CHAPTER 8

Information Technologies and Testing: Past, Present, Future

Contents

Boxes

Figures

CHAPTER 8

Information Technologies and Testing: Past, Present, Future

Highlights

- **Information and data processing technologies have played a critical role in making existing modes of testing more efficient. The combination of the multiple-choice item format and machine scoring technologies has made it possible for massive numbers of students to be tested all through their educational careers.**
- **By and large, computers and other information technologies have not been applied toward fundamentally new ways of testing. However, advances in computers, video, and related technologies could one day revolutionize testing.**
- **Computer-based testing and computer-adaptive testing can have several advantages over conventional paper-and-pencil tests. They are quicker to take and score, provide faster feedback, and reduce errors due to human scoring and administration. Some computerized tests can hone in on students' achievement levels much more quickly and accurately than conventional tests.**
- **Cutting-edge technology could push tests well beyond the existing paper-and-pencil formats. Structuring and presenting complex tasks, tracking student cognitive processes, and providing rapid feedback to learners and teachers are promising avenues for continued research and development.**
- **Computerized testing also has drawbacks. It may introduce new types of measurement errors, place students who lack familiarity with computers at a disadvantage, make it harder for students to** skip or **review questions, raise new privacy issues, and create questions of comparability when students take essentially "personalized" tests.**
- **Realizing the full potential of new testing technologies will require continued research, and better coordinated research, in the fields of learning theory, computer science, and test design.**

Information and data processing technologies have had a powerful influence on educational testing. The invention of the multiple-choice item format, coupled with advances in machine scoring, made possible the efficient testing of millions of children at all stages of their education. But these efficiency attributes of machine-based scoring and reporting also raised serious concerns: from the earliest days of application of these technologies, critics lamented the loss of richness in detail that had been a feature of open-ended questions scored by human judges, and contended that machine-scored tests encouraged memorization of unrelated facts, guessing, and other distortions in teaching and learning.

Multiple-choice items and machine scoring of tests brought a revolution in student assessment. And, not surprisingly, once the technology became an entrenched feature of school life, there began a 70-year period of gradual evolution: as information and data processing technologies become more powerful and sophisticated, they continued to influence educational testing, but the applications have principally improved automation of the basic test designs initiated at the turn of the century. There has been relatively little exploration of how the technology might open altogether new approaches to student assessment. Today, however, some experts believe a new revolution is in the making: they contend that the increasing power and flexibility of personal computers, video, and telecommunications could move testing well beyond what paper-and-pencil testing can accomplish.

The purpose of this chapter is to examine the state of the art of information technologies in testing, consider policy initiatives that could foster better uses of current technology, and explore the possibilities for wholly new paradigms of student assessment, The chapter is divided into four sections. The first provides a brief historical synopsis of technology in testing, focusing on the combined effects of multiple-choice and electromechanical scoring.

The second section is concerned with applications of computers and video-related technologies to conventional models of educational assessment. It

addresses issues such as test design and construction, scoring and analysis of test results, item banking, computer-adaptive testing, and new video and multimedia applications.

The third section of the chapter describes the gap between current and future models of testing, and explores ways in which computers or other technologies could advance the development and implementation of new models.

Finally, the fourth section examines key policy issues in developing new models of testing.

Historical Synopsis

Multiple choice made its debut in 1915 with the Kansas Silent Reading Test, produced by Frederick Kelly at the State Normal School in Emporia. With modifications by psychologist Arthur Otis, multiple choice ''. . . soon found its way . . . from reading tests to intelligence tests, ' and made possible the administration of the Army Alpha and Beta tests to millions of draftees during the First World War.[1] Clerks scored each test by hand, using stencils superimposed on answer sheets. This new method of testing transformed Alfred Binet's individually administered test format (called by some authors the ''methode de luxe"[2]) into a format amenable to group administration and the development of group norms. According to one chronicle of this technological change:

> . . . the multiple choice question [was] . . . an invention ingenious in its simplicity . . . [an] indispensable vehicle for the dramatic growth of mass testing in this country in the span of a few years. It had not existed before 1914; by 1921 it had spawned a dozen group intelligence tests and provided close to two million soldiers and over three million schoolchildren with a numerical index of their intelligence; it was also about to transform achievement testing in the classroom.[3]

It was the Iowa testing program, under the leadership of E.F. Lindquist, that was instrumental in turning the twin concepts of group testing and the multiple-choice item format into a streamlined process for achievement testing of masses of school children.[4] Lindquist took the first hand-scored tests and designed a scoring key that could be cut into strips, each strip fitting a test page, with the answers positioned on the key to match the pupil's responses on the page. Later, Lindquist pursued his dream of mechanical, and later electronic, scoring. IBM's prototype photoelectric machine encouraged Lindquist, who built his own analog computer in the 1940s. During the 1950s, he embarked with Professor Phillip Rulon of Harvard in an effort to design an electronic scoring machine. Their basic innovation has since become a staple of the testing industry:

> . . . a specially designed answer sheet would pass under a row of photo tubes in such a manner that each photo tube would sense a mark in one of the boxes on the answer sheet when illuminated by a light source, and the pulses from this sensing would trigger a counter cumulating a total raw score for each test on the answer sheet; the raw score would be converted to a standard score in a converter unit; the standard score would be recorded by an output printer geared to the scoring device.[5]

The first 'Iowa machine' went into production in 1955, and cost close to $200,000 (nearly three times more than planned).[6] Continuing refinements through 1957 led Lindquist to boast that the machine was living up to virtually all expectations. It could now, in a single reading of an answer sheet, obtain up to 14 separate raw scores; convert these into 20 different standard scores, percentile ranks, or converted totals of the converted scores; obtain simultaneously as many totals and/or subtotals as the desired combinations of counters would permit; print and punch scores simultaneously; print or punch both names and scores simultaneously; and

---

[1]Franz Samelson, "Was Early Mental Testing (a) Racist Inspired, (b) Objective Science, (c)A Technology for Democracy, (d) The Origin of Multiple Choice Exams, (e) None of the Above? Mark the RIGHT Answer," *Psychological Testing and American Society, 1890-1930,* M. Sokal (cd.) (New Brunswick, NJ: Rutgers University Press, 1987), pp. 113-127. See also ch. 3 of this report for discussion and ch. 1 for a reproduction of the cover of the 1915 Kansas test.

[2]Rudolf Pintner, cited in Samelson, op. cit., footnote 1, p. 116.

[3]Samelson, op. cit., footnote 1.

[4]For a comprehensive discussion of the history of the Iowa program, see Julia J. Peterson, *The Iowa Testing Program* (Iowa City, IA: University of Iowa Press, 1983,) For discussion of the principal roles of Lewis Terman, Edward Thorndike, Robert Yerkes, and others in the birth of the group-administered intelligence and achievement testing movement, see, e.g., Paul Chapman, *Schools as Sorters: Lewis M. Terman, Applied Psychology, and the Intelligence Testing Movement, 1890-1930* (New York, NY: New York University Press, 1988); also see ch. 3 of this report.

[5]Peterson, op. cit., footnote 4, p. 91.

[6]Ibid., p. 89.

do a number of "interesting tricks" it was not originally intended to do.

A new era of testing in American schools had dawned. Here is how one test publisher, whose experiences date from the earliest days of this new era, describes the transition:

> . . . [before machine scoring] most standardized tests were hand-scored by the teachers. . . . Under that system, tests corrected and scored by the teacher provided opportunity for careful pupil analysis by the teachers. In turn that analysis, pupil by pupil and class by class, provided meaningful measures for individualizing pupil instruction, improving instruction, reassessing the curriculum, and making appropriate textbook selections. Furthermore, and by no means should this be overlooked, it gave the teacher support beyond his or her undocumented human judgment of pupils that by no means goes unchallenged by many parents and, for that matter, pupils. As the machine-scoring movement grew, the activities related to testing changed. Certainly, the scoring activity left the classroom and often as not the school system itself. Test results moved increasingly into the hands of the administrative staff. Test specialists were employed who were interested in an ever broader array of derived scores to be used for many purposes . . . the hands-on dimension for teachers receded and in due course disappeared almost entirely.[7]

Current Applications of Computers in Testing[8]

Design and Construction of Tests

Item Writing

Computers have many capabilities that can aid test publishers in the efficient design and construction of standardized tests. In addition, basic word processing, graphics, and spreadsheet programs make it possible for State and district school personnel, as well as individual teachers, to create their own items or to edit items developed by others. Editing the text of test items, selecting specific items from a collection stored in memory, and sequencing the test items are all substantially easier with basic desktop computers and generic tool software.

Increasingly, however, dedicated item writing and test construction packages have become available. These go beyond the capacity of generic word processing software and are intended specifically for writing tests. For example, they can contain item templates and special notations such as mathematical symbols not usually available with commercial word processing software. Once the test is created on the computer, it can then be printed out, reproduced, and administered to students who fill in the responses in the traditional paper-and-pencil format.

Using computers to construct items is not a new concept. Researchers in the 1960s had attempted to develop software to facilitate the construction of sentence completion and spelling items, but the software was not adopted by test constructors.[9] This is explained in part by the feeling among some experts that item writing for educational and psychological testing is more art than science, and that computer technology routinizes what ought to be a more fluid and creative process. Most item-writing efforts for standardized achievement tests involve an interplay between content specialists (teachers in the content areas) and psychometric experts who identify item-writing flaws and examine the match between items and objectives of the test.[10]

Item Banking

Increases in computer memory capacity have made ''item banks' an important enhancement in test construction. Large collections of test items are organized, classified, and stored by their content and/or their statistical properties, allowing test developers or teachers to create customized tests. Item banks in use today consist almost exclusively of multiple-choice or true-false questions, although there is some research under way on the use of CD-ROM technology to store longer open-ended items.[11]

---

[7] Harold Miller, former chairman of the Board, Houghton Mifflin Co., Inc., personal communication, Dec. 14, 1990.

[8] This section draws on C.V. Bunderson, J.B. Olsen, and A. Greenberg, ''Computers in Educational Assessment,' OTA contractor report, December 1990.

[9] Tse-chi Hsu and Shula F. Sadock, *Computer Assisted Test Construction The State of the Art* (Washington, DC: ERIC Clearinghouse on Tests, Measurement, and Evaluation, American Institutes for Research, November 1985), p. 5.

[10] Gale H. Reid, ''Item Writing and Item Banking by Microcomputer: An Update, ' *Educational Measurement Issues and Practice, vol. 8, No. 3,* fall 1989, p. 18.

[11] I see, e.g., Judah Schwartz and Katherine A. Viator (eds.), *The Price of Secrecy: The Social, Intellectual, and Psychological Costs of Current Assessment Practice: A Report to the Ford Foundation* (Cambridge, MA: Harvard Graduate School of Education, September 1990).

A variant on the item-bank concept is one in which testing *objectives are* **stored in the form of algorithms that can be used to create individual test items. The algorithm draws on stored data to produce a vast number of variations on an objective. Instructors choose the objective and specify the number of different problems, and the computer provides the appropriate test items (see figure 8-l). One item bank currently on the market covers mathematics objectives, from basic mathematics through calculus.[12] If the teacher wishes to test a student on adding two two-digit numbers, the objective is represented as A + B, where A and B are whole numbers greater than 9 and less than 100. The computer would then insert random numbers for A and B, so that literally thousands of different items sharing a similar measurement function can be produced. The system can be customized to meet the objectives of States, districts, or even specific textbook or curriculum objectives.**

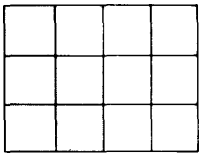
Constructing standardized tests to meet the elaborate and detailed test specifications of school districts and States is a complex and time-consuming task. Computers can help speed and streamline this task by selecting test questions for use in a test form to match detailed statistical and content specifications. After the computer selects test questions for the first draft of a test form, these items can be reviewed by test development staff, and possibly field tested.[13] Computing power greatly speeds up this process and makes it possible for States and local education authorities to create their own standardized tests as well as varying forms of the same test for multiple administrations.

