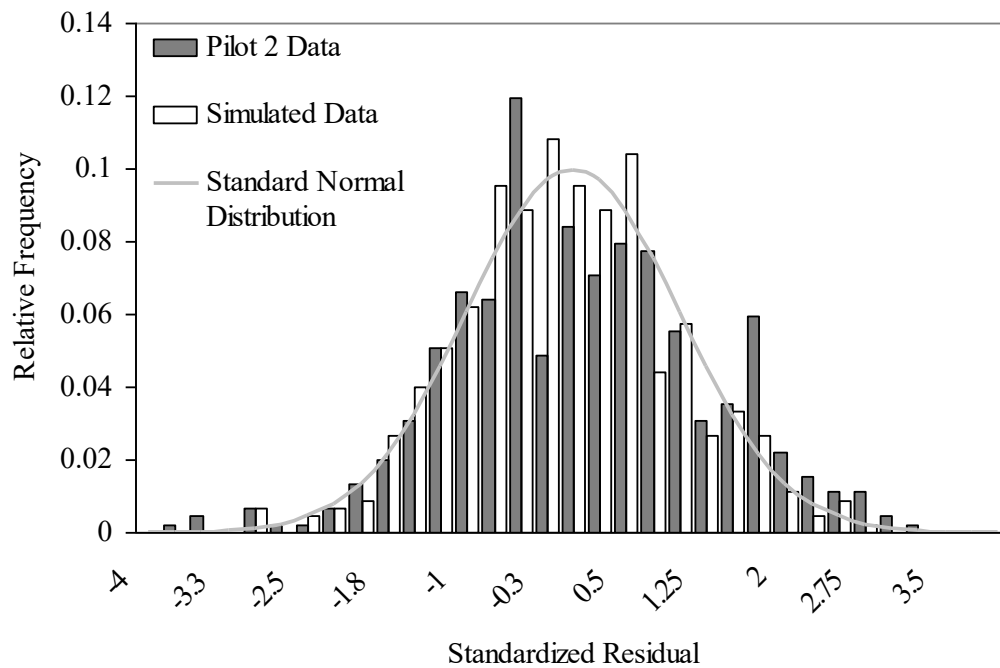


Figure VII.8

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 3 Operational MAPT for Reading Items

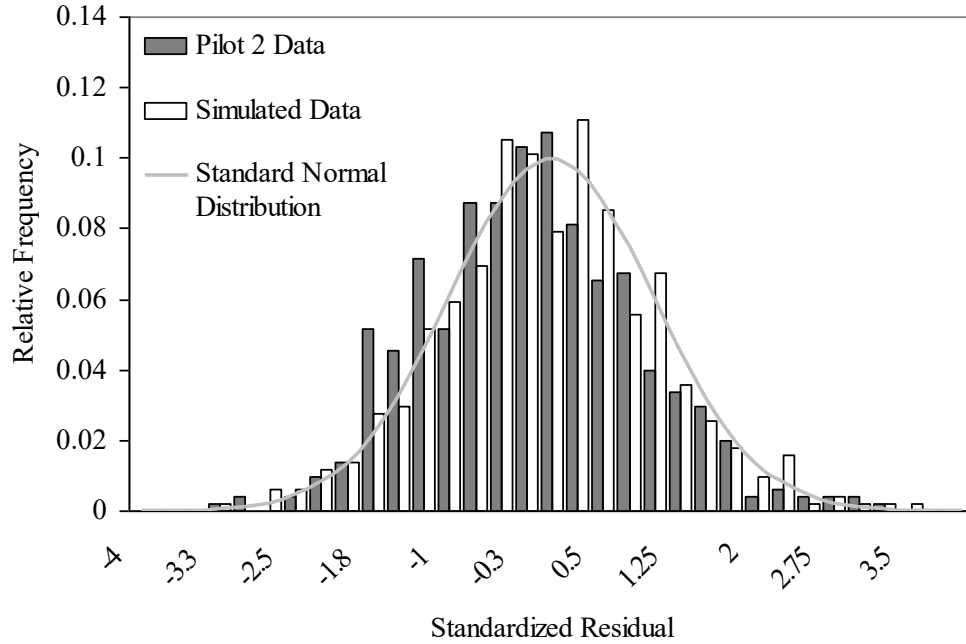


Figure VII.9

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 4 Operational MAPT for Reading Items

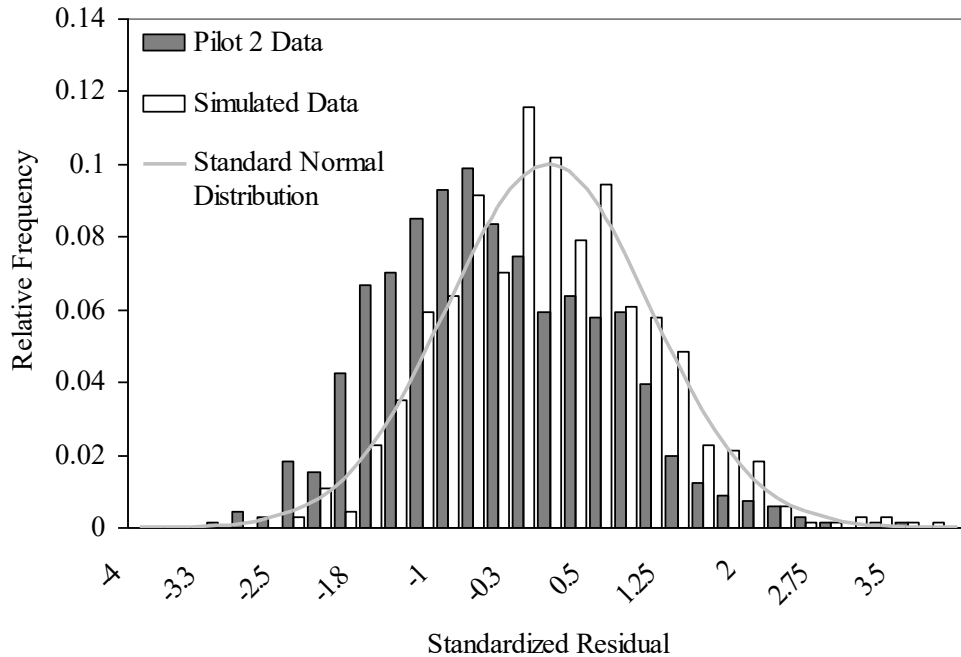


Figure VII.10

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 2 Operational MAPT for Math and Numeracy Items

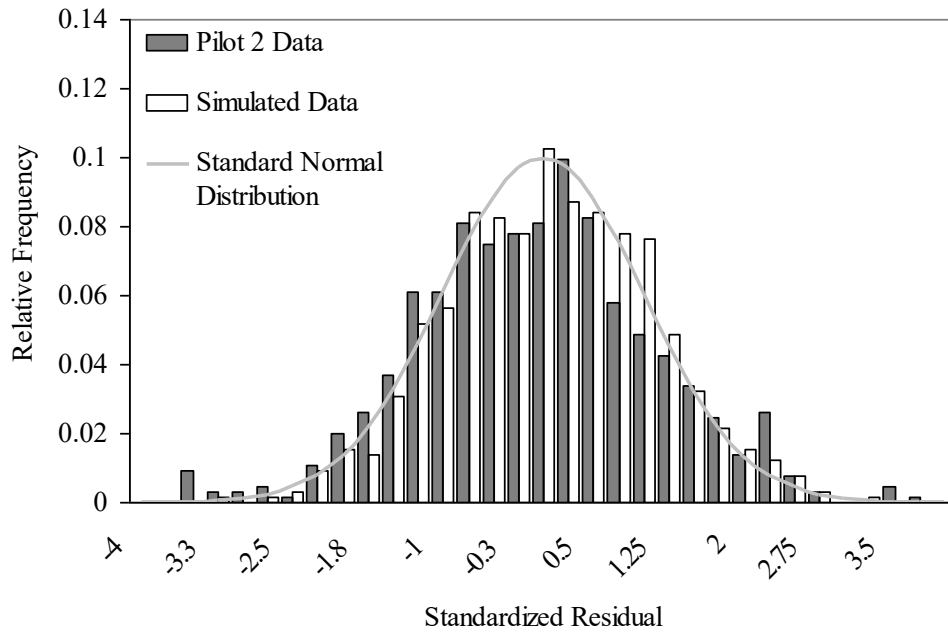


Figure VII.11

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 3 Operational MAPT for Math and Numeracy Items

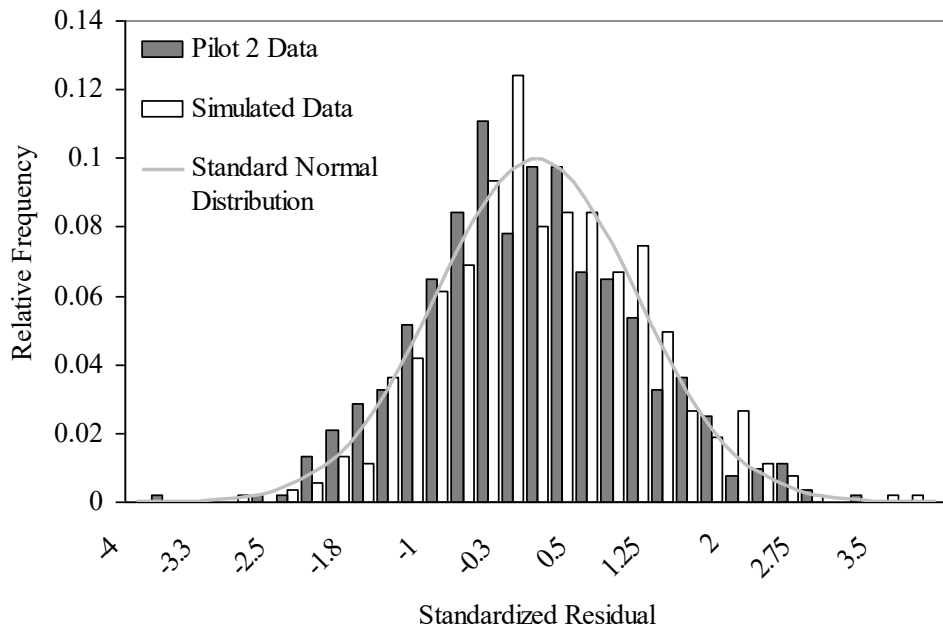


Figure VII.12

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 4 Operational MAPT for Math and Numeracy Items

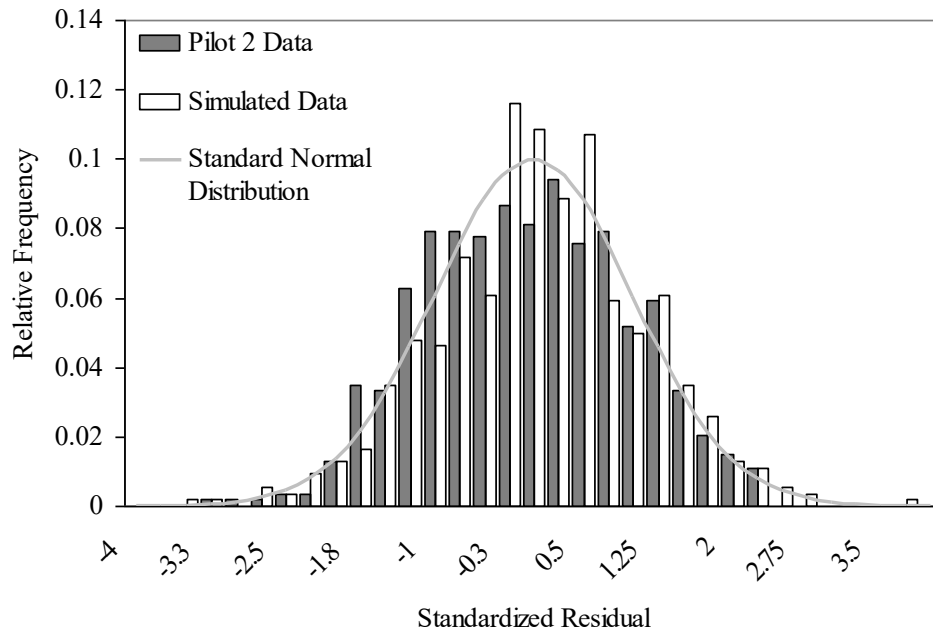
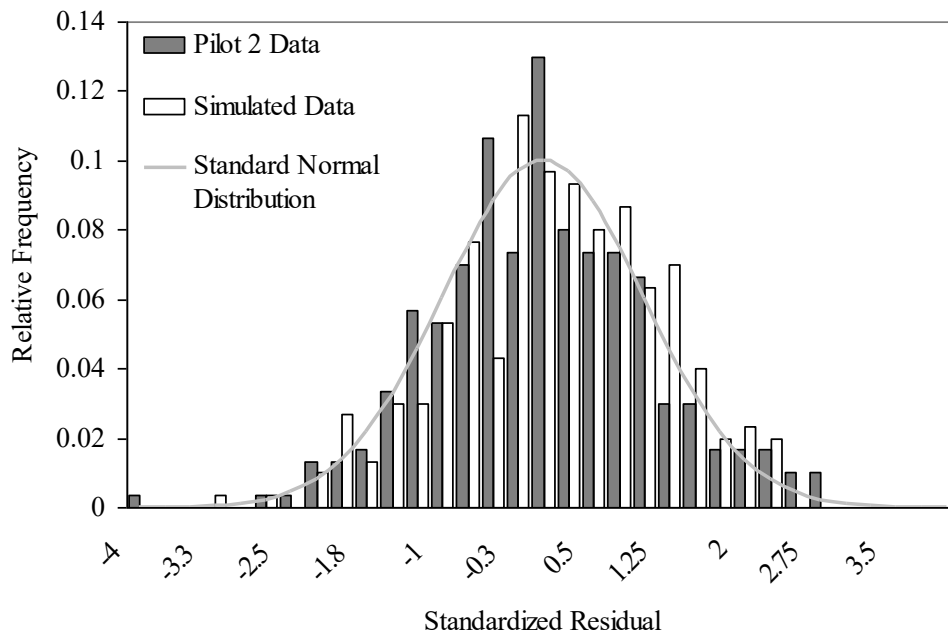


Figure VII.13

Frequency Distribution of Standardized Residuals for Pilot 2 Data and
Simulated Data: Level 5 Operational MAPT for Math and Numeracy Items



4. Validity Evidence Based on Relations to Other Variables

Studying the relationships among test scores and other measures of the construct tested aids in understanding the degree to which test scores support their intended uses. As described in Chapter I, the MAPT is designed to measure ABE learners' achievement with respect to the ABE curriculum frameworks in math and reading developed by the Massachusetts Department of Education. There are no independent measures of learners' achievement relative to these frameworks, which makes criterion-related validation studies problematic at best. While the MAPT were being developed, many ABE learners in Massachusetts were required to take the Test of Adult Basic Education (TABE). Although it has been documented that the TABE has poor overlap with the Massachusetts Curriculum Frameworks (Massachusetts Adult Basic Education Performance Accountability Working Group, 2002), the relationship between MAPT and TABE scores is of some interest. If two tests are measuring similar, but not identical, constructs we would expect a positive relationship among the scores from learners who took both tests.

Given that many ABE learners who participated in our pilot tests also took a TABE test, we created a data file that included these two sets of scores and computed correlations between the TABE math test and the MAPT for Mathematics and Numeracy, and between the TABE reading test and the MAPT for Reading. This file contained 1,870 learners who had both TABE and MAPT math test scores, and 2,570 learners who had TABE and MAPT reading scores. The MAPT scores came from Pilot Test 2, which occurred in the fall of 2005. The TABE test scores came from fiscal year 2006 (July 2005, through May 2006).

Scatter plots displaying the relationships among the test scores are presented in Figures VII.14 and VII.15 for math and reading, respectively. The correlation for math was .71 and for reading was .66. These results are as expected. They indicate the two tests are measuring similar, but distinct constructs, which supports the concurrent validity of MAPT scores.

At this juncture, no other data are available for conducting concurrent or predictive validity studies on the MAPT. Other data that could be useful for such studies include teacher evaluations of learners' performance in math and reading. We recommend gathering such data in a standardized format and evaluating their relationship with MAPT scores on a periodic basis.

