****

**Developing Next-Generation Assessments for Adult Education:**

**Test Administration and Delivery**

**Mike Russell, Boston College**

**Larry Condelli, American Institutes for Research**

In September 2014 the Office of Career Technical and Adult Education (OCTAE) convened a panel of assessment experts to discuss Next Generation Assessments for Adult Education. OCTAE asked the panel to discuss three topics related to development of new assessments for adult education that would help comply with the requirements for measuring skill gains under the Workforce Innovation and Opportunity Act (WIOA) and address the need for improved assessments in adult education. The topics were:

* Approaches to Assessment for Accountability in Adult Education
* Characteristics and Approaches for Next Generation Assessments
* Promoting Development of Next Generation Assessments

This paper presents a summary of the discussion on the first two topics. We briefly touch on the third topic at the conclusion of this paper. We hope to stimulate discussion and interest in the development of new types of assessment that will have greater relevance and validity to adult education programs and students.

**The Need for New Assessments**

A key component of the National Reporting System (NRS), the federal adult education program’s accountability system under the Workforce Investment Act (WIA), is educational gain, which measures improvements in adult learners’ knowledge and skills following a period of participation in an adult learning program. Educational gain is measured by a pre- and post-test administration of an approved test and is reported as the percent of learners who experience a gain in an educational functioning level (EFL), to which the test is benchmarked[[1]](#footnote-1). EFLs are defined through descriptors which provide illustrative literacy and mathematics skills that students at each level are expected to have.

While the approach to quantifying the positive impact of an adult learning program on learning has been employed for many years, recent developments created the need for significant improvements in the EFLs and assessments. These developments include the reauthorization of the federal adult education program by the Workforce Innovation and Opportunity Act (WIOA) with its requirements for states to align content standards for adult education with state-adopted standards. There also have been changes in testing technology and administration methods in recent years that create the opportunity for more efficient administration and greater validity in assessment.

In 2013, OCTAE released the College and Career Readiness (CCR) Standards for Adult Education[[2]](#footnote-2). These standards include a subset of the Common Core State Standards in English language arts/literacy and mathematics that are most appropriate for adult education. Many states are adopting these standards for their adult education programs to meet the new requirement in WIOA to align its adult education standards with the state adopted standards for k-12. In turn, OCTAE is revising the EFL descriptors for Adult Basic Education and Adult Secondary Education to reflect the CCR standards.

WIOA also requires new performance indicators for accountability, including a new indicator on measureable skill gains. OCTAE plans to use the same pre- and post-test approach to educational functioning gain to measure this indicator. Consequently, federally-funded adult education programs will need assessments to match the EFL descriptors. These changes, along with improvements in technology and test delivery, create a need for the development of a new generation of assessments in adult education.

**Technology for Test Administration and Delivery for Next-Generation NRS Assessment**

The technology of testing has advanced rapidly over the past decade. Among these advances are widespread use of computer-based test administration, embedding and using digital accessibility supports, expanding applications of adaptive testing, increasing use of automated scoring, introducing a larger variety of item types, and introducing scalable open-source test development and administration components. In addition, there is growing interest in and use of formative and diagnostic assessments to more closely link assessment with instruction. These advances provide important opportunities to enhance current approaches to assessment of adult learners. Below, these advances are described in greater detail and their potential application to adult learning assessments is explored.

*Computer-based Test Administration*

Since the turn of the century, the use of computer-based technologies to deliver assessments has expanded rapidly, particularly in the K-12 arena. A decade ago, only a small number of states had begun exploring computer-based administration. Beginning in spring 2015, nearly all states will administer a portion, if not nearly all, of their tests on-line. In addition, whereas desktop computers were the primary tool used to deliver computer-based tests a decade ago, testing programs now rely on a wide variety of desktop, laptop, and tablet-based devices for test administration.

Interest in computer-based testing is driven by at least four factors:

* The delivery of tests in a digital form reduces (or eliminates) several costs associated with printing, shipping, and scanning paper-based testing materials. While there are still substantial costs associated with computer-based test delivery, the time and effort required to manage paper-based materials is largely eliminated.
* Computer-based testing allows student responses to be scored in a more efficient and timely manner. This increased efficiency allows score information to be returned to educators more rapidly, allowing them to take action with students based on test results.
* As is described in greater detail below, computer-based testing opens up the possibility of using a wider array of test items to measure student achievement.
* Also described below, computer-based testing can provide a larger array of accessibility supports in a more standardized and individualized manner.