Among the many applications of the item-bank concept, a large-scale effort begun in West Virginia in 1988 offers some useful lessons.[14] As part of a larger effort to restructure financing in the State and to assess learning outcomes for students, the State purchased 1,200 copies of the testing software, one for every school in the State, Reflecting a bottom-up strategy, the system allows teachers to select items, construct their own tests, print them out, copy them, and administer them in the traditional paper-and-pencil format. Score results can be analyzed and student progress tracked through the use of instructional management software. A pilot test of the system highlighted the fact that teachers needed training on how to use the hardware and software and that the existing infrastructure of computers for teachers was inadequate. Among the benefits noted were the ease in generating tests for many uses and the advantages of relieving teachers of some of the "busy work" of test construction and administration.

The West Virginia system deals with traditional subject areas. Note, however, that in its request for proposals for a computer system, the State sought a system capable of storing item types other than multiple choice and true-false, with software available in both IBM and Apple formats.

Figure 8-l-Three Questions Created by One Algorithm

1. What fraction of this figure is shaded?

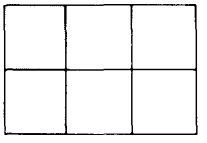


A. 5/7 **B. 5/12**
C. 7/12 **D. 5**

2. What fraction of this figure is shaded?

A. 3/10 **B. 3**
C. 3/7 **D. 7/10**

3. What fraction of this figure is shaded?



A. 2/3 **B. 3**
C. 1/3 **D. 1/2**

SOURCE: ips Publishing, *Exam in a Can* (brochure) (West Lake Village, CA: 1990).

[12] ips Publishing, *Exam in a Can* (computer software) (Westlake Village, CA:1990).

[13] Mark D. Reckase, director, Development Division, Assessment Innovations, American College Testing Program, personal communication, September 1991. See also Dato N.M. de Gruijter, "Test Construction by Means of Linear Programming," *Applied Psychological Measurement,* vol. 14, No. 2,1990, pp. 175-182; and Ellen Beokkooi-Timminga, " The Construction of Parallel Tests From IRT-Based Item Banks, " *Journal of Educational Statistics, vol. 15, No. 2,* 1990, pp. 129-145.

[14] John A. Willis, "Learning Outcome Testing Program: Standardized Classroom Testing in West Virginia Through Item Banking, Test Generation and Curricular Management Software," *Educational Measurement: Issues and Practice,* vol. *9, No. 2, s*ummer 1990, pp. 11-14.

Scoring, Reporting, and Analyzing Test Results

Computers are now vital to large-scale testing programs. They allow for fast and efficient scanning and processing of answer sheets, computation of individual and group scores and subscores, and storage of score data for later analysis. Item analysis and item-response theory statistics can be calculated across large numbers of test takers, and the item and test statistic files can be automatically updated using only a few simple commands. Archival copies of test scores can also be easily made. Computers provide a wide range of individual and group reports that can be printed from the resulting test scores and profiles. Computerized interpretative reports are also prepared for an increasing number of educational and psychological tests.

Large mainframes or computers are used to process and analyze test data and to prepare printed reports for individual students or groups of students. These mainframes and computers are typically located at centralized test development, publication, and scoring service centers run by test publishers.

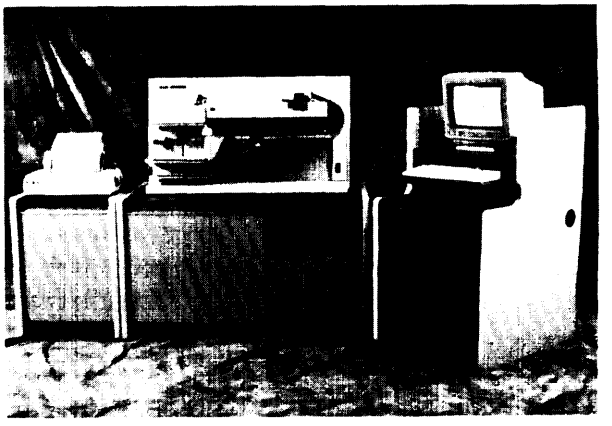


Photo credit: Courtesy of National Computer Systems, Inc.

Using machines like the National Computer Systems' Opscan 21, 10,000 tests can be scored in 1 hour.

Taking Tests on the Computer

In addition to their role as workhorses to aid in test construction, recordkeeping, and analysis and reporting of results, computers can also be the medium on which tests are administered. This report defines computer-based testing (CBT) as applications in which students respond to questions directly on the computer, via keyboard, keypad, mouse, or other data-entry device. Test booklets, fill-in-the-bubble answer sheets, and other traditional paper-and-pencil testing techniques are not used.[15]

Classroom Testing With Networks and Integrated Learning Systems

Much of the available computer software designed for instruction includes questions throughout the program designed to check on a student's understanding of the material. Responses can be printed out for the teacher to gauge student progress and identify problem areas. Many schools have linked the computers they have in laboratories and classrooms; networks generally consist of 15 to 25 computers linked through a central file server. With these local area networks (LANs), the same software can be shared among many computers, easing the logistics of administration for the teacher. Through computers connected by a networked system, programs and data can be shared and then sent to common peripheral devices such as a printer, hard disk, or videodisc. Each computer on the LAN can operate independently, using different pieces of software for each student, or share software among several or all students, enhancing the teacher's ability to manage and individualize instruction and testing for each child.[16]

One of the greatest selling points of networks is the added tracking and reporting capabilities that become possible when all student data are stored on a single storage device such as a hard disk. Stand-alone computers with individual floppy disks do not have sufficient storage capacity for all of the student records in a class or school. In contrast, networked systems make it possible to collect extended reports on student progress. In large part because of the appeal of these assessment features, the number of districts with network installations has grown steadily over the past 3 years, from just over 1,500 in 1988-89 to over 2,800 in 1990-91.[17]

[15] Paper and pencils may be used as backup tools, such as scratch pads or worksheets, but they are not the form of entry of final answers to test questions.

16 For further discussion of how school computers can be networked, see, e.g., U.S. Congress, Office of Technology Assessment, *Power On! New Tools for Teaching and Learning, OTA-SET-379* (Washington, DC: U.S. Government Printing Office, September 1988).

[17] Quality Education Data, "Technology in Schools: 1990-91 School Year," *Market intelligence* (Denver, CO: 1991), p. T-7.

Photo credit: Steve Woit

Computers are a key feature at the Saturn School of Tomorrow. A Mac Lab is available at all times for students to do word processing and publishing.

Integrated learning systems (ILSs) are LANS with a comprehensive instructional management system. Courseware is typically published and sold by the ILS vendor, and spans part or all of a curriculum (e.g., K-6 language arts). It is possible to add additional software in some ILSs. As in other networked systems, instruction is controlled and managed through the central computer, which may be connected to printers, modems, videodiscs, or other peripheral devices.

Because of their close linkages between instruction and testing, both of which can be matched to district curricula, ILSs have become increasingly popular. Although fewer schools have ILSs than networks, their number has been growing rapidly (from about 3,300 in 1989-90 to almost 7,000 in 1990-91).[18] The vast majority of ILS use is at the elementary level, with more than 80 percent of ILS usage in reading/language arts and mathematics.[19]

With an ILS testing is an integral part of instruction. The testing part of the system highlights what to teach, and the instructional part is designed for easy assessment of student performance. Some critics fear this focus on test-based skills reinforces a linear and limited approach to learning. Others,

Photo credit: Steve Woit

At the Saturn School of Tomorrow, students work independently on integrated learning systems.

however, suggest it could help bridge assessment and instruction. The importance of networks/ILSs is heightened by the fact that continued demand for these technologies could create opportunities for testing-software developers to collaborate with suppliers of these products.

ILS vendors include Computer Curriculum Corp., Education Systems Corp., ICON, PLATO, Wasatch, WICAT, and the Jostens Learning Corp. For example, Jostens' Instructional Management System is intended to allow teachers to deliver a customized sequence of lessons to each student; direct and monitor student progress; adopt the sequence of the embedded curricula and prescribe lessons from third-party materials; branch students to appropriate remedial or enrichment activities; generate criterion-referenced pre- and post-tests; create, maintain, and update instructional records on each student; and electronically transfer records within and between schools.

Although networks and ILSs offer a promising way to bring computerized testing into the schools, their focus is primarily on classroom instruction. The growth in the installed base of networks and ILSs in schools suggests the potential for their expanded application in testing. It is important to note that these centralized systems place software and test items under the control of one person (usually the teacher).

---

[18] Ibid., p. T-4.

[19] Charles L. Blaschke, "Integrated Learning Systems/Instructional Networks: Current Uses and Trends," *Educational Technology, vol. 30, No.* 11, November 1990, p. 21.

Computers and Testing: Beyond the Classroom

Computer-based testing is not commonly used for system monitoring or student selection, placement, and certification in elementary and secondary schools. Few schools have enough computers to implement a large-scale testing program via computer.[20] Even where adequate hardware exists, the demand for computerized standardized tests has, in the past, been low. Today's standardized paper-and-pencil tests are a well-entrenched technology and practice. Students, teachers, and the public are familiar with test books and 'bubble' answer sheets and the technology is easy to use, score, and administer. There is also a well-developed and longstanding support system underpinning this type of testing.

In their most basic form, CBT takes existing paper-and-pencil tests and administers them on a computer: test items, format, and procedures remain the same as for paper-and-pencil, and the computer's role is that of an "automated answer sheet. ' '[21] Computers offer capabilities that make even these limited applications more flexible, powerful, and efficient.

Tests other than those of academic achievement have also become the subject of research in CBT. Examples are various psychological tests and tests used for admissions, placement, and certification at the postsecondary level. The Educational Testing Service (ETS) has been pilot testing computer-based versions of the Graduate Record Examination (GRE); both ETS and the American College Testing Program (ACT) have developed computer testing packages for college placement testing and are currently conducting research to verify comparability of scores from the computerized and paper-and-pencil tests. Finally, there is growing interest in the use of computerized tests for professional certification, in the military, and in industry for selection and placement purposes.

To date, research on comparability between computer-based and conventional paper-and-pencil tests has had mixed results. Most studies have found that students score slightly, but not significantly, higher on paper-and-pencil tests than on computer-based tests. Although it was hypothesized that computer inexperience and computer anxiety might exacerbate score differences between testing models, this has not been found to be significant. It has been suggested, however, that earlier forms of CBT, which did not allow examinees to skip items and go back and answer them later in the test, or to review and change responses of items already answered, may have accounted for lower scores on computer-based tests.[22] Because of this concern, the American Psychological Association *Guidelines* recommends that test publishers perform separate equating and/or norming studies when computer-based versions of standardized tests are introduced.[23] It should be noted that current forms of CBT usually allow students to skip items, return to them later, and change their answers just as they would in a paper-and-pencil test,

Computerized Adaptive Testing

An innovation in testing that applies the computer's rapid processing capability to an advanced statistical model is called ''computerized adaptive testing' or CAT. In conventional testing all examinees receive the same set of questions, usually in the same order. But with CAT the computer chooses items to administer to a given examinee based on that examinee's responses to previous test items. Thus, not all examinees receive the same set of test items .[24]

The advent of "item-response theory" in the 1960s led to the realization that relative performance of students could be assessed more efficiently if test items were selected and sequenced with specific reference to individual student ability. Instead of presenting a broad range of items to all students, some of which are too difficult and some too easy, item-response theory allows the range of difficulty

[20] James B. Olsen, Apryl Cox, Charles Price, Mike Strozeski, and Idolina Vela, ''Development, Implementation, and Validation of a Computerized Test for Statewide Assessment, " *Educational Measurement: Issues and Practice, vol. 9, No. 2,* summer 1990.

[21] Isaac 1. Bejar, ''Speculation on the Future of Test Design+ *Test Design: Developments in Psychology and Psychometrics,* S.E. Embretson (cd.) (Orlando, FL: Academic Press, 1985), p. 280.

[22] Steven L. Wise and Barbara S. Plake, "Research on the Effects of Administering Tests Via Computers, " *Educational Measurement: Issues and Practice*, vol. 8, No. 3, fall 1989, p. 7.