Figure VII.14

Scatter Plot of MAPT for Math and TABE Math Test Scores (n=1,870, $r=.71$)

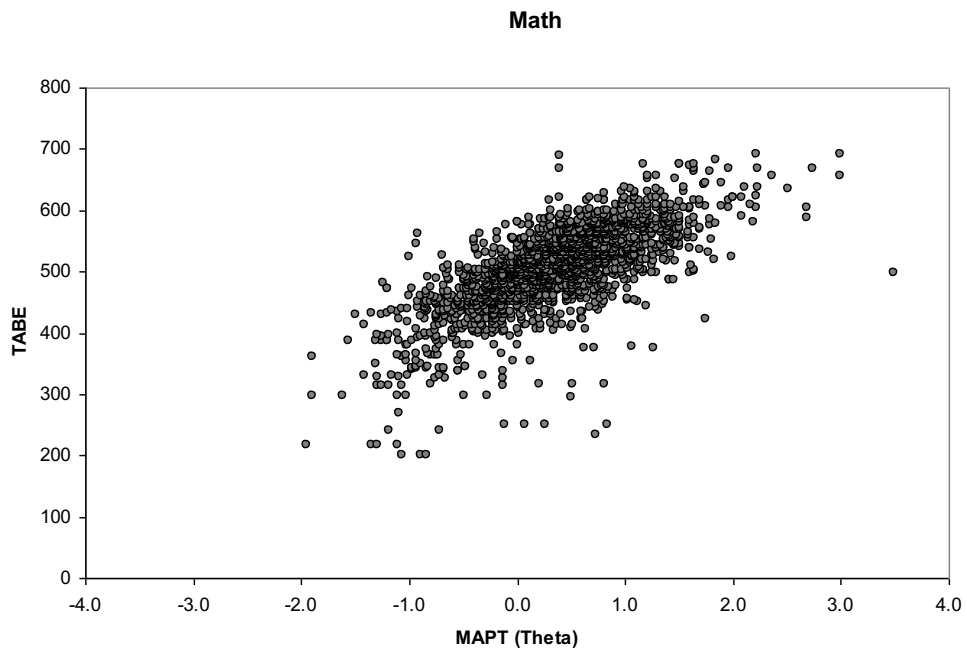
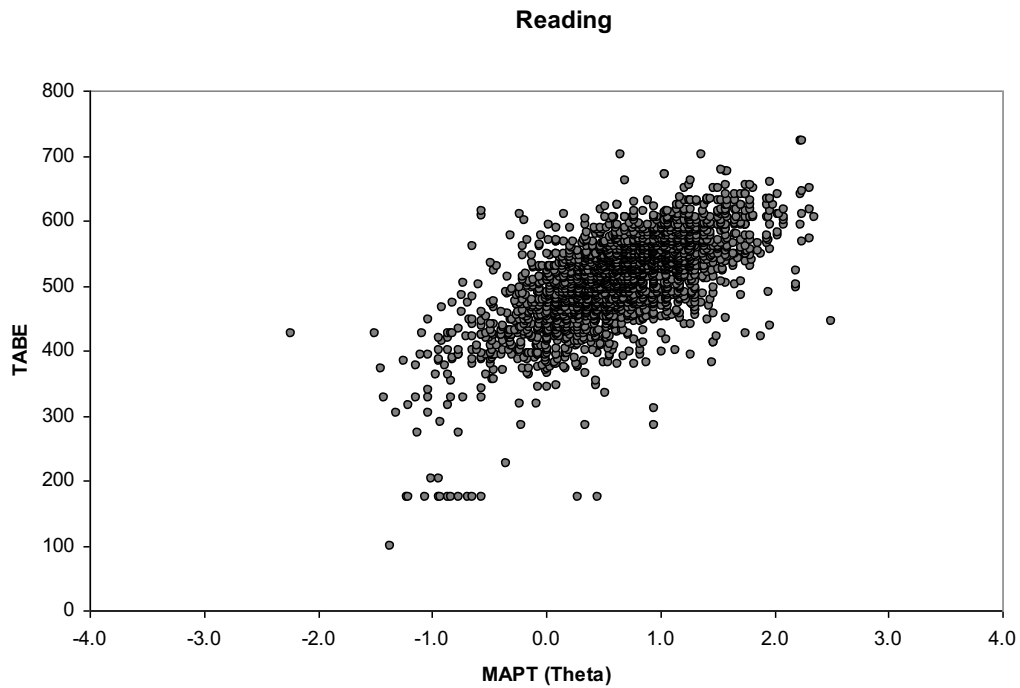


Figure VII.15

Scatter Plot of MAPT for Reading and TABE Reading Test Scores (n=2,570, $r=.66$)



5. Validity Evidence Based on Consequences of Testing

As described earlier, an evaluation of the positive and negative consequences of a testing program can provide valuable information regarding the degree to which a test is fulfilling its intended goals and the degree to which the scores are appropriately used. One of the underlying principles of the MAPT is that it will improve instruction. Thus, this principle is a positive, intended consequence of the testing program. The logic behind this principle is that if a test were aligned to the curriculum frameworks, it would encourage teachers to teach to those frameworks.

At the time of this writing, it is too early to directly study whether the test is affecting instruction in a positive way, since it has been operational for only a very short time (operational testing commenced July 21, 2006 when most programs are not open full-time). However, we already see many positive consequences such as teachers becoming more familiar with using computers in their classroom, familiarizing their students with how to interact with a computer, and experiencing professional development workshops pertaining to assessment issues for adult learners.

MAPTs' Influence on Professional Development for ABE Teachers and Staff

As stated in Chapter I, our vision for the MAPT was to have the exam essentially created by adult educators in Massachusetts. By doing so not only would we maximize the link between instruction and assessment, we would also enable these educators to be part of the process and have some ownership over the exam. In accomplishing this goal we consequently provided considerable professional development opportunities for the ABE field. For example, we delivered over a dozen workshops across the state over a two-year period. About half of these workshops focused on item writing, the others focused on understanding educational tests (e.g. "Assessment 101") and understanding the characteristics of the forthcoming MAPT. In all the professional development workshops, the curriculum frameworks from which the MAPT are derived were discussed. Additional workshops were provided to train teachers how to deliver the pilot tests, and eventually the MAPT, at their programs. We also provided a three-credit graduate course, which was free of charge to ABE teachers and staff throughout the state, although they were required to create test items to be considered for use on the MAPT as a course requirement. Fourteen teachers participated in the course. We also offered a 15-hour introduction to assessment course for the System for Adult Basic Education Support (SABES) curriculum and assessment coordinators throughout the state. The idea behind that course was to "train the trainers" so that the assessment literacy among ABE teachers and staff throughout the state would be elevated. Thus, a positive consequence of the MAPT testing program was that hundreds of ABE educators throughout Massachusetts were trained in assessment literacy, learned valuable item writing and editing skills, and became familiar with the learning standards in the Massachusetts ABE Curriculum Frameworks.

Surveys of Learners Participating in the Pilot Tests

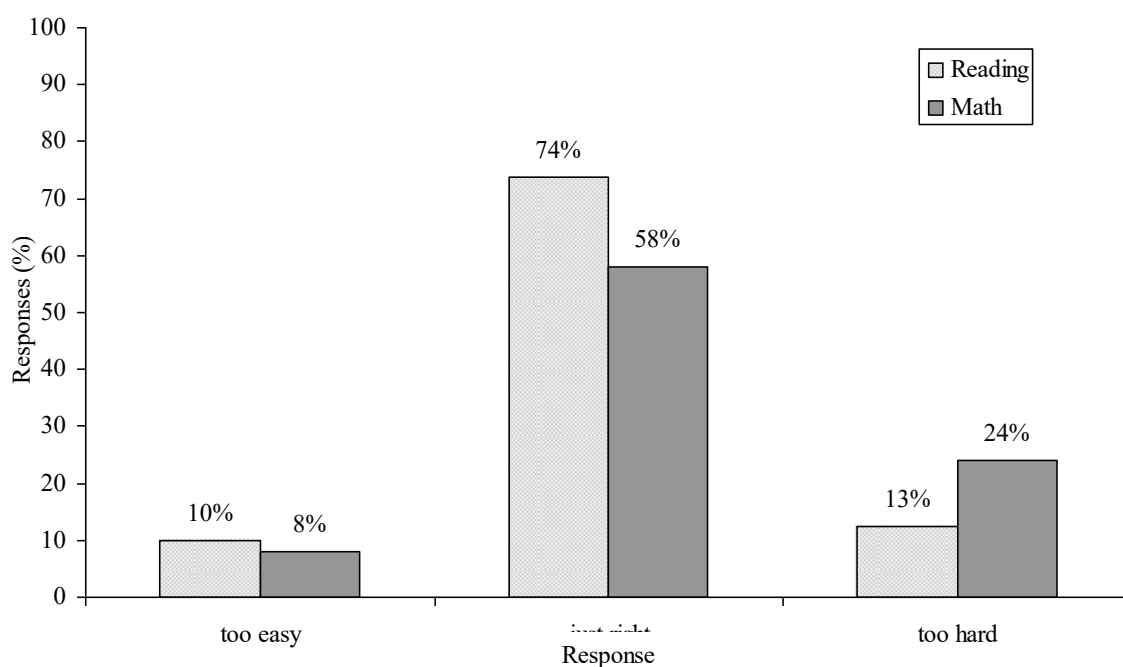
Evaluating the consequences of a testing program should involve investigation of students' experiences taking the tests and ensuring irrelevant factors do not impede their performance. Although we were concerned about the amount of time it might take learners to complete the pilot tests, we thought it was important to ask them a few questions to help us

understand their test taking experience and background. Three survey items were presented to learners at the conclusion of Pilot Test 1. Two of these items were asked again during Pilot 2. These questions inquired about learners' (a) perception of test difficulty, (b) opinion about the computer-based test delivery (Pilot 1 only), and (c) whether they were native English speakers. With respect to native language, about 1/3 of the Pilot 1 learners were not native speakers of English. This percentage increased to about 40% for Pilot 2.

With respect to perceived test difficulty, the majority of the learners thought the Pilot 1 tests were "about right" (see Figure VII.16). The results for Pilot 2 (Figure VII.17) were very similar for Reading, but for Math, the percentage of learners rating the test "too hard" increased. This finding may be explained by the addition of more difficult items to the math item pool. About 68% of Pilot 1 learners and 57% of Pilot 2 learners found the test to be "just right." Overall, these results suggest that the majority of learners felt comfortable with the difficulty level of the tests, but some perceived the tests as being too easy or difficult. Measuring these perceptions during operational testing is recommended.

Figure VII.16

Pilot 1 Examinee responses to "How difficult was this test for you?"



In Pilot Test 1, learners were asked, if they had a choice, would they prefer to take the test on a computer or in a paper-and-pencil format. The results are summarized in Figure VII.18. About 77% of the Pilot 1 sample responded they would prefer to take the test on the computer. The percentage preferring the computer was noticeably higher for Reading than for Math, and learners taking the Math exam had a noticeably higher omit rate for this question. Interestingly, of those who found Pilot 1 too difficult, only 28% (5% of the total sample) indicated they would

prefer a paper-and-pencil test. This suggests that the computer delivery model was not a major contributor to perceived test difficulty.

Figure VII.17

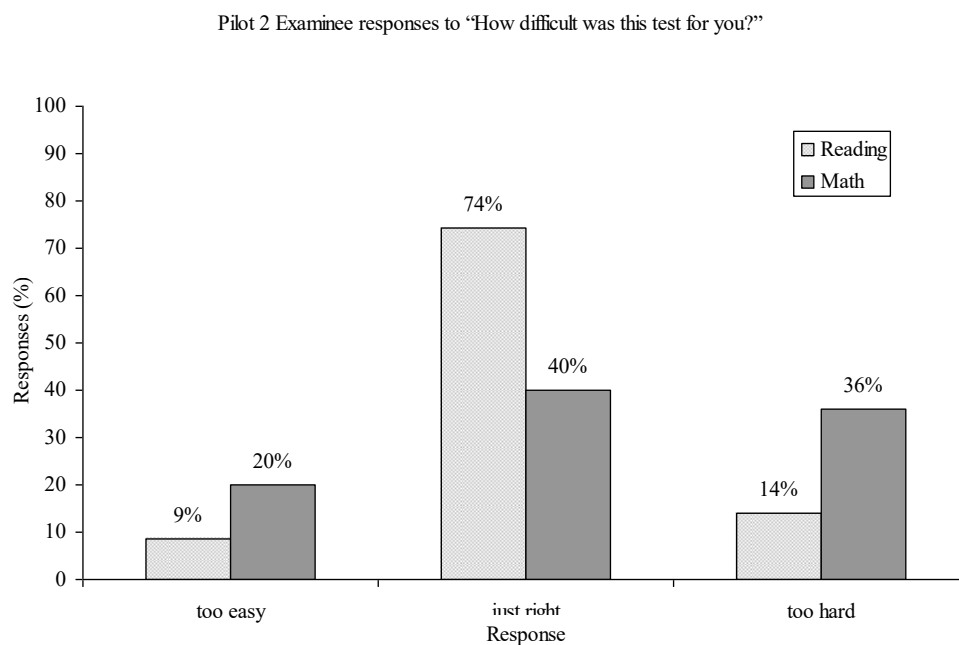
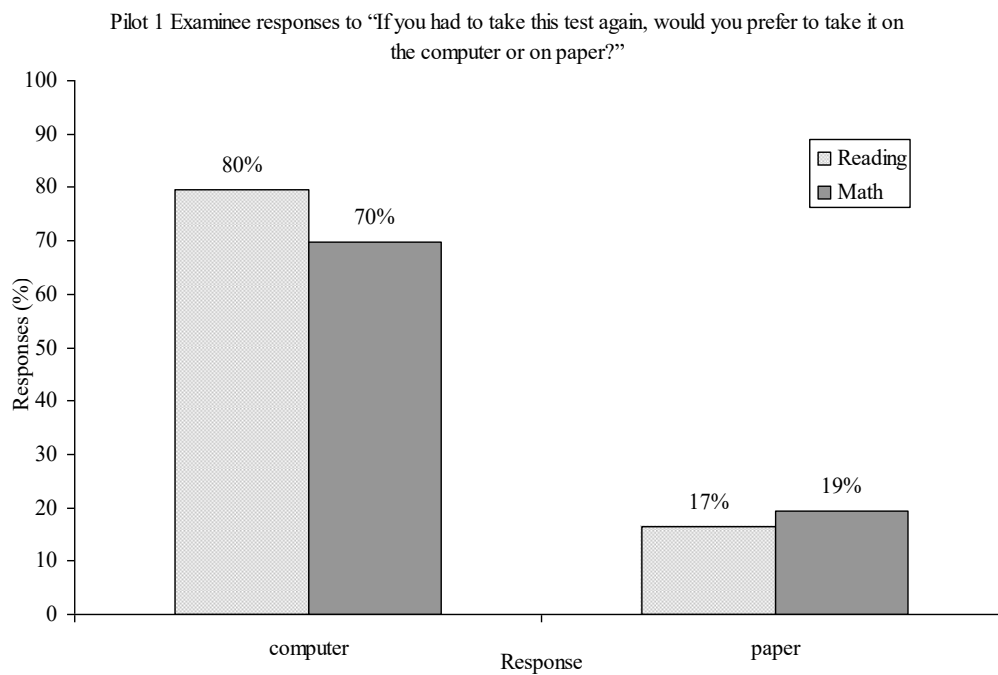


Figure VII.18



6. Validity Evidence Supporting Evaluation of Educational Gain

As discussed in Chapter V, under the NRS, learners' educational gains are measured by comparing their initial EFL at intake with their EFL at some later point in time after instruction. Specifically, for the purposes of the NRS, any learner who advances from one EFL to a higher EFL in the course of a given fiscal year is classified as demonstrating education gain. Any learner failing to advance in this manner, is considered to have *not* shown educational gain.

