Using Web-Based Testing For Large-Scale Assessment

By
Laura S. Hamilton
Stephen P. Klein
and
William Lorié

RAND
EDUCATION

Building on more than 25 years of research and evaluation work, RAND Education has as its mission the improvement of educational policy and practice in formal and informal settings from early childhood on.

Acknowledgments

This work was supported by the National Science Foundation, Division of Elementary, Secondary, and Informal Education, under grant ESI-9813981. We are grateful to Tom Glennan for suggestions that improved this paper.

Table of Contents

I. INTRODUCTION

Efforts to improve the quality of education in the United States increasingly emphasize the need for high-stakes achievement testing. The availability of valid, reliable, and cost-effective measures of achievement is critical to the success of many reform efforts, including those that seek to motivate and reward school personnel and students. Accurate measurement of student achievement is also important for initiatives, such as vouchers and charter schools, that require publicly available information about the academic performance of schools. This paper begins with a brief discussion of the context surrounding large-scale (e.g., statewide) achievement testing. We then describe a new approach to assessment that we believe holds promise for reshaping the way achievement is measured. This approach uses tests that are delivered to students over the Internet and are tailored ("adapted") to each student's own level of proficiency.

We anticipate that this paper will be of interest to policymakers, educators, and test developers who are charged with improving the measurement of student achievement. We are not advocating the wholesale replacement of all current paper-and-pencil measures with web-based testing. However, we believe that current trends toward greater use of high-stakes tests and the increasing presence of technology in the classroom will lead assessment in this direction. Indeed, systems similar to those we describe in this report are already operational in several U.S. school districts and in other countries. Furthermore, we believe that although web-based testing holds promise for improving the way achievement is measured, a number of factors may limit its usefulness or potentially

lead to undesirable outcomes. It is therefore imperative that the benefits and limitations of this form of testing be explored and the potential consequences be understood. The purpose of this paper is to stimulate discussion and research that will address the many issues raised by a shift toward web-based testing.

After presenting a brief background on large-scale testing, we describe the new technology of testing and illustrate it with an example. We then discuss a set of issues that need to be investigated. Our list is not exhaustive, and we do not provide answers to the many questions we raise. Instead, we hope that this discussion reveals the critical need for cross-disciplinary research to enhance the likelihood that the coming shift to an emphasis on web-based testing will truly benefit students.

II. The Context of Large-Scale Testing

Testing is closely linked with the current emphasis on standards-based reform and accountability. Nearly every state in the United States has adopted academic standards in four core subjects—English, mathematics, science, and social studies. Most states assess achievement toward these standards in at least reading and mathematics at one or more grade levels. The grade levels at which tests are administered and the number of subjects tested continue to increase. Some states also are exploring the use of open-ended test items (e.g., essays) rather than relying solely on the less-expensive multiple-choice format. As a result of these trends, the cost of assess-

Testing is closely linked with the current emphasis on standards-based reform and accountability.

ment for states has doubled from approximately $165 million in 1996 to $330 million in 2000, in constant dollars (Achieve, Inc., 1999). California alone tested over four million students in 1999 using the commercially available Stanford 9 tests.

How Tests Are Used

State and district test results are used to make decisions about schools and teachers. For example, in recent years, state policymakers have instituted high-stakes accountability systems for schools and districts by tying various rewards and sanctions (e.g., extra funds for the school or reassignment of staff) to student achievement. These accountability systems typically involve disseminating results to the public. This puts pressure on lagging schools to improve their performance. In 1999, 36 states issued school-level "report cards" to the public (Achieve, Inc., 1999).

Test scores are also used to make important decisions about individual students. Several states (including New York, California, and Massachusetts) are developing high school exit examinations that students must pass to earn diplomas. Many of the nation's large school districts have adopted policies that tie promotion from one grade to the next to performance on district or state tests. The use of test scores for tracking, promotion, and graduation is on the rise and suggests that the need for valid and reliable data on individual student achievement will continue to grow (National Research Council, 1998).

The use of test scores for tracking, promotion, and graduation is on the rise and suggests that the need for valid and reliable data on individual student achievement will continue to grow.

Teachers also develop and adopt tests that they use for instructional feedback and to assign grades. In this paper, we focus on externally mandated, large-scale tests rather than these classroom assessments, though both forms of measurement have a significant impact on students.

How Testing Is Done

Many large-scale testing programs purchase their tests from commercial publishers. The three largest companies are Harcourt Educational Measurement, which publishes the Stanford Achievement Tests; CTB/McGraw-Hill, which publishes the Terra Nova; and Riverside Publishing, which publishes the Iowa Tests of Basic Skills (ITBS). In some cases, these publishers have adapted their materials to accommodate the needs of particular testing programs, such as by adding items that are aligned with a state's content standards. Typically, however, states and districts purchase these "off-the-shelf" materials and use them as is, even when it is evident that their curricula are not aligned especially well with the test's content.

The other main approach to large-scale testing is the use of measures developed by states or districts. For example, several states have developed tests that are designed to reflect their own content and performance standards, and others have plans to do this in the future. Some of the state-developed tests include constructed-response or performance-based items that are intended to measure important aspects of curriculum standards that are difficult to assess well with multiple-choice tests. Although there are some commercially available constructed-response tests, most of the states that use this type of item have

developed their own measures, typically with the help of an outside contractor.