[23] Ibid.

'Ibid., p. 5.

of items to be determined by the test-taker's responses to previous items:

> Adaptive testing . . . seeks to present only items that are appropriate for the test taker's estimated level of skill or ability. Questions that are too easy or too difficult for the candidate contribute very little information about that person's ability. More specifically, each person's first item on an adaptive test generally has about medium difficulty for the total population. Those who answer correctly get a harder item; those who answer incorrectly get an easier item. After each response, the examinee's ability is estimated, along with an indication of the accuracy of the estimate. The next item to be posed is one that will be especially informative for a person of the estimated ability, which generally means that harder questions are posed after correct answers and easier questions after incorrect answers. The change in item difficulty from step to step is usually large early in the sequence, but becomes smaller as more is learned about the candidate's ability. The process continues until there is enough information to place the person on the ability scale with a specified level of accuracy, or until some more pragmatic criterion is achieved.[25]

The concept of adaptive testing is not new; most individually administered tests have some adaptive features, and in some group testing in a paper-and-pencil format there may be a form of pretest to determine student ability and to narrow the range of items presented on the main test. However, the enormous superiority of the computer in terms of storage capacity and processing speed has made adaptive testing much more efficient.

Computerized adaptive tests can be used for instructional feedback, system monitoring, or selection, placement, and certification functions. One example is the College Board Computerized Placement Tests, developed jointly by the College Entrance Examination Board and ETS, for use by 2- and 4-year colleges to assess the readiness of entering students for college-level work in English, reading, and mathematics, and to determine their need for additional preparatory courses. These tests have been used since the mid-1980s at approximately 80 colleges across the United States.[26]

The Portland (Oregon) school district has developed a CAT system linked to its districtwide testing program. The Portland Achievement Level Testing (PALT) program, a combined norm-referenced and criterion-referenced test battery developed by the district, has been the district's principal evaluation instrument since 1977. It has been expanded and refined regularly to keep up with changes in curricula and instructional priorities. All students in grades three to eight take the PALT paper-and-pencil tests in reading and mathematics twice yearly; eighth graders are expected to meet the district's minimum competency levels, and if they fail they must repeat the test periodically through high school in order to graduate with a standard diploma. Roughly 40,000 students (out of a total K-12 enrollment of 55,000) are tested twice yearly.

The CAT version of the test, known as Computerized Adaptive Reporting and Testing (CARAT), was initially developed over the 5-year period 1984 to 1989 with annual support from the Portland School Board of $250,000 or more. It is expected to be implemented districtwide by 1992 under a 3-year $1 million grant from the school board. It is available for students to work on any time during the year.

CARAT consists of items drawn from the PALT item banks. CARAT tests can count for placement in special programs (talented and gifted, or Chapter 1). However, at present students must take the paper-and-pencil test on its electronic equivalent-not the adaptive version—in order to be certified for graduation.

CARAT began on **a pilot basis in six** schools in 1985-86, and has since been implemented in all Chapter 1 schools in the district. Computer adaptive tests have been used for more than 5,000 students for Chapter 1 evaluation and for assessing competency in mathematics and reading, grades three through eight, since the program was begun.

District officials hope to have CARAT installed in every school by the 1992-93 school year, and eventually to shift the entire testing program to CARAT They believe that CARAT:

- makes it possible to test students as soon as they enter the district, in order to place them in appropriate instructional programs;

[25]Bert F. Green, R. Darrell Bock, Lloyd G. Humphreys, Robert L. Linn, and Mark D. Reckase, ''~tic/d Guidelines fOr Assessing Computerized Adaptive Tests,'' *Journal of Educational Measurement,* vol. 21, No. 4, winter 1984, pp. 347-348.

[26]Bunderson et al., op. cit., footnote 8, p. 22.

- makes possible more continuous assessment of student progress during the school year than would be possible from the fall and spring testing alone;
- is available at all times, providing access to students alone or in groups at any time and at any site;
- provides ready access to longitudinal test data on any designated group of students in the school;
- allows for the shortest possible tests (a CARAT test takes about 20 minutes) with known measurement properties; and
- offers enhanced test security, since students rarely get the same questions and since test questions can be changed regularly .27

The Northwest Evaluation Association has marketed the Portland adaptive testing system, including the item banks and computerized software, to other districts in Oregon, at a cost of approximately $16,000. Currently about 15 districts, including some other large systems, use PALT-based paper-and-pencil tests and CAT.

Computerized Mastery Testing

One application of CAT, known as computerized mastery tests, includes cut scores (the decision point separating masters from nonmasters) to assess whether the test taker has achieved "mastery' in a field.[28] Students pass or fail the test depending on how many items they answer correctly. If the responses do not provide a clear enough picture, additional items of similar difficulty are presented until mastery is determined. These tests typically require only one-half of the questions administered in the conventional paper-and-pencil format to reach the same reliability levels. Reliability is high around the cut score. As in the case of Portland, computerized mastery testing can be used for minimum competency testing.

Occupational competency testing has also been a target of new technological applications. Although assessments such as the one designed for the National Board of Medical Examiners (see box 8-A) serve quite different functions than tests in the elementary and secondary school years, they offer some important lessons for the capability of computers and simulation software. (See also below, under "New Models of Assessment and the Role of Technology.")

Taking Tests on the Computer: Pros and Cons

Computer-based testing can improve the efficiency of standardized test administration and provide administrative benefits when compared to standardized paper-and-pencil testing. But like any new technology, benefits need to be weighed against potential drawbacks.

Advantages of CBT

Because questions are presented together with the response format (as opposed to a separate answer sheet), *it is* ***faster to take a computer-administered test. One*** study showed that CBTs and CATS are between 25 and 75 percent faster than paper-and-pencil tests in producing otherwise comparable results (see figure 8-2).[29]

A greater variety of questions can be included in the test-builder's tool kit.[30] Constructed response items and short answers involving words, phrases, or procedures can also be scored relatively easily by matching them to the correct answer (or answers) stored in the computer. Voice synthesizers can be used for spelling or foreign language examinations. Computer graphics and video can make possible other novel item types or simulations.

Computers allow new possibilities for items that require visualization of motion or complex interdependencies. For example, a conventional physics examination might require long and complex syntax or a series of static diagrams to depict motion. On a computerized test, motion can be more simply and clearly depicted using either a high-resolution graphic or video display. A computerized version of the item gives a purer measure of the examinee's understand-

[27]District officials note, however, that Computerized Adaptive Reporting and Testing test items can appear on the paper-and-pencil Version Of the test that counts. The extent of overlap, which could affect test validity, has not been measured.

[28]David J. Weiss and G. Gage Kingsbury, "Applications of Computerized Adaptive Testing to Educational Problems, " *Journal of Educational Measurement,* vol. **21, winter 1984, pp. 361-375.**

[29]James B. Olsen, **"The Four Generations of Computerized Testing**: Toward Increased Use of AI and Expert Systems, " *Educational Technology,* vol. *30, No. 3,* March 1990, p. 37.

[30]Howard Wainer, ''On Item **Response Theory** and Computerized Adaptive 'lksts,' *The Journal of College Admissions,* vol. *27, No. 4,* April 1983, p. 15.

Box 8-A-Certification Via Computer Simulations: The National Board of Medical Examiners

A 65-year-old man arrives at the Emergency Department of a major teaching hospital, complaining of respiratory distress and sharp chest pains. He appears to be in acute distress, moaning and holding his hands over the left side of his chest The emergency medical technician who brought the patient in says he has a history of asthma and emphysema. You area medical student, and must diagnose and treat the patient. The entire spectrum of modern medicine is at your fingertips, but time is of the essence in this potentially life-threatening condition of respiratory or cardiovascular distress. What do you do?[1]

This is an example of 1 of 25 patient simulations in a Computer Based Exam (CBX) that has, since 1988, been used at 75 medical schools in the United States and Canada. The ultimate objective for these simulations is use in the certification examination of the National Board of Medical Examiners (NBME), required of physicians in training before they can become licensed.

Medical schools have long been concerned that the examinations used to test students are heavy on the recall of factual information, but may not adequately test other important indicators of a candidate's readiness to practice medicine. One of these characteristics is the ability to employ the skills needed in clinical care--evaluating patient symptoms, conducting the appropriate procedures, ordering and evaluating tests, bringing in other experts for consultation-in order to accurately and quickly diagnose patient problems and diseases. In the NBME's CBX, the examinee is provided a simulated clinical environment in which cases are presented for actual patient management. Through a blank entry screen that automatically processes free-text orders, the examiner can request more than 8,500 terms representing over 2,300 diagnostic studies, procedures, medications, and consultants, and can move the patient among the available health care facilities. As the examinee proceeds, the computer records the timing of all actions taken. These actions are compared with a codified description of optimal management based on the judgments of expert doctors, and scoring is based on how well the examinee follows appropriate practice.

An examinee's management of the case presented above might proceed as follows (see figure 8-Al):

The results suggest a diagnosis of spontaneous pneumothorax (a collapsed lung), a possibly life-threatening disease process. The patient's low blood pressure suggests some degree of cardiovascular difficulty, indicating immediate decompression of the patient's left hemithorax (one-half of the patient's chest cavity). Pressing F1 allows a review of tests on order. It is currently 16:03; the chest x-ray result will not be available until 16:20 and the examinee must decide whether to treat the patient now or wait until x-ray results are available. She decides to perform an immediate needle thoracostomy (insertion of a needle into the chest cavity to evacuate the air) and the computer simulates the process and results:

The rush of air confirms the diagnosis, but suddenly another message appears on the screen: "Nurses Note: The patient's pain is more severe." More action is required. The examinee orders placement of a chest tube; once the patient is stabilized, she orders blood to be drawn and additional medical history to be taken. The examination continues until, at 16:37, the examinee completes the workup, admits the patient to the ward, and leaves orders for followup procedures. At 16:50 the message appears on the screen: "Thank you for taking care of this patient."

In this example, the simulated case time was 50 minutes; it took the student 17 minutes in real time to complete the case simulation. Cases can last for months of simulated time; examinees typically are allowed about 40 minutes, but usually take 20 to 25 minutes.

NBME computer-based testing is being phased in in stages. In Phase I, results from a 1987 field study were reviewed by an external advisory panel of experts in medicine, medical education, medical informatics, and psychometrics; they concluded the following:[2]

- CBX succeeded in measuring a quality (reasonably assumed to be related to clinical competence) not measured by existing examination formats.
- NBME should continue its current level of developmental activity directed at the ultimate use of the CBX in the NBME examination sequence for certification.

---

[1]This example excerpted from K. Cotton and D.M. Durinzi, *Software I-Phase II* (Philadelphia, PA: National Board of Medical Examiirs, 1990).

[2]S.G. Clyman and N.A. Orr, "Status Report of the NBME's Computer-Based Testing," *Academic* vol. 4, April

- Examinations should be delivered through a system that incorporates collaborations with medical schools.
- A phased approach should be taken: Phase I would entail distribution of software so that students and faculty could familiarize themselves with the format and participate in collaborative research; Phase II would entail formal field studies; Phase III would entail extended intramural testing services; Phase IV would entail introduction in the certification examination(s).

For the first phase of testing, the case simulations, an evaluation of each student's management of the case is offered in the form of qualitative "case-end feedback," derived from a scoring key developed by interdisciplinary committees of expert clinicians. The record of action is preserved by the computer and becomes the basis for computer grading of performance. Actions are evaluated in several item categories:[3]

Benefit: considered appropriate and useful in the management of the patient;

Neutral: representing acceptable actions that do not necessarily differentiate one student from another;

Risk: not required and may result in morbidity;

Inappropriate: represent nonharmful actions that are not indicated in the management of the patient;

Flag: indicate that the student did not successfully fulfill the testing objective or subjected the patient to unacceptable risk or poor probable **outcome, through errors of omission or commission.**

Additional data provided include itemized charges for services and tests, and a transaction list of actions taken.

[3]Stephen G. Clyman, M.D., project director for Computer Based Exam, National Bored of Medical Examiners, personal communication, November 1991.