Because of measurement error, there will always be some number of learners that show educational gain despite an insufficient increase in their true proficiency. Likewise, there will always be some number of learners that fail to show educational gain, even though their true gain is enough to be classified into the next EFL. When a learner is classified as exhibiting gain, but the gain is really due to measurement error, it is called a *false positive* error. When a learner truly advances an EFL, but due to measurement error is classified as not exhibiting gain, it is called a *false negative* error. Our estimates of decision accuracy were described in Chapter III. In this chapter, we describe a simulation study that was conducted to evaluate the sensitivity of the MAPT to capture educational gain. This is an important validity study because one of the primary uses of the MAPT is to classify learners' gains according to the NRS, which defines gains in terms of moving from a lower EFL to a higher EFL. It is important to note, however, that learners' gains are aggregated for accountability purposes, which is likely to be the most common use of MAPT scores. Thus, our simulation study also considers aggregate gains.

Study Overview and Conditions

For each subject, Reading and Mathematics and Numeracy, and each study condition, 25,000 simulated learners (simulees) were randomly assigned an initial true proficiency between 200 and 700. Each simulee was then administered a MAPT (administration 0), starting at the appropriate level given their true proficiency. Responses were drawn at random, given the item parameters and the simulees' true proficiency. For example, if the probability of a correct response, given a particular item and a particular true proficiency, were .70, 70% of the time a correct response would be drawn and 30% of the time an incorrect response would be drawn. Simulees' observed (i.e., estimated) scores on the initial (administration 0) MAPT were not used except to use to determine the appropriate starting level for the first recorded administration, administration 1.

The remaining administrations, beginning with administration 1, were conceptualized as constituting a single evaluation period, which in practice corresponds to the instruction that occurs within a given fiscal year. Over the course of an evaluation period, each simulee then takes two administrations and simulee proficiency increases by some amount, which is also varied as a study condition (in 10-point increments).

At the conclusion of the evaluation period, each simulee was classified as having demonstrated gain or not with respect to their true gain and their observed gain, where the observed gain is what we expected from the operational MAPT administration. Gain was demonstrated by moving up at least one level (i.e., EFL) between pretest and posttest. Classification success is defined here as the correspondence between the observed classification and the true classification for each simulee. For illustrative purposes, Table VII.7 considers the 8

possible score trends for simulees taking 3 administrations. These individual score trend examples warrant further discussion because understanding them is critical for understanding the study results.

Simulees 1 to 4 share the same true scores: all increase 40 points during the evaluation period (220, 240, 260); however because all the scores fall between 200 and 299, no true gain occurs (i.e., simulees 1-4 all remain in EFL 2). For simulee 1, the observed scores all remain between 200 and 299 and thus no observed gain occurred either. Therefore, simulee 1 is correctly classified. Simulee 2 has the same true scores as simulee 1 but the simulee 2 observed scores are all > 300. Clearly, there is considerable error in simulee 2's observed scores. Still, even though simulee 2's classification is based on erroneous data, the classification itself is nevertheless correct: both the true and observed classifications show no gain, because this simulee was misclassified as a Level 3 learner after Administration 1. Simulee 3, again with the same true scores, shows yet another pattern of observed scores: the observed scores start at level 3 but then move down to level 2. Since simulee's true score is increasing, the decline in observed score can only be attributed to measurement error. Nevertheless, the classification again remains correct: both the true and observed classifications show no gain. Simulee 4, again with the same true scores, shows yet another possible score trend. Simulee 4 shows gain by scoring at level two and then subsequently scoring at level 3. However, the true gain for simulee 4 was below the threshold for gain under the NRS, and is therefore an example of a false positive error.

Simulees 5 to 8 share the same true scores: 270, 290, and 310 across administrations 1, 2 and 3, respectively, and so these four simulees should be classified as exhibiting gain under the NRS, because they move from EFL 2 to EFL 3. Simulee 5's observed scores demonstrate gain and therefore simulee 5's classification is correct. Simulee 6 and 7, on the other hand, fail to show gain in their observed scores, remaining in levels 2 and 3 respectively, and therefore are examples of false negatives. Simulee 8 is also a false negative because while the true score moves up a level, the observed score moves down a level.

In the example from Table VII.7, simulees gained 40 points but for the simulation study, the number of points varied from 0 to 100 in increments of 10. The study was conducted for both Reading and Mathematics and Numeracy. For each administration, 25,000 simulees were created and assigned an initial proficiency drawn from a uniform distribution between 200 and 700 (inclusive). Given the 11 levels of gain and 2 subject areas, the simulation study involved 550,000 simulees.

Table VII.7

Simulee		Administration 1		Administration 2		Administration 3		Classification		Observed
		true	observed	true	observed	True	observed	true	observed	Classification Success
<i>No True Gain</i>	1	220	212	240	237	260	272	no gain	no gain	correct
	2	220	303	240	314	260	319	no gain	no gain	correct
	3	220	306	240	261	260	273	no gain	no gain	correct
	4	220	235	240	255	260	304	no gain	gain	false positive
<i>True Gain</i>	5	270	262	290	288	310	317	gain	gain	correct
	6	270	254	290	278	310	296	gain	no gain	false negative
	7	270	332	290	344	310	361	gain	no gain	false negative
	8	270	308	290	326	310	294	gain	no gain	false negative

Results

Figure VII.19 presents a summary of the results for the MAPT for Reading. The figure shows the proportion of simulees that correctly showed (or correctly failed to show) gain, erroneously showed gain (false positive error), and erroneously failed to show gain (false negative). Proportions are shown as a function of score gain. There are two features of this figure that deserve special emphasis. First, the upper dashed line indicates the proportion of simulees correctly classified. It is desirable that this line be as high as possible. For example, if all simulees gained 40 points over the course of the evaluation period, one would only expect about 56% to be correctly classified. This is perhaps an uncomfortably low percentage, however, when aggregating examinee performance (e.g., when reporting state results), the magnitude of this percentage is tempered by the second feature of the figure: the ratio of false negatives to false positives. The ratio of false negatives to false positives is of interest because upon aggregating results, false positives and false negatives cancel out. For example, given a sample of 100 examinees, all of whom gained 40 points over the of the evaluation period, the expected number of examinees correctly classified would be about 56 (56%), the expected number of false positives would be about 22 and the expected number of false negatives would also be about 22. However, note that if 22 examinees erroneously show gain and 22 erroneously fail to show gain, the *net* error is zero. That is, for this sample of 100, one could perfectly describe examinee classification despite the unfortunate fact that 44% of the examinees were wrongly classified! For this reason, it is especially desirable that the discrepancy between false positives and false negatives be minimized.

Figure VII.20 presents the results for the MAPT for Mathematics and Numeracy. The proportion of correctly classified simulees is higher than that observed for Reading, as is the discrepancy between false positives and false negatives. These findings are consistent with the results from the decision consistency simulation study discussed in Chapter III, wherein it was found that Mathematics and Numeracy simulees were expected to be more accurately classified (into an EFL) for a given administration than were Reading simulees.

Figures VII.21 and VII.22 show the *net* proportions of correct and incorrect simulee classifications versus the magnitude of score gain per evaluation period for Reading and Math, respectively. It can be seen that the net effects are similar to the simulated true values, which indicates the MAPT appears to be appropriate for evaluating gain at the aggregate (state, program) level.

Figure VII.19

Proportion of Correct, False Positive, and False Negative Classifications Vs. Magnitude of Score Gain Per Evaluation Period
(Reading; 2 Administrations Per Evaluation Period)

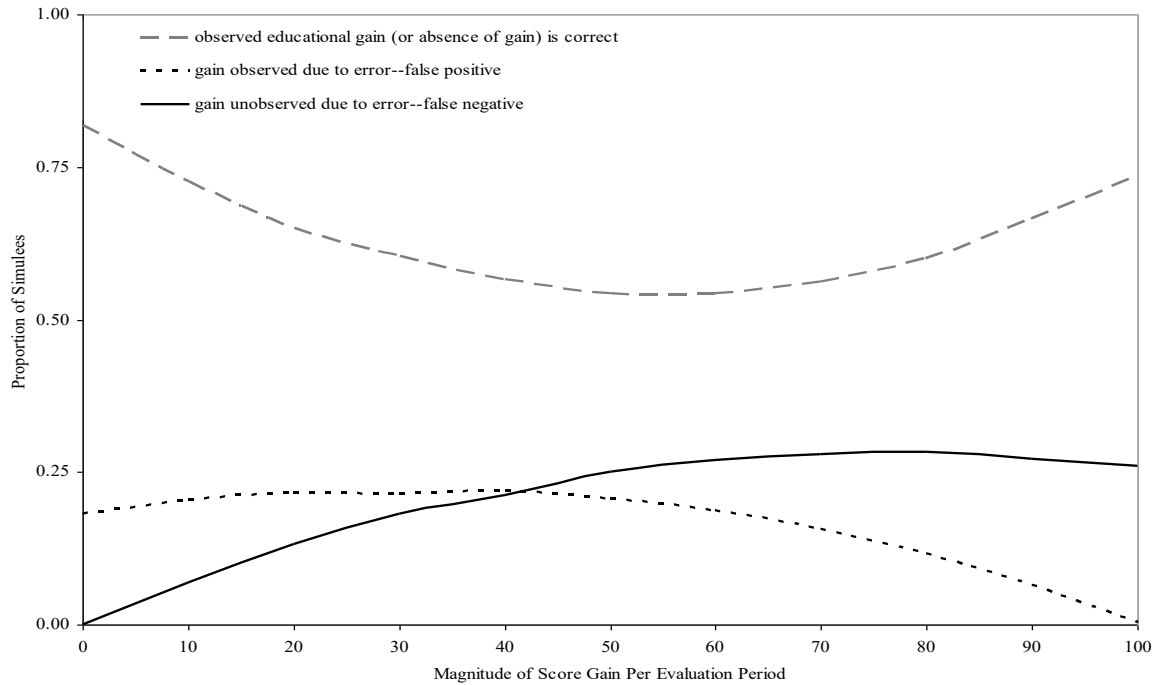
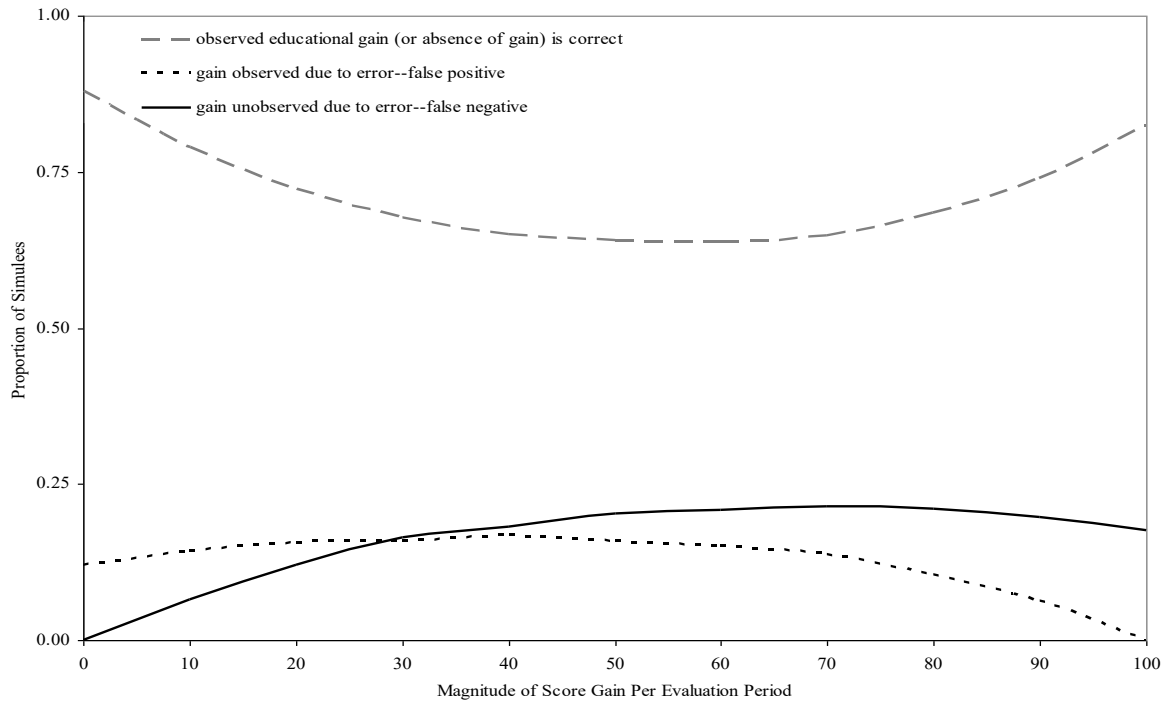


Figure VII.20

Proportion of Correct, False Positive, and False Negative Classifications Vs. Magnitude of Score Gain Per Evaluation Period
(Mathematics and Numeracy; 2 Administrations Per Evaluation Period)



Figures VII.21

Proportion of Net Correct, Net False Positive, and Net False Negative Classifications Vs. Magnitude of Score Gain Per Evaluation Period (Reading; 2 Administrations Per Evaluation Period)

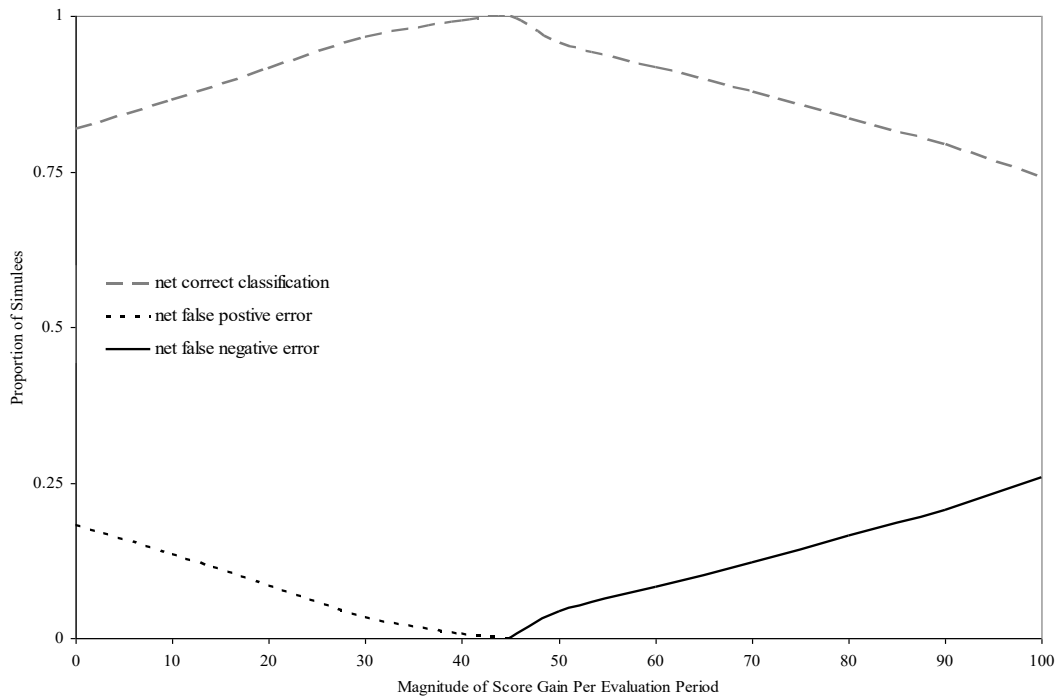
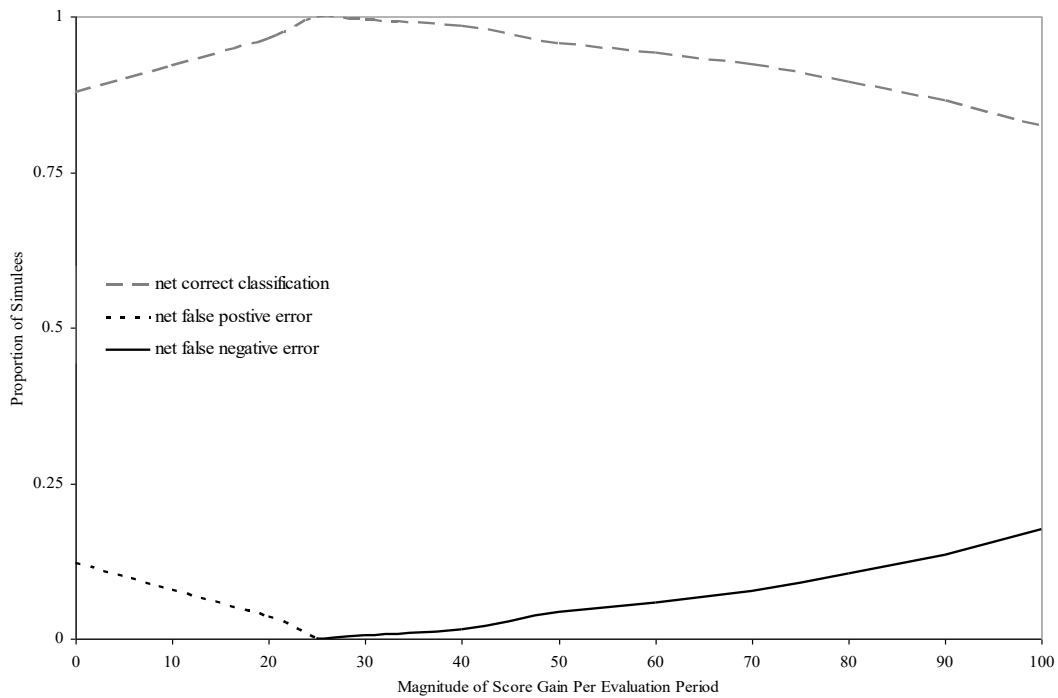


Figure VII.22

Proportion of Net Correct, Net False Positive, and Net False Negative Classifications Vs. Magnitude of Score Gain Per Evaluation Period (Mathematics and Numeracy; 2 Administrations Per Evaluation Period)



Discussion and Limitations of the Simulation Study

Our simulation study indicated that for individual simulees, accurate identification of the gain (or its absence) occurred about 63% of the time for Reading and about 72% of the time for Mathematics and Numeracy, when averaged across all possible score gains. However, the net (aggregate) results indicated very close agreement between the true percentages of learners who exhibited gain under the NRS conditions, and the observed gains expected based on the MAPT.

