A POLICY INFORMATION PERSPECTIVE

Reinventing Assessment

by Randy Elliot Bennett

SPECULATIONS ON THE FUTURE OF LARGE-SCALE EDUCATIONAL TESTING

POLICY INFORMATION CENTER
Research Division
Educational Testing Service
Princeton, New Jersey 08541-0001

Table of Contents

Preface iii

Acknowledgments v

Introduction 1

First-Generation Computer-Based Tests 3

Next-Generation Electronic Tests 5

Generation “R” 11

Conclusion 15

References 17

PREFACE

Reform in elementary and secondary education remains in the forefront of the public's mind, and therefore also remains high on the agenda of elected officials and educators generally. And the demand for performance and accountability is spreading to higher education as well. Most of these reforms rely on testing — testing to show increased rigor of school curricula, testing to determine if students advance and graduate, testing to judge the effectiveness of schools and teachers, and testing to compare districts, states, and nations.

Despite the increased demands on testing, Randy Bennett points out that large-scale assessment has little changed over the last two decades. And that cannot continue. Bennett closes this brief report saying: "…large-scale assessment must change in the most fundamental ways, for nothing short of reinvention will prepare it to meet the dramatically different demands it will soon face." Bennett tells us what these demands are and might come to be. And he gives us a glimpse into the future of the testing enterprise at different stages of change and reinvention.

Paul E. Barton
Director
Policy Information Center

Acknowledgments

This report is based on an invited paper presented to the National Research Council's Board on Testing and Assessment, Orlando, February 1997. I appreciate the helpful comments of Paul Barton, Isaac Bejar, Nancy Cole, Larry Frase, and Bob Mislevy on an earlier version of the manuscript. The positions presented are, however, my own and not necessarily those of the reviewers, the National Research Council, or ETS. I am also grateful for the production contributions of Al Benderson, Carla Cooper, and Kelly Gibson, without whom this Policy Information Center publication would not have been possible.

INTRODUCTION

Large-scale educational assessment consists of those tests administered to sizable numbers of people for such purposes as placement, course credit, graduation, educational admissions, and school accountability. It includes group-administered, standardized tests used most often in the secondary through postsecondary years.

Although there has been much recent intellectual ferment and experimentation in educational assessment, the practice of large-scale testing is much the same today as it was 20 years ago. Most large-scale tests still serve only institutional purposes, are administered to big groups in single sittings on a few dates per year, make little use of new technology, and are premised on a psychological model that probably owes more to the behaviorism of the first half of this century than to the cognitive science of the current half.

There are good reasons to believe that this situation is about to change. Perhaps most important is the emergence of a global economy, which has created a general business climate of intensified competition (Sherman, 1995). Many domestic companies now find themselves not only competing on the home front with new foreign rivals but, at the same time, having to expand to offshore markets where they may encounter more competition. This environment has encouraged successful U.S. businesses to emphasize, among other things, innovation that responds to rapidly shifting market needs, productivity, and customer service as keys to competitive advantage (Treacy & Wiersema, 1995, p. 9). To foster these innovation, productivity, and customer-service goals, businesses are capitalizing upon rapid advances in hardware, software, and communications technology. Finally, our changing demographics — especially, the growth of minority populations — is making responsiveness to diverse customers yet another path to competitive edge.[1]

These market forces are having substantial impact on our society generally, so it is understandable that they are affecting large-scale educational assessment too. The testing enterprise has grown dramatically during the past several decades, attracting a larger corporate presence and fueling competition. (Witness the purchase of testing and test-preparation companies by big corporations, and the partnership of nonprofit agencies with for-profit ones.) At the same time, constituents are beginning to expect the same things from testing agencies that they get in everyday commerce: innovation, productivity (as reflected in competitive pricing), and customer service. Increasingly, they are expecting adaptation to diversity too.

[1]Diversity may belong under market needs; however, it is such an important issue in educational testing that I have categorized it independently.

This paper offers a scenario for how educational assessment might change in response to these forces. In doing so, the paper seeks to stimulate thinking about the future of large-scale testing. This thinking should include the development of alternative scenarios that can be played off against one another to help define valued characteristics for future assessments.

The scenario divides into three generations distinguished by purpose of testing, test format and content, test delivery location, and the extent to which testing capitalizes on new technology. The first generation lays the basic infrastructure for electronic testing. In the second generation, large-scale tests undergo qualitative change, but their purposes and delivery mechanisms remain essentially the same. The last generation brings a rethinking of the purposes and mechanisms of large-scale assessment.

First-Generation Computer-Based Tests

The early effects of innovation, customer service, and (to a lesser extent) productivity can be seen in the first generation of computer-based tests just emerging.