Figure 8-A1--CBX Case Computer Screen

Day I (Wed) Time 16:03 **Location: Emergency Department**

Vital signs (MD-recorded) **Day I @ 16:03**

Pulse rate (supine)	**I I 8 beats/min**
Systolic (supine)	**98 mm Hg**
Diastolic (supine)	**58 mm Hg**
Respiratory rate	**32/minute**

Chest/lung examination **Day I @ 16:03**

Thorax normal. Breath sounds absent on the left. Hyperresonance to percussion on the left.

Cardiac examination **Day 1 @** 16:03

Heart sounds faint. Radial, brachial, femoral and popliteal pulses weak but equal bilaterally.

***** ********** • ***

SELECT ANY FUNCTION KEY

***** ********** ****

FI-ORDER F2-H&P F3-REVIEW F4-CLOCK F5-PAUSE F6-HELP

SOURCE: K. E. Cotton and D.M.Durlnzi, *Computer Based Examination Software System: /-Phase II Update* **(Philadelphia, PA: National Board of Medical Examiners, 1990).**

Continued on next page

Box 8-A-Certification Via Computer Simulations: The National Board of Medical Examiners-Continued

Phase II entails formal field studies addressing the validity, reliability, utility, and practicality of the system and its derivative scores for use at the level of clinical clerkships. The testing software includes 8 CBX simulations and a 140-item multiple-choice examination. These examinations were administered at the completion of clerkships in surgery, pediatrics, internal medicine, and obstetrics-gynecology. Separate scores were generated for each measure in each discipline for over 1,700 students at 9 schools since 1989. Scores are generated by an automated scoring system that codifies criteria specified by expert clinicians and consist of an ability measure and flag score. The findings to date areas follows:[4]

- Student surveys indicated that students believed that CBX simulations were more representative of the materials in the clerkship and more effective in allowing demonstrations of what was learned in the clerkship than were the multiple-choice questions.
- Reliability of the CBX scorns in which them were large samples ranged from 0.70 to 0.80. These findings have been consistent across subjects, time, examinee level of training, and machine interface changes.
- The validity of the scores in this context is supported by multiple studies in which independent evaluations of average case performance by clinicians show high correlations with the CBX scoring systems.
- Correlations between multiple-choice and CBX scores in the same discipline are more moderate (0.37 to 0.50 corrected for the unreliability of the measures). Assuming the CBX scores are valid, as supported by the above-mentioned rating studies, this indicated that unique measurement information of merit in the evaluation of medical students is provided by both CBX and the multiple-choice questions.
- Analysis of multiple-choice questions compared the computerized versus paper-and-pencil versions. Students were ranked similarly on both versions, although the computerized multiple-choice version appears to be more difficult than the Paper-and-pencil version by about 25 standard score points (@.01), suggesting that use of norm data from the paper-and-pencil tests would be inappropriate for the computer-based version.

Several other research questions are being addressed. They include:[5]

1. Are the CBX scores valid as an interdisciplinary evaluation of senior medical students?
2. What are effective means for weighting the relative importance of items and defining pass-fail standards?
3. How comparable are different sets of simulations in providing equivalent challenges to examinees?
4. Can simulations be "disguised" and reused without jeopardizing test fairness and meaningfulness of scores?

In addition, the Nation Council of State Boards of Nursing has taken the CBX model and is in the process of adapting it to the model of nursing education, and researching its use for possible certification examination.

[4]Unpublished National Board of Medical Examiners data, cited in National Board of Medical Examiners, *CBT Phase II* (Philadelphia, PA: 1991).

[5](11 op. tit., footnote 3.

ing of the **physics concept because it is less** confounded with other skills such as reading level.[31]

Alternate modes of response can be used on the computer. Keyboarding reduces problems in interpreting handwriting, and the use of tablets, mouse, touch screens, light pens, and voice entry can provide new data entry formats. These new sources for data input also open doors for testing students with physical disabilities who may be unable to use traditional paper-and-pencil testing methods.

CBTs allow for improved standardization of test administration. For example, time allowed for any **given item can** be controlled, and instructions to test takers are not affected by variations in **presentation** by human examiners.

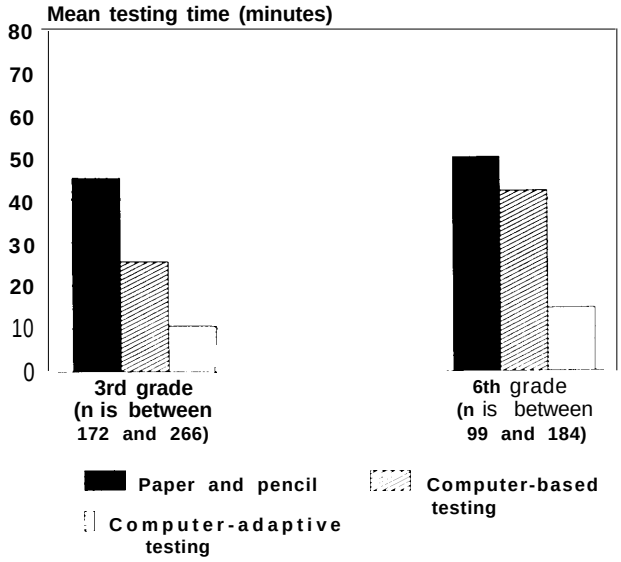
Scheduling of CBTs is more flexible, since not all students have to be tested at the same time.[32]

CBTs are not affected by measurement error due to erasures or stray marks on answer sheets. Young

[31]Wise and Plake, op. cit., footnote 22, p. 6.

[32]See, for example, Gerald Bracey, "Computerized Testing: A Possible Alternative to Paper and Pencil?" *Electronic Learning*, *vol. 9, No. 5,* February 1990, p. 16.

Figure 8-2—Mean Testing Time for Different Testing Formats



SOURCE: James B. Olsen et al., "Comparisons of Paper-Administered, Computer-Administered and Computer Adaptive Achievement Tests," *Journal of Educational Computer Research, vol.5, No. 3, 1989, pp. 311-326.*

children, who may have difficulty connecting an answer with its associated letter on a separate answer sheet, may have less trouble supplying their answer directly on the computer.

Computerized adaptive tests provide greater measurement accuracy at all ability levels than either CBTs or paper-and-pencil tests,[33] because *they can more accurately discriminate using fewer items.*

CBTs allow for immediate scoring and reporting; responses entered directly on the computer can be scored and tabulated in seconds, and scores can be reported back to the examinee and the teacher virtually instantaneously. Rapid feedback of this sort can be particularly important for teachers and more useful than paper-and-pencil tests that can require 6 weeks or more to be scored.

CBT allows for greater integration between instruction and assessment. Students working through lessons on an ILS can be assessed as they progress. Assessment can take the form of pauses in the instructional sequence during which students respond to questions or other prompts; with more sophisticated tracking software the assessment can take place on a continuous basis, providing information to teachers about student strengths and weaknesses as they work.

CBTs can provide more detailed information than paper-and-pencil tests. For example, student response time for any or all items can offer clues to student strengths and weaknesses; tests equipped with this feature can keep track of skipped questions, item-response times, and other possibly relevant data. This information can be useful to test takers as well as teachers.

CBTs provide a more efficient means to pretest new items, which can be inserted unobtrusively into any sequence of questions; faulty items can be eliminated and the computer can adjust its scoring algorithm accordingly.[34]

CBTs are more secure than paper-and-pencil tests. There are no paper copies of tests to be misplaced or stolen, items can be presented in mixed sequences to different students, and the number of items stored in memory is too large for anyone to attempt to memorize. Computerized adaptive tests have a particular security advantage: each test taker gets essentially a unique test.

Finally, CBTs may offer a set of less tangible advantages over paper-and-pencil. Among the issues researchers are exploring are: whether successful handling of the technology itself raises self esteem of students, especially developmental or low-ability students; whether rapid feedback reduces test anxiety; whether students become less frustrated and bored with CBT than with paper-and-pencil tests; and whether students are less embarrassed when results are given by the computer rather than by a teacher.

Disadvantages of CBT

CBTs may introduce new kinds of measurement error or may introduce new factors that compromise the accuracy of the results. For example, results on a mathematics or science test could be skewed if poor screen resolution interferes with the student's decoding of graphs or images; long reading passages requiring the examinee to scroll through many screens could favor students with ability to manipulate computer keys rapidly rather than gauge relative

33 Bunderson et al., op. cit., footnote 8, p. 385.

34 Wainer, op. cit., footnote 30.

reading comprehension proficiency .35 Input devices such as a mouse may be difficult for some students to operate, and current touch screens may not be accurate enough for sophisticated items requiring pointing and drawing. These issues suggest also that the lack of experience or familiarity with computers and keyboarding may put some students at a disadvantage compared to others.

Most CAT software, **because** of its branching algorithms, *prevents examinees from reviewing or changing an answer without changing all of the items following the changed ones. The* effects of this rigid sequencing on response patterns and cognition are not well understood.

Results of CATs are less obviously comparable to one another because each student's test is different in both the questions presented and the time allotted to finish. This may cause a perception on the part of students or others that test scores are somehow not a fair basis for comparisons.[36] These problems are aggravated by the general lack of familiarity with CAT on the part of test takers and the general public.

Ironically, *the computer might provide too much information:* teachers, parents, students, and administrators may be unable to digest the large amounts of data made available from CBTs.[37]

Reliability and validity of CBT generally and CAT specifically are important issues. Some studies have found that CAT can achieve reliability as high as conventional tests with far fewer items.[38] However, potential threats to validity and reliability warrant careful consideration: for example, issues related to content validity, effects of presentation mode on construct validity, potential negative effects on low-ability examinees, different contexts for item presentation, and the uses of data from conventional tests to set parameters of CATS.

Cost Considerations

Cost factors could pose formidable barriers to widespread adoption of CBT. Under current large-scale testing arrangements, when masses of students are tested at the same time, hardware requirements for CBT would be prohibitive. Scheduling students to be tested at different times could provide relief and would not necessarily create security risks, especially if a CAT model is used. But this approach would require drastic organizational changes from existing testing practice. Nevertheless, it may be possible to conduct some large-scale testing activities in shared facilities equipped with the appropriate testing hardware. Today's college entrance examinations are not offered in every school, but in selected sites on preselected dates; ETS is now considering setting up testing sites for administration of the GRE and professional certification examinations that are supplied with sufficient hardware to support CBT. These sites could be in schools or separate testing centers; in either event, the facility would be rented or leased by the test users (e.g., a professional association sponsoring certification examinations) for the time required to conduct the testing. Schools could adopt this shared facilities concept if it were necessary to conduct large-scale testing activities during a set time period.

Test Misuse and Privacy: A Further Caveat

Fully integrated instruction and assessment, hailed by some as the ideal approach to student testing, raises important questions related to test misuse and privacy. In a word, when testing is more closely linked to instruction it may become increasingly difficult if not impossible to prevent test results from being used inappropriately. It is precisely the tremendous recordkeeping and administrative efficiencies of CBT that pose this threat. To illustrate this concern, consider the ethical dilemmas that arise if students do not know they are being tested: as long as the information is used solely as feedback to teachers and students to improve learning, then there would be little objection. But if the results are used in high-stakes decisions such as graduation from grade school or placement into special classes (e.g., gifted or remedial) or made

[35] Research has shown that most people read 30 to 50 percent slower from a computer screen than from paper. Until screen resolution is improved significantly (e.g., 2,000 by 2,000 lines of resolution), this problem may not be resolved. Chris Dede, George Mason University, personal communication, Sept. 3, 1991.

[36] Green et al., op. cit., footnote 25.

[37] Olsen et al., op. cit., footnote 20, argue that too much information was provided to teachers on each child in the Texas pilot study. The solution was finally to print one page of analysis for each child accompanied by an order form for the teacher wanting additional information.