The goal of a simulation study such as this one is to produce results that generalize to operational data. This can be difficult, especially for a new test such as the MAPT because the population characteristics remain unknown and therefore cannot be easily simulated. For example, the proficiency distribution of the population of adult learners in Massachusetts is not known at this time and consequently it was decided to draw simulees' proficiencies from a uniform distribution spanning the reporting range of the MAPT score scale. Very likely, this strategy, in effect, resulted in an over-sampling of proficiency regions at the ends of the score scale, which in practice are likely to have relatively few examinees. Consequently, we gain the benefit of being equally confident in the findings across the score scale, however, to the extent that simulees in these end regions are classified differently from other more populous regions, results may be biased one way or the other. For example, any examinee with an EFL of 2 is less likely to be misclassified because the only error possible is a false positive (false negative is not possible because 200 is the minimum score).

Similarly, the distribution of score gains per evaluation period is not known. The magnitude of score gain plays an important role in the extent and nature of classification error. This dependency highlights the importance of collecting and evaluating operational data to further understand the utility of the MAPT for measuring learners' gains. Finally, it should be noted that all simulated responses were model-based and item parameters were treated as true. In practice, learners do not always answer or conform to expectation and, when scored, are done so based on estimates of item parameters. The extent that learner behaviors and item parameter estimates deviate from the behaviors and values assumed in this study, limits the generalizability of these results.

7. Other Validity Evidence

The preceding sections in this chapter discussed specific studies or aspects of validity that can be categorized under the sources of validity evidence described in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). However, it is difficult to disentangle the various aspects of validity and separate them into distinct components. In fact, the argument that MAPT scores are appropriate for their intended purposes began in Chapter I and permeates this entire manual. The care with which the test specifications were developed, the way in which we evaluated item response and test completion times, the development of the computer basics tutorial, and the extensive training provided to educators throughout Massachusetts are all examples of promoting the validity of inferences derived from MAPT scores. In this section, we describe other activity and data that pertain to the MAPT involving our evaluation of potential threats to valid interpretation of learners' proficiencies.

Reduction of Test Anxiety

Very few students enjoy taking tests and in ABE test anxiety is always a concern. Although it is unlikely we will ever develop tests that all students enjoy taking, we instituted several strategies to reduce ABE learners' anxiety about the forthcoming tests. Our first strategy was to engage teachers in the process and get them excited about the new tests. We believe reducing teachers' test anxiety is an important first step in addressing students' anxiety. We also strived to ensure that test content was appropriate for and familiar to adult learners. A third strategy for reducing test anxiety was the development of a computerized tutorial (described in Chapter III) that introduces learners to the testing system and familiarizes them with the item formats. In addition, we decided to conduct a small-scale "computer proficiency" study to discover students' impressions and concerns regarding taking a test on the computer.

Computer proficiency study

ABE students vary not only in their familiarity with computers in general, but also in their knowledge of how to navigate through a computerized test and complete onscreen actions to answer test questions. To assess this variability, we conducted a small study where we designed a low-stakes computerized assessment and observed ABE learners as they responded to the test questions. The study involved 32 learners of various learning levels (i.e., beginning ABE to GED prep students) from several ABE programs in Massachusetts. These students completed a short series of test questions involving the following item formats: multiple-choice with radio buttons, drag and drop, drop-down menu, typing/fill-in, and multiple-multiple-choice. For further descriptions of these item formats, see Zenisky and Sireci (2002). We investigated several different item formats to evaluate the feasibility of future use of innovative item formats on the MAPT.

Most learners involved in this study exhibited at least a moderate level of familiarity with basic computer actions (e.g., moving the mouse, scrolling, clicking) and many also demonstrated proficiency with more advanced skills (such as drag-and-drop and typing). Upwards of 80% of students agreed that it was easy to use the mouse for each of the given tasks, and somewhat less (about 75%) reported the same for using the keyboard to type answers. However, many students exhibited considerable more difficulty typing responses, suggesting that additional training and

experience in keyboarding may be necessary before extended open-response items could be used. Interestingly, 88% of students involved indicated it was “easy” or “very easy” to see the questions, figures, and answer choices on the computer screen, and 82% reported they were “comfortable” or “very comfortable” answering test questions on a computer.

In general, the students reviewed the item formats favorably. A summary of their responses to the survey questions specific to each of the item formats is presented in Table VII.8. This summary presents only the percentage of students who “agreed” with each statement. Although the multiple-choice format was rated highest overall, the novel item types that took advantage of the computer interface (i.e., drag-and-drop and drop-down menu items) were rated slightly higher than the multiple-choice item with respect to clearness of directions and desire to see that item type on a test. The results suggest that ABE learners would not have difficulty taking a computerized test and that several different item formats are viable for the MAPT.

Table VII.8

Summary of Survey Responses Regarding Computerized Item Formats (n=32)

Survey Item	Item Format (% Agree)				
	Multiple-Choice	Drag-and-Drop	Drop-Down Menu	Typing/Fill-in	Multiple-Multiple-Choice
<i>It was easy to see the words on the computer screen.</i>	100	96.6	93.3	89.3	96.8
<i>It was easy to see the pictures on the computer screen that you needed to use to answer one or both of the questions.</i>	100	89.3	83.3	85.7	89.6
<i>This kind of item was easy to answer.</i>	96.7	93.1	96.6	85.7	93.5
<i>It was easy to use the mouse to answer these questions.</i>	90.3	86.2	93.2	79.4	80.7
<i>The directions were easy to follow.</i>	87.1	89.7	80.0	86.4	83.3
<i>The directions were clear.</i>	80.6	82.8	89.3	89.3	74.1
<i>I would like to see this kind of item on a test.</i>	79.9	82.8	80.0	82.8	74.1
Average % Agree	90.7	88.6	88.0	85.5	84.6

8. Test Security

Another threat to the validity of a testing program is theft of items. From test development through the release of the operational tests, test security has been a high priority. ABE teachers and administrators participated in the test development process at a number of important points, but the security of the item bank was ensured at all times. Prior to participating in the sensitivity review, content validity review, and/or standard setting process all panelists reviewed and signed a non-disclosure agreement. This document specified that all materials and knowledge about the MAPT and specific items were not to be discussed outside of the meeting.

At the conclusion of each meeting all of the materials were collected and at no time were any potential operational items released in electronic form. Through the secure login process, ABE participants in the review process have also never had access to the item pool.

A number of measures have been put in place to address security concerns for the operational MAPT. Each student has a unique login that is requested through ACLS's SMARTT SSL system. The logins are only accessible from 8:00 am until 9:00 pm. Each time a login is used, a record is created in the OWL system. If the MAPT for Reading or MAPT for Math were accessed, the status field will change to show the test has been accessed. Students have a two-hour time period in which to complete the MAPT. Within this time frame students are not allowed to leave the testing location for any reason. If a student needs more time for the test, their teacher must request a time reset and this will be granted within 24 hours. The student can then continue from where they left off but they will not be able to return to any earlier questions.

The students must also complete the test within 15 days of the start date. If this deadline were not met, the student will need to restart the test from the first question and will see a new set of questions. There is also a minimum two-month time interval between pre- and post-testing, to ensure the tests are being used only as intended. In addition, there is a maximum of three test administrations per year. After the test is completed, the data are securely transferred back to the SMARTT system at regular intervals.

Tests are administered by a trained test administrator or teacher, and the test administrator is present throughout the entire testing period. Learners are allowed to use scrap paper, but the test administrators are trained to collect all papers at the conclusion of the testing session. Teachers and test administrators are given a unique login as well, but they are not able to see the MAPT for Math or MAPT for Reading. Their login only provides access to the Sample Questions and the Practice Tests. If any questions arise during the testing process about computer functionality or test content, the test administrator has been trained to use the Send Message feature to request support.

Summary of Validity Evidence

Much of the information presented in this Technical Manual speaks to the validity of interpretations of MAPT scores. Defensibility of test use and score interpretation starts with clear descriptions of the purposes of the test and of what the test is measuring (Chapter II). Quality control and concern for equity in test development (Chapter III) must also be demonstrated, and the characteristics of the score scale should be thoroughly described (Chapter IV). Information regarding standard setting for specific uses of the test should be provided (Chapter V), as should information regarding measurement precision (Chapter VI). Finally, specific data gathered to provide validity evidence should be presented (this chapter). In developing the MAPT and investigating the properties of its scores, we have used various sources of validity evidence, involved external, independent subject matter experts, comprehensively evaluated item response data from ABE learners, and conducted simulation studies to estimate the functioning of the exams in situations where data are not currently available. The end result of these test development and test evaluation endeavors is a strong argument that the MAPT fulfills its intended goals—that is, the MAPT measures the relevant portions of the Massachusetts ABE Curriculum Frameworks and is appropriate for evaluating learners’ gains and for judging accountability under the NRS.

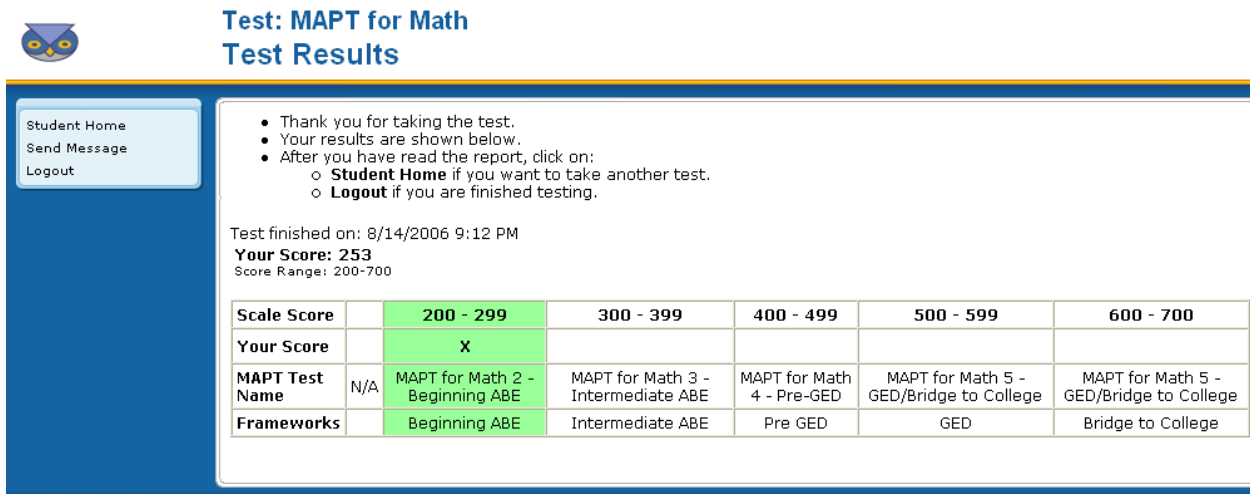
VIII. Interpreting MAPT Test Scores

In Chapter IV, we described how the MAPT score scale was created, and in Chapter V we described how the cut-scores on the MAPT that distinguish among the NRS EFLs were set. In this chapter, we revisit the MAPT score scales and EFLs to ensure consumers of MAPT score information appropriately interpret these scores.

A sample score report for a learner who took both the MAPT for Reading is presented in Figure VIII.2, and sample score report for the MAPT for Mathematics and Numeracy is presented in Figure VIII.3. The first score report (Reading) illustrates a learner who obtained a score of 253 on the Reading test and is placed in the Massachusetts Learner Level “Beginning ABE.” The NRS EFL is not reported to learners, but for NRS accountability purposes, this learner would be classified into the “Beginning Basic” level. S/he and would have to earn a score of 300 or higher upon retesting to be classified as exhibiting gain under the NRS.

Figure VIII.2

Sample Score Report: MAPT for Reading



The image shows a sample score report for the MAPT for Math. It features a blue header with an owl icon and the text "Test: MAPT for Math Test Results". On the left, there is a sidebar with links: "Student Home", "Send Message", and "Logout". The main content area contains a list of instructions, test completion information, and a table of score ranges and frameworks.

• Thank you for taking the test.
• Your results are shown below.
• After you have read the report, click on:
 ◦ **Student Home** if you want to take another test.
 ◦ **Logout** if you are finished testing.

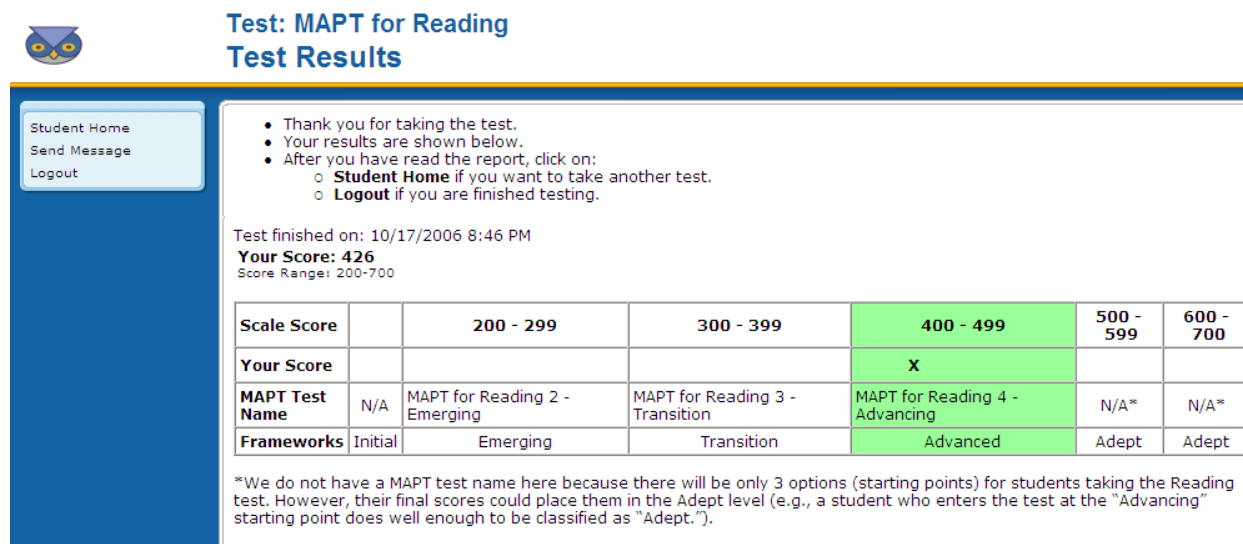
Test finished on: 8/14/2006 9:12 PM
Your Score: 253
Score Range: 200-700

Scale Score	200 - 299	300 - 399	400 - 499	500 - 599	600 - 700
Your Score	X				
MAPT Test Name	N/A	MAPT for Math 2 - Beginning ABE	MAPT for Math 3 - Intermediate ABE	MAPT for Math 4 - Pre-GED	MAPT for Math 5 - GED/Bridge to College
Frameworks		Beginning ABE	Intermediate ABE	Pre GED	GED

The fictitious score report in Figure VIII.3 shows a learner who scored 426 on the MAPT for Mathematics and Numeracy, and is classified into the “Advanced” Massachusetts Learner Level. For NRS accountability purposes, the learner would be classified as “High Intermediate,” and would have to achieve a score of at least 500 upon retesting to be classified as achieving educational gain.