The adult learning population attends classes for a limited time and consequently programs have less time for assessment administration and scoring. Computer-based testing also holds the potential to streamline the delivery of tests without requiring adult learning programs to sort out which paper-based tests to administer. In addition, the increased efficiency in scoring and reporting results holds potential to allow instructors to use test results to tailor instruction based on student needs. The wide range of skill areas which adult students need to improve makes this customization especially attractive and has the potential to increase the validity of test information.

*Embedded Accessibility Supports*

It is widely accepted that many learners can benefit from a variety of accessibility supports. Moreover, the use of these supports during assessment can improve the validity of assessment for many learners. This understanding has led to important changes in the way testing programs view accommodations. Whereas accommodations were once believed to be reserved for those students with defined disabilities and special needs, the modern approach allows a wide array of accessibility supports to be available to all test-takers. As an example, many learners, particularly adults, can benefit from having text-based content presented in a larger size. Whereas the older accommodation-based approach would limit the use of a large-print version of a test to those with low-vision, accessibility policies now allow any user to modify the display of text by either increasing font-size or magnifying the testing environment.

Given the challenges many adult learners have previously experienced accessing the curriculum and/or assessment content, such as low vision ability, the use of embedded accessibility supports holds potential to improve access during assessment and, in turn, increase the validity of test information.

*Adaptive Testing*

Traditional fixed-form tests are often designed to provide a general measure that spans a wide spectrum of achievement levels. To do so, a fixed form test is often composed of items that range widely in difficulty, with some items being relatively easy, some very difficult, and most being of moderate difficulty. While a fixed-form test generally does a good job separating test-takers by their ability, it does not provide detailed information about performance at any one level. It is only after a test-taker completes the full set of items that comprises a fixed form that his/her achievement level is estimated.

In contrast, an adaptive test begins with developing a large pool of items that vary in difficulty. For each level of difficulty, several items are developed. Next, small samples of items that range in difficulty are initially administered to the test-taker. The initial set of items is used to develop a preliminary estimate of achievement. This initial achievement estimate is then used to inform the selection of the next item or set of items. If the initial achievement estimate is high, more difficult items are selected and administered. If the estimate is low, easier items are selected. Then, after each new item or set of items is answered, the estimate of achievement is refined and the process of item selection is repeated. This process continues until a stable estimate of the test-taker’s achievement is obtained. As an analogy, adaptive testing works like an efficient algorithm used for the children’s game of “Guess a number between 1 and 100.”

Adaptive testing has increased in popularity because it is generally a more efficient approach to measuring achievement that typically results in a shorter test that takes less time than a fixed form equivalent, usually with a more accurate and reliable test score. Adaptive testing also can be tailored to meet a variety of conditions, including administering only those items that are deemed accessible for a specific sub-population of students, or tailoring to provide diagnostic information about student understanding.

The wide variation in prior achievement among adult students requires a broad range of items that vary by difficulty and content. Considering the limits in time available for testing in adult learning programs, coupled with this variation in prior achievement, adaptive testing holds the potential to decrease testing time while providing more accurate information for learners across the achievement spectrum. In addition, adaptive testing can be an efficient placement tool for determining the appropriate starting level of measurement for learners new to a program. In effect, just as an adaptive test typically uses a small set of items to develop an initial achievement estimate, an adult learning program could use a small set of items to determine the educational functioning level (EFL) at which to place students and target for further assessment.

However, there are two challenges to implementing adaptive testing: a) the need to develop a large item pool with an adequate number of items at each achievement level; and b) the need to administer tests on computers using an adaptive test delivery system. Developing a large pool of items can be an expensive and time-consuming process. Similarly, developing an adaptive delivery system that meets a given program’s needs requires considerable technical expertise.