This first generation combines advances in psychometrics with technology to deliver large-scale tests adaptively. The computer selects questions based in part on previous responses, tailoring the test to individual skill levels. Depending on the testing program, individuals can register by phone or email; pay by credit card; test by appointment in a relatively small, comfortable center; and receive scores at the conclusion of the session. Testing organizations can electronically exchange questions and examinee responses with test centers, and send scores to institutions in the same way.

In the 1997-98 academic year, relatively few tests were offered in this mode, accounting for roughly a million examinees. Among those tests were the Graduate Record Examinations (GRE) General Test (offered as an alternative to the paper version), the Graduate Management Admission Test, Praxis I (for entry into teacher education programs), the SAT I: Reasoning Test (only for students applying to special precollege programs for the academically talented), ACT's COMPASS, and the College Board's ACCUPLACER (the last two being for placement in first-year college courses). However, in the next several years, volumes will increase dramatically as the Test of English as a Foreign Language comes on line and the GRE General Test eliminates all paper examinations. In addition, the Law School Admissions Council has started exploratory work on a computerized version of the LSAT, and the National Assessment Governing Board has recommended that the next NAEP contractor be required to computer administer the National Assessment in at least one subject at one grade level (National Assessment Governing Board, 1996).

Whereas there is certainly a concerted move toward electronic large-scale tests (especially in educational admissions), there is no question that this assessment mode is still in its infancy. Like many innovations in their early stages, today's computerized tests automate an existing process without reconceptualizing it to realize the dramatic improvements that the innovation could allow. Thus, these tests are substantively the same as those administered on paper: they measure the same skills, use the same behavioral designs, and depend primarily on the same types of tasks. In addition, like many initial implementations, this first generation has increased costs since, at least for high-stakes programs, item pools must be continuously replenished to support ongoing administration.

To become firmly established, computer-based assessment will clearly have to offer much more. Fortunately, this first generation may do just that. In fact, its most important legacy may be the basic infrastructure for a new generation of tests that uses technology more pointedly to innovate in response to education community needs, improve productivity, enhance customer service, and accommodate diversity.

Next-Generation Electronic Tests

The next generation of large-scale electronic tests will steadily incorporate advances in technology, psychometrics, and to a growing extent, cognitive science.[2]

These tests will still be largely aimed at creating value for institutional constituencies, and they will continue to be administered in dedicated centers.

The typical test in this period will be qualitatively different from those of the first generation. This difference will be evident in the test questions (and, in some cases, the characteristics they measure), as well as in development, scoring, and administrative processes. The impetus for this shift, of course, predates electronic testing. It comes from sustained criticism by educators (NCTM, 1989), the measurement community (Mislevy, 1993a; Shepard, 1992), cognitive psychologists (Pellegrino, 1992; Resnick & Resnick, 1990, 1992; Sternberg, 1992), and the public, challenging the relevance of standardized tests.

A major impediment to making these changes has been the cost of delivering alternative forms of assessment (Office of Technology Assessment, 1992, p. 243). This next generation will deliver qualitatively different tests cost-effectively, opening the way to significant change.

Logically, change should occur first in test design which, as noted, today reflects an outmoded psychology. Design, however, will probably not be the first element to change because existing designs may well be adequate for the limited selection purposes served by most tests (Everson, in press; Mislevy, 1996). Also, design change is fundamental and will be a difficult shift for institutional testing programs to make. It *will* occur, but only as testing purposes themselves transform.

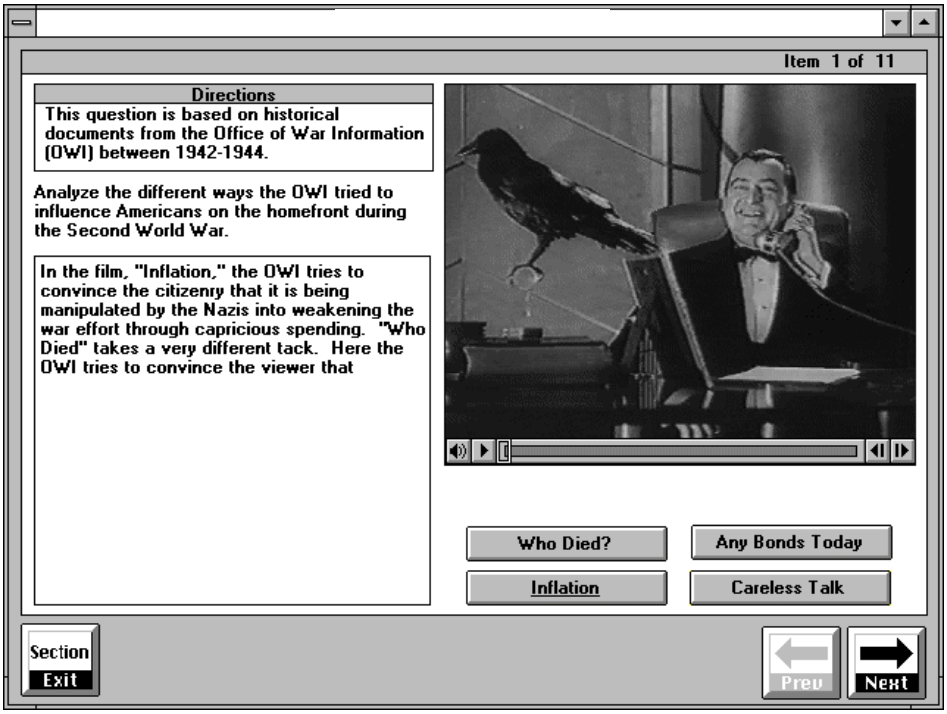
Widespread change will come first in the nature of test questions and the formats for response, and the possibilities both provide for measuring new skills. In conventional testing, feasibility concerns dictate that audio or video be used only when a critical skill cannot be adequately assessed by standard methods (e.g., when measuring foreign-language listening skills). Even under these circumstances the temporary nature of conventional test centers makes using audio or video an administrative burden. However, the advent of permanent computer-based test centers with equipment capable of delivering high-quality multimedia will allow us to