Despite the diversity of tests administered by states and districts, most large-scale testing programs share several common features. All state programs and most district programs currently rely heavily on paper-and-pencil exams, although a few districts also use some hands-on measures, such as in science. Almost all emphasize multiple-choice items, and many rely solely on this format. Tests are typically administered once per year, in the spring, with results generally released in early to late summer. Finally, many programs stagger subjects across grade levels; e.g., math and social studies in grades 4 and 7, reading and science in grades 5 and 8. However, a few states, such as California, test every student every year in almost every core subject, and the general trend is toward increasing the number of grade levels tested. The amount of time and resources devoted to testing is often a point of friction between those who want to measure student progress and those who are concerned about taking class time away from instruction for testing (and for preparing students to take the tests).

There are several limitations to the current approach to large-scale assessment. First, the reliance on paper-and-pencil multiple-choice tests limits the kinds of skills that can be measured. For this reason, many states and districts have experimented with other formats, such as hands-on testing, but these can be very expensive (Stecher & Klein, 1997) and do not necessarily measure the constructs that their developers intended (Hamilton, Nussbaum, & Snow, 1997). A second limitation aris-

. . . the reliance on paper-and-pencil multiple-choice tests limits the kinds of skills that can be measured.

es from the lag time between test administration and score reporting. Answer sheets typically must be sent to an outside vendor for scoring, and the test results must then be linked with school and district data systems. This process takes even longer if items cannot be machine-scored, such as when open-response questions are used. Consequently, students, parents, and teachers generally do not receive scores from tests administered in the spring until the summer or fall, which severely limits the usefulness of the results for guiding instruction. A third problem is the distinct possibility of security breaches undermining the validity of the test when the same questions are repeated across years (Linn, Graue, & Sanders, 1990; Koretz & Barron, 1998). Security also can be a problem when the results are used to make high-stakes decisions for teachers and schools, regardless of whether questions are changed each year (Linn, 2000).

Other limitations to the typical approach to testing relate to the integration between assessment and instruction. Statewide tests are typically administered apart from the regular curriculum, so students may perceive these tests as disconnected from their everyday school experiences. Consequently, they may not be sufficiently motivated to perform their best, which in turn compromises the validity of results. In addition, separating classroom instruction from assessment leads to the perception that students spend too much time taking tests, even if the amount of time devoted to testing is minimal. This occurs in part because the time spent testing is often considered as time that is taken away from instruction. A related concern is the narrowing of curriculum that often occurs as a result of high-stakes testing. There is a tendency for teachers to focus on the topics that are

tested and to neglect those that are not (Kellaghan & Madaus, 1991; Madaus, 1988; Stecher et al., 1998). Thus, although the purpose of the test may be to monitor progress of schools, the test is likely to have a significant influence on instruction if there are high stakes attached to the scores. Finally, the need to develop or adopt assessments that are aligned with state and local standards means that existing tests may not be suitable for many schools. This creates problems for generalizing results from one jurisdiction to another.

Although the traditional paper-and-pencil standardized multiple-choice test continues to be the norm, a few districts have recently experimented with computer-based testing. Advances in psychometrics and information technology are likely to accelerate the adoption of this approach, particularly when the tests are administered (or downloaded into the school) via the Internet (Klein & Hamilton, 1999). We believe that this form of testing may address many of the problems discussed above, though not all. In the next section, we describe this approach and discuss some of its advantages.

III. A New Technology of Testing

The role of information technology in virtually every type of educational enterprise is growing rapidly. Educational assessment is no exception. Several well-known tests are now administered via the computer, including the Graduate Record Exam (GRE), the Graduate Management Admissions Test (GMAT), and the Medical Licensing Examination. The high speed and large storage capacities of today's computers, coupled with their rapidly shrinking

The high speed and large storage capacities of today's computers, coupled with their rapidly shrinking costs, make computerized testing a promising alternative to traditional paper-and-pencil measures.

costs, make computerized testing a promising alternative to traditional paper-and-pencil measures. Although computers are now used widely for large-scale, high-stakes admissions and licensing exams, their use in the K–12 market is still quite limited. In addition, most existing computerized assessments for K–12 students are administered locally on a stand-alone workstation rather than over the Internet. However, we expect this will change within a few years.

The next sections of this paper discuss some relevant technical issues. This is followed by a description of the kind of system we envision and a scenario for how it might be implemented. We then discuss the advantages of this approach. Later we address the issues and concerns that are likely to arise as we move toward this new form of testing.

As mentioned earlier, we are primarily concerned with the type of large-scale testing that is conducted at the district and state levels, rather than with the tests that teachers use for instructional feedback. However, the effects of large-scale tests on instruction must be considered because we know that high-stakes assessment influences what happens in the classroom. Furthermore, information technology may offer opportunities to create a closer link between large-scale assessment and instruction, so it is worth considering these tests' instructional effects as well. The computer and the Internet obviously offer promising new approaches for teacher-made tests, but this is beyond the scope of this paper.

The computerized testing approach we discuss below has three main features. First, items are administered adaptively. Second, the system makes use of several different types of questions, including selected-response (e.g., multiple-choice) and constructed-response items (e.g., short-answer or fill-in-the-blank questions). Third, the assessment is administered via the Internet rather than relying solely on stand-alone workstations.

Adaptive Versus Linear Administration

Most paper-and-pencil tests present items in a *linear* fashion—that is, items are administered sequentially in a predefined order, and all students are asked the same questions within a given "form" (version) of the test. Students may skip items and go back, but the order of presentation is constant. Some computerized tests are also linear. However, technology provides the opportunity to allow examinee responses to influence the difficulty of the questions the student is asked. This is known as *adaptive* testing. In this type of testing, the examinee responds to an item (or set of items). If the examinee does well on the item(s), then the examinee is asked more-difficult items. If the examinee does not answer the item(s) correctly, the examinee is asked easier items. This process continues until the examinee's performance level is determined. Because information about the difficulty of each item is stored in the computer, the examinee's "score" is affected by the difficulty of the items the examinee is able to answer correctly.

