

No Adult Left Behind, Either:  
Creating Large-Scale Computer-Based Tests for Adult Basic Education Students<sup>1</sup>

April L. Zenisky and Stephen G. Sireci

Center for Educational Assessment  
University of Massachusetts Amherst

April 15, 2005

---

<sup>1</sup> Center for Educational Assessment Report No. 563. University of Massachusetts, School of Education. Paper presented at the annual meeting of the American Educational Research Association (Adult Literacy Special interest Group), Montreal, Canada.

## Abstract

Testing to chart student performance and to hold schools and districts accountable is everywhere in K-12 school systems in the United States today. These accountability and student assessment demands are equally strong in adult basic education (ABE), even though such assessments receive relatively less attention from policy makers and psychometricians. In Massachusetts, efforts to accurately and validly measure what adult learners know and can do has led to the process of creating tests that 1) are aligned to the state's ABE curriculum frameworks (i.e., criterion-referenced), 2) contain content appropriate for *adult* learners, and 3) use computer-based testing technology to tailor the tests to adult learning levels and to provide accurate measurement throughout a large achievement continuum. The vision underlying the development of these new tests centers on recruitment and involvement of ABE teachers and administrators throughout the state. This strategy enables teachers and others to take some ownership of the testing program and formalizes the link between assessment and instruction. The purpose of this paper is to describe aspects of the process of developing computerized tests from both psychometric and practical perspectives, with particular discussion of efforts to ensure teacher involvement, curricular relevance, and measurement precision using an innovative computer-based testing design.

## **No Adult Left Behind, Either: Creating Large-Scale Computer-Based Tests for Adult Basic Education Students**

### **Introduction**

The increased use of standardized tests for the purposes of school and district accountability in elementary and secondary education is well known. Less known, however, is the similar emphasis being placed on test results in adult basic education (ABE) in the United States today. With the Workforce Investment Act (WIA) of 1998, state ABE systems must measure and report learners' literacy gains to the United States Department of Education. This must be done using assessments that are reliable and valid for this purpose. However, a recent report from the National Academy of Sciences (Mislevy & Knowles, 2002) pointed out several significant challenges in accurately measuring adult learners' knowledge, skills, and abilities.

One such significant challenge is accurate measurement across the wide range of skills possessed by ABE students. The range of skills in the ABE population typically ranges from preliterate to postsecondary. Thus, a "one-size fits all" testing approach is insufficient. A second significant challenge stems from the amount of time adults spend in ABE classrooms, which is typically much shorter than in K-12 education. Shorter instructional time obviously impacts the amount of possible gains learners may exhibit. In short, the accountability demands in ABE require measurement of very small gains across a wide spectrum of ability. An additional significant challenge for valid assessment of ABE learners is that when it comes to testing this varied population, many adult educators often resort to using tests that are not optimal for their students, either with respect to the level of the content, the match between the curriculum and the test, and/or the intended interpretation of test scores (Mislevy & Knowles, 2002).

Drawing on key suggestions provided in Mislavy and Knowles (2002), the Massachusetts Department of Education decided to develop customized tests for measuring student performance relative to their ABE curriculum. In January 2003, the Massachusetts Department of Education formalized a relationship with the Center for Educational Assessment at the University of Massachusetts Amherst to collaboratively develop reading and math tests for adult learners in Massachusetts. In this paper, we describe the key features of this assessment system and how the features facilitate more valid assessment of adult learners' knowledge and skills. In addition, we summarize the work completed on these new assessments thus far, and the lessons learned in completing this work. Our purpose in communicating these activities and lessons learned is to illustrate how modern advances in test development and administration can lead to improved assessment in ABE for accountability and other purposes.

### Key Features of the Massachusetts Proficiency Tests for Adults

While the process of creating a customized test for adults for use in a large-scale statewide testing context is not one to be taken lightly, several recommendations from the National Academy of Sciences report were particularly influential in guiding our test development efforts. The specific recommendations incorporated were (a) prioritizing assessment goals, (b) fostering collaborations among ABE personnel and psychometricians, (c) taking advantage of computer-based testing technology, (d) using test development to create professional development opportunities for ABE teachers and administrators, and (e) developing benchmarks associated with NRS levels. Our vision for the Massachusetts Tests involves collaboration among psychometricians and ABE teachers and administrators. Table 1 presents a very brief description of the desired features of the testing program and how these features are

incorporated into the system. These and other features of the new system are described subsequently.