[2] Much has already been learned from cognitive science about how to improve ability and achievement testing (Glaser, 1991; Sternberg; 1991). So why is this discipline likely to have limited influence on tests of the next generation? One reason is that, as bureaucracies with multiple competing constituencies, large-scale assessment programs change incrementally. From the perspective of many program directors and test developers, the move to computers is change enough. Even so, the importance of this transition is readily understandable because computers are infusing most other parts of our society. The need to rethink tests to reflect cognitive principles is not as well appreciated and, with few exceptions, serious efforts at conceptual redesign have not yet begun.

administer tests incorporating sound and video more efficiently. Perhaps more importantly, it will permit us to introduce audio, video, and animation in the assessment of other skill areas.

One result of this introduction will be the capability to measure traditional skills more comprehensively (Bennett, Goodman, Hessinger, Ligget, Kahn, Marshall, & Zack, in press). For example, most introductory college history courses include the analysis of artifacts. To reflect this emphasis, the Advanced Placement Examination in U.S. History asks students to use artifacts (e.g., excerpts from diaries, news articles, letters, maps, and political cartoons) as evidence to formulate a written argument in response to a question prompt. Twentieth-century history, of course, is documented by film and broadcasts, as well as by print. Computers can accommodate all the artifacts used on the conventional test as well as historical

FIGURE 1: A QUESTION USING A MULTIMEDIA STIMULUS TO TEST HISTORICAL ANALYSIS SKILLS



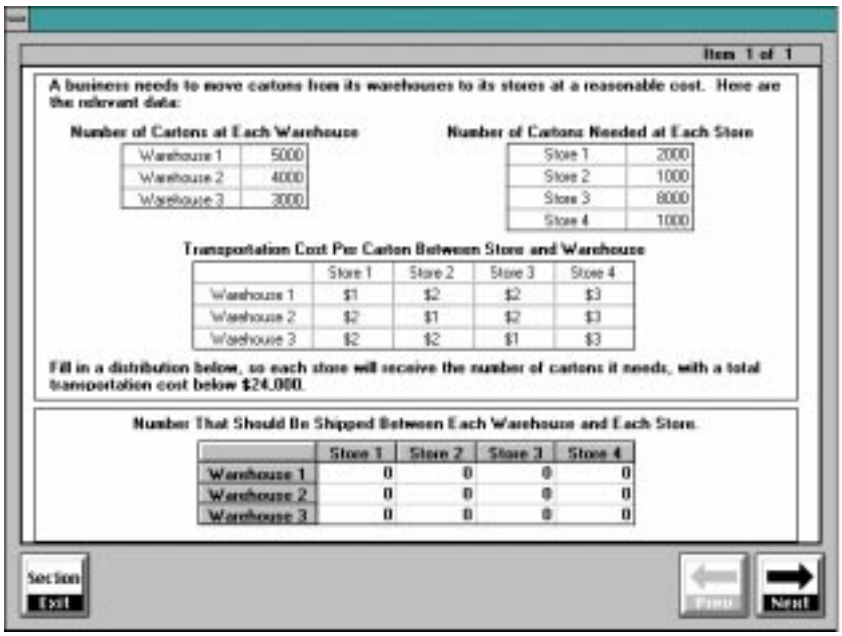
To play a selection the examinee clicks on one of the four labeled buttons under the video window and then on the sideways triangle just above "Who Died?". Responses are typed into the lower-left entry box (which shows a partially complete answer). From *Using multimedia in large-scale computer-based testing programs* (RR-97-3), by R. E., Bennett, M. Goodman, J. Hessinger, J. Ligget, G. Marshall, H. Kahn, & J. Zack, 1997, Princeton, NJ: Educational Testing Service. The scene depicted above comes from the CD-ROM, "Powers of Persuasion: The Art of Propaganda in World War II," produced by Fife and Drum Software, Silver Spring, MD, from records of the National Archives, Washington, D.C.

films, TV, and radio broadcasts, thereby extending the range of source material with which students must be proficient (see Figure 1).

