Using Technology for Assessment in Adult Learning

John P. Sabatini

The topics of assessment and testing make most learners and teachers uncomfortable. The culture of testing in educational settings typically involves routinized/dull standardized procedures, security surrounding test materials, suspicion of examinees' motives and actions, secretive and occult treatment and reporting of results, and critical/punitive consequences. Time-consuming, high-stakes tests such as the Scholastic Aptitude Test (SAT) and the General Education Development (GED) embody all these aspects and do little to improve the public relations of testing. No wonder test anxiety is an important psychological variable considered in test data analysis. At best, tests are recognized as necessary evils for gathering information for student placement, diagnosis, achievement, or program evaluation purposes. As often, they are criticized as a waste of precious instructional time.

The combination of technology and assessment, on the other hand, creates a different culture of testing. Individuals of all types enjoy self-tests of their trivia knowledge or competing for the fastest time or highest score in a simulated performance such as driving a race car or shooting bandits in a saloon. The military uses the motivational benefits of computerized testing and simulation with success throughout their training operations. In fact, the success of video game-style testing suggests a new assessment paradigm: delivery of the GED exam in short segments as part of arcade games in taverns. In addition, if individuals were charged 50 cents every minute to continue to play to improve their score, literacy levels nationwide might begin to climb in a matter of months.

Perhaps I have engaged in too much hyperbole. Still, the promise of technology applied to assessment is profound. From computerized adaptive testing (Wainer 1990) to intelligent measurement (Lesgold, Ivill-Friel, and Bonar 1989), there are new and boundless possibilities for making assessment both tailored to the individual
and more integral to the educational process. In many cases, the
distinction between assessment and instruction is being erased
(Venezky and Sabatini, forthcoming). Even technically simple
computerized tests represent paradigm shifts when compared to the
culture of testing that has grown out of paper and pencil technol-
gy. Combined with constructivist and other learner-centered
theories of instruction, the new technology-based assessment mod-
els can become an integral part of instruction.

This chapter consists of two major sections. In the first, the use of
computerized assessments in the SARA study is described. SARA
can be considered a case study of how technology is changing the
culture of testing such that more targeted information about spe-
cific learners can be obtained and used in instructional settings. In
the second section, a brief review of computerized adaptive testing
is used to introduce the idea of a learner-centered assessment
system.

SARA: An Assessment and Technology Case Study

The Study of Adult Reading Acquisition (SARA) was a longitudi-
nal research project designed to measure the development of read-
ing subskills of adults enrolled in instructional programs. The study
participants were enrolled in a variety of adult education sites and
settings including correctional institutions, community-based pro-
grams, adult basic education (ABE) classes, general educational
development (GED) classes, public assistance job programs, and
tutoring services. The study design serves as a model for using
achievement tests in conjunction with cognitive assessments as part
of a comprehensive assessment strategy. As in traditional as-
sessment programs, achievement tests were administered at inter-
vals of instruction appropriate for observing reliable change. In
addition, SARA added cognitive assessments that were given at
more closely spaced intervals to monitor stability or change in
underlying skills or to decide whether additional diagnostic mea-
sures would be helpful.

Participants were monitored over a 6-7 month period while they
were enrolled in instructional programs. Over the course of the
study participants received a variety of tests over a number of
sessions. Each participant received an initial battery of baseline
tests, then a battery of repeated measures every 5-7 weeks. Follow-
up achievement and diagnostic tests were also administered. Par-
ticipants were also screened for vision and hearing problems and
were interviewed about their reading and leisure habits and educational backgrounds.

Overall SARA Assessment Strategy

A review of existing published paper and pencil achievement tests showed that, although able to rank adults on a single reading ability scale, most fail to capture the complexity of adult learners' literacy development and do not provide instructors with diagnostic information sufficient for making instructional decisions (Sabatini, Venezky, and Bristow 1995; Venezky, Bristow, and Sabatini 1994; Venezky, Sabatini, Brooks, and Carino 1997). Although normed for adult populations, most published tests for adult learners are linked to elementary and secondary school developmental scales such as grade level.

Because the assumptions used to support the validity of published test series at elementary and secondary levels are not necessarily valid for adult learners, the assessment strategy used in the SARA study combined both reading achievement tests and cognitive assessments in instructional settings. The cognitive assessments were computerized, short in duration, and targeted. The total time adults can devote to full-time instruction and learning is constrained and often interrupted by external demands, suggesting the need for finer grain monitoring of skill consolidation, maintenance, and retention than is possible with global achievement tests.

The published achievement tests were administered in a traditional manner; however, the SARA project computerized many of the cognitive assessments given to adult learners.