36 For example, a study of the California version of the Armed Services Vocational Aptitude Battery found that the alternate forms reliability coefficient for a 15-item California test was equivalent to that of a 25-item conventional test. Similar findings have been found in other studies. Wise and Plake, op. cit., footnote 22, p. 8.

available to districts and States for accountability measures, the concept of seamless integration of instruction and assessment becomes less obviously attractive. And, in addition to the ethical problems of using data derived from tests that students did not know were tests, there is also the danger that in the long run students (and teachers) will figure out how their test results are being used, which would lead to distortions in test-taking practice and teaching. "Teaching to the test" and other unintended effects of high-stakes testing (see also ch. 2), could undermine the value of integrated teaching and testing.

Other New Tools for Testing: Video, Optical Storage, Multimedia

Video technologies are the newest tools of instruction. The near ubiquity of videocassette recorders (VCRs) in schools makes the use of video more feasible for testing as well.[39] Furthermore, videodiscs and digital video interactive also offer new possibilities for integrating video capabilities in item presentation for more realistic kinds of tasks. Often new technologies are combined with older formats for innovative testing arrangements. In the Oregon Statewide Assessment test of listening skills, for example, prerecorded videotapes set the scene for questions, which are presented on traditional paper-and-pencil multiple-choice tests. Developers believe that the visual stimuli presented on the tape is more realistic and better than having questions read aloud from text. The system was first used as an element of the statewide assessment in the spring of 1991.[40]

A more sophisticated optical storage device now also coming into use in some schools is the videodisc: a large silver platter (resembling a long-playing record) that uses analog technology to store text, data, sound, still images, and video. Computer branching algorithms can be used to manage and sequence the vast amounts of information stored on videodisc; this coupling of optical storage and computing technology has already resulted in some powerful instructional applications, either in the form of enrichment materials or for courseware, some of which contain built-in testing and evaluation components. Researchers in this field anticipate new testing applications of videodisc in the future, given the capacity of the technology to store large amounts of multimedia items and integrate them with testing programs residing in the computer. Roughly one-fifth of American schools already own videodisc players.[41]

An application of videodisc to certification testing is the prototype developed by ETS to assess teaching and classroom management skills as part of the new National Teachers Examination. The experimental program presents filmed dramatizations of classroom management problems that typically occur in an elementary school classroom, and prompts the viewer to respond to each vignette. For example, after watching a scene the viewer may be asked to choose the teacher's next course of action; the choice activates a branch in the computer algorithm and displays the consequences of the choice.

Cost Considerations

As with many other instructional technologies, high costs of software development coupled with uncertainty and fragmentation on the demand side have slowed the development of innovative applications. However, if videodisc technology becomes a more common instructional tool in classrooms, software developers will face better prospects for return on their development investments. Without some sort of public intervention, it is unlikely the private market will produce the kinds of videodisc or other high-end technological innovations that could make a real difference in schools.[42] There is already some evidence that State education policies could stimulate growth in this market. For example, the decision of the Texas Board of Education to allow videodisc purchases with textbook funds is expected to lead to increased videodisc use in Texas schools, and. because of the large percentage of the school market that Texas represents, this policy is likely to spur increased videodisc development and use.[43]

---

[39] As of the 1990-91 school year, 94 percent of all schools have one or more videocassette recorders, Quality Education Data, op. Cit., footnote 17, p, T-8,

[40] Evelyn Brezinski, Interwest (Oregon), personal communication Jan. 3, 1991.

[41] Quality Education Data, op. cit., footnote 17, p. T-10.

[42] For analysis of the instructional software market and discussion of public policy options see Office of Technology Assessment, Op. cit., footnote 16, especially ch. 4.

[43] Peter West. "Tex. Videodisc Vote Called Boon to Electronic Media," *Education Week*, vol. 10, No. 13, Nov. 28, 1990, p. 5.

New Models of Assessment and the Role of Technology[44]

Most current uses of computer and information technology in large-scale testing make the conventional test format faster and more efficient than paper-and-pencil methods. The computer technologies have not, to date, created real alternatives to standardized multiple-choice tests.[45] Rather, the focus of computer applications has been on the familiar psychometric model, with enhancements that adapt the number, order, difficulty, and/or content of standard assessment items to the responses.[46]

There are two possible consequences that may spring from this replication. First, such a concentration may reinforce existing test and item formats by disguising them in the trappings of modem technology, creating a superficial air of advancement and sophistication. Moreover, these technical advances could make it even harder to break the mold of current testing practices, ignoring advances in test theory.

Using Information Technologies to Model Learning

How could computers and computer-related information technologies make possible enhancements to the current models of testing? How could these technologies be applied toward assessments of a broader range of human ability, cognition, and performance? Recent developments in cognitive psychology point to fruitful avenues for research and development (R&D).

First, human cognition and learning are now seen as *constructive processes: seeing,* hearing, and remembering are themselves acts of construction. Learners are viewed not as blank slates, passively recording and recalling bits of information, but as active participants who use the fragmentary cues permitted them by each of their senses to construct, verify, and mod@ their own mental models of the outside world.

Assessment procedures consistent with this view of cognition as an active, constructive activity are not limited to simply judging responses as correct or incorrect, but take into account the levels and types of understanding that a student has attained. Imaginative new types of test items are required to accomplish these ends, along with new techniques for scoring items that permit construction of dynamic models of the levels and types of learner understanding. Most if not all of these new techniques will require the use of computers. This work could lead to measures of human cognition and performance that are at present only dimly perceived, because of limited access and inexperience in measuring them.[47]

Second, some research on cognition holds that all learning is *situated* within "webs of distributed knowledge."[48] Cognitive performances in real-world settings are supported by other people and knowledge-extending artifacts (e.g., computers, calculators, texts, and so forth). This concept challenges traditional views of how to determine students' competence. If knowledge is tied in complex ways to situations of use and communities of knowers, then lists or matrices of abstracted concepts, facts, procedures, or ideas are not adequate descriptors of competence. Achievement needs to be determined by performances or products that interpret, apply, and make use of knowledge in situations. It follows from this view that estimates of learner competencies are inadequate if they are abstract or without context.

Computer-related technologies may be able to help integrate what is known about how children learn into new methods of assessment. This could include: diagnosing individualized and adaptive learning; requiring repeated practice and performance on complex tasks and on varying problems, with immediate feedback; recording and scoring multiple aspects of competence; and maintaining an

[44]Much of this discussion is based on Bank Street College, Center fOr Children and Technology, "Applications in Educational Assessment: Future Technologies," OTA contractor report, 1990.

[45]Walter Haney and George Madaus, "Searching for Alternatives to Standardized Tests: Whys, Whats, and Whithers," *Kappan,* vol. 70, No. 9, May 1989, p. 686.

[46]Dexter Fletcher, Institute for Defense Analyses, "Military Research and Development in Assessment Technology," unpublished report prepared for O'IA, May 1991.

[47]Ibid., p. A-2.

[48]Bank Street College, op. cit., footnote 44.

efficient, detailed, and continuous history of performances. There are four specific areas in which computer technology has begun to demonstrate the potential for significant enrichments to assessment.

Tracking Thinking Processes

Computers enable certain kinds of process records to be kept about students' work on complex tasks as the work evolves and is revised. They allow the efficient capturing of views of students' problem-solving performances that would otherwise be invisible, evanescent, or cumbersome to record. For example, it is possible to keep records of whether students systematically control variables when testing a hypothesis, to look at their metacognitive strategies, to determine what they do when they are stuck, how long they pursue dead ends, and so forth.[49]

Learning With Immediate Feedback

Because students can be put into novel learning environments where the feedback is systematically controlled by the computer, it is possible to assess how well or how fast different students learn in such environments, how they use feedback, and how efficiently they revise.

Structuring and Constraining Complex Tasks

Computer environments can structure and constrain students' work on complex tasks in ways that are otherwise difficult to achieve. In simulations, dynamic problems that may have multiple outcomes can be designed, and student progress toward solutions can be automatically recorded, including time, strategy, use of resources, and the like. The tasks can be designed to record students' abilities to deal with realistic situations, like running a bank, repairing broken equipment, or solving practical problems that use mathematics. They can show how students sift, interpret, and apply information provided in the computer scenarios, making it possible to measure students abilities in understanding situations, integrating information from different sources, and reacting appropriately in real time.

Using Models of Expertise

In more advanced assessment systems, models of expertise can be programmed and used to guide and gauge students' development of understanding in a subject area or domain. In this case, learning and its monitoring occur simultaneously as the expert system diagnoses the student's level of competence. This makes it possible to record the problem-solving process and compare the student's process with that of experts in the field.

Hardware and Software

Many types of hardware and software configurations apply to these concepts of assessment. *Telecommunications,* for example, is an important tool for sharing information about alternative assessment tasks. Vermont is using a computer network to share information on student portfolios that are now used for statewide accountability in mathematics and writing. Teachers will be able to share examples of work to help develop common standards of grading the portfolios, as well as to discuss teaching strategies and other concerns over the statewide electronic bulletin board.[50] As shown in box 8-B, another example is the use of technology in support of the demonstrations of mastery (' 'exhibitions' required of students in the Coalition of Essential Schools (see also ch. 6).

There are many examples of attempts to adapt *generic software tools* **to** assessment: word processors, database software, spreadsheets, and mathematics programs for statistical reasoning. These tools can be modified in order to record information in a sequence of work sessions and provide snapshots of students' processes in solving a problem or task. A word processor can record the stages of development of an essay; a spreadsheet program can record the steps taken in the solution of a multistage problem in mathematics. Because technology-based environments support accumulation and revision of products over time, they are well suited to portfolio models of assessment (see also ch. 6).

As teachers use these tools in teaching, it is appropriate that they be employed in testing situations as well. For example, when writing is taught as a process using a word processor, students develop

[49]This represents an extension of basic concepts such as the ''audit trail,' already in use in some instructional software, to assessment. For discussion of intelligent tutoring and related concepts, see Office of Technology Assessment, op. cit., footnote 16, ch. 7.

[50]Harry Miller, New England Telephone, personal communication% September 1991.

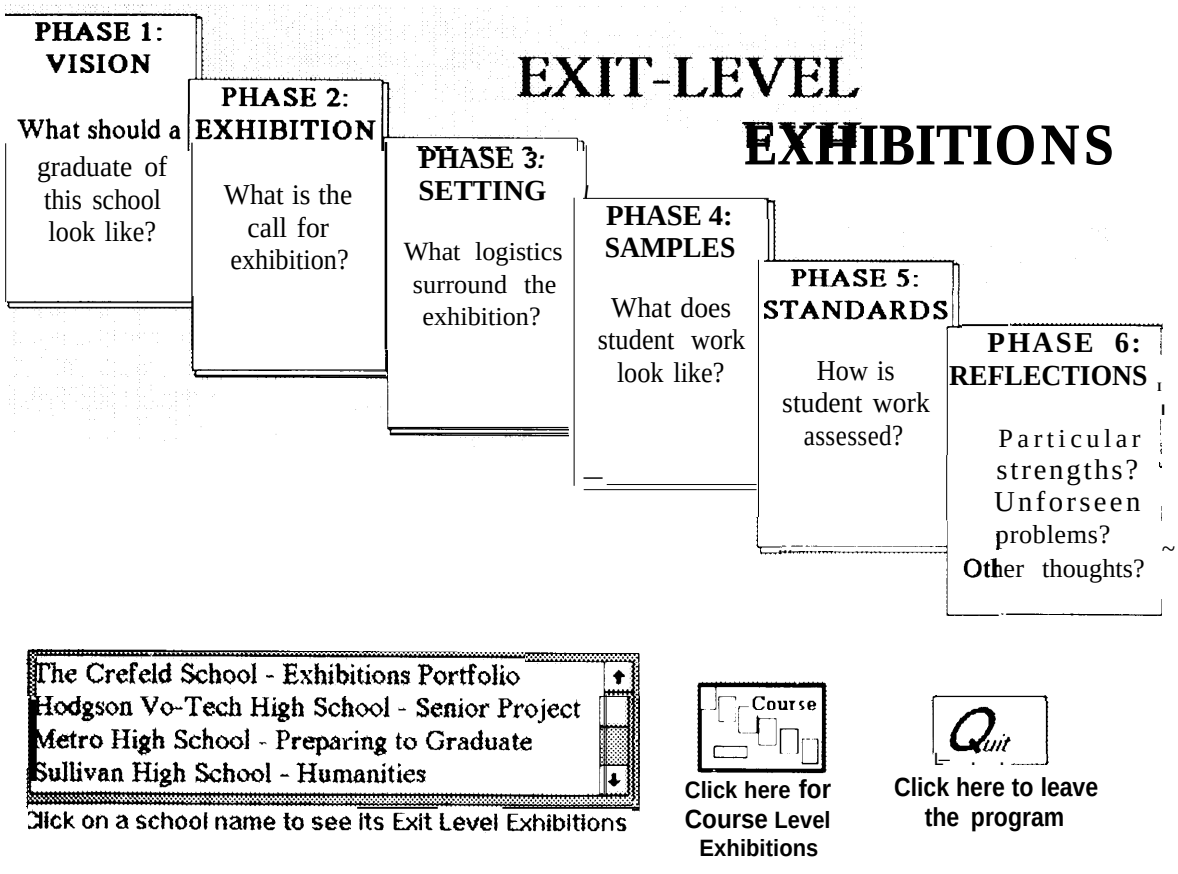
Box 8-B—The IBM/Coalition of Essential Schools Project: Technology in Support of "Exhibitions of Mastery"[1]