Computers permit this type of interactive testing to be conducted in a rapid and sophisticated manner. Because items can be scored automatically and imme-

diately, the computer can select the next item almost instantly after an examinee responds to the previous one. Current computerized adaptive testing systems, or CATs, use item response theory (IRT) to estimate examinee proficiency and determine item selection. The length of a CAT may be specified in advance or it may be based on the examinee's responses. With the latter type of CAT, the computer stops administering items once the examinee's proficiency has been estimated to some prespecified degree of precision.

Item Format

Currently, most CAT systems rely on multiple-choice items; i.e., questions in which the examinee selects one choice from among four or five alternatives. These *selected-response* items are commonly used in large-scale testing because the answers to them can be machine-scored, which minimizes costs. Computers also can accommodate selected-response items that vary from the standard four- or five-option multiple-choice item. For example, examinees might be asked to select a subset of choices from a list.

Many large-scale testing programs that traditionally have relied on selected-response items are now exploring the use of items that require examinees to generate their own answers—these are called *constructed-response* items. Computers can accommodate a wide variety of such items. For example, students may be asked to move or organize objects on the screen (e.g., on a history test, put events in the order in which they occurred). Other tests may involve students using the comput-

Many large-scale testing programs . . . are now exploring the use of items that require examinees to generate their own answers.

er to prepare essay answers or other products. Some constructed-response items may be machine scored, particularly if the responses are brief and straightforward (e.g., a numerical response to a math problem). Researchers are exploring the use of computerized essay scoring, and it is likely that future generations of examinees will take tests that involve automatic scoring of extended responses. Currently, however, most constructed responses must be scored by trained readers.

Possible Scenario for Web-Based Test Administration

Clearly, computers offer the possibility of radically changing the way students take tests. There are many ways this could happen. We discuss below one scenario for computerized testing, involving the delivery of assessment items adaptively and the collection of student response data over the Internet. It is important to recognize that the adoption of this or other computer-based assessment systems is likely to be gradual: It will evolve over a period of time during which old and new approaches are used simultaneously. Thus, there will be efforts to ensure the comparability of results from web-based and paper-and-pencil systems. We return to this problem in a later section.

In the scenario we envision, a large set of test items is maintained on a central server. This "item bank" contains thousands of questions per subject area, covering a wide range of topics and difficulty levels. For example, a math item bank would include items covering numbers and operations, algebra, geometry, statistics, and other topics. Within each area, questions

range from extremely easy to very difficult. The questions also are drawn from a wide range of grade levels.

On the day of testing, the entire item bank (or a large portion of it) for the subject being tested (e.g., science) is downloaded to a school from the central server. Students take the test in their classrooms or in the school's computer lab. The items are administered adaptively, so each response leads to a revised estimate of the student's proficiency and a decision either to stop testing or to administer an additional item that is harder or easier than the previous one. The student's final score is computed almost instantly and is uploaded to a centralized data file. Scores may be given to students that same day so they know how well they did. Students complete the testing program several times a year rather than the "spring only" approach that is typical of current statewide testing programs. Each student has a unique identifier, so students' progress can be monitored even if they change classrooms or schools.

The scores can be used to inform several decisions. Policymakers and staff in district or state offices of education may use the results to monitor achievement across schools and provide rewards or interventions. Results also may provide evidence regarding the effectiveness of various educational programs and curriculum materials.

Most large-scale assessments are used for external monitoring and accountability purposes. They are rarely if ever used for instructional feedback. Nevertheless, the greater frequency of administration and the prompt availability of results from a computer-based system may enable teachers to use the scores to assign grades, modify their instruction in response to common problems or misconceptions that arise, and

provide individualized instruction that is tailored to student needs. Similarly, principals may use the results to monitor student progress across different classrooms.

Potential Advantages over Paper-and-Pencil Testing

A computerized-adaptive testing system that is Internet-based offers several advantages over paper-and-pencil multiple-choice tests. Some of these benefits arise from the use of computers, and adaptive administration in particular, whereas others derive from delivering the tests over the Internet. These benefits are discussed in turn below.

Benefits of Computerized Adaptive Testing

One of the major advantages of CAT is decreased testing time. Because the difficulty of the questions a student is asked is tailored to that student's proficiency level, students do not waste time on questions that are much too easy or too difficult for them. It takes many fewer items to achieve a desired level of score precision using CAT than using a standard multiple-choice test (see, e.g., Bunderson et al., 1989). This not only saves time, but it may minimize student frustration, boredom, and test anxiety. Similarly, this method reduces the likelihood of ceiling and floor effects that occur when a test is too easy or too hard for a student, thereby providing a more accurate measurement for these students than is obtained when the same set of questions is administered to all students. The use of computers may also reduce costs because the hardware can be used for other instructional purposes rather than being dedicated solely to

the testing function. Whether such savings are realized depends on a number of factors that we discuss later.

Another potential benefit is improved test security. Because each student within a classroom takes a different test (i.e., one that is tailored to that student's proficiency level) and because the bank from which the questions are drawn contains several thousand items, there is little risk of students being exposed to items in advance or of teachers coaching their students on specific items before or during the testing session. As software to increase test security becomes more widely available, the risk of unauthorized access to test items and results should diminish, thereby further increasing the validity of test results.