Table 1. Summary of Key Features of Massachusetts Proficiency Tests for Adults

| Desired Feature  | Implementation Strategies   |
|--|---|
| Content aligned with state ABE curriculum                        | Test specifications derived from curriculum frameworks.<br>Teachers trained to write test items   |
| Content representative of classroom instruction                  | Test specifications derived from curriculum frameworks.<br>Teachers trained to write test items<br>Innovative (computer-based) test items |
| Accurate measurement across wide proficiency spectrum            | Computerized adaptive test administration<br>Item response theory scaling   |
| Professional development for teachers                            | Fundamentals of Testing Workshops<br>Item writing workshops<br>Test administration workshops<br>Development of curricula material         |
| Reduction of test anxiety  | Appropriate content for adult learners<br>Computer proficiency study<br>Computerized testing tutorial<br>Computerized pilot tests         |
| Provision of reliable and valid data for accountability purposes | Vertical score scales<br>Standard (achievement level) setting<br>Recommendations regarding criteria for gains at the program level        |

Following some background on the new tests being developed, the remaining sections of this paper briefly describe some of the unique steps and challenges faced throughout the process of creating computerized tests. Important lessons we have learned thus far are also shared.

### Background on the New ABE Tests

After the implementation of the WIA, a committee of ABE teachers and administrators was formed in Massachusetts to work together to develop recommendations for performance accountability in ABE programs. Among its many tasks, this group, called the Massachusetts

ABE Performance Accountability Working Group (PAWG) reviewed all available standardized assessments for use for WIA reporting purposes, finding that most existing products in use exhibited a mediocre level of overlap at best with the Massachusetts ABE curriculum frameworks. This issue of curriculum-test alignment is not trivial. Scores from a test that a) do not adequately align with the information students are learning and/or b) assess knowledge of content students did not have the opportunity to learn do not lend themselves to a meaningful representation of what those students know and are able to do.

The Massachusetts PAWG found that the commercially available test with the greatest level of overlap with the ABE math curriculum framework covered only about 56% of the framework. The other tests in that same battery exhibited even less overlap with the English Language Arts framework. This instrument, a standardized test from a national testing agency, was recommended for testing ABE students at a grade-equivalent of performance at 2.0 and above in Math and Reading on an interim basis until a customized system of both standardized tests and performance-based assessments that are fully aligned with these frameworks could be developed (for the final PAWG report, see <http://www.doe.mass.edu/acls/pawg/report0902.pdf>).

Shortly after these recommendations were made, the process of developing new tests was initiated, with the Center for Educational Assessment at the University of Massachusetts Amherst as the contracted test developers. A primary goal of the new test development process was the optimization of test-curriculum alignment. Given limited resources, math and reading were identified as the most important subject areas for test development. The scores derived from these tests will be criterion-referenced, and will provide students, teachers, and the state with scale scores and performance classifications for individual students as well as scores that can be aggregated at the program and state levels.

### **Implementing the Desired Features of the New Tests**

The desired features of the new reading and math tests, summarized in Table 1, were influenced by several sources including the National Academy of Sciences committee report (Mislevy & Knowles, 2002), the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education, 1999), and conversations with the Adult and Community Learning Services office at the Massachusetts Department of Education. Essentially, these desired features are designed to provide content-valid tests that are useful and appropriate for ascertaining students' proficiencies and providing information useful for evaluating educational gains at the program level. In this section, we discuss each of these desired goals and explain how they were (or will be) implemented.

#### **Aligning the Tests to the ABE Curriculum**

The development of the test specifications for the Reading and Math tests involved formation of committees of content experts comprised of accomplished ABE teachers and administrators. Convened to meet over the course of several months, their charge was to review the curriculum frameworks and develop test specifications identifying the percentage of items at each level that come from each combination of content strands and cognitive levels. The Math (Sireci, Baldwin, et al., 2004) and Reading (Sireci, Li, et al., 2004) test specification reports are available at the Massachusetts Department of Education of Education website ( see <http://www.doe.mass.edu/acls/mailings/2004/0709/mathtestspecs.pdf> and <http://www.doe.mass.edu/acls/mailings/2004/0709/readtestspecs.pdf> for math and reading specifications, respectively).