"Planning backwards' '-that is the term for how schools in the Coalition of Essential Schools determine what knowledge they want their students to possess, and what skills they want them to be able to demonstrate when they graduate. At Sullivan High School in Chicago, every member of the school community reads and participates in seminars discussing the works of great men and women, from Aristotle to Martin Luther King, in order to demonstrate their abilities to analyze and interpret works of original text. Seniors at Walbrook High School in Baltimore spend 1 year researching a specific question like "Is the city water safe to drink?" and must present findings, answer questions, and defend their positions before a panel of teachers and students, much like a Ph.D. student defending a dissertation. At Thayer High School in Winchester, New Hampshire, the faculty work in teams of four with a group of students for 3½ hours each day on a set of interdisciplinary "essential questions" chosen by the teaching team, allowing the students to show the connections among multiple disciplines.

These new teaching approaches require new assessment approaches. What is perhaps unique is how technology is being considered from the start as a tool for facilitating the restructuring that such "planning backwards" requires. IBM has committed $900,000 to the Coalition project at Brown University, along with equipment and technical

[1]Material for this box is from The Brown University News Bureau, "IBM and Brown University Select Five High Schools for National 'Exhibitions of Mastery' Project,'' news release, June 26,1991, and David Niguidula, Coalition for Essential Schools, Providence, RI, personal communication December 1991.

Figure 8-B1--Menu for Coalition of Essential Schools' Exit-Level Exhibitions



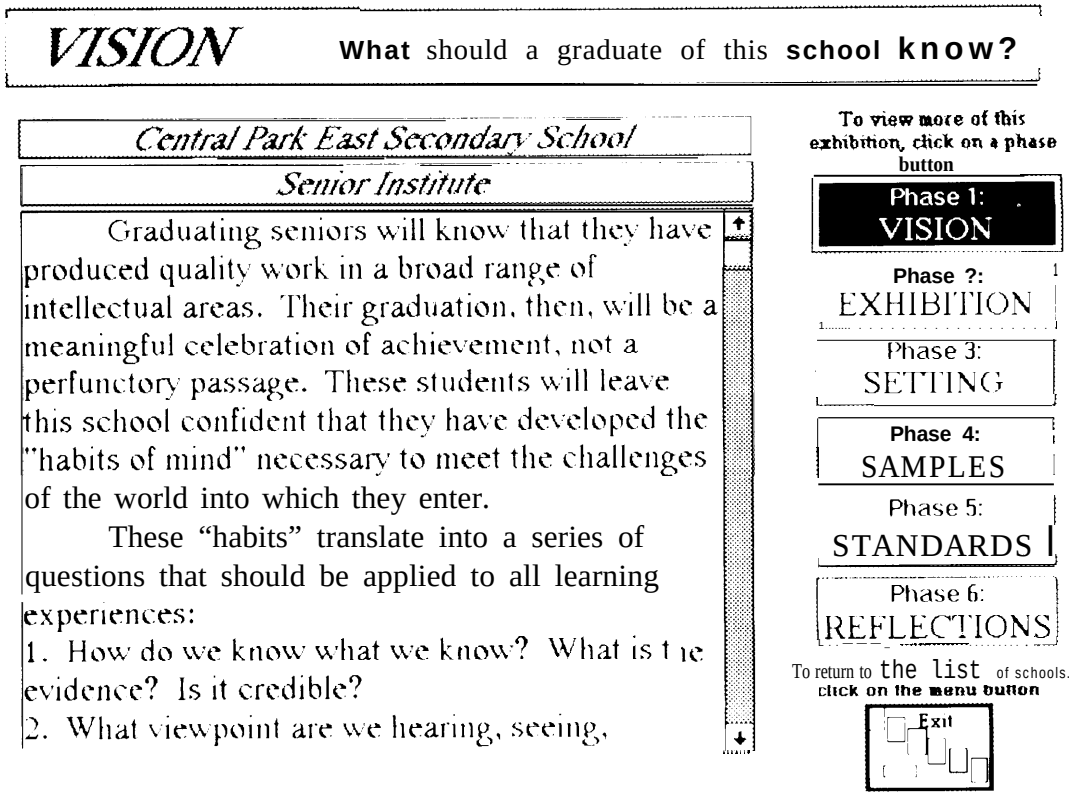
SOURCE: Coalition for Essential Schools, Brown University, Providence, RI.

support, to work with these schools and two others (Eastern High School, Louisville, and English High School, Boston) to examine how technology can facilitate the planning, development, and evaluation of the "exhibitions of mastery' assessment procedures at these schools. Technology is expected to be used in the following ways:

- Research: CD-ROM, videodiscs, computer databases, and telecommunications will be used for accessing and keeping track of information the teachers need for their teaching and the students need for their exhibitions.
- Student-Teacher Communications: Electronic mail will make it possible for information to be shared between students and their teachers both within a school and among sister exhibition schools. Project management will be tracked on the computer networks, and file transfers will be made so teachers can "red-pen" student drafts in progress.
- Performances: Tools such as word processing, desktop publishing, and multimedia will be used for creating student products.
- Assessment: Electronic portfolios of work in progress and records of student activity throughout the exhibition project will be created. Telecommunications will be used for assessing exhibitions within and among schools.

An electronic exhibitions resource center has been established by the 110+ member schools of the Coalition for Essential Schools. They are all contributing to this library of practical ideas, methods, and materials, which will be available on-line to help Coalition member schools create their own exhibitions. The exhibition resource center will provide a forum for discussing exhibitions and receiving updated information (see figures 8-B1 and 8-B2).

Figure 8-B2--Sample Screen When "Visions" is Selected From Menu



SOURCE: Coalition for Essential Schools, Brown University, Providence, RI; example from Central Park East Secondary School, New York, NY.

the skills of freewriting, drafting **ideas, writing a draft, revising, moving ideas around, editing-using all the tools of creation and revision provided by today's word processing software. To then test these writing skills using a paper-and-pencil examination would be as inappropriate as teaching a pilot to fly a jet and then testing his skills in a hang glider. Similarly, students taught to use calculators as mathematical tools should be tested on their ability to use these tools to carry out mathematical calculations.**

The tests under development for certifying architects provide an interesting example of how advanced tools available on computers can enrich test design and scoring. Examinees use the computer tools that allow them to draw, measure, calculate, change the size and scale of objects, and extract information from databases embedded within the testing software (see box 8-C).

Another category of software includes *simulations and modeling programs* **that** create highly realistic problem-solving contexts. Examples can be found in most domains, both in and out of school, and are available for computers in the schools. They enable students to observe, control, and make decisions about scientific phenomena that would otherwise be difficult or impossible to observe. For example, with *Physics Explorer,* students can conduct and observe a series of experiments that simulate the behavior of objects and phenomena under different conditions.[51] For example, a student can compare the upward acceleration of an object under different conditions of gravity. The assessment includes onscreen records of various experiments that are conducted; printouts of steps taken by the student in the form of note cards, experimental parameters, and sequences of decisions; and video recordings of students interacting with software and explaining their work. Scoring is based on understanding of interactions among parameters, appropriateness of experiments conducted, systematic approach to testing of variables, use of different information sources, nature of predictions and hypotheses, interpretation of experiments, and quality of group collaboration.

Other computer simulations enable students to carry out complex actions by simulating decision-

Photo credit: MECC

Wagon Train 1848, **created by MECC, is an example of an educational simulation program.**

making activity in the sciences, social science, history, and literature. For example, *Rescue Mission* is a simulation that allows elementary school students to navigate a ship to rescue a whale trapped in a net by learning the mathematics and science required to read charts, plot a course, and control navigation instruments.[52]

One of the most promising aspects of simulation software for education is the fact that this software is already in use and popular in schools today, and can be supported on relatively inexpensive computers. Simulation and modeling programs can provide multiple complex tasks and record how students go about solving them. They provide opportunities for assessing students' skill in such problem-solving activities as formulating the relationships between variables, troubleshooting or diagnosing problems, and integrating multiple types of information in decisionmaking.

Video and multimedia systems are **a third category** of technology with applications to new concepts of student assessment. **VCRs can** record the interactions of students in groups, and the ways they use aspects of their social and physical environment in accomplishing tasks. Video technologies can record continuing activities, products at various stages of development, explanations, and presentations in rich detail. The video record can be analyzed in minute

[51] Bank Street College, op. cit., footnote 44.

[52] Ibid.

detail over time, much as one would review a written record of performance.

The electronic integration of different media (video, graphics, text, sound) has made possible new multimedia opportunities for instructional environments and new, but relatively unexplored opportunities for assessment. These developments allow multiple forms of media to be stored and orchestrated on a single disk, simplifying the ease of use.

Although the technology for some of these projects is currently too expensive for average classroom use, costs are expected to drop as more powerful computers enter classrooms.[53] Some schools have begun to experiment with multimedia applications. The *Jasper Woodbury Series,* for example, presents a story through dramatic video segments, and enlists the student in solving problems using information provided through multiple linked databases (see box 8-D). *Jasper,* which is still in R&D, is being integrated into the science and mathematics programs in a number of schools that have expressed their willingness to experiment.[54]

Performance assessments often call for student-created productions or projects over time as a basis for evaluation, and multimedia systems can provide rich composition tools to meet this goal. In some systems, students can make use of the information (in graphic, text, or video formats) available within a multimedia system as they compose their own projects or productions. This makes new kinds of student products available for assessment purposes. Since students create these productions from within these ' 'closed' ' systems, traces of their creative composition process in choosing and composing information can be recorded.

Finally, *intelligent tutoring systems (ITSs), originally* conceived as instructional systems, have recently begun to be adapted to assessment. ITSs are based on principles of artificial intelligence and expert systems.[55] They combine models of what

Photo reedit: IBM Corp.

Ulysses, created for IBM Corp. by And Communications Inc., is an example of an advanced interactive educational program combining video, graphics, text, and sound.

constitutes expertise within a field or domain with models of the learners' own technique--diagnosing, evaluating, and guiding student performance compared to expert performance. Responses of students throughout the learning process can be aggregated and interpreted in relation to representation of expert problem solving. The systems offer the opportunity to understand student performance not simply in terms of correct answers, but in sequences of responses that can reveal how a student learns.

There are very few ITSs available today and their focus is typically on instruction, not assessment. They are extremely expensive to develop and require a higher level of computer technology than most schools own. The few in place cover circumscribed parts of the curriculum, and concentrate on the domains where computational power has the most leverage and where skills and content are more narrowly defined (e.g., science, mathematics, and computer science). It is unclear how feasible they would be in other areas that are more open-ended, such as history or literature.