Figure VIII.3

Sample Score Report: MAPT for Math



Limitations of MAPT Scores

MAPT scores represent an estimate of learners' proficiency in math or reading according to objectives listed in the Massachusetts ABE Curriculum Frameworks. As we have stressed throughout this technical manual, educational tests, like all psychological measures, are imperfect. Therefore, it is possible that learners' true proficiencies may be above or below the estimate provided by MAPT. Furthermore, the learner levels and EFLs associated with MAPT scores are very general descriptions of proficiency. Although these descriptions are useful for understanding learners' performance on the test and for a general description of their skills, other information should be used to describe the strengths and weaknesses of individual learners. For example, the MAPT does not provide diagnostic information within either subject area. Such information would require a much longer exam to reliably measure subcomponents of the math and reading domains. At this juncture, use of the MAPT for placement into instructional programs is also not recommended, although used in conjunction with other information about learners' knowledge, skills, and abilities, programs may find MAPT scores helpful for such placement, particularly when both pretest and posttest scores are available.

Educators who help adult learners interpret their MAPT scores should stress that these scores are approximations of where the learners stand on a continuum of content taught in various ABE classes. They should also stress that any score gains on the MAPT scale are indicators of progress on this continuum. In those cases where a learner's scale score decreases from pretest to posttest, potential reasons for such score decline should be discussed with the learner.

IX. Looking Forward: Next Steps for the MAPT

Educational test development and test evaluation are dynamic processes. As more and more adult learners take the MAPT, we will learn more about its utility for instructional and accountability purposes. Even at this early point in this new testing program, our research and experience points to several areas that could be investigated. These areas include:

- Revisiting the MAPT for Reading test specifications to see if the Purposes for Reading need to be revised or clarified, or if the Skill Area of Word Recognition needs clarification.
- Embellishment of the item pools for both subject areas to keep the content domain fresh and minimize the likelihood that learners who take the test more than twice will see items they have taken before. Developing and piloting new items, particularly more difficult items on the Reading test, is likely to be a priority for the near future.
- New, non-multiple-choice, item formats could be piloted to see if they improve measurement of learners' knowledge and skills. At this point, some innovative math items have been developed and so it is possible future MAPT exams will include these new item formats.
- Although learners throughout the Commonwealth have successfully taken MAPT exams, this is not the case in the Massachusetts prison system where prisoners are not allowed access to the Internet. An important task for the near future should be to provide access to the MAPT for incarcerated learners.
- Currently, there are no MAPT exams for learners at the lowest levels (pre-literate). Development of Level 1 tests that can be integrated into the current MAPT is another priority area.
- The MAPT could also be extended into other subject areas such as Writing, Civics, and Computer Technology. The cost of developing tests in these areas is formidable, but the advantages of the MAPT design would improve assessment of these proficiencies. Assessment of English proficiency is another area in which new MAPT exams could be extremely useful.
- To better evaluate the utility of the MAPT, the opinions and experiences of ABE educators and learners throughout the Commonwealth should be systematically solicited. Surveys, interviews, and focus groups are some means through which the effects of MAPT could be evaluated.
- The use of the MAPT for evaluating educational gain should be further evaluated by studying typical gains made by learners, and by establishing criteria for gains within EFLs that represent meaningful amounts of achievement.

The utility of the MAPT is just starting to become understood. These tests represent tight relationships among curriculum, instruction, assessment, and accountability. By keeping the process collaborative among the field of ABE educators, the Massachusetts Department of Education, the test developers, and the learners, these exams will continue to meet the assessment needs of adult educators throughout the Commonwealth.

References

- Adult and Community Learning Services (2005a). *Massachusetts Adult Basic Education Curriculum Framework for Mathematics and Numeracy*. Malden: MA: Massachusetts Department of Education. Available at <http://www.doe.mass.edu/acsl/frameworks>.
- Adult and Community Learning Services (2005b). *Massachusetts Adult Basic Education Curriculum Framework for Reading*. Malden: MA: Massachusetts Department of Education. Available at <http://www.doe.mass.edu/acsl/frameworks>.
- 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.
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990). *Standards for teacher competence in educational assessment of students*. Washington, DC: American Federation of Teachers.
- Anastasi, A. (1988). *Psychological testing* (6th edition). New York: Macmillan.
- Baldwin, P. & Sireci, S. G. (2003, December). Preliminary specifications for ACLS Mathematics and Numeracy Proficiency Tests. *Center for Educational Assessment Research Report No. 510*. Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.
- Bloom, B. S., et al. (Eds.). (1956). *Taxonomy of educational objectives: Handbook I, cognitive domain*. New York: D. McKay.
- Brennan, R. L. (2002, October). Estimated standard error of a mean when there are only two observations. *CASMA Technical Note Number 1*. Iowa City, IA: Center for Advanced Studies in Measurement and Assessment, University of Iowa.
- Cizek, G. J. (1996). Setting passing scores. [An NCME instructional module]. *Educational Measurement: Issues and Practice*, 15 (2), 20-31.
- Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice*, 20 (4), 19-27.
- Cronbach, L. J. (1971). Test Validation. In R.L. Thorndike (Ed.). *Educational measurement* (2nd ed., pp. 443-507). Washington, DC: American Council on Education.
- Curry, D., Schmitt, M. J., & Waldron, S. (1996, July). *Framework for Adult Numeracy Standards: The Mathematical Skills and Abilities Adults Need to Be Equipped For the Future*, Boston, MA: The Adult Numeracy Practitioners Network.
- Division of Adult Education and Literacy (2006, July). *Implementation guidelines: Measures*

and methods for the national reporting system in adult education. Office of Vocational and Adult Education, U.S. Department of Education, Washington: DC. Available at http://www.nrsweb.org/foundations/implementation_guidelines.aspx.

Downing, S. M., & Haladyna, T. M. (1997). *Test item development: Validity evidence from quality assurance procedures.* *Applied Measurement in Education*, 10, 61-82.

Ferrara, S., Perie, M., & Johnson, E. (2002, April). Setting performance standards: The item-descriptor matching procedure. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Ferrara, S., Perie, M., & Johnson, E. (in press). Matching the judgmental task with standard setting panelist expertise: The item-descriptor matching procedure. *Journal of Applied Testing Technology*.

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, 6, 427-439.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items.* Hillsdale: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item writing rules. *Applied Measurement in Education*, 2, 37-50.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. Linn (Ed.), *Educational Measurement (3rd ed.)*; pp. 147-200. New York: Macmillan.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Hillsdale, NJ: Erlbaum.

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage.

Hambleton, R. K., Zaal, J. N., & Pieters, P. (1991). Computerized adaptive testing: Theory, applications, and standards. In R. K. Hambleton & J. N. Zaal (Eds.), *Advances in Educational and Psychological Testing* (pp. 341-366). Norwell, MA: Kluwer.

Han, K. T., Baldwin, P., & Sireci, S. G. (2006). Analyses of differential item functioning of potential MAPT items. *Center for Educational Assessment Research Report No. 60?*. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

- Huff, K. L., & Sireci, S. G. (2001). Validity issues in computer-based testing. *Educational Measurement: Issues and Practice*, 20 (3), 16-25.
- Jenkins, J. G. (1946). Validity for what? *Journal of Consulting Psychology*, 10, 93-98.
- Jodoin, M. L., Zenisky, A. L., & Hambleton, R. K. (2006). Comparison of the psychometric properties of several computer-based test designs for credentialing exams with multiple purposes. *Applied Measurement in Education*, 19, 203-220.
- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, 112, 527-535.
- Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, 64, 425-461.
- Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 53-88). Mahwah, NJ: Erlbaum.
- Karantonis, A., & Sireci, S. G. (2006). The bookmark standard setting method: A literature review. *Educational Measurement: Issues and Practice*, 25 (1), 4-12.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th edition). Upper Saddle River, NJ: Prentice-Hall.
- Livingston, S. A., & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. *Journal of Educational Measurement*, 32, 179-197.
- Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.
- Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard setting methods. *Applied Measurement in Education*, 2, 121-141.
- Luecht, R. (2005). Some useful cost-benefit criteria for evaluating computer-based test delivery models and systems. *Journal of Applied Testing Technology* (available at http://www.testpublishers.org/Documents/JATT2005_rev_Criteria4CBT_RMLuecht_Apr2005.pdf).
- Martone, D., Goodridge, B., Moses, M., & Titzel, J. (2004, September). Refinements to ABE mathematics standards for assessment purposes. *Center for Educational Assessment Research Report No. 548*. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

- Massachusetts Adult Basic Education Performance Accountability Working Group (2002, September). *PAWG final report*. Malden, MA: Massachusetts Department of Education. Available at <http://www.doe.mass.edu/acls/pawg/report0902.pdf>.
- Massachusetts Department of Education (2005). 2004 MCAS technical report. Malden, MA: Author. Available at <http://www.doe.mass.edu/mcas/2005/news/04techrpt.doc>.
- Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement*, (3rd ed.) (pp. 13-103). Washington, DC: American Council on Education.
- Mislevy, R. J., & Knowles, K. T. (Eds.) (2002). *Performance assessments for adult education: Exploring the measurement issues*. Washington, DC: National Academy of Sciences Press.
- Nunnally, J. C. (1978). *Psychometric theory*. New York: MacGraw-Hill.
- O'Neil, T., Sireci, S. G., & Huff, K. L. (2004). Evaluating the consistency of test content across two successive administrations of a state-mandated science assessment. *Educational Assessment*, 9, 129-151.
- Popham, W. J. (1992). A tale of two test-specification strategies. *Educational Measurement: Issues and Practice*, 11(2), 16-17, 22.
- Ramsey, P. A. (1993). Sensitivity review: The ETS experience as a case study. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 367-388). Hillsdale, New Jersey: Lawrence Erlbaum.
- Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.) (1997). *Computerized adaptive testing: From inquiry to operation*. Washington, DC: American Psychological Association.
- Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education*, 19, 405-450.
- Sireci, S. G. (1998a). Gathering and analyzing content validity data. *Educational Assessment*, 5, 299-321.
- Sireci, S. G. (1998b). The construct of content validity. *Social Indicators Research*.
- Sireci, S. G. (2003). Content validity. *Encyclopedia of psychological assessment* (pp. 1075-1077). London: Sage.
- Sireci, S. G. (2003). Validity. *Encyclopedia of psychological assessment* (pp. 1067-1069). London: Sage.
- Sireci, S. G. (2004). Computerized-adaptive testing: An introduction. In J. Wall and G. Walz (Eds.). *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 685-694). Greensboro, NC: CAPS Press.

- Sireci, S. G., Baldwin, P., Keller, L. A., Valle, M., Goodridge, B., Leonelli, E., Schmitt, M. J., Tamarkin, K., & Titzel, J. (2004). *Specifications for the ACLS Mathematics and Numeracy Proficiency Tests*. Center for Educational Assessment Research Report No. 513. Amherst, MA: Center for Educational Assessment, University of Massachusetts.
- Sireci, S. G., Li, S., Martone, A., Valle, M., Bayer, J., Kelly, J., Hanley, S., Greene, S., Royer, J. M., & Schwerdtfeger, J. (2004). *Specifications for the ACLS Reading Proficiency Tests*. Center for Educational Assessment Research Report No. 514. Amherst, MA: Center for Educational Assessment, University of Massachusetts.
- Sireci, S. G., & Mullane, L. A. (1994). Evaluating test fairness in licensure testing: the sensitivity review process. *CLEAR Exam Review*, 5 (2) 22-28.
- Sireci, S. G., Robin, F., Meara, K., Rogers, H. J., & Swaminathan, H. (2000). An external evaluation of the 1996 Grade 8 NAEP Science Framework. In N. Raju, J.W. Pellegrino, M.W. Bertenthal, K.J. Mitchell & L.R. Jones (Eds.), *Grading the nation's report card: Research from the evaluation of NAEP* (pp. 74-100). Washington, DC: National Academy Press.
- Sireci, S. G., Robin, F., & Patelis, T. (1999). Using cluster analysis to facilitate standard setting. *Applied Measurement in Education*, 12, 301-325.
- Stein, S. (2000). *Equipped for the Future Content Standards: What Adults Need to Know and Be Able to Do in the 21st Century*. Washington, DC: National Institute for Literacy. ED Pubs document EX0099P.
- Thompson, S., & Thurlow, M. (2002, June). Universally designed assessments: Better tests for everyone! *Policy Directions, Number 14*. Minneapolis, MN: National Center on Educational Outcomes.
- Thompson, S., Thurlow, M., & Malouf, D. B. (2004). *Creating better tests for everyone through universally designed assessments*. *Journal of Applied Testing Technology*, 6(1). Available at <http://www.testpublishers.org/jattart.htm>.
- Toops, H. A. (1944). The criterion. *Educational and Psychological Measurement*, 4, 271-297.
- Valle, M., & Sireci, S. G. (2003, December). Developing specifications for ACLS reading proficiency tests. *Center for Educational Assessment Research Report No. 511*, Amherst, MA: Center for Educational Assessment, University of Massachusetts Amherst.
- Valle, M., Sireci, S. G., Bayer, J., Kelly, J., Hanley, S., & Greene, S. (2004). Proposed refinements to ABE Reading Test Specifications: Revisions to the Reading Standards for Assessment. *Center for Educational Assessment Research Report No. 555*. Amherst, MA: Center for Educational Assessment, University of Massachusetts.

Wainer, H. (Ed.) (2000). Computerized adaptive testing: A primer (2nd edition). Hillsdale, NJ: Lawrence Erlbaum.

Zimowski, M., Muraki, E., Mislevy, R. J., & Bock, R. D. (2002). BILOG-MG [Computer software]. Lincolnwood, IL: Scientific Software International.