CATs are particularly useful for evaluating growth over time. Progress can be measured on a continuous scale that is not tied to grade levels. This scale enables teachers and parents to track changes in students' proficiency during the school year and across school years, both within and across content areas. Students take different items on different occasions, so scores are generally not affected by exposure to specific items. Thus, the test can be administered several times during the year without threatening the validity of the results.[1] This offers much greater potential for the results to have a positive influence on instruction than is currently available in the typical one-time-only spring test administration schedule. CATs can also accommodate the testing of students who transferred into the school during the year, those who may have been absent on the scheduled testing date,

[1]Item exposure is a topic of growing interest among psychometricians and test publishers, and its effects need to be considered when developing item banks and testing schedules. New methods have been devised to control exposure so that CATs can be administered multiple times to the same examinees, though none eliminates the risk completely (see, e.g., Revuelta & Ponsoda, 1998).

and those with learning and other disabilities who may require additional time, large type, or other testing accommodations. Several school systems serving special populations, such as the Juvenile Court and Community Schools in Los Angeles, have adopted CAT systems to address the widely varying ability levels and extremely high mobility rates of their students.

Finally, computer-based testing offers the opportunity to develop new types of questions, especially those that can assess complex problem-solving skills. For example, students can observe the effects on plant growth of various amounts of water, types of fertilizer, and exposure to sunlight in order to make inferences about the relationships among these factors. Several development efforts are currently under way to utilize technology by developing innovative tasks, including essay questions that can be machine-scored, simulations of laboratory science experiments, and other forms of constructed-response items that require students to produce, rather than just select, their answers. Many of these efforts have sought to incorporate multimedia technology to expand the range of activities in which students can engage. Bennett (1998) describes some of the possibilities offered by computer technology, such as the use of multimedia to present films and broadcasts as artifacts for analysis on a history test.

. . . computer-based testing offers the opportunity to develop new types of questions, especially those that can assess complex problem-solving skills.

Computerized assessments are especially appropriate for evaluating progress in areas where computers are used frequently, such as writing. Russell and Haney (1997), for example, found that students who were

accustomed to using computers in their classes performed better on a writing test when they could use computers rather than paper and pencil. Students using computers wrote more and organized their compositions better. As instruction comes to depend more heavily on technology, assessment will need to follow in order to provide valid measurement that is aligned with curriculum (Russell & Haney, 2000).

Benefits of Web Administration

Web-based tests offer efficient and inexpensive scoring. Scoring is done on-line, eliminating the need for packaging, shipping, and processing of answer sheets. Students could be given their results immediately after completing the tests. A web-based system would allow all records to be stored automatically at a central location, facilitating the production of score summaries. Norms could be constantly updated, and analyses of results could be done quickly and efficiently. Teachers would have results in time to incorporate the information into their instruction. Teachers could also use results for assigning grades, so that students would be motivated to do well on the tests.

There are clear benefits to maintaining the testing software and item banks in a central location so they can be downloaded onto school computers via the Internet. Economies of scale would be achieved by refreshing the item bank from a central location. New questions could easily be inserted into existing tests to gather the data on them for determining their difficulty and whether they would be appropriate for operational use. Updating of software is done centrally rather than locally, so there is no need for expensive hardware and software at the school site. Moreover, down-

loading the bank (or a portion of it) onto a local server and uploading the results daily avoids the delays in computer response times that might otherwise arise if students were connected directly to the main server. Thus, administering tests via the web addresses several logistical and technical problems associated with both paper-and-pencil testing and computer-based testing on local workstations.

IV. Issues and Concerns with Computerized Adaptive Testing

Despite the many potential advantages of CATs, a number of issues must be resolved before a computerized testing program can be implemented on a large scale. This section discusses several of these. In the next section, we discuss some of the additional issues associated with administering CATs over the web. We do not attempt to resolve these issues. Instead, this discussion is intended to help formulate a research agenda that will support the future development of web-based CATs.

Psychometric Issues Related to CATs

The use of CATs raises a host of psychometric concerns that have been the focus of intense discussion in the psychometric community over the last decade. Some of the major issues that have been examined are summarized below, but the discussion is not exhaustive.

Does the Medium Matter?

Do CATs function differently from paper-and-pencil tests that contain the same items? CATs reduce cer-

tain kinds of low-frequency error associated with paper-and-pencil testing, such as answer sheet/item number mismatches, distractions from other items on the printed page, and errors made by scanners. However, CATs may introduce other kinds of errors. For example, a reading comprehension item that has the passage and questions on separate screens might measure different skills than the traditional one with the passage and questions on the same or facing pages. Using multiple screens places a heavy emphasis on an examinee's ability to recall a passage, or part of one, presented on a previous screen (Bunderson et al., 1989). CATs also may place heavier demands on certain skills, such as typing, thereby changing the nature of what is tested.

In general, studies have shown that computer-based and paper-and-pencil tests tend to function similarly, at least in the multiple-choice format (Bunderson et al., 1989; Mead & Drasgow, 1993; Perkins, 1993; Segall et al., 1997; Zandvliet & Farragher, 1997). The two methods produce similar statistical distributions (means, standard deviations, reliabilities, and standard errors of measurement) and are comparable in their predictive validity. However, there are still possible threats to comparability.

The dependence of test scores on keyboarding speed for one medium of test administration but not another is one potential threat. A few studies have examined relationships between specific types and amounts of experience with computers and performance on tests. For example, Russell (1999) found that keyboarding speed was a good predictor of students' scores on open-ended language arts and science tests taken from the Massachusetts Comprehensive Assessment System and the National Assessment

of Educational Progress (NAEP). However, controlling for keyboarding speed, computer experience was not related to test performance. In contrast, scores on an open-ended math test were only weakly predicted by keyboarding speed, probably because the answers to math items rely mainly on less frequently used number and symbol keys. Students with computer experience are not much faster at identifying and using these keys than are students without computer experience.