---

[53] The digitial video interactive product *Palenque*, which allows users to ''eXp]Ore" the Mayan archaeological site via computer and screen, and to consult a variety of visual databases to gather additional data along the way, requires a hardware/software system costing approximately $20,000. It is currently being used in several science museums around the United States. See ibid., p. 26; and Office of Technology Assessment, op. cit., footnote 16.

[54] Jasper and other similar systems attempt t. capitalize on students' ever-increasing familiarity (and comfort) with television and video, and promotes the development of their skills in analyzing and using information provided via video format.

[55] 'Artificial intelligence asks the questions: what *is* the fundamental nature of intelligence and how can we make computers do the @S that we consider intelligent? . . . An expert system is an automated consultant. Given a problem, it requests data relevant to the solution. After analyzing the problem, it presents a solution and explains its reasoning. Expert systems are relevant to education because they can represent problem-solving expertise and explain to students how to use it. " See Henry M. Halff, ' 'Instructional Applications of Artificial Intelligence, " *Educational Leadership*, March 1986, pp. 24-26.

Box 8-C—Computer Technology for Professional Certification Testing: National Council of Architectural! Registration Boards

It is not surprising **that perhaps the most ambitious research on the use of computer technology for professional certification examinations is found in the field of architecture: architects** often look for creative solutions and new ways to solve problems, using the most advanced technologies. At the same time, because only one-half the States require architects to have a college education and only 60 percent of the candidates who sit for the architectural boards have a professional degree in architecture, the examination has traditionally played an important gatekeeping role, i.e., assuring that candidates who receive national certification meet high standards of skills and knowledge. Furthermore, since the number of candidates who seek certification is relatively small (each year only 4,500 candidates begin the examination process), field testing is more manageable than in other professions. [1] Several other professional groups are following this research with great interest before developing their own technology-based testing for professional certification.

Since 1965, all architecture candidates have been required to pass a multipart uniform paper-and-pencil national examination developed by the National Council of Architectural Registration Boards (NCARB). This examination, which has been revised periodically based on task analyses of the profession, currently consists of nine parts, seven of which are traditional multiple-choice tests of discrete knowledge in various architectural fields. Two sections require candidates to draw solutions to design vignettes; one section involves solving six discrete site design problems, while the other entails a comprehensive building design. These sections are scored by juries of practicing architects, similar in process to the scoring of Advance Placement examinations (see ch. 6).

Since 1985, NCARB has been working with the Educational Testing Service (ETS) in a joint research project to develop computer-administered examinations. The first phase of the research entailed converting four of the seven multiple-choice sections to Computer Mastery Tests. [2]

The computer mastery model uses item-response theory to select questions from the full item bank, reorganizing them into "testlets," each of which provides a collection of questions, which offers precise measurement of a candidate's ability. The items within a testlet are presented on the computer. When the candidate answers enough questions to determine that a passing or failing score has been achieved, testing ceases. If the outcome is unclear, more questions are presented until a clear pass-fail determination has been made. The computer mastery tests were pilot tested between 1988 and 1990. They successfully met the desired psychometric standards; the computerized tests achieved the same or better accuracy of measurement at the pass-fail point as that provided by the current tests, using as few as one-third as many test items as are needed in the paper-and-pencil version. However, because the computer tests were offered as an option to paper-and-pencil testing but were more expensive ($75 per subject as **compared to $35 per** subject for the paper-and-pencil format), not enough candidates opted for the computer version to make it economically feasible, Since 1990, only the paper-and-pencil version has been offered.

NCARB plans to switch over to computer-administered testing for all seven of the discrete knowledge sections in 1997, dropping the paper-and-pencil option altogether. At that point, the second research activity will also be put into place. This project involves administering the test (examinees use a mouse or other pointing or drawing device to design directly on the computer screen), and scoring the discrete site design vignettes directly on the computer. Field testing has shown that design problems that take an average of 20 minutes on the paper-and-pencil version require only 5 minutes to complete on the computer, because of the ease of erasing, redrawing, and adjusting drawings. As a result of this research, NCARB expects to be able to present candidates with up to 15 vignettes to solve, compared to the existing 6, in the same period of time (see figure 8-Cl).

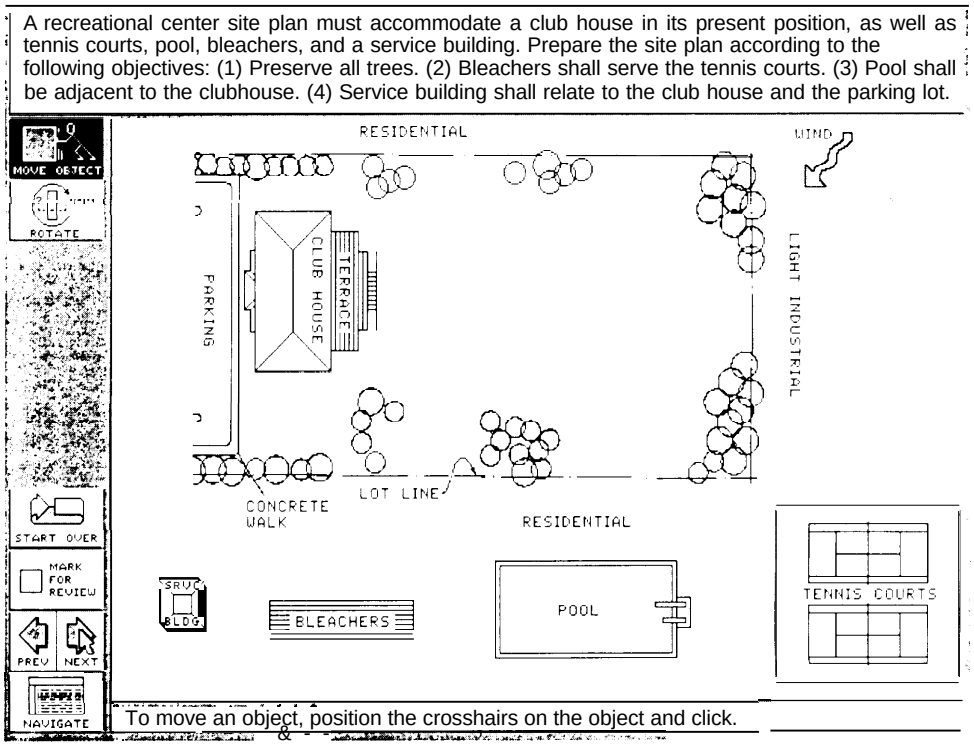
Finally, the comprehensive building design problem is being converted to a computer-administered examination as well. In this case, each candidate will use two computers, one which presents and serves as the ‘‘answer sheet’ for a candidate's design solutions to comprehensive, multistep design problems; the other monitor provides the “model architect's office,” containing all the design tools, resources, and reference manuals needed

---

[1] Jeffrey F. Kenney, director ‘‘ Examinations Development National Council of Architectural Registration Boards, personal communication October 1991,

[2] ‘Breakthrough Development in Computerized Testing offers Shorter Test, More Precise Pass-Fail Decisions,” *ETS Developments*, vol. 33, Nos. 3 and 4, winter/spring 1988, pp. 3-4.

Figure 8-CI—Example of an NCARB Site Design Vignette



KEY: NCARB - National Council of Architectural Registration Boards.
SOURCE: Educational Testing service, 1991.

to complete each task. Each of the substeps in the comprehensive design problems will be presented as separate sections and scored separately. For example, a candidate may be asked to design a library that meets certain site and client requirements. In the first step, the "bubble diagrams' that relate rooms to one another would be drawn. A second section would require taking a block diagram and relating it to the site requirements in terms of light, ground contours, zoning, and other constraints. Each of these individual predesign tasks will be scored separately, making it possible to give a candidate partial credit, instead of scoring the building problem as a whole, as is done in the existing paper-and-pencil format.

In 1989, six different item types were developed for the simulations and pilot tested at an architectural firm and the NCARB annual meeting. It is anticipated that the computer format will permit more reliable assessment of candidates' abilities. Whereas it now takes candidates up to 12 hours to complete 1 comprehensive problem (typically **4** hours to come up with a design and another 8 to put it down on paper), using the computer simulations broken down into subtasks, up to 10 design samples can be presented over a period of 5 to 6 hours. It is anticipated that perhaps 3 comprehensive building problems, with a total of some 20 to 30 subtasks can be administered for this portion of the examination over the same time period, **giving the State board examination** a fuller and more reliable picture of an architect's design skill and ability to meet the necessary health, safety, and environmental standards.

Researchers are encouraged by the progress made in the design of the computer interfaces; indeed, erasers, drafting tools, measuring tapes, calculators, and other design tools that make it possible to move and adjust drawings are available in many computers today, as are the appropriate data storage and retrieval capabilities. The hardware required is Windows-based 386 machines with approximately 4 megabytes of memory. Advances in object-oriented

Continued on next page

Box 8-C-Computer Technology for Professional Certification Testing: National Council of Architectural Registration Boards-Continued

programming make it possible to use icons for frequently used architectural components (e.g., corridors, doors, walls, and windows). It is the development of scoring psychometrics that poses the largest research challenge. In order to develop scoring protocols, each solution must be decomposed or broken into component parts. Seasoned practitioners list the characteristics of an appropriate solution to a particular problem and these judgments are programmed into the computer. The computer, functioning as an expert system, evaluates the examinee's response. In cases where the expert system is unable to make a clear "right or wrong" judgment (similar to the case with live panels of judges when two scorers disagree), a master scorer will be brought into make a final determination.

Although the original target goal was 1997,[3] the NCARB/ETS team has moved along further than originally anticipated and, if progress continues as the same rate, implementation of a fully computer-administered and scored examination system could be possible in 1996.[4]

[3]Richard DeVore, senior examiner, Center for Occupational and Professional Assessment, Educational Testing Service, personal communication, Oct. 15, 1991.

[4]William Wiese II, "License Exams by Computer," vol. *179, No. 7,* July 1991, p. 80.

One of the **greatest** concerns with ITSs is that, like all testing activities, they may gravitate toward promoting the skills that they measure best. These skills tend to be algorithmic and routine. At the same time, educators are concerned that we may not be focusing our efforts on developing in students those thinking skills dependent on complex knowledge. The skills required for understanding a written passage, writing a composition, solving a problem that has many steps and approaches, interpreting the results of an experiment, or analyzing an argument are not so easily broken into discrete components. Furthermore, attempts to segment these skills may result in analysis that fails to capture the overall picture of what makes up true competence. Creativity may be neither recognized nor rewarded in existing ITS models.

Toward New Models of Assessment: Policy Issues

A main finding of this chapter is the gap that separates current applications of information technology in testing from a vision of fundamental reform in the assessment of human learning and educational achievement. In sum, computers and other data processing equipment that have made possible a "mass production" testing technology could become essential in the design and implementation of new testing paradigms.

Computers and related technologies have proven indispensable to research on human cognition, and lessons from this research are, in turn, being applied-also with the help of sophisticated computer-based systems—to the design of educational assessments that correspond to the growing body of research on learning. The research community, though still fragmented, has begun to coordinate the efforts of cognitive theorists, computer scientists, subject matter experts, and educators. These early efforts have led to particularly promising breakthroughs in the application of technology to improved classroom diagnosis and instructional feedback. Whether these efforts will eventually also contribute to the creation of tests that can be used for other functions, such as system monitoring or student placement and certification, remains to be seen. In any event, it is not clear that these latter functions of testing require the diagnostic specificity of computer-based learning and assessment tools. Overall, most experts would agree that applications of computer technology to new forms of assessment are still at a very rudimentary stage. The road ahead is a long one.