### Linking Test Content to Classroom Instruction

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 students work on in the classrooms. 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. In addition to maximizing test-curriculum alignment, this strategy encouraged teachers to be part of the testing 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 over a two-year period. 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. Five workshops were offered in the five regional divisions of the state over the course of a year, and two other workshops were offered at statewide conferences. Through these workshops, about 100 teachers were trained to write test items targeted to the test specifications. In addition, 15 other teachers participated in a two-week

graduate course in test development at the University of Massachusetts. Through the workshops and test development course, over 1,000 items were written by ABE teachers for the new tests.

#### Accurate Measurement Across the ABE Learner Continuum:

As mentioned earlier, a key feature of our vision for quality ABE test development was accurate measurement across a wide spectrum of proficiency. As pointed out in the National Academy of Sciences report (Mislevy, & Knowles, 2002), computerized adaptive testing technology is appropriate, and necessary, to accomplish that goal. Although adaptive testing technology is one of the most sophisticated developments in contemporary psychometrics, merely computerizing a test for adult learners is a challenge in itself.

#### Computerization and Test Development

When considering computerizing tests for adult learners, the array of possibilities seems endless. Primary issues to consider are test delivery options (e.g., Web versus standalone application), linear versus adaptive administration, and practical matters such as registration procedures and operating systems. Critical concerns include ease of administration and the extent to which the interface can be understood and navigated by adult learners. Furthermore, the need for an adaptive test requires that the test delivery system be flexible enough to handle all the mathematics of and constraints associated with adaptive testing. Significant attention must also be paid to the specific delivery system due to the fact that the test will be administered numerous times a day at numerous adult education programs throughout the state. Is it better to use a Web-based browser for delivery, or is it preferable to create a standalone, clickable Windows-based application? The concern in the former case is ensuring that all test

administration sites invest in connecting to the Internet at a minimum level of reliability and speed, while in the latter case is that all programs must invest in a certain level of operating system (e.g., Microsoft Windows 98 or higher, or perhaps XP).

These are just a few of the decisions that face the developer of a computer-based test for adult learners. Some of the considerations central to the Massachusetts experience include the need to evaluate the computer facilities of adult basic education programs across the state with respect to the type and number of computers, the operating systems of those computers, and the level of familiarity with computers among both ABE students and program staff. Some focus on linking a computerized test delivery system to the data reporting system used in the state was also essential.

In this section, we describe specific steps in our test development process that were particularly challenging. A review of these steps and issues should prove illuminating to others considering development of computer-based tests for adults. We begin our list of critical steps with the selection of test delivery vendors.

Vendor selection. O'Neil, Zenisky, and Sireci (2003) reviewed the available vendors for computer-based test delivery. In this review, several trends became apparent. First, more and more vendors are marketing products for Web-based testing using Web browsers rather than executable files that must be installed on individual machines. Second, a wide range of features and levels of customer support are available among vendors, which necessitates careful consideration of the needs of the testing population as well as the technical support needs of the test development and contracting agencies.

In comparing the services of vendors, two main criteria are suggested. One involves the level of measurement expertise a vendor provides with respect to the customer (either in terms of

the software they offer or technical support). The second is the level of design flexibility a vendor offers. By this we mean in terms of features like the delivery models supported (non-adaptive to adaptive CBT), delivery options available (PC, LAN, Internet), item creation flexibility (item formats, appearance, editing capabilities, etc.), item banking (storage, revision, exposure controls, etc.), whether classical test theory and/or item response theory is supported, level of item/test analysis (expert review, incorporated into software), and quality of examinee data-collection (general and test performance), management, and analytic software tools.