Some studies have found that differences in mean scores due to administration medium can be explained by test type. Van de Vijver and Harsveld (1994) conducted a study with 326 applicants to the Royal Military Academy in the Netherlands. One group took the computerized version of the General Aptitude Test Battery (GATB), and the other took a paper-and-pencil form of this test. From this study and that of others, Van de Vijver and colleagues tentatively concluded that cognitively simple clerical tests are more susceptible to medium effects than are more complex tasks. They speculated that these differences might be related to previous computer usage and should disappear with repeated administration of the computerized version.

Some medium effects have been observed with open-response tests. As discussed earlier, experimental studies have shown that students who are accustomed to writing with computers perform better on writing tests that allow them to use computers than they do on standard paper-and-pencil writing tests (Russell & Haney, 1997; Russell, 1999). The medium can also affect the way that responses are scored. Powers et al. (1994) discovered that essay responses printed from a computer are assigned lower scores

than the same essays presented in handwritten format. Additional research is needed to examine medium effects across the range of subject areas, grade levels, and item formats that are likely to be included in a large-scale testing system, and to identify implications of these effects.

Additional research is also needed to examine differences in medium effects on tests in different subjects as well as for different examinee groups. This research is especially important in contexts in which there is a need to compare the results from the two approaches to testing—e.g., if computerized testing is phased in gradually, perhaps starting with older children and working backwards to younger ones (or vice versa).

How Should an Adaptive Test Be Implemented?

As discussed above, there are advantages to making tests adaptive. In particular, adaptive tests are usually shorter (i.e., fewer items are needed to obtain a given level of reliability) than standard tests. This occurs because item difficulty is more closely tailored to each examinee's proficiency level. In a meta-analysis of 20 studies, Bergstrom (1992) found that mode of administration (adaptive or non-adaptive) did not affect performance, regardless of test content or examinee age.

Still, questions remain about how to implement adaptivity. For example, there are three approaches to determining when to stop an adaptive test. Stopping rules can dictate fixed-length, variable-length, or mixed solutions to this problem. On a fixed-length test, the *number* of items to be administered is fixed in advance of the administration (but not the specific

items). On a variable-length test, items are administered until an examinee's proficiency level is estimated to within a prespecified degree of precision. A mixed solution to the stopping rule problem combines some aspects of the fixed-length and variable-length strategies. Researchers have found that fixed-length tests can perform as well as variable-length tests in terms of the level of uncertainty about the final proficiency estimate (McBride, Wetzel, & Hetter, 1997), but the decision about stopping rules needs to be informed by a number of factors related to the context of testing.

Another question pertains to item review, or the practice of allowing examinees to change their answers to previously completed questions. This is often cited as a way to make CATs more palatable to examinees who are accustomed to paper-and-pencil testing. When asked, examinees often state a preference for this option (Vispoel, Rocklin, & Wang, 1994). Some research suggests that when item review is permitted, examinees who change earlier answers improve their scores, but only by a small amount (Gershon & Bergstrom, 1995). Although the ability to change answers is common on paper-and-pencil tests, researchers and test developers have expressed concern that CATs might be susceptible to the use of item review to "game" the test. For example, examinees might deliberately answer early questions wrong or omit these questions so that subsequent questions are easier, and then go back and change their initial wrong answers, which could result in a proficiency estimate that is artificially high (Wainer, 1993). Use of this strategy does sometimes result in higher scores, but its effects depend on a number of factors including the statistical estimation method used and the examinee's true

proficiency level (Vispoel et al., 1999). To mitigate the risk of using item review to "game" the test, it may be limited to small timed sections of the test (Stocking, 1996), and techniques to identify the plausibility of particular response patterns may be used to evaluate results (Hulin, Drasgow, & Parson, 1983).

Item Bank Development and Management

Successful implementation of a CAT system requires a sufficiently large bank of items from which the testing software can select questions during test administration. Some of the issues related to item bank management are psychometric. For example, how large must the banks be to accommodate a particular testing schedule or test length? How many times can an item be reused before it must be removed from the bank to eliminate effects of overexposure? What is the optimal distribution of item difficulty in a bank? What is the most effective way to pretest items that will be used to refresh a bank? These and related topics have been the focus of recent research on CATs. For example, Stocking (1994) provided guidelines for determining the optimal size of an item bank and for making one bank equivalent to another. Most of the research on this problem has focused on multiple-choice questions, which are scored either correct or incorrect and which can be completed quickly by examinees. Use of constructed-response items raises additional questions, particularly because fewer items can be administered in a given amount of time.

Several organizations have produced item banks appropriate for K–12 testing. The Northwest Evaluation Association (NWEA), for example, has published CATs in several subjects and has item banks

that are appropriate for students in elementary through secondary grades. Many of these testing systems offer the possibility of modifying the assessment so that it is aligned with local or state content standards. However, it is not always clear what methods have been used to determine this alignment or whether curtailing the types of questions that can be asked would change their characteristics.

Developing new test items is time-consuming and expensive. Although items may continue to be produced by professional test publishers, new ways of generating tests should be studied. It is becoming increasingly possible for computers to generate items automatically, subject to certain constraints. This type of item generation, which can occur in real time as examinees take the test, has great potential for saving money and for reducing test security problems. However, it may lead to certain undesirable test preparation strategies.