At the same time as multimedia is used to extend measurement of traditional skills, it will be used to measure new skills. In both paper and computerized tests, we often assess skill in getting information from print. We do this assessment because we consider reading critical to success in school, in most jobs, and in activities of daily living. The importance of electronic media in communicating information is clearly growing. (Witness the fact that most Americans get their news from TV and also note the rapid ascent of the World Wide Web.) Consequently, we will increasingly expect students to be able to process information from a variety of sources. Given this expectation, perhaps we should evaluate not only how effectively people handle print but how well they reason with information from film, radio, TV, and computers.

In addition to changes in the nature of test questions, response formats will shift dramatically (Bennett, 1993). Some items on these next-generation tests will be much the same type of short, single-best-answer problem as found on today's multiple-choice tests (but without the response options). Other questions will be somewhat lengthier and perhaps less well-determined — the kind of problem in which one is looking not for the best answer, but only for a reasonable one given certain constraints (see Figure 2). Still

FIGURE 2: A CONSTRUCTED-RESPONSE PROBLEM WITH SEVERAL REASONABLE ANSWERS



Responses are typed into the bottom matrix. Fully creditable answers must meet all stated conditions.

other problems will be completely open-ended; for instance, the education community's insistence on testing writing performance will compel most programs to include an essay component that may, as students move universally to word processing, be offered only on computer. Tasks calling for other types of performance — oral presentation, sign language, the display of an artifact —will also be administered routinely, with responses captured digitally through computer-controlled microphones and TV cameras (Bennett, in press-a). Short simulation exercises will also appear.[3]

Along with redefining the test question will come the reengineering of test development, scoring, and administrative processes, all driven primarily by the need to improve productivity. With continuous testing, item pools for high-stakes programs have to be replenished constantly, otherwise security may be compromised. Replenishing the pool involves not only writing new questions but also pretesting them to generate the statistical calibrations used in adaptive testing.

To help create and calibrate items at the required rate, new tools will emerge (Singley & Bennett, 1995). These tools will allow developers to construct question templates (or select them from large libraries) and then vary both surface elements tangential to solution as well as deeper structural elements. Variants of questions will be created based on theories of item difficulty drawn from cognitive psychology (e.g., Kirsch & Mosenthal, 1990; Sebrechts, Enright, Bennett, & Martin, 1997). Using these theories, item parameters will be estimated precisely enough to reduce significantly the sample sizes needed for pretesting (Mislevy, Sheehan, & Wingersky, 1993; Sheehan & Mislevy, 1994).

Item generation tools may also eventually help in specifying test designs — the type, organization, number, and parameters of tasks needed to achieve a specific result. Once a test design is specified, these tools could automatically suggest which question templates to use. A variety of standard designs derived from cognitive principles might exist from which the developer could select or create a new design, thereby pushing tests toward formulations based more solidly on cognitive theory (e.g., see Bennett & Sebrechts, 1997).

At some point the tools could well become sophisticated enough to generate items without human intervention. Adaptive tests would no longer be built by creating pools from which items were selected. Items would be created on the fly, each one fashioned to a particular specification at the moment of administration (Bejar, 1993, 1996).

Test design will also be the focal point for responding to diversity (though the most far-reaching design changes for this purpose

[3] Why not long domain-based simulations? The economics of large-scale educational testing in dedicated centers won't permit it. Such simulations will become a standard part of occupational and professional assessment. This field already employs lengthy tests, sometimes taking days instead of hours, and can more easily charge the fees required to cover the considerable costs that simulations incur.

will be yet to come). The effects of different test designs on minority group members, females, and older examinees will be routinely simulated in deciding what skills and which task formats to use in large-scale assessments (Willingham & Cole, 1997). (The recent addition of an essay to the PSAT-NMSQT is an early, if post hoc, example.)

Also, as the complexity of responses required by computer-delivered performance tasks increases, test designers will find it particularly challenging to build examinee interfaces that are equally easy for all (Bennett, in press-b). The need to achieve such a result will make systematic interface planning and evaluation, as well as the development of more extensive preparation materials, central to equitable test design.

The need to develop more extensive preparation materials to compensate for increasing response complexity will, itself, force innovation. Test preparation software will incorporate intelligent tutoring methods (Wenger, 1987), that enable test takers to master new interfaces and test content more rapidly. The former purpose will recede in prominence as more natural ways of interacting with computers become conventional. Employing sophisticated technological tools to teach educational substance, however, will play an important role in the reconceptualization of large-scale testing described in the next section.