SARA and the Advantages of Computerized Assessment

Cognitive assessments such as the ones used in SARA are measures that are based upon models of the cognitive processing skills or components critical to skilled reading. The battery of tests in the SARA project were used to monitor the levels of three cognitive processing components central to reading acquisition: decoding, word recognition, and sentence processing. The tests included both accuracy and rate measures so that both overall reading performance abilities and efficiency of skill application could be assessed.

Although it is not necessary to computerize such assessment tasks and tests, the computerized administration of these tests in the SARA project and for future uses highlight the advantages of
computerized testing and assessment in adult education. These advantages include—

- Test administration, which can be standardized with technology. Programmed tutorials with audio assistance like the ones used in SARA lead students through instructions, demonstration items, practice items, and test items.

- Learners have control of the test situation. They are active participants, not passive test takers.

- Learners can work at their own pace, yet performance rate information can be collected and analyzed. The information is crucial for assessing fluency and automaticity.

- Learners gain a sense of privacy. They are given a private space to work, with headsets to listen to instructions. Mechanical responses can be collected as keystrokes on the computer, and verbal responses through a microphone on a headset can be recorded for later scoring.

- Most data are collected electronically and can be scored automatically by analysis programs. Feedback can also be immediate, though in the SARA research project analyses were performed later.

- Finally, it saves time. In the SARA project each component skill test took from 5-15 minutes including instructions and tutorials. A computerized testing session including a battery of three or four component skills tasks took about 20 minutes.

**A Glimpse into the Future**

The SARA project had many findings but none more important than these: By using technology to computerize testing, it is possible to reduce the anxiety and boredom inherent in traditional testing settings, give the learner more control, and collect better information.

SARA demonstrates that computerized testing offers a number of advantages over traditional testing models. Computerized Adaptive Testing (CAT) offers even more potential advantages, if its power can be harnessed for the learner, not simply to make more efficient the large-scale data collection efforts of agencies. Wainer (1990) describes adaptive testing as follows:
The basic notion of an adaptive test is to mimic automatically what a wise examiner would do. Specifically, if an examiner asked a question that turned out to be too difficult for the examinee, the next question asked would be considerably easier... An adaptive test first asks a question in the middle of the prospective ability range. If it is answered correctly, the next question asked is more difficult. If it is incorrectly answered, the next one is easier. This continues until we have established the examinee's proficiency to within some predetermined level of accuracy. (p. 10)

Noncomputerized adaptive tests include Binet's ability tests. However, the cost of trained administrators as well as the inconsistency across examiners make the widespread use of such tests problematic (Hulin, Drasgow, and Parsons 1983). A number of mass administered versions of adaptive tests have been developed including the two-stage, flexi-level, and branching tests (see Lord, 1980 and Weiss 1983). However, paper and pencil adaptive tests pose a number of technical problems, mostly associated with standardizing their administration. With the development of powerful microcomputers, work on noncomputerized, adaptive tests has become all but obsolete.

Computerized adaptive tests in development are more likely to be used to maximize efficiency in assessing group differences by summing over or correcting for individual differences, rather than applying the new technology to the benefit of the individual. Following is a description of how a learner-centered assessment system might operate and how CAT could be used to facilitate its adoption.

**Learner-Centered Computer Assessment Systems**

A learner-centered computer assessment system will maximize the control, relevance, and value of the assessment information for the examinee or learner (and the instructor when applicable), as well as provide valid information for higher-level users (school administrators, employers, local and state policy makers, public officials, the media, and general public). In the past, psychometric theory and research has been disproportionately concerned with providing information for use by these higher-level users, who in turn make inferences and decisions about and for individuals or programs. The psychometric models used maximize efficiency in assessing group differences by summing over or correcting for individual differences. This practice continues today, even though new
technologies have been developed that could be applied to the benefit of the individual.

Even though test packages offer various score scales and individual profile analysis forms in an attempt to link performance levels to instruction, the assumption is that tests must first be highly valid and reliable on a large scale before they can be useful in the classroom. This implies standardized administration, sufficient items to obtain reliable individual scores, extensive investment in item development and consequently security to protect the pool of items or questions.

**How CAT Changes Test Administration and Test Setting**

Following are some attributes of computerized testing that have the potential to change the culture of testing for the learner. CAT provides test administration and test setting with—

C **Truly Adaptive Tests.** A computerized adaptive test begins with an estimation of the examinee’s ability on the trait being measured. If no prior information is known about the examinee, an average ability is assumed. Note that a developmental profile of the person could be stored in the system, to be used to establish an initial ability estimation in the next testing session. Next, items are selected according to an algorithm such that the item difficulty matches the estimated ability of the examinee at that point. After the response to the item is scored (correct, incorrect, or partially correct), a new proficiency level is estimated and a new item chosen from the item pool. The test ends when the target level of precision in estimating the examinee’s ability is reached. Several researchers have studied and tested the benefits and problems associated with different algorithms for item selection and stopping rules (Hulin, Drasgow and Parsons 1983; Samejima 1983).