*Automated Scoring*

Efforts to score written responses began in the 1960s and have matured considerably in the last decade. Today, several testing programs use automated scoring methods (sometimes termed “artificial intelligence” or “AI”) to score both short-response and essay-type items. While some observers have raised concerns about the accuracy with which a computer can score a written response, research generally shows that automated scoring engines are equally, if not more, accurate and reliable than pairs of human scorers.

Automated scoring provides a testing program with at least two advantages. First, the scoring is performed very efficiently and, in some cases, can be performed immediately after a student completes a response. In contrast, human scoring of responses typically takes several weeks to complete. Second, once trained, an automated scoring engine experiences no variation in the consistency with which it scores responses. In contrast, human scorers often experience drift in their scoring, sometimes becoming more lenient over time and sometimes more difficult. Similarly, scoring at different periods of time always yields consistent scores with automated scoring whereas the use of different human scorers at different times can produce inconsistencies.

Automated scoring, however, does require some effort to prepare. Specifically, a set of responses must be scored by humans and then used to “train” or calibrate the scoring engine. Once calibrated, a second set of responses is needed to verify the calibration was accurate. Once calibrated, however, an automated scoring engine typically requires no further refinement.

A second challenge results from a small set of responses that may not be interpretable or do not align with any calibration responses. These responses are typically flagged and must be scored by human readers. Thus, for both calibration purposes and for responses that cannot be interpreted, automated scoring still requires some amount of human scoring.

For adult learning programs, automated scoring has distinct advantages. An automated scoring engine allows a test to use open-response items without increasing the burden on instructors to score responses or requiring the use of human scoring centers. Automated scoring increases the speed with which test scores can be calculated and returned, which is particularly important if test results are to be used for instructional purposes. Use of automated scoring also increases the objectivity and the reliability of scores over time.

Like adaptive testing, introduction of automated scoring would require the development or use of an automated scoring engine. In addition, calibration of the engine would require collection and human scoring of samples responses.

*New Item Types*

In the past five years, there has been growing interest in the use of new item types that capitalize on digital technologies. Technology-enabled items employ multimedia as part of the stimulus and/or response options. For example, rather than presenting a passage from a speech that is followed by an item that focuses on the content of the speech, the assessment would present either an audio recording or a video recording of the speech.

Technology-enhanced items allow students to produce responses to items in ways other than selecting from a set of response options (i.e., multiple-choice) or by producing text. As an example, to demonstrate understanding of the order of events, a test-taker might arrange a set of events in the order in which they occurred. Or, to demonstrate understanding of graphing linear functions, a test-taker might be presented with a coordinate plane and asked to create a graphical representation of a given function by drawing a line on that plane. Responses to technology enhanced items can then be scored automatically.

Technology-enabled and technology-enhanced items are believed to increase student engagement with a test and to produce situations that are more authentic. They allow students to produce evidence of understanding in ways that are more directly aligned with the knowledge and skills being measured. Many adult learners have not had positive experiences taking tests in the past. Use of new item types has the potential to increase their engagement with the test and create a sense that they are better able to demonstrate their knowledge and skills. New item types also may allow programs to measure skills and knowledge measured poorly or missed by multiple-choice tests.

*Open-Source Delivery Platform*

A major impediment to computer-based assessments is the need for a platform to deliver tests. In the K-12 market, state assessment programs have traditionally contracted with test vendors to provide assessment services, including the provision and use of computer-based test delivery systems. This approach often results in the use of high-quality test delivery systems that are the responsibility of the vendor to support and maintain to assure proper operation. However, this approach often results in challenges customizing a platform to meet a program’s specific needs and is believed to increase costs due to the price of using a vendor’s proprietary system.