Shifts in test design, and especially changes in the nature of test questions, will cause scoring processes to alter considerably too. Many constructed-response tasks that are now graded manually will be processed automatically (e.g., Bennett, Steffen, Singley, Morley, & Jacquemin, 1997; Page & Petersen, 1995; Sebrechts, Bennett, & Rock, 1991). The character of the responses will be simple at first — mathematical expressions, equations, and graphs; words, phrases, and single sentences — but will eventually encompass more complex performances, including essays (e.g., Burstein et al., 1998). Responses that cannot be graded automatically will be handled collaboratively by humans and computers (Bennett, in press-a).

Obviously, human judges will also retain purview over those tasks that are not computer deliverable — some important skills won't soon be amenable to assessment in this medium (e.g., in the performing arts) — but even those tasks that can be electronically administered may not be delivered that way until all the requisite technological pieces are in place. Where responses can be recorded on paper, they will be scanned and digitized. Other responses will be digitally recorded by computer-controlled camera or microphone. Once in digital form, the encoded results will be sent electronically to human judges who will grade the performance on screen. Judges may be at the same general location as the examinee or elsewhere. If they are elsewhere,

they may work at the same site or at separate locations. If they work at separate sites, they may do their grading simultaneously or at different times. Software will train judges in the scoring rules, make work assignments, facilitate the judges' interactions with one another and with supervisors, introduce pre-calibrated benchmarks to maintain grading standards, and monitor any drift in standards that might occur.

As scoring and development processes evolve, there will be an inevitable linkage of the two. Software tools will help developers specify and test automatic scoring keys as questions are composed. Automatic item generation will bring with it automatic key generation: the templates used to spawn variants of questions and estimate their parameters will simultaneously define the features of correct responses that scoring engines will need to do their grading.

The last notable change will be administrative. The establishment of international systems of test centers will cause a shift from the dedicated electronic networks of the first generation to the Internet. This shift will be motivated by cost efficiencies, improvements in Internet data security, and improved transmission capability, allowing huge amounts of data to be exchanged and manipulated quickly. Competition among testing agencies will lead to alternative test-center networks, which should help fuel innovation, productivity, customer service, and greater response to diversity.[4]

One result is that test makers will forge new public-private partnerships, opening school-based centers to supplement those already in commercial establishments. A second result is that customer interactions with testing companies will be entirely electronic: students will register, get preparation materials, practice, explore institutions of higher education, apply, receive scores, send scores and performance samples, and resolve service problems all via the Internet.

In sum, several key transformations will define this next generation of large-scale tests. These transformations will be in the character of questions, development and scoring processes, test design, and test center networks. Most notably, the ability to deliver multimedia questions, capture and score complex constructed responses, create tasks efficiently, and move and manipulate large amounts of data electronically will make performance assessment a vital, if not principal, element of large-scale testing. Finally, although the administrative aspects of testing will improve for test takers, in overall purpose the enterprise will still be institutionally driven.

[4] In planning or already established are test center networks run by ETS/Sylvan Learning Systems, National Computer Systems/VUE, and Harcourt-Brace Educational Measurement/Assessment Systems, Inc.

Generation "R"

In this third generation, testing will reinvent itself, breaking radically with tradition in several ways.

One hallmark will be the emergence of interactive environments that facilitate individual growth in addition to serving the accountability functions normally fulfilled by large-scale tests. To achieve this end, large-scale assessment will join with instruction, first at "arm's length" but in time commingling totally. Along with other forces, instructional integration will have profound effects, producing the eventual decline of conventional, one-time, center-administered examinations.

The need for new tests will be driven by an increasingly competitive global economy in which continuous learning becomes central to success for much of our population; the establishment of the Internet as a pivotal commercial and social structure, making the delivery of quality education more cost-effective; and the imperative to formulate a testing system that actively helps the nation's expected non-White majority succeed.

The quickly escalating cost of a traditional college education coupled with the convenience of electronic networks will establish distance learning as a dominant force in higher education. Distance learning will be international, permitting foreign students to enroll at U.S. institutions without leaving home and U.S. students to take courses abroad. Once established, this approach will migrate down to secondary school, permitting more students to learn in neighborhood institutions from world-class teachers and curricular materials, move to a new community without interrupting their schooling, receive a home education, or return to pursue a General Equivalency Diploma (Owston, 1997).

Large-scale electronic distance examinations will play a key role in this scenario. Computerized assessments based on conventionally recognized standards (e.g., NCTM, 1989), will be embedded in the school curriculum and occur frequently. Sometimes an assessment will be made known in advance; at other times — with informed consent — it will simply be embedded seamlessly in the distance-learning session and be indistinguishable from the instructional components of that session. Decisions like certification of course mastery, graduation eligibility, and school effectiveness will no longer be based largely on one examination given at a single time but will also incorporate information from a series of measurements (Bennett, in press-a). College

admissions and placement may follow suit as the assessments used in standards-based high school courses like the College Board's Pacesetter program and its Advanced Placement program, migrate to this curriculum-embedded, distance model. (See Bunderson, Inouye, and Olsen, 1989, for a more elaborated, though not distance-based, conceptualization of continuous curriculum-embedded measurement.)