Research Support

Policymakers face a formidable dilemma: reaching the as-yet uncharted territory of *new* assessment models requires investments in technologies that have uses in the current paradigm of testing and that render that paradigm ever more efficient. Increased efficiency encourages reliance on old models of testing. This problem is manifest in the arena of funded research: much of the research on test theory and new technology is funded by commercial test companies, which face strong incentives to reinforce

Box 8-D—The Jasper Series: A Story is Worth a Thousand Questions

The National Council of **Teachers of Mathematics has suggested that the mathematics curriculum should:**

> . . . engage students in problems that demand extended effort to solve. Some might be group projects that require students to use available technology and to engage in cooperative problem solving and discussion. For grades 5-8 an important criterion of problems is that they be interesting to students.[1]

The Jasper Woodbury Problem Solving Series is **a video-based adventure series designed to help students develop their skills in solving mathematical problems.[2] Each of the six video segments is from 14 to 18 minutes long and presents a dramatic adventure featuring Jasper and his friends. Students are motivated to solve the problem posed at the end of each segment to see how the story ends. (There is a solution shown on the video that students see only after they have solved the problems themselves.) Although the problems are complex and require many steps, all the data needed to solve the problems are contained as a natural part of the story.**

For example, the adventure "Rescue at Boone's Meadow" begins with Jasper's friend Larry flying his ultralight airplane. Larry teaches Emily the principles she needs to know in order to fly solo in the plane: fuel capacity, speed, payload limits, how the shape of the wing produces lift, and so on. After Emily's maiden solo flight, she, Larry, and Jasper celebrate at a local restaurant. They discuss Jasper's upcoming fishing trip, and his plan to hike 15 miles into the woods at Boone's Meadow. **Details presented as a part of the unfolding adventure become important data that students will later need to use in solving the problem. The next scene shows Jasper alone in the deep woods, peacefully fishing, when a shot rings out. He runs in the direction of the sound, finds an eagle that has been** seriously wounded, and radios for help on his CB radio. Emily receives his message, contacts a local veterinarian, and is told that time is of the essence in rescuing the eagle. The story ends with Emily posing the question: "What's the fastest way to rescue the eagle and how long will it take?" The students, no longer passive watchers, have to put themselves in the role of Emily and solve the problem using data contained in the video.[3]

Researchers, working with teachers and students in 9 States, have found that students become extremely engaged in the problem-solving tasks. Teaching strategies vary, but most teachers begin with large group activities and then move into smaller cooperative learning groups, guiding the students to consider a variety of solutions. In the episode summarized above, for example, if the students contemplate using the ultralight plane as a rescue vehicle, they must take into account landing area, fuel consumption, payload limitations, speed, and other information that can be reviewed by going back into the videodisc. Groups typically spend a minimum of two 1 -hour class periods working out their solution, and then must present and defend their plan to the entire class.

One of the research goals has been to create new ways to assess the learning that occurs in solving problems presented in the series. One-on-one interviews with students were found to be much too time consuming. Paper-and-pencil tests were developed, asking students to list and explain the kinds of subproblems that Jasper and his friends needed to consider to solve each problem. Transfer problems, similar to the problems in the series but involving new settings and data, were also given. Although the paper-and-pencil assessments showed that learning occurred, there was one problem: teachers and students hated **them! Teachers said: "My kids, as much as they liked** Jasper, as much as they begged for Jasper, finally told **me: 'If** I have to take another test on Jasper I don't want to see another Jasper' "; or ''it seems to me that we're really asking kids to do something strange when we've introduced this wonderful technology and we've gotten them involved in the video experience. . . . Then you give them this test that's on paper."[4]

How then **should the students be tested? One approach has been to explore ways technology can be used in the assessment process. In May of 1991 the researchers produced an experimental teleconference, the** *Challenge Series,* **a game** show format featuring three college students as contestants, each of whom claimed to be an expert

---

[1]National Council of Teachers of Mathematics, *Curriculum Standards* (Reston, VA: March 1989).

[2]The series is a research and development project of the Cognition and Technology Group at Vanderbilt University, supported by the James S. McDonnell Foundation the National Science Foundation, and Vanderbilt University.

[3]Cognition and Technology Group at Vanderbilt University, ''The Jasper Experiment: An Exploration of Issues in Learning and Instructional Design,'' July 26, 1991, p. 7 (forthcoming in Michael Hannafii and Simon Hooper (eds.), *Development*, special

[4]Cognition and Technology Group at Vanderbilt University, "The Jasper Series: A Generative Approach to Improving Mathematical Thinking,'' pp. 11-12 (forthcoming in American Association for the Advancement of Science, *in*

Continued on next page

Box 8-D—The Jasper Series: A Story is Worth a Thousand Questions-Continued

on flight and on the Jasper adventure 'Rescue at Boone's Meadow. ' While the contestants all answered questions correctly on the first round, by the fourth round everyone except the true expert had made some erroneous arguments. Would the students be fooled by actors, or could they identity the real expert? They called in their votes and 85 percent of the students correctly identified the true expert. Enthusiasm for this form of "testing" was sky high.

Other ideas building on the teleconference motif are being considered for each of the Jasper adventures. There are also plans to help teachers engage in formative evaluations of student learning following each Jasper adventure with video-based ''what if'' analogs like the ones used to prepare for the *Challenge Series* teleconference. Spinoff vignettes that connect with other parts of the curriculum (e.g., an exploration of Lindbergh's historic flight from New York to Paris) are also in progress. Finally, the researchers are designing a prototype set of computer-based "students" or "tutees." The students must teach the "tutees" how to solve Jasper problems, and their progress is tracked by the computer. This approach maybe linked with the teleconferences. For example, the students could teach computer-based tutees, who would then compete in a game show where the tutees become game show contestants. The class that did the best job teaching its tutees wins.

The seven design principles underlying the Jasper Series, and their hypothesized benefits, are summarized in table 8-D1.

Table 8-D1--Seven Design Principles Underlying the Jasper Adventure Series

Design principle	Hypothesized benefits
1. Video-based format	a. More motivating. b. Easier to search. c. Supports complex comprehension. d. Especially helpful for poor readers yet it can also support reading.
2. Narrative with realistic problems (rather than a lecture on video)	a. Easier to remember. b. More engaging. c. Primes students to notice the relevance of mathematics and reasoning for everyday events.
3. Generative format (i.e., the stories end and students must generate the problems to be solved)	a Motivates students to determine the ending. b. Teaches students to find and define problems to be solved. c. Provides enhanced opportunities for reasoning.
4. Embedded data design (i.e., all the data needed to solve the problems are in the video)	a Permits reasoned decisionmaking. b. Motivates students to find. c. Puts students on an "even keel" with respect to relevant knowledge. d. Clarifies how relevance of data depends on specific goals.
5. Problem complexity (i.e., each adventure involves a problem of at least 14 steps)	a. overcomes the tendency to try for a few minutes and then give up. b. Introduces levels of complexity characteristic of real problems. c. Helps students deal with complexity. d. Develops confidence in abilities.
6. Pairs of related adventures	a. Provides extra practice on core schema b. Helps clarify what can be transferred and what cannot. c. illustrates analogical thinking.
7. Links across the curriculum	a. Helps extend mathematical thinking to other areas (e.g., history, science). b. Encourages the integration of knowledge. c. Supports information finding and publishing.

SOURCE: Cognition and Technology Group at Vanderbilt University, "The Jasper Experiment: An Exploration of Issues in Learning and Instructional Design," Juty 26, 1991 (forthcoming in Michael Hannafin and Simon Hooper (eds.), *Education Technology Research Development*, special issue).

the economic and educational advantages of the conventional test paradigm. This is in contrast to the test development process in other countries, which is usually undertaken or supported wholly by the government. Just how far the commercial research community will go in experimenting with nontraditional test designs, without external support, is uncertain.

It is important to recall, however, that Federal intervention frequently played a critical role in the history of research, development, and implementation of new testing technology: perhaps the best example is the Army testing program during World War I (see also ch. 3), which provided the most fertile ground imaginable for proving the feasibility of new forms of testing, such as group administration, as well as statistical models based on normative comparisons and rankings.

Indeed, the military has since then remained a major player in the development of personnel selection and placement tests, assessments of basic job skills, and experimentation with a variety of models of performance assessment. Some of these advances have spilled over into the civilian arena.[56] In addition, there is the more recent example of National Science Foundation (NSF) support for research leading to development of tasks used in the 1988 National Assessment of Educational Progress (NAEP) science assessment. Not only were these items viewed as important innovations in NAEP, but many of them were then adopted by New York State for its statewide fourth grade hands-on science assessment. Similarly, Department of Education funding for NAEP has supported research into constructed response items and innovative testing formats. Thus, while federally funded research on assessment has not been large, it has been an important complement to the large R&D projects financed privately-such as those by ETS, ACT, the National Council of Architectural Registration Boards, the National Board of Medical Examiners, and computer companies such as IBM and Apple--or financed by States and districts, such as in California and Portland, Oregon.

The history of testing in the United States teaches that the Federal Government can be a catalyst for reform, through support for expansion of existing technologies and through support for basic research leading to new technologies. The Federal Government could continue to support basic research and applied development of a wide range of new models of testing. Specific options include:

- earmarking resources in programs like Chapter 1 for research into how advanced technologies can improve testing;
- continuing to fund educational laboratories and centers for school-based research on assessment;
- providing grants to independent researchers, States, and school districts through NSF or other existing programs;
- coordinating the efforts of the many research players both within and outside the Federal Government's research network, i.e., Federal laboratories, the National Diffusion Network, NSF Net, and Star School Programs in support of improvements in testing; and
- supporting the exchange of data among the many States and districts involved in pioneering theoretical and practical research.

Infrastructure **Support**

If computers, video, and telecommunications technologies are to play a significant role in assessment, a combined "technology-push/market-pull' strategy will be necessary .[57] Technology-push in this context focuses on the technology of software, and is shorthand for software development support that could lead to increased demand for computer-based instructional and assessment systems in schools. The market-pull side of the equation refers to direct investments in hardware: increasing the installed base of technology in the schools could lead to increased demand for good software, which could in turn create improved economic incentives for software developers and entrepreneurs.

To make inroads in this interrelated system, the Federal Government could support investments in CBT facilities that could be shared among schools within and across districts. This could entail investments in communications technologies to link hardware already in place, along with software and training. Another approach would be for schools to

[56] The flow has gone in the other direction to assessment techniques developed for educational institutions have been adopted by the military.

[57] See also Office of Technology Assessment, Op. Cit., footnote 16, for discussion of this approach to fostering improved instructional software development.

lease their computer facilities to the Federal Government for use in its large education and training programs, or to other outside users (adult education, business, professional groups). The idea is to utilize the capacity of the hardware that exists in schools now, or the hardware that could be installed in the schools, during nonschool hours, and to reinvest the revenues in testing-related hardware or software technologies. Federal support for purchase of multipurpose computer and video technologies for testing **activities under existing Federal programs, such as Chapter 1,** Magnet Schools, and Bilingual Education could **build up the infrastructure of testing technologies.**

Continuing Professional Development for Teachers

Teachers are the most important link between instructional or testing technologies and the students whose achievement and progress those technologies are intended to affect. The problem is that few teachers have adequate preparation in the theory and techniques of assessment. This gap in teacher education is not limited to the arcana of psychometrics, but extends even to the design and interpretation of classroom-based tests.[58] At the same time, many teachers have not yet come ''online' with computer use.[59] While teachers may be learning about computers faster than about testing and assessment, most teachers have not been exposed to continuing professional development aimed at helping them master the implications of matching technology and new approaches to testing.

Federal support for teacher development could have two benefit streams: first, it could result in greater acceptance of new testing and assessment technologies, which would in turn lead to heightened demand for innovative software products; and second, it could involve teachers in the early stages of testing technology development, which could make the technologies that much more relevant.

Photo credit: Educational Testing Service

Teachers need help in learning to use teaching technology for testing purposes.

Leadership

In 1990, the President and the Governors adopted ambitious education goals to be met by the year 2000, and there has been much discussion on developing new tests to measure success in meeting these goals. The Federal Government has the opportunity to provide guidance in a time that has been marked by many suggestions for improvement and much accompanying confusion. Congress could take a leadership position in guiding, shaping, and supporting a vision of education that links learning with assessment in a rich, meaningful, engaging, and equitable fashion.

[58] See, e.g., John R. Hills, "Apathy Concerning Grading and Testing," *Phi Delta Kappan*, vol. 72, No. 7, March 1991.

[59] See, e.g., Bank Street College, op. cit., footnote 44.