Although several vendors appeared viable for delivering our tests, we were fortunate to be able to partner with the instructional technology resources of our University. Thus, the new ABE tests are being developed using the Online Web-based Learning (OWL) system that is a product of the Center for Computer-Based Instructional Technology at UMass Amherst. The OWL system is an Internet-based learning tool that is used by dozens of departments and programs at UMass to administer homework, tests and quizzes to thousands of UMass undergraduates every semester. This system is entirely Web-based, and can be accessed with either the Internet Explorer or Netscape Navigator browser. Our decision to deliver the tests through a University system was motivated by factors such as cost and flexibility. We are able to work directly with the system administrators as they customize the system for our needs, and we hope that ultimately, the University will partially support this endeavor as part of its service mission to the state.

Item banking and test development software. At the outset of test creation the capabilities of the OWL system for item banking were substantial. The system permits use of multiple item formats (including multiple-choice, drop-down menus, free-text entry, matching, and hot-spot clicking on graphics). Each item is defined within OWL by multiple attributes

including content area, content substrand measured, cognitive skill required, specific curriculum framework standard, item author, ready status (e.g., new, pretest, operational), review status (under content review, under sensitivity review, etc.), test level/difficulty, and keywords.

An XML upload of the curriculum framework allows for the every element of the curriculum framework to be browse-able for linking individual items to the frameworks. At some point in the near future it is envisioned that this electronic version of the framework could not only be a place to link test items, but also a curricular resource for teachers to link in instructional resources across the larger Massachusetts (and perhaps national) ABE community for sharing of lesson plans and ideas that are keyed directly to standards in the curriculum. Other features of the system in terms of test development include reports that are used internally by UMass staff for reviewing test items and an extensive system of grading reports that can be used to monitor student progress as well as manage data from field-tests.

As OWL is a Web-based system, members of UMass' test development staff have gained greater familiarity with the basic functions of hyper-text markup language (HTML) and use that approach for rendering graphics and passages for test items. Graphics such as charts, tables, and pictures are handled as image file objects that are uploaded to the server and can be called for use in any question, and text passages are similarly created as files of type text are marked in HTML and can be associated with single or multiple items as needed (and items can be denoted as belonging to sets).

Field-testing. Field-testing items is a critically important step in test development. Through field-testing, student response data allows for analyses to evaluate the statistical quality of items, to confirm subject matter experts' judgments of the level-appropriateness of items, to identify potential mis-keyed items or those with more than one answer, and to carry out tests of

differential item functioning. Furthermore, the field test represents an opportunity for students and ABE teachers to develop a familiarity with the OWL test administration system and to identify aspects that may need to be changed for operational testing.

Although field-testing items on a statewide basis is important, it presents several significant logistical challenges. Among these:

- Logins for all students needed to be created so that students could access the OWL system via the Internet (also, these logins and passwords needed to be communicated to teachers and students);
- Teachers at over 100 programs across the state had to be trained on the system to administer the field test to their students; and,
- Sets of test items at different difficulty levels needed to be assembled, and multiple forms at each difficulty level had to be folded in using a round-robin algorithm to allow for more items at each level to be tried out.

Each of these challenges will be discussed in turn.

The OWL system is created for open access, as it is built on a platform that allows only those users with active accounts on the system to access content. To carry out some aspects of item analysis, it was necessary to be able to link the previous standardized test scores of students in the Department of Education's database to their performance on the field-test items for validation purposes.

To solve this problem a numerical login was created for each student using a four-digit code that corresponded to a student's program and a six-digit code that identified each individual student in the statewide database (example: 7723-838394). Working with the Department of Education, we uploaded all students in the state's database into the OWL system, so that every

student would have access to the system to take a test. Every night, a supplementary upload is performed to ensure that logins are created for any new students joining an ABE program in Massachusetts within twenty-four hours of the students being entered into the state's system. Each program in the state has access to the state's database, and can print class lists with the students' logins as a field in that report. A common password was selected and told to teachers, to minimize at least one step in the complexity of the login process.

Training teachers for test administration. To train teachers across the state to administer this computerized test, computer labs at multiple programs across Massachusetts were used as sites for training sessions. At least two training sessions were scheduled in each region, and the state Department of Education required a minimum of two teachers per program to attend one training session. At each training session, teachers were provided with some background information on the new tests, as well as an overview of the testing system. Each teacher was provided with an individual teacher login which gave them access to the system and allowed them to go into OWL and experience the test for themselves. Teachers could take these logins and passwords back to their programs and use them to develop their own familiarity with the system and to train other teachers at their programs to administer field tests as well.