Such a model has profound implications for accommodating population diversity, in particular for addressing group differences in test performance. In the current model, where large-scale tests are too often divorced from schooling in both content and delivery, it is easy to become fixated on questioning the veracity of assessment. And, indeed, enormous energy has been invested — with relatively little return — in debating how well current tests reflect the skills of different population groups. By virtue of moving assessment into the curriculum, the locus of the debate over performance differences must logically shift from the *accuracy of assessment* to the *adequacy of instruction*.

By this generation, the influence of cognitive science will be more strongly evident, driving course and test design, and making possible closer articulation of assessment and instruction. For assessment in particular, design will be "theory-based," resting squarely on fundamental conceptions of the nature of subject-matter expertise and the structure of intellectual abilities (Everson, in press). These conceptions will help designers organize tests around descriptions of what skills are important to proficiency in a given field, how those skills are composed and interrelated, and how they might be trained.

This merger of assessment and instruction will be realized in some significant part through the use of electronic learning tools. These tools will implement a range of instructional methods. For present purposes, however, their most salient characteristic will be in leaving an electronic record of student activity that might contribute almost incidentally to summative decision making.

For example, intelligent tutors (Wenger, 1987), microworlds (Shute & Glaser, 1990), and simulations (Schank & Cleary, 1995, chap. 5) exemplify environments that could be used for delivering instruction and certifying competence in specific skill areas (much the way flight simulators are used for these dual purposes in aviation today). With intelligent tutors particularly, student knowledge will be dynamically modeled using cognitive and statistical approaches capable both of guiding instruction on a highly detailed level and of providing a general summary of overall standing (e.g., Mislevy & Gitomer, 1996). Instruction will be adapted not only to the multiple dimensions that characterize standing in a broad skill area, but to personal interests and background, allowing more meaningful accommodation to diversity than was possible with earlier approaches.

While these tools will typically be used by individuals in isolation, other tools will be oriented toward collaborative work undertaken in electronic learning communities (Gordin, Gomez, Pea, & Fishman, 1996). Such communities should permit students and teachers from different schools to exchange ideas and jointly carry out projects. Mentoring may also be provided by practitioners who share expertise via the Internet to help inculcate the methods, mores, and social organization of a field. For summative assessment purposes, a common framework might be developed that allows considerable latitude in the choice of group projects and, at the same time, comparable scores across the differing results (Mislevy, 1993b; Myford & Mislevy, 1995). Assuming the ability to grade even the most complex projects remotely and assignment of the same grade to all members of a project team, work in learning communities might incidentally serve large-scale assessment purposes too.[5]

The tasks associated with electronic learning and assessment tools will be very different from those of prior eras. For instance, simple multimedia exercises will give way to virtual reality simulations. These simulations will model complex environments — science labs, field experiences — giving students a chance to learn and be assessed under conditions similar to those encountered by practitioners. (Again, today's flight simulators suggest the type of educational environment that could develop.)

In addition to improving the realism of questions, Generation "R's" computers will let individuals respond more naturally. For example, they may respond directly to the student's physical actions in virtual reality simulations. Speech also will be accurately understood, as the additional information in lip movements is employed to reduce ambiguity (Gates, 1996). This understanding will further aid diversity by allowing instruction and its embedded assessment to be delivered through multiple presentation modes (e.g., by presenting the same information in text and through audio), so that all students (but especially those with disabilities) can use computers more easily.

What skills will these new learning and assessment tasks target? The proliferation of sophisticated information and communication technologies will undoubtedly influence what skills are valued, taught, and assessed. Many valued skills will be the same as those we consider important today, but some will be new. With continuous learning required almost universally, interest may reemerge in

[5] Work in virtual communities may someday come to dominate education, as well as the world of work. But collaboration will not eliminate the need to assess individuals. Our political and economic systems value individual accomplishment, reward, and accountability — even in the presence of group assignment (e.g., note the common distribution of individual awards and sanctions in team sports). As long as those ingrained values remain, assessing the individual will continue to be important in educational decision making.

cognitive modifiability, or how effectively an individual benefits from instruction (Lidz, 1987). Also, the growth of electronic learning and work communities will make remote collaborative skills essential. Finally, in a knowledge-rich environment, we will have easy, cheap, and almost instantaneous access to information. The ability to pose the right questions and find, analyze, and organize relevant knowledge may take on increased importance. Intelligent agents will be widely available to help us do some of these tasks (Gilbert, undated), making one's deftness in deploying virtual assistants a critical skill.

Where will this distance learning and assessment occur? It will certainly occur in the home and for the foreseeable future in schools, commercial learning establishments, libraries, businesses, and community centers. To the extent that the Internet becomes universal, some of these institutions may transform to virtual entities. Electronic networks will likely reduce the need for physical libraries (Lesk, 1997) and, perhaps in the long term, for schools as we know them.