Appendices

Appendix A

Lists of MAPT Committee Members

Membership of MAPT for Mathematics and Numeracy

Test Specification Development Committee

Barbara Goodridge
Lowell Adult Education Center

Esther Leonelli
Notre Dame Education Center

Mary Jane Schmitt
TERC

Kenny Tamarkin
System for Adult Basic Education Support

Judith Titzel
World Education

Jane Schwerdtfeger
Massachusetts Department of Education

Center for Educational Assessment Staff:

Stephen G. Sireci
Peter Baldwin
Lisa A. Keller
Mercedes Valle

Membership of MAPT for Reading
Test Specification Development Committee

Jeri Bayer
Northern Essex Community College

Sylvia Greene
Cambridge Community Learning Center

Stephen Hanley
WAITT House

Janet Kelly
Read/Write/Now Adult Literacy Center, Springfield

James M. Royer
University of Massachusetts Amherst

Jane Schwerdtfeger
Massachusetts Department of Education

Center for Educational Assessment Staff:

Stephen G. Sireci
Shuhong Li
Andrea Martone
Mercedes Valle

Membership of the MAPT for Mathematics and Numeracy

Item Reviewers

Barbara Goodridge
Lowell Adult Education Center

Marilyn Moses
Brockton Adult Education Center

Judith Titzel
World Education

Membership of the MAPT for Reading

Item Reviewers

Jackie Coelho
QCC Continuing Education Program, Worcester

Sylvia Greene
Cambridge Community Learning Center

Janet Kelly
Read/Write/Now Adult Literacy Center, Springfield

Deborah McGill
The Immigration Learning Center, Malden

Membership of the Sensitivity Review Committee

Bob Chulu
University of Massachusetts Amherst

Saul Augustine
Catholic Charitable/Haitian Multiservice Center

Ernest Best
Mass ALL

Papa Diop
Perkins Community Center and WAITT House

Sylvia Greene
Cambridge Community Learning Center

Maria Kallifanou
QCC/College at CitySquare

Amy Park
Asian American Civic Association

Victor Tavares
SER-Jobs for Progress (SER Family Learning Center)

Membership of the MAPT for Mathematics and Numeracy

Content Validity Panel

Jeanne Almanzar
Valley Opportunity Council

Barbara Goodridge
Lowell Adult Education Center

Martha Mangan
Lowell Adult Education Center

Suzanne Martin
Brockton Adult Education Center

Marilyn Moses
Brockton Adult Education Center

Lee Wertheimer
Valley Opportunity Council

Membership of the MAPT for Reading

Content Validity Panel

Jon Chapman
Crittenton

Sylvia Greene
Cambridge Community Learning Center

Helen Haller
Jamaica Plain Community Center

Stephen Hanley
WAITT House

Laurie Hartwick
Lawrence Adult Learning Center

Andrea O'Brien
Lawrence Adult Learning Center

Jane Schwerdtfeger
Massachusetts Department of Education

Membership of the MAPT for Mathematics and Numeracy

Standard Setting Panel

Jeanne Almanzar
Valley Opportunity Council

Donna Cornellier
Massachusetts Department of Education

Patricia Donovan
World Education

Veronica Kell
Mount Wachusett Community College

Raymunda Lagmay-Carter
Cambridge Community Learning Center

Julianne Leger-Fullerton

Esther Leonelli
Notre Dame Education Center

Shirley Lyon
Northern Essex Community College

Judith Titzel
World Education

Lee Wertheimer
Valley Opportunity Council

Membership of the MAPT for Reading

Standard Setting Panel

Jeri Bayer
Northern Essex Community College

Jon Chapman
Crittenton

Nancy Coffey
Operation Bootstrap, Lynn

Janet Fischer
Northern Essex Community College

Helen Haller
Jamaica Plain Community Center

Stephen Hanley
WAITT House

Laurie Hartwick
Lawrence Adult Learning Center

Janet Kelly
Read/Write/Now Adult Literacy Center, Springfield

Andrea O'Brien
Lawrence Adult Learning Center

Jane Schwerdtfeger
Massachusetts Department of Education

David Stearns
Jamaica Plain Community Center

Appendix B

Links to Additional Resources

UMass Center for Educational Assessment Website

<http://www.UMASS.edu/remf/remf7.htm>

ACLS General Website

<http://www.doe.mass.edu/acls/default.htm>

ACLS Assessment Website

<http://www.doe.mass.edu/acls/assessment/>

The above URL links to:

MAPT link (for pre- and post-testing for learner educational gain)

MAPT Practice Tests (20 Items)

MAPT Sample Questions (4 items)

MAPT Computer Basics

MAPT Learner Information Sheet

MAPT for Math Test Results Blank Page

MAPT for Reading Test Results Blank Page

List of Program Site Names Used by MAPT to Identify Programs

MAPT Student Home Status Column Tip Sheet

Testing--Standardized and Otherwise--and the Massachusetts ABE Pilot Test:

A Curriculum Unit for ABE Classes

ACLS Curriculum Frameworks

<http://www.doe.mass.edu/acls/frameworks/>

NRS Website

<http://www.nrsweb.org>

Education Functioning Level Descriptors

http://www.nrsweb.org/reports/EFL_Descriptors.doc

MAPT for Mathematics and Numeracy Test Specification Report

<http://www.doe.mass.edu/acls/mailings/2004/0709/mathtestspecs.pdf>

MAPT for Reading Test Specification Report

<http://www.doe.mass.edu/acls/mailings/2004/0709/readtestspecs.pdf>

Massachusetts ABE Curriculum Framework for Mathematics and Numeracy

<http://www.doe.mass.edu/acls/frameworks/mathnum.doc>

Appendix C

MAPT for Reading Standards by Level¹⁰

10 November 1, 2004 version. Revised by the Reading Test Review Committee. See Valle, et al., (2004): “Proposed Refinements to ABE Reading Test Specifications: Revisions to the Reading Standards for Assessment.” Center for Educational Assessment Research Report No. 555, University of Massachusetts Amherst.

Test Level 2 (GLE 2-3.9)

SKILL AREA	STANDARDS
Word Recognition	This skill area involves recognizing controlled words, sight words, and visual cues. At this test level, the following standards are tested:
Standard #	Standard Description
2.W.1	Identify short vowels with single consonants (one syllable words) (CVC).
2.W.2	Identify short vowels with two letter initial and/or final consonant blends (CCVC, CVCC, CCVCC).
2.W.3	Identify short vowels with initial and final consonant digraphs (“ship,” “mash,” “thin,” “path”).
2.W.4	Identify short vowels with three letter consonant blends (CCCVC, CCCVCC).
2.W.5	Identify short vowels with final “ff,” “ll,” “ss,” and “ck” (“egg,” “cliff,” “clock”).
2.W.6	Identify short vowels with final “ch” and “tch” (“branch,” “match”).
2.W.7	Identify vowel, consonant, silent “e” pattern and use it to decode one-syllable words.
2.W.8	Identify vowel digraphs in one- and two-syllable words (e.g., ai, ay, ee, ea, oa, oe, ui, ue).
2.W.9	Identify vowel diphthongs (e.g., ou, ow, au, aw, oi, oy) and use to decode one- and two-syllable words.
2.W.10	Identify “r-controlled” vowel sounds (e.g., er, ir, ur, ar, or).
2.W.11	Identify plural -es (“dishes”) and plurals of nouns ending in “y” or “f” (“babies,” “lives”).
2.W.12	Identify and use silent consonants (e.g., “kn,” “gh”).
2.W.13	Identify verb endings: “s,” “ed,” “ing”.
2.W.14	Identify soft “c” and “g” words (“cent,” “gem”).
2.W.15	Identify “y” as a vowel with its long “i” and long “e” sounds (“cry,” “party”).
2.W.16	Identify two- and three-letter consonant blends and digraphs (e.g., br-, sp-, cl-, -nd, -sk, str-, spl-, thr-) in initial, medial, and final word positions and use them to decode one-syllable words.
2.W.17	Identify words with other endings (e.g., -er, -est, -ful, -less, -ness, -ment).
2.W.18	Identify and use contractions and be able to match them to the two words being replaced (e.g., I’m for I am).
2.W.19	Identify long vowels at the end of words (“be,” “go”).
2.W.20	Identify closed syllables (“des”) and open syllables (“de”).
2.W.21	Identify and use compound words.
2.W.22	Identify, count, and divide syllables (syllabicate) in multi-syllabic words, using rules for syllabication: (1) vccv (e.g., nap-kin) (2) vcv (e.g., tu-lip, lev-el) (3) vcle (e.g., can-dle)
2.W.23	Identify the schwa sound in unaccented syllables (e.g., “away”) and use to decode simple words.
2.W.24	Read a minimum of 80% of sight words from an appropriate level word list.
2.W.25	Use a dictionary to locate the meaning of words used in a statement.
2.W.26	Locate specific items in an alphabetical listing (e.g., class list, phone directory, dictionary) or in a topical listing (e.g., picture dictionary, table of contents).
2.W.27	Arrange words in alphabetical order.

Test Level 2 (GLE 2-3.9) - Continued

SKILL AREA		STANDARDS
Vocabulary & Meaning	This skill area involves the ability to understand word meanings and the ability to infer word meanings from context.	
Standard #	Standard Description	
2.V.1	Recognize synonyms, antonyms, homonyms, and homophones (e.g., dear-deer) for identified vocabulary words presented in isolation or within a group of words.	
2.V.2	Recognize the correct meaning of multiple-meaning words (e.g., bill, train) when presented in text.	
2.V.3	Use context clues to determine the meaning of an unknown word.	
2.V.4	Use structural analysis to determine the meaning of words (e.g., prefixes, suffixes, and root words).	
Comprehension	Comprehension refers to the ability to extract meaning from connected discourse that is read.	
2.C.1	Locate pertinent information in simple familiar materials (e.g., want ads, job listings, schedules, signs, food packages) including who, what, which, when, where.	
2.C.2	Respond to instructional level text by making simple inferences.	
2.C.3	Read and interpret simple and compound sentences in a short paragraph containing familiar vocabulary.	
2.C.4	Extract information from simple charts, graphs, labels, and payroll stubs to answer questions.	
2.C.5	Use comparison and contrast to draw conclusions in a story.	
2.C.6	Make predictions about content based on titles, pictures, and type of material.	
2.C.7	Identify the author's purpose in a selection when it is not explicitly stated.	
2.C.8	Read pertinent information to perform a task.	

Test Level 3 (GLE 4-5.9)

SKILL AREA	STANDARDS
Word Recognition	This skill area involves recognizing controlled words, recognizing sight words, and visual cues.
Standard #	Standard Description
3.W.1	Read a minimum of 80% sight words from an appropriate level word list.
3.W.2	Identify words with the hard and soft sounds of c and g.
3.W.3	Identify words where ci, si and ti sound like /sh/.
3.W.4	Identify words where ea has the short e sound.
3.W.5	Identify the three sounds of –ed.
3.W.6	Identify the three sounds of y as a vowel.
3.W.7	Identify words where final -que sounds like /k/.
3.W.8	Identify root words, prefixes (e.g., pre-, un-, dis-), and suffixes (e.g., -less, -ly, -ment) and use them to decode multi-syllabic words.
3.W.9	Identify words with common suffixes, such as -ous, -able and -ible, -ture, -ive, and -age.
3.W.10	Apply the rules of syllabication to unfamiliar words from the content areas (social studies, science, and literature) and be able to know what syllable should be accented. (1) vccv (e.g., nap-kin) (2) vcv (e.g., tu-lip, lev-el) (3) vcle (e.g., can-dle)
Vocabulary & Meaning	This skill area involves the ability to understand word meanings and the ability to infer word meanings from context.
3.V.1	Use prefixes, suffixes, and root words to determine meaning of unfamiliar words.
3.V.2	Use context clues to determine meaning of unfamiliar words, including increasingly sophisticated multi-syllabic words originating from Latin and Greek prefixes, suffixes, and roots.
3.V.3	Identify the meaning of frequently used synonyms, antonyms, and homonyms.
3.V.4	Use a dictionary to locate the meaning of unfamiliar words and a thesaurus to find antonyms, synonyms, and to expand vocabulary.
Comprehension	Comprehension refers to the ability to extract meaning from connected discourse that is read.
3.C.1	Locate explicitly stated information in functional reading (e.g., invitations, bulletins).
3.C.2	Distinguish between fact and opinion in a short paragraph.
3.C.3	Use comparison and contrast to draw conclusions in a story.
3.C.4	Make predictions about content based on titles, pictures, and type of material.
3.C.5	Identify the topic sentence in each paragraph and the main idea of the passage.
3.C.6	Identify relevant supporting details in a passage.
3.C.7	Follow simple written multi-step instructions.

Test Level 3 (GLE 4-5.9)- Continued

SKILL AREA	STANDARDS
Comprehension	This skill area involves the ability to understand word meanings and the ability to infer word meanings from context.
Standard #	Standard Description
3.C.8	Locate pertinent information in multi-paragraph passages.
3.C.9	Locate pertinent information in print materials (e.g., ads, labels, pay stubs, public signs) to perform a task or answer a question.
3.C.10	Use comparison and contrast to determine the best purchase of an advertised item.
3.C.11	Locate information in simple graphic materials such as charts, pictures, maps, signs, diagrams, tables, or graphs.
3.C.12	Locate information on a given topic in several types of reference materials (e.g., dictionary, atlas, encyclopedia, newspaper).
3.C.13	Draw conclusions and make inferences about short passages (such as by identifying correct multiple choice answers or by writing short answers to questions).
3.C.14	Identify the four main ways a paragraph can be organized--time order, simple listing, comparison and contrast, and cause and effect--and recognize the signal words that indicate each pattern.
3.C.15	Identify and apply appropriate strategies to aid comprehension (e.g. graphic organizers, outlining).

Test Level 4 (GLE 6-10.9)

SKILL AREA	STANDARDS
Vocabulary & Meaning	This skill area involves the ability to understand word meanings and the ability to infer word meanings from context.
Standard #	Standard Description
4.V.1	Recognize and comprehend the meaning of moderately complex vocabulary using word, sentence and paragraph clues to determine meaning.
4.V.2	Use a dictionary to locate the meaning of unfamiliar words.
4.V.3	Use a thesaurus to find antonyms, synonyms, and to expand vocabulary.
4.V.4	Identify idioms and their use in passages.
4.V.5	Recognize and understand common abbreviations, acronyms, and shortened words (e.g., exam, examination).
4.V.6	Demonstrate understanding of specialized vocabulary from student' chosen occupational field (e.g., electronics manual or nursing text).
Comprehension	Comprehension refers to the ability to extract meaning from connected discourse that is read.
4.C.1	Read and interpret information in common manuals and legal forms (e.g., driver's manual, rental agreement).
4.C.2	Read and interpret expository writing on common topics in newspapers, periodicals, and non-technical journals.
4.C.3	Discriminate between the implied main idea and the supporting details from an instructional level passage.
4.C.4	Determine the sequence of events in a written passage.
4.C.5	Predict probable outcomes from knowledge of events obtained from a reading selection.
4.C.6	Determine author's purpose (e.g., to entertain, inform, persuade) from a variety of written pieces (e.g., newspaper article, travel brochure, store catalog, ad).
4.C.7	Distinguish factual information from opinion or fiction.
4.C.8	Determine the meaning of persuasive language and propaganda used in functional text (e.g., advertisement).
4.C.9	Identify the inferred main idea from a multi-paragraph passage and restate it in your own words.
4.C.10	Draw conclusions and make predictions and inferences from graphic materials such as charts, pictures, maps, signs, diagrams, tables, or graphs.
4.C.11	Identify the central idea or theme of a literary work (e.g., short story, drama, poetry).
4.C.12	Identify and interpret common figurative language (e.g., simile, metaphor) and other poetic devices such as alliteration, puns, personification and hyperbole found in a literary work.
4.C.13	Interpret information in real-life contexts (e.g., medical, occupational, parenting) and apply that information to a new situation.
4.C.14	Identify relationships between similar documents (e.g., order form and invoice) and compare information for accuracy.

Test Level 4 (GLE 6-10.9)- Continued

SKILL AREA	STANDARDS
Comprehension	Comprehension refers to the ability to extract meaning from connected discourse that is read.
Standard #	Standard Description
4.C.15	Draw conclusions and make inferences about increasingly sophisticated passages.
4.C.16	Make generalizations from implicit ideas (e.g., first paragraph of The Declaration of Independence).
4.C.17	Make inferences about a character based on their actions and dialogue in a literary work.
4.C.18	Identify the style of a passage (e.g., formal vs. informal, use of active voice vs. passive voice).
4.C.19	Determine the tone (feeling that the writer is trying to convey) of a variety of writing samples (e.g., editorials, magazine articles, literary texts).
4.C.20	Summarize the main idea of increasingly sophisticated texts when it is not explicitly stated.