Teachers were also provided with a sequence of six lessons plans about standardized testing that they could use in part or as a whole sequence. These lessons were developed by three ABE teachers in Massachusetts to introduce the topic of testing and specifically the activity of field-testing to students to help students understand their roles in the process. These lesson plans were 1) having students discuss and write about their prior experiences with testing, 2) reading short stories (targeted to a GLE of about 3.5) of ABE students' experiences taking tests, 3) introducing the definition of standardized tests, their purposes, and pros and cons, 4) an

overview of test-taking strategies in which students identify things that they could do to prepare for taking high-stakes tests, 5) discussion specific issues related to the field-tests and their participation, and 6) a post-field test discussion where students talk or write about their experiences and refine the test-taking strategies they identified in Lesson 4. At the trainings, teachers were urged to relay to test developers and staff from the Department of Education their observations and those of their students.

Developing pilot-test forms. In addition to creating logins and training teachers, behind the scenes it was necessary to develop some modifications to the OWL system to ensure that as many items as needed from a psychometric perspective could be tried out. Of course, each student cannot take hundreds of tryout items. Given limits of time for field-testing at programs, each student taking a field test receives just thirty test items. This required the use of multiple “forms” of the test, assembled to reflect rudimentary content balancing within difficulty levels, but some overlap of items across difficulty levels and across forms within a difficulty level was necessary to ensure that each items being tried out could be evaluated relative to all other tryout items on the same mathematical scale.

Test developers at UMass created multiple forms in this way to try out items. In any given difficulty level of the test, there are perhaps 10 different forms. Specifically, we developed five forms at each learner level, but had two different orderings of test items within each form so that items at the end of the test would not have higher omit rates. An algorithm for how to assign test forms to students was also developed so that all forms would receive adequate exposure and students within a program would not all take the same form.

Multi-stage test delivery model. To meet our goal of getting accurate measurement of ABE learners’ proficiencies, it was clear that we needed some form of “tailored” or adaptive

testing technology. *Computerized adaptive testing* 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 (CAT) reduces testing time by shortening the length of the test without losing information about examinee proficiency.

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 each time he or she answers a test item, and a new item is selected based on the updated estimate. When the proficiency estimate is calculated, an estimate of the amount of uncertainty in the estimate (i.e., an estimate of the error of measurement) is also calculated. As more and more items are administered, the degree of uncertainty diminishes. This reduction explains why accurate measurement of ABE learners can be accomplished using fewer items than is typically required.

Although CAT technology is efficient, there were two immediate problems in applying it to ABE assessment. The first problem was deciding on a starting point for each learner on the

wide ABE proficiency continuum. A second problem was ensuring representation of the content specifications of the test. In considering these problems, we devised a three-stage testing model that represents a variation on the traditional CAT method:

- Stage 1: Background Positioning Questionnaire
- Stage 2: 20 Computerized Adaptive Items
- Stage 3: Final Fixed Set of 20 Items.

The purpose of the first stage is to determine the most appropriate starting point for each examinee. The purpose of the second stage is to find the best testing level for the examinee. The purpose of the third stage is to accurately measure each examinee with the desired content at that level. The three-stage design is explained in greater detail in Table 2 and Figure 1.

Having students answer several short academic background questions in Stage 1 gives the computer a rough idea of performance level, which determines where it will begin administering Stage 2 items. Stage 2 is an item-level adaptive test. At the conclusion of Stage 2, the proficiency estimate for a student is mapped into a specific testing level that occurs at Stage 3. Stage 3 represents a traditional, non-adaptive (fixed-form) test that completes the task of ensuring accurate measurement of each student's proficiency and complete content representation. To our knowledge, this type of multi-stage, partially-adaptive testing system has not been implemented to date.