Over the past four years, the Race to the Top Assessment program has spurred interest in the development and use of open-source test delivery systems. While there are a variety of licensing agreements that may accompany open-source technologies, the core idea with open-source code is that it can be used freely and/or extended to meet the needs of end-users. To date, the Smarter Balanced Assessment Consortium (SBAC) has worked with the private sector to create a test delivery system that is open-sourced. In addition, both the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the National Center and State Collaborative (NCSC) have committed to extending the Open Assessment Technology TAO platform. TAO is an open-source platform that has more than a decade-long track record of development and use, and which is supported by a vibrant open-source development community.[[3]](#footnote-3)

Concerned about creating a dependency on a proprietary system and recognizing the unique assessment needs for the adult learning community, an open-source platform may provide the adult learning assessment programs with several benefits. They include, for example, greater control over the functionality of the system, the accessibility supports provided by the system, the variety of item types supported by the system, and the scoring and reporting features of the system. One challenge, however, to the use of open-source technologies is the need to host, maintain, support, and enhance open-source software. While the open-source community generally shares in maintenance and enhancement of open-source software, participation in that community would require a commitment of at least one person to be actively involved. In addition, resources would need to be available to support on-going contributions to code maintenance and development. While these on-going costs are expected to be lower than those associated with licensing and customizing proprietary technologies, it is important for a program to plan and budget accordingly.

*Summary*

The adult education program has measured educational gain to meet accountability requirements through pre- and post-testing with fixed form, paper-based, off-the-shelf tests for many years. During this time, several advances in the technology of testing have occurred. Other changes, including the passage of WIOA with its requirement for the adoption of adult education content standards aligned to state adopted k-12 standards and the development of new EFL descriptors on which assessments have been based, have created a need for a new generation of assessments for adult education. Should adult learning assessments be upgraded, careful consideration should be given to how to incorporate and benefit from recent advances in assessment technology.

Adoption of adaptive computer-based testing that incorporates embedded accessibility supports and employs a wider variety of item types may improve the validity of measures for many adult learners. It also will provide more detailed information about achievement across a broader spectrum of knowledge and skills. In addition, should there be a desire to expand assessment to writing or to incorporate open-response items, the use of automated scoring may increase the speed with which results are returned and increase the stability of measures. Finally, the use of open-source technologies may provide the program with greater control over the features of its assessment program and allow for reporting that provides timely information that can be used to inform instruction. In short, advances in the technology of testing open a variety of new opportunities for high-quality and instructionally relevant assessment of adult learners.

**Next Steps: Promoting Development of Next Generation Assessments**

While it is clear that there is a need and opportunity for a new generation of assessments for adult education, the field faces several significant challenges to developing and implementing them.

* The number of adult education programs and participants is relatively small compared to the number of K-12 students and schools.
* Funding for adult education programs to purchase tests is small compared to K-12 educational programs and to other testing programs (such as college admissions and certification testing programs).
* Current policies that allow states to independently select tests from an approved list to use to evaluate their effectiveness segments a small market even further.

Together, these factors provide little incentive for test developers to produce tests that are specifically designed for adult education programs or to improve their current tests. The lack of financial incentive results in tests that may be less sensitive to achievement across the full spectrum of adult learners and may not be well aligned with the standards and content that is addressed by adult education programs. In addition, the use of different tests across programs makes it difficult to directly compare outcomes among programs or to aggregate impact across programs.

Another challenge for adaptive delivery systems is the need to administer tests on computers. Many adult education programs lack computers and the space and infrastructure to administer computer-based tests on a large scale. The field’s challenge is to support development of next generation assessments in this environment. Some issues to consider include:

* How can we learn from other approaches to assessment development that might help (e.g., large scale K-12 assessments)?
* What sources of funding are available?
* Are there ways OCTAE and the states can support development of new assessments?
* What resources and other incentives can states or OCTAE provide to local programs to promote development of modern, computer-based assessment environments?
* What other obstacles to implementing computer adaptive assessments exist at the local level? How can we resolve them?
* How can we make the results of assessments most useful to inform instruction and accountability for local staff programs?
* What provisions can be made for learners lacking basic computer skills?

The success of developing high quality assessment systems depends on how the field can resolve these issues.

1. For further information about the NRS and adult education accountability requirements see http://www.nrsweb.org. [↑](#footnote-ref-1)
2. Pimentel, S. (2013). College and Career Readiness Standards for Adult Education. Office of Career, Technical and Adult Education, U.S. Department of Education. http://lincs.ed.gov/publications/pdf/CCRStandardsAdultEd.pdf [↑](#footnote-ref-2)
3. See <http://www.taotesting.com/> for more information. [↑](#footnote-ref-3)