Teachers are another promising source of items, and including teachers in the item generation process may significantly enhance their acceptance of this type of large-scale testing program. The inclusion of teachers may be especially valuable in situations where a test is intended to be aligned with a particular local curriculum.

Regardless of the source of the items, several practical concerns arise. These include copyright laws and protections, means of safeguarding data on item characteristics, and arrangements for selling or leasing item banks to schools, educational programs, and parents.

V. Issues and Concerns with Web-Based Testing

Administering CATs via the Internet rather than on stand-alone machines adds another layer of complexity. However, some of the problems discussed below, such as those related to infrastructure, would need to be addressed even in a system that used locally administered CATs. We address them here because the requirements that must be in place for successful web-based test administration are often closely linked with those that are necessary for any computerized testing. The section begins with a discussion of infrastructure, including equipment and personnel, followed by a brief discussion of costs. Finally, we bring up some issues related to reporting of assessment results.

Infrastructure

The feasibility of delivering assessment over the Internet depends on both the material infrastructure and the human capital already in place in schools. Recent data indicate that the availability of computers and the frequency and quality of Internet access are becoming sufficient to support this form of testing. In 1999, 95 percent of public schools had some form of Internet access, and fully 65 percent of instructional classrooms in the nation's public schools were connected to the Internet (U.S. Department of Education, 2000). In earlier years, schools with large proportions of students living in poverty were less likely to be connected than were wealthier schools, but this difference had disappeared by 1999. The percentage of instructional rooms that are connected, in contrast, does vary by socioeconomic status. However,

this gap is likely to shrink due in part to the E-rate program, which requires telecommunications companies to provide funds for Internet access in schools (U.S. Department of Education, 1996).

The quality of Internet connectivity has also improved over the years. For example, the percentage of schools connecting using a dedicated line has increased significantly over a recent two-year period. The use of dial-up networking declined from nearly 75 percent of schools in 1996 to 14 percent in 1999 (U.S. Department of Education, 2000). Most of these were replaced with speedier dedicated-line network connections. However, in contrast to the number of schools with Internet access, gaps between poor and wealthy schools in the quality of connection persist. For example, 72 percent of low-poverty schools (those with fewer than 11 percent of students eligible for free or reduced-price lunches) and 50 percent of high-poverty schools (those with 71 percent or more eligible students) had dedicated lines in 1999 (U.S. Department of Education, 2000).

Although the improvements in connectivity are encouraging, some important questions remain. For example, what is the quality of the hardware available to students? Are existing allocations of computers to classrooms sufficient to support several rounds of testing each year? In addition, the figures cited above illustrate that schools serving large numbers of students who live in poverty may have the basic equipment but lack the features that are necessary for effective implementation of a large-scale web-based assessment system. True equality of access is clearly an important consideration. More-detailed surveys of infrastructure and a better understanding of what types of schools are in need of improvements will

help determine the feasibility and fairness of implementing the kind of system we have been describing.

There are additional questions related to how computers are used. Are some machines dedicated solely to testing or are they used for other instructional activities? What are the test and data security implications of using the computers for multiple purposes? Placement of computers throughout the school building is another important consideration. Do computers need to be in a lab or can a web-based assessment system be implemented in a school that has only a few computers per classroom? Answers to these and related questions have implications for cost estimates. For example, testing costs would be reduced if the computers that were used for assessment activities were also being used for other instructional purposes.

Human Capital

Data suggest that most schools have staff with computer knowledge sufficient to supervise student computer use and that students are increasingly familiar with computers (U.S. Department of Education, 1999). NAEP data from 1996 indicate that approximately 80 percent of fourth and eighth grade teachers received some professional training in computer use during the preceding five years (Wenglinsky, 1998), and it is likely that this number has grown since then. However, it is not clear what the quality of this training was or how many of these staff members have the skills needed to address the unexpected technical problems that will inevitably arise. Research indicates that teachers' willingness to use computers is influenced by the availability of professional development opportunities and on-site help (Becker, 1994),

so an investment in training and support will be necessary for any large-scale instructional or assessment effort that relies on technology. Equity considerations also need to be addressed, since teachers in more-affluent school districts tend to have easier access to professional development and support.

Students' increasing familiarity with the use of computers increases the feasibility of implementing a CAT system in schools. Data from 1997 showed that a majority of students used computers at home or school for some purpose, including school assignments, word processing, e-mail, Internet browsing, databases, and graphics and design (U.S. Department of Education, 1999). Of these activities, the majority of computer use was for school assignments. Some differences in computer use were observed across racial/ethnic and socioeconomic groups, with larger differences reported for home computer use than for school use. For example, in the elementary grades, 84 percent of white students and 70 percent of black students used computers at school in 1997. The corresponding percentages for home computer use were 52 percent and 19 percent, a much larger difference (U.S. Department of Education, 1999). The gap in school computer use is likely to shrink as schools continue to acquire technological resources. However, the difference in access to computers at home, along with variation in the quality of technology in the schools (which is not currently measured well), raises concerns about the fairness of any testing system that requires extensive use of computers.

Because schools differ in the degree to which they are prepared for large-scale computerized testing, it may be necessary to implement a system of testing that supports both paper-and-pencil and computer-

ized adaptive testing simultaneously. Such a system is only feasible if there is sufficient evidence supporting the comparability of these two forms of testing. Although much of the research discussed earlier suggests a reasonable degree of comparability, the differences that have been observed, particularly on open-response tests and those that require reading long passages, suggest that caution is warranted in making comparisons across testing approaches.