Table 2. Multi-Stage Test Design

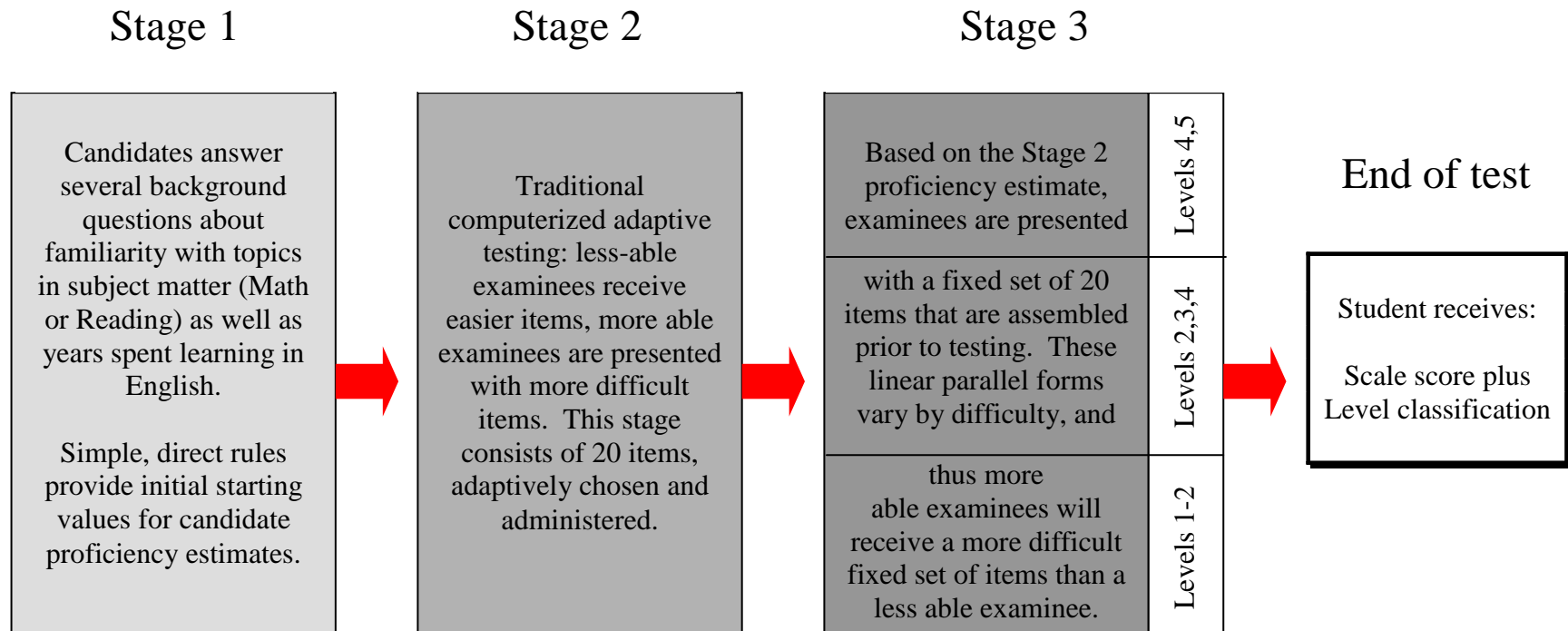
| Stage 1  | Stage 2   | Stage 3  |
|--|---|--|
| <p>In Stage 1, Examinees are presented with 2-5 “can do”-type questions, relevant to the content area (Math or Reading)</p> <p>Simple direct rules determine starting values for proficiency estimation in Stage 2</p> <p>“How many years of schooling did you have?” or “What is the highest grade level you achieved?”</p> | <p>Stage 2 is a traditional CAT. Consists of 20 items chosen adaptively.</p> <p>Proficiency estimate obtained at the conclusion of Stage 2 determines Level for Stage 3 items to be administered.</p> | <p>Stage 3 is composed of multiple fixed forms varying by approximate difficulty, 20 items per form.</p> <p>At each difficulty level, for each examinee a form is randomly selected from 3-5 parallel forms</p> <p>Proficiency estimate from Stage 2 identifies level of Stage 3 form for each examinee.</p> <p>Some overlap of items across levels will allow for proficiency estimation for all examinees on the same scale.</p> |

---

Note: Content specifications are met across Stages 2 and 3.