Dedicated test centers also may be on the endangered list. To the extent that some form of embedded assessment works effectively as a basis for summative decision making, the large-scale, one-time, test-center-delivered examination may diminish. The latter model may be weakened further as miniaturized, wireless broadcast technologies are used to break test security (Colton, 1997). These technologies will pose such new threats as the electronic pilfering of items from the radio frequency signals emitted by screen displays and the real-time coaching of test takers by expert compatriots who remotely read screens and immediately broadcast the correct answers back to examinees.[6] The almost continuous (and often unobtrusive) nature of assessment in the curriculum-embedded model should make such high-tech gamesmanship considerably less practical.

What does a curriculum-embedded, distance assessment model assume? Certainly, it assumes a common set of content standards for educational attainment. Such standards have been promulgated in most subject areas (e.g., NCTM, 1989). Second, it requires setting performance standards and developing measurements that are largely coterminous with instruction. The Advanced Placement program, Pacesetter, and the New Standards Project represent early, paper-based moves in this direction; the University of the State of New York's Regents College, a virtual university that certifies knowledge and grants course credits and degrees through distance assessment is a post-secondary-level example; and work on electronic learning tools that can incidentally provide global proficiency estimates is a complementary development. Finally, it demands an electronic infrastructure for delivery and management, the foundations of which are already emerging.

[6] The basic technology for CRT emanation monitoring already exists. van Eck (1985) wrote the classic scientific paper in the field and Behar (1997, pp. 66, 70) gives a more recent popular description.

Conclusion

This paper presented one scenario for the future of large-scale educational assessment. The scenario divides into three generations (summarized in Table 1). In recent years, large-scale assessment has changed relatively little in purpose, administration mode, use of technology, and scientific grounding. This situation is about to alter because competition will force test makers to satisfy new market needs through innovation, improve productivity, enhance customer-service, and address population diversity. In response, large-scale assessment will come to serve both summative and formative purposes, be curriculum-embedded and performance-based, occur at a distance, and measure new competencies as the skills valued by society change. New technology — along with advances in cognitive and measurement science — will be the chief catalyst in reaching these goals.

Although it is attractive to emphasize technology and the possibilities it engenders, our focus should remain on large-scale educational assessment and the needs it must satisfy. Right now, large-scale assessment runs the risk of falling out of synch on multiple counts, including its relevance to educational decision-making and its ability to accommodate diversity. Change, unfortunately, does not always come easily or as quickly as it should to well-established institutions. But large-scale educational assessment must change in the most fundamental ways, for nothing short of reinvention will prepare it to meet the dramatically different demands it will soon face.[7]

[7] Some readers may find the scenario presented in this report implausible. But what may seem incomprehensible today can quite abruptly become reality. As an example, take global geo-politics, which is almost certainly far more complex than large-scale testing. Had one asserted 15 years ago that the Berlin Wall would fall, the Soviet Union collapse, Germany reunite, and the Cold War end, the reaction would have been disbelieving, to say the least!

Table 1: Three Generations of Large-Scale Educational Assessment	
Generation	**Key Characteristics**
First-Generation Computer-Based Tests (Infrastructure Building)	1. Primarily serve institutional needs 2. Measure traditional skills and use test designs and item formats closely resembling paper-based tests, with the exception that tests are given adaptively 3. Administered in dedicated test centers as a "one-time" measurement 4. Take limited advantage of technology
Next-Generation Electronic Tests (Qualitative Change)	1. Primarily serve institutional needs 2. Use new item formats (including multimedia and constructed response), automatic item generation, automatic scoring, and electronic networks to make performance assessment an integral program component; measure some new constructs 3. Administered in dedicated test centers as a "one-time" measurement 4. Allow customers to interact with testing companies entirely electronically
Generation "R" Tests (Reinvention)	1. Serve both institutional and individual purposes 2. Integrated with instruction via electronic tools so that performance is sampled repeatedly over time; designed according to cognitive principles 3. Use complex simulations, including virtual reality, that model real environments and allow more natural interaction with computers 4. Administered at a distance 5. Assess new skills

References

Behar, R. (1997, February 3). Who's reading your e-mail? *Fortune, 135(2)*, 56-70.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksen, R. J. Mislevy, & I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-357). Hillsdale, NJ: Erlbaum.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium* (RR-96-13). Princeton, NJ: Educational Testing Service.

Bennett, R. E. (1993). On the meanings of constructed response. In R. E. Bennett & W. C. Ward (Eds.), *Construction vs. choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 1-27). Hillsdale, NJ: Erlbaum.

Bennett, R. E. (in press-a). An electronic infrastructure for a future generation of tests. *Journal of Learning and Evaluation*.