Costs and Charges

A number of costs are associated with this form of testing, including development of item banks, development of software with secure downloading and uploading capabilities, and acquisition of the necessary hardware. How these costs compare with the cost of traditional paper-and-pencil testing is not known, and the cost difference between these two options is likely to change as technology becomes less expensive. A critical area of research, therefore, is an analysis of the costs of alternative approaches. This analysis would need to consider the direct costs of equipment and software, as well as labor and opportunity costs. For example, if web-based testing can be completed in half the time required for traditional testing, teachers may spend the extra time providing additional instruction to students. In addition, cost estimates should consider the fact that the computers and Internet connections used for assessment are likely to be used for other instructional and administrative activities.

It will also be important to investigate various ways of distributing costs and getting materials to schools. Should schools pay an annual fee to lease

access to the item bank or should they be charged for each administration of the test? The advantages and limitations of alternative approaches must be identified and compared.

An additional source of expenses is the scoring of responses, particularly essays and other constructed-response questions. Student answers to open-ended questions are usually scored by trained raters. However, the technology required for computerized scoring of such responses is developing rapidly. Data about the feasibility of machine scoring of open-ended responses will be critical for determining the degree to which an operational web-based CAT system can incorporate novel item types.

Cost analysis of assessment is complicated by the fact that many of the costs and benefits are difficult to express in monetary terms, or even to quantify. Furthermore, costs and benefits vary depending on the context of testing and how scores are used. For example, if high school graduation is contingent on passing an exit examination, some students who would have been given diplomas in the absence of the test will be retained in school. Although the cost to the school of providing an extra year of education to a student can be quantified, the opportunity costs for the student and the effects of the retention on his or her self-concept and motivation to learn are much more difficult to express in monetary terms. We also do not have sufficient evidence to determine whether, and to what degree, the extra year of high school will increase productivity or result in increased dropout rates and therefore increase social costs (Catterall & Winters,

. . . costs and benefits vary depending on the context of testing and how scores are used.

1994). This is just one example of a potential outcome of testing, and to the degree that web-based CATs improve the validity of this type of decision (e.g., by increasing test security), a cost-benefit analysis needs to reflect this improvement. The current assessment literature is lacking in studies of costs and benefits of different approaches to, and outcomes of, large-scale assessment.

Reporting Results

As discussed earlier, two of the anticipated benefits of a web-based CAT system are its capabilities for instant scoring and centralized data storage. These features have the potential for improving the quality and utility of the information obtained from testing and making results more timely and useful for students, parents, teachers, and policymakers. Current assessment programs do not share these features. Thus, we have little experience to guide decisions about how to generate, store, and distribute results. For example, what types of information should be produced? Should the database of school, district, state, or national norms be continuously updated and used to interpret individuals' results? Should students receive their scores immediately, or is it better to have score reports distributed by teachers some time after the testing date?

. . . the ease of access to results may have both positive and negative consequences for students and teachers, and it is imperative that we anticipate possible misuse of scores.

In today's high-stakes testing environment, the ease of access to results may have both positive and negative consequences for students and teachers, and it is imperative that we anticipate possible misuse

of scores. For example, although the possibility for principals and superintendents to have timely information about the progress of students by classroom may help them provide needed assistance and encouragement, this information could also be used in ways that unfairly penalize teachers or students. It may also increase public pressure to raise test scores, creating unrealistic demands for rapid improvement. These concerns underscore the need to examine the incentive structures and policy context surrounding any assessment system, including the kind of system we have described here.

Transmitting test questions and possibly student responses over the Internet raises a host of additional concerns. How will security of the tests and results be ensured? Who will have access to what data and when? Who will have access to the system? For example, can students practice for the tests by accessing some or all of the item banks from their homes? Will parents be able to see how well their children are doing relative to the typical performance of students in the same or other classrooms? The answers to these and related questions will play a large part in determining the validity, feasibility, fairness, and acceptability of web-based testing.

VI. Conclusion

Web-based testing, and especially the computerized adaptive version of it, will soon be competing with and possibly replacing the paper-and-pencil tests that are now relied upon for large-scale K–12 assessment programs. This new type of testing will use the Internet to deliver tests to students in

their schools. Test developers and policymakers will need to prepare for this transition by embarking on a comprehensive program of research that addresses a number of critical web-based testing issues.

A system in which adaptive assessments are delivered to students at their schools over the Internet has several advantages, including decreased testing time, enhanced security, novel item types, and rapid reporting. However, before this system can be put in place, a number of issues need to be considered, such as the psychometric quality of the measures, methods for maintaining item banks, infrastructure, human capital, costs, comparability with paper-and-pencil measures across subject matter areas, and reporting strategies. Research is needed in all of these and other related areas to provide the foundation for a smooth and effective transition to web-based testing.

Several issues and questions that were not mentioned in this paper undoubtedly will arise concerning the implementation of web-based testing in general and CATs in particular. However, we anticipate that the foregoing discussion provides some of the broad brush strokes for framing a research agenda to investigate the validity, feasibility, and appropriateness of this form of assessment. A program of research that is designed to address these and other related important questions will require the expertise of researchers from a wide range of disciplines, including psychometrics, psychology, sociology, information sciences, political science, and economics, as well as input from educators, test developers, and policymakers. Information technology has permeated nearly all aspects of our lives, including education, and it is only a matter of time before large-scale testing will utilize current technologies, including the Internet. It

is critical, therefore, that we begin a program of research now, while there is still time to steer this process in a direction that will improve education and benefit students.

REFERENCES

Achieve, Inc. (1999). *National Education Summit briefing book* (available at http://www.achieve.org).