Appendix D

Sensitivity Review Guidelines

Appendix E

Math Standard Setting Meeting – Agenda, Description, Rating Form, Survey

Massachusetts Adult Proficiency Test (MAPT) for Math

National Reporting System (NRS) Standard Setting Meeting

May 31, 2006

Agenda

8:00	Coffee, Muffins, Donuts, and Casual Conversation
8:30	Welcome and Introductions
8:45	Overview of MAPT Development Purpose of the Meeting
9:00	Description and Discussion of NRS Achievement Levels
10:15	Break
10:30	Introduction to Item Matching Task and Practice Ratings
11:30	Panelist Ratings—Booklet 1
12:15	Lunch
1:00	Panelist Feedback/Group Discussion
2:00	Panelist Ratings—Booklet 2
3:00	Break
3:15	Panelist Feedback/Group Discussion
4:00	Discussion
4:30	Survey and Assignments
5:00	Adjourn

Math Performance Level Descriptions (NRS Functioning Literacy Levels)

Beginning Literacy • Individual has little or no recognition of numbers or simple counting skills • May have only minimal skills, such as the ability to add or subtract single digit numbers	
Beginning Basic • Individual can count, add, and subtract three digit numbers • Can perform multiplication through 12 • Can identify simple fractions • Perform other simple arithmetic operations	Notes
Low Intermediate Basic • <i>Individual can perform with high accuracy all four basic math operations using whole numbers up to three digits</i> • <i>Can identify and use all basic mathematical symbols</i>	
High Intermediate Basic • Individual can perform all four basic math operations with whole numbers and fractions • Can determine correct math operations for solving narrative math problems • Can convert fractions to decimals and decimals to fractions • Can perform basic operations on fractions	
Low Adult Secondary • Individual can perform all basic math functions with whole numbers, decimals, and fractions • Can interpret and solve simple algebraic equations, tables, and graphs • Can develop own tables and graphs • Can use math in business transactions	
High Adult Secondary • Individual can make mathematical estimates of time and space • Can apply principles of geometry to measure angles, lines, and surfaces • Can also apply trigonometric functions	

Massachusetts Adult Proficiency Test for Mathematics and Numeracy

Standard Setting Rating Sheet (Sample)

Thank you for taking the time to review these MAPT for Math items. As we discussed this morning, your review is very important for determining the cut-points (standards) along the MAPT for Math that will correspond to the NRS Educational Functioning Levels helping us improve the quality of our test. Please review each item and circle your answer. Then check your answer and indicate the NRS level you think best measures the skills required to answer the item by placing a check mark in the appropriate column. If you have any comments you would like to make about a problem with the item, please mark it directly on the item booklet, and mark your page with a post-it.

#	ID	Strand	Correct Answer	NRS Educational Functioning Levels (check one)				
				Beginning Basic	Low Intermediate	High Intermediate	Low Adult Secondary	High Adult Secondary
1	2212	Geometry & Measurement	c					
2	1020	Number Sense	c					
3	535	Number Sense	a					
4	1407	Patterns, Functions, & Algebra	c					
5	2504	Statistics & Probability	c					
6	297	Patterns, Functions, & Algebra	d					
7	2210	Statistics & Probability	c					
8	616	Geometry & Measurement	b					
9	2203	Patterns, Functions, & Algebra	b					

10	2453	Geometry & Measurement	b					
----	------	---------------------------	---	--	--	--	--	--

Massachusetts Adult Proficiency Test (MAPT) NRS Standard Setting Meeting

Panelist Demographic Questionnaire

Name: _____

Current occupations relevant to ABE: _____

Teaching site: _____

Average class demographics: _____

Approximately how many years experience do you have teaching ABE learners? _____ Years

Please indicate the number of years experience you have in teaching each of the following subject areas:

Reading: _____ Years Math: _____ Years

Please list any teaching certificates or awards related to teaching that you may have received:

The NRS Educational Functioning Levels are listed below. Our standard setting meeting will involve the levels *Beginning Basic Education* through *High Adult Secondary Education*. Please indicate the learner levels that you are most comfortable discussing by rank-ordering each of the 5 relevant levels by circling a number for each level (where 1=the level you would most prefer discussing).

NRS Educational Functioning Level	Your Preference (circle one, 1=1 st , 5=5 th)				
Beginning ABE Literacy	N/A				
Beginning Basic Education	1	2	3	4	5
Low Intermediate Basic Education	1	2	3	4	5
High Intermediate Basic Education	1	2	3	4	5
Low Adult Secondary Education	1	2	3	4	5
High Adult Secondary Education	1	2	3	4	5

MAPT for Reading Standard Setting End-of-Meeting Survey

Thank you once again for agreeing to participate in this study to help set the NRS Educational Functioning Levels on the MAPT for Reading. At this point we would like to get your **anonymous** opinions regarding the study. Please take some time to let us know your impressions of the meeting today. You do not need to put your name on this survey. Your anonymous answers to the following questions will be used to evaluate the total process for establishing the NRS Educational Functioning Levels on this exam and to suggest revisions to the process, where necessary.

1. **Optional:** What are your primary responsibilities?

_____ teaching _____ administration _____ both _____ other

2. How clear was the goal of this meeting?

_____ Very clear _____ Clear _____ Somewhat Clear _____ Not Clear

3. How clear were the descriptions of the NRS Educational Functioning Levels?

_____ Very clear _____ Clear _____ Somewhat Clear _____ Not Clear

4. How would you judge the length of time that was allocated for discussing the NRS Educational Functioning Levels?

_____ About right _____ Not enough time _____ Too much time

5. How would you judge the length of time for making your item ratings (matching items to NRS levels) *for the first booklet* of items?

_____ About right _____ Not enough time _____ Too much time

6. How would you judge the length of time for making your item ratings (matching items to NRS levels) *for the second booklet* of items?

_____ About right _____ Not enough time _____ Too much time

7. How would you describe your understanding of the item matching task?

____ Understood Completely ____ Mostly Understood ____ Partially understood
____ Didn't understand at all

8. How confident are you that your item ratings will be useful for helping us set cut-points on the MAPT score scale to demarcate the NRS Educational Functioning Levels?

_____Very confident _____Confident _____Unsure _____Not Confident

9. Were you comfortable in expressing your opinions during the group discussions?

_____Yes _____No

10. When matching items to the NRS levels, which factors influenced your item ratings?
(check all that apply)

_____The NRS Educational Functioning Level definitions

_____The difficulty of the exam items

_____The group discussions

_____The knowledge and skills measured by the exam items

_____My experience in working with students

_____Other factors (please explain):

11. Please provide any other comments you would like to tell us about the standard setting study:

Thank You for Participating in this Study and Taking the Time to Complete This Survey!

Reading Standard Setting Meeting – Agenda, Description, Rating Form, Survey

Massachusetts Adult Proficiency Test (MAPT) for Reading

National Reporting System (NRS) Standard Setting Meeting

May 30, 2006

Agenda

8:00	Coffee, Muffins, Donuts, and Casual Conversation
8:30	Welcome and Introductions
8:45	Overview of MAPT Development Purpose of the Meeting
9:00	Description and Discussion of NRS Achievement Levels
10:15	Break
10:30	Introduction to Item Matching Task and Practice Ratings
11:30	Panelist Ratings—Booklet 1
12:15	Lunch
1:00	Panelist Feedback/Group Discussion
2:00	Panelist Ratings—Booklet 2
3:00	Break
3:15	Panelist Feedback/Group Discussion
4:00	Discussion
4:30	Survey and Assignments
5:00	Adjourn

Reading Performance Level Descriptions (NRS Functioning Literacy Levels)

Beginning Literacy • <i>Individual has no or minimal reading skills</i> • <i>Can recognize and read letters and numbers but has a limited understanding of connected prose and may need frequent re-reading</i>	
Beginning Basic • Individual can read simple material on familiar subjects • <i>Comprehend simple and compound sentences in single or linked paragraphs containing a familiar vocabulary</i>	Notes
Low Intermediate Basic • Individual can read text on familiar subjects that have a simple and clear underlying structure (e.g., clear main idea, chronological order) • Can use context to determine meaning • Can interpret actions required in specific written directions	
High Intermediate Basic • Individual is able to read simple descriptions and narratives on familiar subjects or from which new vocabulary can be determined by context • Can make some minimal inferences about familiar texts • Compare and contrast information from such texts but not consistently	
Low Adult Secondary • Individual can comprehend expository writing and identify spelling, punctuation, and grammatical errors • Can comprehend a variety of materials such as periodicals and nontechnical journals on common topics • Can comprehend library reference materials • Can identify the main idea in reading selections • Use a variety of context issues to determine meaning	
High Adult Secondary • Individual can comprehend, explain, and analyze information from a variety of literacy works, including primary source materials and professional journals • Can use context cues • Higher order processes to interpret meaning of written material	

Massachusetts Adult Proficiency Test for Reading

Standard Setting Rating Sheet (Sample)

Thank you for taking the time to review these MAPT for Reading items. As we discussed this morning, your review is very important for determining the cut-points (standards) along the MAPT for Reading that will correspond to the NRS Educational Functioning Levels helping us improve the quality of our test. Please review each item and circle your answer. Then check your answer and indicate the NRS level you think best measures the skills required to answer the item by placing a check mark in the appropriate column. If you have any comments you would like to make about a problem with the item, please mark it directly on the item booklet, and mark your page with a post-it.

Page	ID	Skill Area	Correct Answer	NRS Educational Functioning Levels (check one)				
				Beginning Basic	Low Intermediate	High Intermediate	Low Adult Secondary	High Adult Secondary
1	2620	Vocabulary	b					
2	1309	Word Recognition	b					
3	2611	Word Recognition	d					
4	251	Vocabulary	b					
5	2589	Vocabulary	c					
6	1891	Word Recognition	c					
7	160	Comprehension	d					
8	1982	Word Recognition	b					
9	1904	Vocabulary	b					
10	1882	Word Recognition	d					

Massachusetts Adult Proficiency Test (MAPT) NRS Standard Setting Meeting

Panelist Demographic Questionnaire

Name: _____

Current occupations relevant to ABE: _____

Teaching site: _____

Average class demographics: _____

Approximately how many years experience do you have teaching ABE learners? _____ Years

Please indicate the number of years experience you have in teaching each of the following subject areas:

Reading: _____ Years Math: _____ Years

Please list any teaching certificates or awards related to teaching that you may have received:

The NRS Educational Functioning Levels are listed below. Our standard setting meeting will involve the levels *Beginning Basic Education* through *High Adult Secondary Education*. Please indicate the learner levels that you are most comfortable discussing by rank-ordering each of the 5 relevant levels by circling a number for each level (where 1=the level you would most prefer discussing).

NRS Educational Functioning Level	Your Preference (circle one, 1=1 st , 5=5 th)				
Beginning ABE Literacy	N/A				
Beginning Basic Education	1	2	3	4	5
Low Intermediate Basic Education	1	2	3	4	5
High Intermediate Basic Education	1	2	3	4	5
Low Adult Secondary Education	1	2	3	4	5
High Adult Secondary Education	1	2	3	4	5

MAPT for Reading Standard Setting End-of-Meeting Survey

Thank you once again for agreeing to participate in this study to help set the NRS Educational Functioning Levels on the MAPT for Reading. At this point we would like to get

your **anonymous** opinions regarding the study. Please take some time to let us know your impressions of the meeting today. You do not need to put your name on this survey. Your anonymous answers to the following questions will be used to evaluate the total process for establishing the NRS Educational Functioning Levels on this exam and to suggest revisions to the process, where necessary.

1. **Optional:** What are your primary responsibilities?

_____ teaching _____ administration _____ both _____ other

2. How clear was the goal of this meeting?

_____ Very clear _____ Clear _____ Somewhat Clear _____ Not Clear

3. How clear were the descriptions of the NRS Educational Functioning Levels?

_____ Very clear _____ Clear _____ Somewhat Clear _____ Not Clear

4. How would you judge the length of time that was allocated for discussing the NRS Educational Functioning Levels?

_____ About right _____ Not enough time _____ Too much time

5. How would you judge the length of time for making your item ratings (matching items to NRS levels) *for the first booklet* of items?

_____ About right _____ Not enough time _____ Too much time

6. How would you judge the length of time for making your item ratings (matching items to NRS levels) *for the second booklet* of items?

_____ About right _____ Not enough time _____ Too much time

7. How would you describe your understanding of the item matching task?

_____ Understood Completely _____ Mostly Understood _____ Partially understood
_____ Didn't understand at all

8. How confident are you that your item ratings will be useful for helping us set cut-points on the MAPT score scale to demarcate the NRS Educational Functioning Levels?

_____ Very confident _____ Confident _____ Unsure _____ Not Confident

9. Were you comfortable in expressing your opinions during the group discussions?

____ Yes ____ No

10. When matching items to the NRS levels, which factors influenced your item ratings?
(check all that apply)

- ____ The NRS Educational Functioning Level definitions
- ____ The difficulty of the exam items
- ____ The group discussions
- ____ The knowledge and skills measured by the exam items
- ____ My experience in working with students
- ____ Other factors (please explain):

11. Please provide any other comments you would like to tell us about the standard setting study:

Thank You for Participating in this Study and Taking the Time to Complete This Survey!

Appendix G

Math Content Validity Meeting – Agenda, Description, Rating Form, Survey

Content Validity Panel Meeting

May 9th, 2006

University of Massachusetts Amherst

Agenda

8:30-8:45	Continental Breakfast and Light Conversation
8:45-9:00	Introductions and Completion of Paperwork
9:00-9:30	Description of the MAPT for Math Purposes Development Progress to date Next steps
9:45-10:15	Introduction to Content Rating Tasks Review of test specifications Description of tasks Training
10:15-10:30	Break
10:30-Noon	Review and Rating of Test Items
Noon-1:00	Lunch
1:00-4:00	Review and Rating of Test Items
4:00-5:00	Survey Completion and Group Discussion
5:00	Adjourn

Massachusetts Adult Proficiency Test (MAPT) for Mathematics and Numeracy

Cognitive Level Descriptions

For Content Validity Meeting, May 9, 2006

There are three cognitive levels on the MAPT for Mathematics and Numeracy. Distinctions between the cognitive levels should be based on the complexity of the task required to answer the question not on the difficulty of the task. It is important to focus on type of thinking required rather than the probability a student will get the task correct.

Knowledge and Comprehension: Questions at this cognitive level test recall of information and require a rote response. These questions test a most basic understanding and ask the student to perform straight calculations. These questions are the lowest level of understanding and test students' ability to comprehend information. Questions at this level are usually not in a context.

Application: Questions at this cognitive level test skills applied to a situation. Such questions require more in-depth thinking than a rote response. In application questions, students need to make decisions about what is required to solve the question and then implement their understanding.

Analysis, Synthesis and Evaluation: Questions at this cognitive level require the most complex thinking. In these questions, students break down the information into component parts to understand the steps required, or relate different ideas together to form a common understanding. These questions might also require students to evaluate and draw conclusions based on an understanding of the components or the situation. Students might also be asked to explain or infer findings based on the context of the question or the data provided. These questions might also ask students to recommend possible approaches or solutions to a problem.

Massachusetts Adult Proficiency Test (MAPT) for Mathematics and Numeracy

Content Strand Descriptions

Number Sense

Number Sense is the foundation of numeracy. Sound number sense enables us to interpret and represent the world in which we live. It is evident in all we do, whether in complex examples such as the Gross National Product, basic issues such as the family budget, or as personal as a blood pressure reading. Mathematical intuition grows with a strong basic understanding of numbers and, with that, our ability to do mathematical problem solving.

To be efficient workers or consumers in today's world, adults must have a strongly developed conceptual understanding of arithmetic operations, as well as the procedural knowledge of computation and number facts. They must be able to perceive the idea of place value and be able to read, write, and represent numbers and numerical relationships in a wide variety of ways. Simple paper-and-pencil computation skills are not enough. Adults must be able to make decisions regarding the best method of computation (mental math, paper-and-pencil, or calculator/computer) to use for a particular situation. Knowledge of numbers, operations and computation must include both a well-developed number sense and the ability to use basic mathematics-related technologies.

Number sense promotes accuracy in estimation and flexibility and efficiency in mental math. While calculators and computers are used to do most of the complex computations in today's world, the ability to estimate is critical for lifelong learners. Adults use informal measurements in life skill activities such as cooking, shopping, buying clothes, or estimating the time required for daily tasks. Estimation is a valuable skill for checking the reasonableness of computation or accuracy in problem solving, and is an aid in timed-test situations such as the GED. It builds on adult experience and knowledge. Good estimators use a variety of strategies and techniques for computational estimation that can be explored and shared by learners.

Learners engage in problem solving within adult contextual situations by communicating, reasoning, and connecting to:

- Standard N-1. Represent and use numbers in a variety of equivalent forms in contextual situations,
- Standard N-2. Understand meanings of operations and how they relate to one another, and in
- Standard N-3. Compute fluently and make reasonable estimates.

Patterns, Functions, and Algebra

Mathematics has been defined as the study of patterns. Learning to recognize, analyze, describe, and represent patterns and number relationships connects math to the world and helps us to appreciate fully the intrinsic value of such pleasures as poetry, art, music, and science. Math concepts formerly taught only in basic algebra courses are increasingly part of the culture and vocabulary of modern life. Headlines and news reports speak of exponential growth of the national debt, a variable rate mortgage, or a balanced budget, while medical literature uses terms like “HIV-positive,” or “RH-negative.”