---

Figure 1. Computerized Test Design for New ABE Tests in Massachusetts



### Reduction of Test Anxiety

Very few students enjoy taking tests and in ABE test anxiety is a common issue. 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 below) 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 students of various learning levels (i.e., beginning ABE to GED prep students) from several ABE programs. 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, multiple-choice. For further descriptions of these item formats, see Zenisky and Sireci (2002).

Most students 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.

A summary of these students’ responses to the survey questions specific to each of the item formats investigated is presented in Table 3. This summary presents only the percentage of students who “agreed” with each statement. In general, the students reviewed the item formats favorably. 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.

Table 3. Summary of Survey Responses Regarding Computerized Item Formats (n=32)

| <u>Survey Item</u>   | 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                     |

Development of computerized test tutorial. While the small-scale observations reported above indicated that a moderate level of computer familiarity was present among some ABE students, it was also considered crucial to provide all students with an opportunity to become familiar with basic computer response actions relative to using a computer to enter answers to test questions. Thus, a tutorial was developed by the test developers in collaboration with the Massachusetts Department of Education and several ABE educators. It, as is the case with the test, is accessible via the Web (<http://owl.cs.umass.edu/departments/ctredassess/howto/b1.htm>).

The current tutorial is comprised of two portions: first, the basics of using a mouse, clicking, scrolling, and answering a question using a radio button. Each page in this section gives students an opportunity to practice each of those skills. Upon completion of that part, students then move on to several web pages in which the information needed to navigate the actual testing system is provided. This second section at this time is not interactive. As it is

intended to show students how to move from question to question, it seemed necessary to tutorial developers to use screen captures to illustrate the process in a static way before giving students access to the system. The tutorial can also be printed out and used as a paper reference.

### Professional Development for Teachers

As indicated in Table 1, one of our goals was to make professional development for ABE teachers and staff a natural consequence of the test development process. At this point, our professional development efforts are probably already understood. These efforts included over a dozen workshops delivered across the state over the course of 18 months. 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 tests. The three-credit graduate course offered at UMass is another example, as are the numerous workshops offered to train teachers how to deliver the pilot tests at their programs. In addition, we offered a 15-hour introduction to assessment course for the curriculum and assessment coordinators in the state. The idea behind this course was to “train the trainers” to increase assessment literacy among ABE teachers and staff throughout the state.

### Provision of Reliable and Valid Data for Accountability Purposes

The last desirable test feature listed in Table 1 is perhaps the most important since it refers to the degree to which the scores from our tests will provide valid and reliable data for the purpose for which the tests are developed. All our test development steps aim toward that goal. Specifically, carefully developed test specifications and quality items facilitate content validity, and an IRT-scaled CAT facilitates accurate measurement along the wide continuum of ABE

learners' proficiency. What we have not described at this point is how standards (achievement levels) will be set on this wide continuum. We are currently reviewing several standard setting options (including modifications of the Angoff method, the bookmark method, and the direct consensus method, see Cizek, 2005) and we will implement those methods prior to or soon after the tests become operational (July 2006). Further future activities planned for the near future that are designed to support the use of test scores for accountability purposes include analysis of test score data to evaluate program gains and providing options to the Massachusetts Department of Education for aggregating test scores to evaluate gain at the program level. We leave discussion of these important issues for a subsequent paper.

### **Current Reactions from the Massachusetts ABE Community**

Over the course of March 2005 when many of the field-test training sessions were being held, ABE instructors across Massachusetts were being presented with the first extended look at the new tests and how they would work. Staff from the Massachusetts Department of Education and the Center for Educational Assessment who were providing the training sessions encouraged teachers to provide feedback via email on the process and the OWL system. Several themes to their comments can be identified.

First, a large proportion of teachers who communicated with us by email expressed appreciation regarding our efforts to create a test based on the Massachusetts ABE curriculum and that provides information useful to the classroom teacher. At the same time, we learned that many teachers feel the process of administering this field test (and indeed, standardized testing in general) is time-consuming, and takes time away from instruction by requiring teachers to attend training sessions and students to complete field-tests.

The idea of testing students via a computer has raised a great deal of interest as well as many concerns and questions. One significant concern involves computer familiarity and the potential for anxiety among students. Some programs have inquired about the use of touch screens for entering student responses rather than using a mouse, indicating that some students may be more at ease being able to interact with the screen directly rather than having to acclimate to using a computer mouse. We are currently exploring that option, which is less expensive than it sounds.