Becker, H. J. (1994). How exemplary computer-using teachers differ from other teachers: Implications for realizing the potential of computers in schools. *Journal of Research on Computing in Education, 26,* 291-320.

Bennett, R. E. (1998). *Reinventing assessment.* Princeton, NJ: Educational Testing Service.

Bergstrom, B. A. (1992). Ability measure equivalence of computer adaptive and pencil and paper tests: a research synthesis. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. ERIC ED377228.

Bunderson, C. V., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3rd Ed., pp. 367-407). New York: Macmillan.

Catterall, J. S., & Winters, L. (1994). *Economic analysis of testing: Competency, certification, and "authentic" assessments* (CSE Technical Report 383). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Gershon, R., & Bergstrom, B. (1995). Does cheating on CAT pay: NOT! ERIC ED392844.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10,* 181-200.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory: Applications to psychological measurement.* Homewood, IL: Dow Jones-Irwin.

Kellaghan, T., & Madaus, G. F. (1991). National testing: Lessons for America from Europe. *Educational Leadership, 49,* 87-93.

Klein, S. P., & Hamilton, L. (1999). *Large-scale testing: Current practices and new directions.* Santa Monica: RAND.

Koretz, D., and Barron, S. I. (1998). *The validity of gains on the Kentucky Instructional Results Information System (KIRIS).* Santa Monica: RAND.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29* (2), 4-16.

Linn, R. L., Graue, M. E., & Sanders, N. M. (1990). Comparing state and district test results to national norms: The validity of claims that "everyone is above average." *Educational Measurement: Issues and Practice, 9,* 5-14.

Madaus, G. F. (1988). The influence of testing on the curriculum. In L. Tanner (Ed.), *Critical issues in curriculum* (pp. 83-121). Chicago: University of Chicago Press.

McBride, J. R., Wetzel, C. D., & Hetter, R. D. (1997). Preliminary psychometric research for CAT-ASVAB: Selecting an adaptive testing strategy. In *Computerized adaptive testing: From inquiry to operation* (pp. 83-95). Washington, DC: American Psychological Association.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

National Research Council, Committee on Appropriate Test Use (1998). *High stakes: Tests for tracking, promotion, and graduation*. Washington, DC: National Academy Press.

Perkins, B. (1993). Differences between computer administered and paper administered computer anxiety and performance measures. ERIC ED355905.

Powers, D., Fowles, M., Farnum, M., & Ramsey, P. (1994). Will they think less of my handwritten essay if others word process theirs? Effects on essay scores of intermingling handwritten and word-processed essays. *Journal of Educational Measurement, 31,* 220-233.

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35* (311-327).

Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives, 7* (20), available at http://olam.ed.asu.edu/epaa/.

Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Education Policy Analysis Archives, 5* (3), available at http://olam.ed.asu.edu/epaa/.

Russell, M., & Haney, W. (2000). Bridging the gap between testing and technology in schools. *Educational Policy Analysis Archives, 8* (19), available at http://olam.ed.asu.edu/epaa/.

Segall, D. O., Moreno, K. E., Kieckhaefer, W. F., Vicino, F. L., & McBride, J. R. (1997). Validation of the Experimental CAT-ASVAB System. In Sands, W. A., Waters, B. K., & McBride, J. R. (Eds.) *Computerized adaptive testing: From inquiry to operation* (pp. 103-114). Washington, DC: American Psychological Association.

Stecher, B. M., Barron, S., Kaganoff, T., & Goodwin, J. (1998). *The effects of standards-based assessment on classroom practices: Results of the 1996-97 RAND survey of Kentucky teachers of mathematics and writing* (CSE Technical Report 482). Los Angeles: Center for Research on Evaluation, Standards, and Student Testing.

Stecher, B. M., & Klein, S. P. (1997). The cost of science performance assessments in large-scale testing programs. *Educational Evaluation and Policy Analysis, 19,* 1-14.

Stocking, M. L. (1994). Three practical issues for modern adaptive testing pools. ETS Report ETS-RR-94-5. Princeton, NJ: Educational Testing Service.

Stocking, M. (1996). Revising answers to items in computerized adaptive testing: A comparison of three models. ETS Report Number ETS-RR-96-12. Princeton, NJ: Educational Testing Service.

U.S. Department of Education (1996). *Getting America's students ready for the 21st century: Meeting the Technology Literacy Challenge.* Washington, DC: U.S. Government Printing Office.

U.S. Department of Education (1999). *The condition of education 1999.* Available at http://www.nces.ed.gov/pubs99/condition99/index.html.

U.S. Department of Education (2000). *Internet access in U.S. public schools and classrooms: 1994-99* (NCES 2000-086). Washington, DC: U.S. Government Printing Office.

Van de Vijver, F. J. R., & Harsveld M. (1994). The incomplete equivalence of the paper-and-pencil and computerized versions of the General Aptitude Test Battery. *Journal of Applied Psychology, 79* (6), 852-859.

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 53,* 53-79.

Vispoel, W. P., Rocklin, T. R., Wang, T., & Bleiler, T. (1999). Can examinees use a review option to obtain positively biased ability estimates on a computerized adaptive test? *Journal of Educational Measurement, 36,* 141-157.

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement, Issues and Practice, 12,* 15-20.

Wenglinsky, H. (1998). *Does it compute? The relationship between educational technology and student achievement in mathematics* ETS Policy Information Report. Princeton, NJ: Educational Testing Service.

Zandvliet, D., & Farragher, P. (1997). A comparison of computer-administered and written tests. *Journal of Research on Computing in Education, 29* (4), 423-438.