Being able to see and use patterns has been identified as a fundamental skill needed for developing mathematical understanding. The Patterns, Functions, and Algebra strand is positioned after the Number Sense strand because of the importance of building pre-number skills such as patterning which, in turn, enable adult learners to learn multiplication tables and number relationships necessary for efficient and fluent computation skills. The strand also encompasses skills that are necessary for developing concepts in the Data and Geometry and Measurement strands.

Algebra serves as a bridge between arithmetic and more broadly generalized mathematical situations. These generalizations can be expressed in words, tables and charts, the notation of formulas, and graphs. Life experience has afforded adult basic education learners with a broad base of real-world ties that can be readily linked to the concepts of equation, function, variable, and graph. From baby formulas to chemical formulas, algebra offers a succinct way to define real-world situations that can aid adults in the home and in the workplace.

Algebra impacts the competency of workers, parents and citizens, and algebraic thinking skills are crucial if adults are to compete in the global economy. Workplace skills requiring competencies in “information,” “systems,” and “technology” stress the need for organizing, interpreting and communicating information and employing computers as a tool for those tasks, as well as the ability to “discover a rule or principle underlying the relationship between two or more objects and apply it in solving a problem.” Identifying and expressing pattern, relation and function are the algebraic skills imbedded within these competencies.

Learners engage in problem solving within adult contextual situations by communicating, reasoning, and connecting to:

- Standard P-1. Explore, identify, analyze, and extend patterns in mathematical and adult contextual situations,
- Standard P-2. Articulate and represent number and data relationships using words, tables, graphs, rules, and equations,
- Standard P-3. Recognize and use algebraic symbols to model mathematical and contextual situations, and
- Standard P-4. Analyze change in various contexts.

Statistics and Probability

The Statistics and Probability strand links numeracy and literacy learning. Numbers, logical reasoning, and texts interweave to describe phenomena visually, numerically and verbally in what we term *data*, which is the heart of this strand.

Data is a wide-ranging topic that touches on many areas of academic study and tells us much about our world. For instance, we learn about preferences, predilections and group characteristics when we read and interpret data. We learn about the power of evidence as we develop the skills to make statements and evaluate arguments based on data. We learn the power of the question and the framer of the question when we collect and represent data, and we learn that sometimes true, sometimes false, pictures are created when we compress data into statistics. Data is a powerful descriptive tool.

So powerful is data that agencies of authority often use it to generate, promote and, sometimes, evaluate decisions. Citizens, therefore, must understand the ways of data in order to exercise their collective and individual intelligence by responding to the expanding presence of this dense expression of information.

The learning standards in the Statistics and Probability strand provide adult learners with the tools for dealing with data.

Learners engage in problem solving within adult contextual situations by communicating, reasoning, and connecting to:

- Standard S-1. Collect, organize and represent data,
- Standard S-2. Read and interpret data representations,
- Standard S-3. Describe data using numerical descriptions, statistics and trend terminology,
- Standard S-4. Make and evaluate arguments or statements by applying knowledge of data analysis, bias factors, graph distortions and context, and
- Standard S-5. Know and apply basic probability concepts

Geometry and Measurement

Geometry and measurement help us represent in an orderly fashion what we see in our world. Whether we are cooking or cartooning, shopping or shipping, painting a canvas or a wall, designing an addition for a house or a play yard for preschool, we continually bump up against these mathematical organizers. Lifelong learners should know and understand these interconnected and symbiotic mathematical domains.

Adult learners who attend basic mathematics classes at any level share a wealth of pragmatic experience surrounding geometric and spatial concepts. They have probably built a bookcase, laid out a garden, applied wallpaper or tiled a floor, all the while discovering informally the rules which formally govern the study of geometry itself

Geometry and measurement often spark a renewed interest in mathematics for those students who have been turned off for some reason or heretofore have felt unsuccessful with mathematics learning. Investigating problems that involve geometry and measurement broadens all students' mathematical understanding and engages them as they explore mathematical ideas.

Hands-on, interactive investigations using nonstandard and standard units help adult basic education students develop an understanding of the many measurable attributes of physical objects. Measurement sense including length, time, temperature, capacity, weight, mass, area, volume, and angle will benefit from this approach. This realistic approach helps build an accessible measurement vocabulary and a meaningful comprehension of what it means to measure.

Learners engage in problem solving within adult contextual situations by communicating, reasoning, and connecting to:

- Standard G-1. Use and apply geometric properties and relationships to describe the physical world and identify and analyze the characteristics of geometric figures,
- Standard G-2. Use transformations and symmetry to analyze mathematical situations,
- Standard G-3. Specify locations and describe spatial relationships using coordinate geometry and other representational systems,
- Standard G-4. Understand measurable attributes of objects and the units, systems, and processes of measurement and apply appropriate techniques, tools and formulas to determine measurements.

Massachusetts Adult Proficiency Test for Mathematics

Content Validity Review Sheet (Sample)

Thank you for taking the time to review these MAPT for Math items. As we discussed this morning, your review is very important for helping us improve the quality of our test. Attached is an information sheet describing the Content Strands and Cognitive Areas measured on the test. Please review each item and indicate (a) the Content Strand *and* (b) the Cognitive Area you think the item is measuring. If you have any comments you would like to make about a problem with the item, please mark it directly on the item booklet, and mark your page with a post-it.

Packet	ID	Content Strand (Choose One)				Cognitive Area (Choose One)		
		Number Sense	Patterns, Functions, & Algebra	Stats. & Prob.	Geometry & Measure.	Knowledge & Comp.	App.	Analysis, Synth., and Eval.
1	1829							
1	2169							
1	2210							
1	2144							
1	1353							
1	1008							
1	2504							
1	1780							
1	992							
1	535							
1	822							
1	2226							
1	1446							
1	947							
1	1389							

MAPT for Mathematics and Numeracy Content Validity Panel

Participant Survey

Thank you for completing the item ratings! We are interested in your overall impressions of the test items you reviewed as well as in your impressions of the meeting today. Please take a few moments to answer the following questions. Your responses will be anonymous. Attach additional paper if necessary.

1) Do you think the item ratings you provided today will be useful for evaluating the degree to which the **Math Content Strands** are adequately represented by MAPT items?

☐ Definitely Yes ☐ Probably Yes ☐ Not Sure ☐ Probably No ☐ Definitely No

2) Do you think the item ratings you provided today will be useful for evaluating the degree to which the **Cognitive Areas** are adequately represented by MAPT items?

☐ Definitely Yes ☐ Probably Yes ☐ Not Sure ☐ Probably No ☐ Definitely No

3) We would like your impressions of the meeting today. Please indicate the degree to which you agree with each of the statements below using the rating scale provided.

Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The descriptions of the MAPT for Math <i>Content Strands</i> were clear.	SA	A	N	D	SD
The descriptions of the MAPT for Math <i>Cognitive Areas</i> were clear.	SA	A	N	D	SD
The directions regarding the item rating task were confusing.	SA	A	N	D	SD
In general, items on the MAPT for Math do NOT represent the Math <i>Content Strands</i> listed in the test specifications.	SA	A	N	D	SD
In general, items on the MAPT for Math do NOT represent the <i>Cognitive Areas</i> listed in the test specifications.	SA	A	N	D	SD
I enjoyed participating in this meeting.	SA	A	N	D	SD
It was hard for me to express my opinions at the meeting today.	SA	A	N	D	SD
The group discussions regarding what the items were measuring were very helpful to me in making my item ratings.	SA	A	N	D	SD
The data gathered today will be useful for evaluating the content quality of the MAPT	SA	A	N	D	SD

for Math.					
If another content validity study were to be conducted, I would like to participate again.	SA	A	N	D	SD

4) Given the test items you reviewed, do you think the content of this test is appropriate for measuring ABE learners' educational gains in Mathematics and Numeracy? ☐ Yes ☐ No

Comments:

5) Are there important mathematic skill areas that are *NOT* measured by the set of items you reviewed? If so, what are they?

6) What is your opinion regarding the range of *difficulty* of these test items? Do you think there is an appropriate range of difficulty for learners at levels Beginning Basic Education through High Adult Secondary Education?

7) Please comment on the clarity of the **Math Content Strand** definitions. Were these definitions clear? Are there any ambiguities? Are there other Math Content Strands that should be added to the test?

8) Please comment on the clarity of the **Cognitive Area** definitions. Were these definitions clear? Are there any ambiguities? Are there other cognitive areas that should be added to the test?

9) Please feel free to add any other comments regarding the content of the MAPT for Mathematics and Numeracy.

Thank you for your thorough review! Your comments are greatly appreciated!!

Appendix H

Reading Content Validity Meeting – Agenda, Description, Rating Form, Survey

Massachusetts Adult Proficiency Test (MAPT) for Reading

Content Validity Panel Meeting

May 3, 2006

University of Massachusetts Amherst

Agenda

8:30-8:45	Continental Breakfast and Light Conversation
8:45-9:00	Introductions and Completion of Paperwork
9:00-9:30	Description of the MAPT for Reading Purposes Development Progress to date Next steps
9:45-10:15	Introduction to Content Rating Tasks Review of test specifications Description of tasks Training
10:15-10:30	Break
10:30-Noon	Review and Rating of Test Items
Noon-1:00	Lunch
1:00-4:00	Review and Rating of Test Items
4:00-5:00	Survey Completion and Group Discussion
5:00	Adjourn

MAPT Reading Skill Areas

Reading is a complex proficiency. To define the reading proficiencies measured on the Massachusetts Adult Proficiency Test (MAPT) for Reading, the Test Specifications Development Committee specified three reading “Skill Areas.” A brief description of each of these areas appears below.

Word Recognition

This skill area involves recognizing controlled words, sight words, and visual cues. Word recognition skills are not comprehension skills. They simply allow the learner to read the word off the page. Examples of MAPT for Reading standards that fall under this skill category include (a) Identify short vowels with single consonants (one syllable words); (b) Identify words with the hard and soft sounds of “c” and “g;” and (c) Identify words with common suffixes, such as -ous, -able and -ible, -ture, -ive, and -age.

Vocabulary and meaning

This skill area involves the ability to understand word meanings and the ability to infer word meanings from context. Examples of MAPT for Reading objectives that fall under this skill area include (a) Recognize synonyms, antonyms, homonyms, and homophones (e.g., dear-deer) for identified vocabulary words presented in isolation or within a group of words; (b) Use context clues to determine meaning of unfamiliar words, including increasingly sophisticated multi-syllabic words originating from Latin and Greek prefixes, suffixes, and roots; (c) Identify idioms and their use in passages; and (d) Recognize and comprehend the meaning of moderately complex vocabulary using word, sentence, and paragraph clues to determine meaning.

Comprehension

Comprehension refers to the ability to extract meaning from connected discourse that is read. Examples of MAPT for Reading objectives that fall under this skill area include (a) Locate pertinent information in simple familiar materials (e.g., want ads, job listings, schedules, signs, food packages) including who, what, which, when, where; (b) Distinguish between fact and opinion in a short paragraph; (c) Determine author’s purpose from a variety of written pieces; and (d) Draw conclusions and make inferences based on passages read.

Definitions of Purposes of Reading

The Purposes of Reading dimension of the MAPT test specifications refers to specific reasons why learners may need reading skills. This dimension was used as one source for selecting material to be included on the MAPT for Reading. A brief description of each of these four purposes is provided below.

Reading to Acquire Basic Word Analysis Skills

This “purpose for reading” area is seen as part of the necessary skill development process to become a proficient reader. This purpose includes a desire to develop an understanding of how phonemes, or speech sounds, are connected to print, as well as the ability to decode unfamiliar words.

Reading for Literary Experience

To read for literary experience is to listen to voices not one’s own and to gain access to new experiences and perspectives. The encounter with literature is likely to engage the heart and the senses, as well as the mind. This encounter may be challenging but also offers the possibility of great pleasure. In many cases, the purpose behind Reading for Literary Experience is to entertain oneself (i.e., read for pleasure).

Reading to Gain Information

Reading is one of the primary means of gathering the information that is required to make appropriate choices and to analyze and appreciate the complexity of life. A wide variety of texts offer information, from government documents to medical advisories, from newspaper articles to notes sent home from a child’s school. In Reading to Gain Information, the purpose is to learn about something.

Reading to Perform a Task

On a daily basis, people need to be able to do things: for example, operate a machine, open a bank account, cook a meal, or vote in an election. The challenge to the reader is to understand the purposes of the task, the vocabulary specific to the task, and the activity required to successfully complete the task.

Massachusetts Adult Proficiency Test for Reading

Content Validity Review Sheet (Sample)

Thank you for taking the time to review these MAPT for Reading items. As we discussed this morning, your review is very important for helping us improve the quality of our test. Attached is an information sheet describing the Skill Areas and Purposes for Reading measured on the test. Please review each item and indicate (a) the Skill Area *and* (b) the Purpose for Reading you think the item is measuring. If you have any comments you would like to make about a problem with the item, please mark it directly on the item booklet, and mark your page with a post-it.

Packet	Item ID	Skill Areas (Choose One)			Purposes for Reading (Choose One)			
		Word Recognition	Vocabulary and Meaning	Comp.	Acquire Basic Skills& Word Analysis	Literary Exp.	Gain Info.	Perform A Task
1	1881							
1	805							
1	340							
1	2491							
1	1940							
1	1552							
1	2570							
1	2612							
1	2611							
1	1937							
1	807							
1	2489							

MAPT for Reading Content Validity Panel

Participant Survey

Thank you for completing the item ratings! We are interested in your overall impressions of the test items you reviewed as well as in your impressions of the meeting today. Please take a few moments to answer the following questions. Your responses will be anonymous. Attach additional paper if necessary.

1) Do you think the item ratings you provided today will be useful for evaluating the degree to which the **Reading Skill Areas** are adequately represented by MAPT items?

☐ Definitely Yes ☐ Probably Yes ☐ Not Sure ☐ Probably No ☐ Definitely No

2) Do you think the item ratings you provided today will be useful for evaluating the degree to which the **Purposes for Reading** are adequately represented by MAPT items?

☐ Definitely Yes ☐ Probably Yes ☐ Not Sure ☐ Probably No ☐ Definitely No

3) We would like your impressions of the meeting today. Please indicate the degree to which you agree with each of the statements below using the rating scale provided.

Statement	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
The descriptions of the MAPT for Reading Skill Areas were clear.	SA	A	N	D	SD
The descriptions of the MAPT Purposes for Reading were clear.	SA	A	N	D	SD
The directions regarding the item rating task were confusing.	SA	A	N	D	SD
In general, items on the MAPT for Reading do NOT represent the Reading Skill Areas listed in the test specifications.	SA	A	N	D	SD
In general, items on the MAPT for Reading do NOT represent the Purposes for Reading areas listed in the test specifications.	SA	A	N	D	SD
I enjoyed participating in this meeting.	SA	A	N	D	SD
I was hard for me to express my opinions at the meeting today.	SA	A	N	D	SD
The group discussions regarding what the items were measuring were very helpful to me in making my item ratings.	SA	A	N	D	SD
The data gathered today will be useful for	SA	A	N	D	SD

evaluating the content quality of the MAPT for Reading					
If another content validity study were to be conducted, I would like to participate again.	SA	A	N	D	SD

4) Given the test items you reviewed, do you think the content of this test is appropriate for measuring ABE learners' educational gains in reading? ☐ Yes ☐ No

Comments:

5) Are there important reading skill areas that are *NOT* measured by the set of items you reviewed? If so, what are they?

6) What is your opinion regarding the range of *difficulty* of these test items? Do you think there is an appropriate range of difficulty for learners at levels 2 through 5?

7) Please comment on the clarity of the **Reading Skill Area** definitions. Were these definitions clear? Are there any ambiguities? Are there other Reading Skill Areas that should be added to the test?

8) Please comment on the clarity of the **Purposes for Reading** definitions. Were these definitions clear? Are there any ambiguities? Are there other purposes that should be added to the test?

9) Please feel free to add any other comments regarding the content of the MAPT for Reading.

Thank you for your thorough review! Your comments are greatly appreciated!!