Regarding the user interface, several teachers requested a simpler interface. The OWL interface was designed for university students and has several sidebars that may distract ABE students. Some ABE teachers requested that these advanced functions be removed from the ABE version of the interface to minimize screen clutter and any potential for confusion. Furthermore, OWL navigation is visual, meaning that students move from item to item using a numbered sequence of icons at the top of the screen where each item on the test is represented by an icon. Depending on whether a student has completed a question, the icon is either a question mark (not answered) or a pencil (answered). Some teachers thought this approach is awkward, while others suggested it allows students to check if they responded to all questions on the test. To this end, several teachers have suggested alternate ideas for navigation, and these comments are encouraged. As field-testing continues, these comments will be compiled and will form the basis for future decisions about the appearance and functioning of the new tests.

### Steps Remaining in Test Development

Though field-testing is well underway (and several hundred students at sites across Massachusetts have already completed sequences of tryout items), much work clearly remains to

be done. Additional forms of tryout items will be spiraled into the system as field-testing continues throughout the spring. As student response data are compiled, we will carry out statistical analyses to evaluate all items, and revise, accept, or discard as warranted. Based on comments from the field gathered throughout the field-testing process, some modifications to the user interface may be implemented. In addition, development work on the adaptive features of the operational test will occur over the summer and will also need to be field tested in the fall to ensure correct functioning.

### **Closing Remarks**

On the one hand, we cannot believe that ABE learners in Massachusetts are currently taking tests tied to the Massachusetts ABE curriculum frameworks on computer. The fact that several hundred items are currently being administered by OWL throughout the state is truly amazing. On the other hand, we realize there are many obstacles still to be overcome, including winning over many ABE teachers and staff, who may be frustrated with this new system. Nevertheless, we remain optimistic and believe we are well on our way to incorporating our vision for quality, 21<sup>st</sup> century assessment of ABE students.

To many members of the public and even the educational research community, the process by which tests are developed and validated for high-stakes uses is often unfamiliar. For test developers, the creation of new tests can be a significant opportunity to involve interested parties such as teachers and students in testing activities and help advance understanding of assessment more generally. In addition, documenting and disseminating information about test development experiences is critical in terms of validity evidence and ensuring a measure of quality control. Ultimately, as testing becomes more common in adult education, test developers

and adult educators will work more closely together to ensure that ideas are exchanged and the test functions as intended. We hope the collaborations among university psychometricians and adult educators in Massachusetts extend to other parts of the United States and beyond.

## References

- 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 the educational assessment of students. *Educational Measurement: Issues and Practice*, 9(4), 29-32.
- 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., & Zenisky, A. (2003). Advances in criterion-referenced testing methods and practices. In C. R. Reynolds & R. W. Kamphaus (Eds.), *The handbook of psychological and educational assessment* [2<sup>nd</sup> ed., pp. 377-404]. New York: The Guilford Press.
- Mislevy, R. J. & Knowles, K. T. (Eds.) (2002). *Performance assessments for adult education: Exploring the measurement issues*. Washington, DC: National Academy of Sciences Press.
- O'Neil, T., Zenisky, A. L., & Sireci, S. G. (2003). Viability analysis of current computer based testing options for adult basic education in the state of Massachusetts. *Center for Educational Assessment Report No. 536*. Amherst, MA: University of Massachusetts, School of Education.
- Popham, W.J., Baker, E.L., Berliner, D.C, Yeakey, C.C., Pelligrino, J.W., Quenemoen, R.F., Roderiquez-Brown, F. V., Sandifer, P.D., Sireci, S.G., & Thurlow, M.L. (2001, October). *Building tests to support instruction and accountability: A guide for policymakers*. Commission on Instructionally Supportive Assessment. Available at [www.asa.org](http://www.asa.org), [www.naesp.org](http://www.naesp.org), [www.principals.org](http://www.principals.org), [www.nea.org](http://www.nea.org), and [www.nmsa.org](http://www.nmsa.org).
- 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.
- Zenisky, A.L., & Sireci, S.G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education*, 15, 337-362.