C **Test Length and Reliability.** The research suggests that in many cases, adaptive tests require fewer items to estimate the ability of an examinee at the same reliability as longer paper and pencil tests. Theoretically, by administering items that are focused on the appropriate range of difficulty for an examinee, approximately half the items as a conventional paper and pencil test are needed to achieve approximately the same or better reliability (Lord 1980). This has been confirmed for
verbal ability items in a military aptitude test, although improvements in

statistical validity shows a less clear-cut advantage (McBride and Martin 1983), in certification exams (Sykes 1991), in achievement tests (Olsen 1990), and in mastery tests (Kingsbury and Weiss 1983). Increased reliability may even be expected because there is less fatigue and more challenge in taking items of the appropriate difficulty.

C Motivation Issues. Shorter tests, with more items aimed at a difficulty level one can handle, are assumed to improve motivation to perform well on the test. Although several authors have suggested this result (Stanfel 1996; Wainer 1990), few empirical studies have been done to corroborate the assumption. These same authors, however, have also pointed out that shorter tests may be perceived by some as unfair in not providing them a sufficient opportunity to demonstrate their true ability.

Other artifactual features of the test setting that may have worked to the detriment of standardized test administration in the past can be and are being altered. The examinee has less probability of making systematic errors in selecting answers on the computer than in marking a score sheet. Most examinees will take approximately the same length of time to complete a test, eliminating perceptions of ability relative to others who work faster or slower. Still, there will also be more time for an individual to work through the items on a test. More privacy can be provided by placing computer terminals in cubicles. Finally, the examinee can trust that human scoring errors are even less likely.

C Current Applications. Wainer (1990) uses the metaphor of the Gedanken CAT (GCAT) to describe the test administration site of the future. In it two examinees of different abilities have generally positive experiences with computerized tests that are tailored to their individual differences. Security issues are enhanced, since the setting is a specially designed test site. All examinees are monitored electronically by a benevolent proctor, whose most pressing interest is to detect aberrant answer patterns that may reveal a problem the individual is having with a particular set of test items, although, of course, cheating patterns are also detectable.
Research Findings. Research continues to validate the assumptions that adaptive tests will be quicker to administer, more reliable, and more motivating to the examinee. However, much more needs to be done. The benefits to the individual of a kinder, gentler test setting are considered more as side effects of adaptive testing than central issues of concern. Much more effort has been put into comparing the validity of paper and pencil versus CAT tests than has been put into understanding the reactions of the examinees. (Six of 19 recent empirical studies found in an ERIC database search concerned this latter issue, whereas only two mention the former.)

Application to an Examinee-Centered CAT System. Two issues that are theoretically well-developed but practically ignored by test constructors in applications of CAT systems are feedback and examinee control. Various features of item response theory technology allow the test constructor to detect aberrant response patterns, including special knowledge, cheating, or cultural biases. There are also a variety of opportunities for examinees to tailor the test to fit their own learning needs, as well as providing the best estimate of their abilities, if the learner were provided feedback and permitted limited control. These are issues awaiting future research.

One application to note is a demonstration project in the Waterford School in Provo, Utah (Miller 1985). Waterford integrated CAT systems within a computerized instructional management system, which enables instructors and students to regulate learning development in curriculum areas represented. More ongoing assessment and student-teacher interaction were made available. The report suggests that achievement scores measured in traditional paper and pencil methods also rose after the system was implemented.

The Future of CAT

Sandals (1983) describes two further generations of computer based assessment strategies beyond computerized testing (CT) and computerized-adaptive testing (CAT): continuous measurement (CM) and intelligent measurement (IM). A CM system keeps track of the information collected on an individual, creating a developmental profile. The CM systems he cites are also linked to specific curricula or instructional programs, usually in a mastery learning
Lesgold, Ivill-Friel, and Bonar (1989) describe an intelligent testing system consisting of four components: instruction and test planning knowledge, curriculum knowledge, treatment knowledge, and domain expertise. Their system diagnoses the level of knowledge of the learner through an assessment strategy, implements an instructional sequence, and retests the learner to update its model of the learner's knowledge. Experimentation with these systems are ongoing, and practical applications are not far behind. As learners continue to seek more control of their own learning, such systems will find their place in the learning environments of the future.
USING TECHNOLOGY FOR ASSESSMENT