Bennett, R. E. (in press-b). Computer-based testing for examinees with disabilities: On the road to generalized accommodation. In S. Messick (Ed.), *Assessment in higher education: Issues of access, student development, and public policy*. Hillsdale, NJ: Erlbaum.

Bennett, R. E., Goodman, M., Hessinger, J., Ligget, J., Marshall, G., Kahn, H., & Zack, J. (in press). Using multimedia in large-scale computer-based testing programs. *Computers in Human Behavior*.

Bennett, R. E., & Sebrechts, M. M. (1997). Measuring the representational component of quantitative proficiency. *Journal of Educational Measurement, 34*, 62-75.

Bennett, R. E., Steffen, M., Singley, M. K., Morley, M., & Jacquemin, D. (1997). Evaluating an automatically scorable, open-ended response type for measuring mathematical reasoning in computer-adaptive tests. *Journal of Educational Measurement, 34*, 163-177.

Bunderson, C. V., Inouye, D. K., Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn, *Educational measurement* (3rd Ed.). New York: ACE/MacMillan.

Burstein, J. C., Braden-Harder, L., Chodorow, M., Hua, S., Kaplan, B., Kukich, K., Lu, C., Nolan, J., Rock, D., & Wolff, S. (1998). *Computer analysis of essay content for automated score prediction* (RR-98-15). Princeton, NJ: Educational Testing Service.

Colton, G. D. (1997, March). *High-tech approaches to breaching examination security*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Everson, H. T. (in press). A theory-based framework for future college admissions tests. In S. Messick (Ed.), *Assessment in higher education: Issues of access, student development, and public policy*. Hillsdale, NJ: Erlbaum.

Gates, B. (1996). *The road ahead*. New York: Penguin.

Gilbert, D. (undated). *IBM intelligent agents* [On-line]. Available: http://www.networking.ibm.com/iag/iagwp1.html.

Glaser, R. (1991). Expertise and assessment. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 17-30). Englewood Cliffs, NJ: Prentice-Hall.

Gordin, D. N., Gomez, L. M., Pea, R. D., & Fishman, B. J. (1996). Using the World Wide Web to build learning communities in k-12. *Journal of Computer-Mediated Communications* [On-line], 12. Available: http://www.usc.edu/dept/annenberg/vol2/issue3/gordin.html.

Kirsch, I. S., & Mosenthal P. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 15*, 5-30.

Lesk, M. (1997). *Practical digital libraries*. San Francisco: Morgan Kaufmann.

Lidz, C. S. (1987). *Dynamic assessment: An interactional approach to evaluating learning potential*. New York: The Guilford Press.

Mislevy, R. J. (1993a). Foundations of a new test theory. In N. Frederiksen, R. J. Mislevy, and I. Bejar (Eds.), *Test theory for a new generation of tests*. Hillsdale, NJ: Erlbaum.

Mislevy, R. J. (1993b). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: Policy Information Center, Educational Testing Service. (ERIC # ED-353-302)

Mislevy, R. J. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*, 379-416.

Mislevy, R. J., & Gitomer, D. H. (1996). The role of probability-based inference in an intelligent tutoring system. *User-Modeling and User-Adapted Interaction*, 5, 253-282.

Mislevy, R. J., Sheehan, K. M., & Wingersky, M. S. (1993). How to equate tests with little or no data. *Journal of Educational Measurement, 30*, 55-78.

Myford, C. M., & Mislevy, R. J. (1995). *Monitoring and improving a portfolio assessment system* (Center for Performance Assessment Research Report). Princeton, NJ: Educational Testing Service.

National Assessment Governing Board. (1996). *Policy statement on redesigning the National Assessment of Educational Progress*. Washington, D.C.: National Assessment Governing Board.

National Council of Teachers of Mathematics (NCTM). (1989). *Curriculum and evaluation standards for school mathematics*. Reston, VA: National Council of Teachers of Mathematics.

Office of Technology Assessment. (1992). *Testing in America's schools: Asking the right questions*. Washington, D.C.: Office of Technology Assessment, Congress of the United States.

Owston, R. D. (1997). The World Wide Web: A technology to enhance teaching and learning? *Educational Researcher, 26*(*2*), 27-33.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading. *Phi Delta Kappan, 76*, 561-565.

Pellegrino, J. W. (1992). Commentary: Understanding what we measure and measuring what we understand. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 275-300). Boston: Kluwer.

Resnick, L. B., & Resnick, D. P. (1990). Tests as standards of achievement in schoools. In J. Pfleiderer (Ed.), *Proceedings of the 1989 ETS Invitational Conference: The uses of standardized tests in American education* (pp. 63-80). Princeton, NJ: Educational Testing Service.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 37-75). Boston: Kluwer.

Schank, R. C., & Cleary, C. (1995). *Engines for education*. Hillsdale, NJ: Erlbaum.

Sebrechts, M. M., Bennett, R. E., & Rock, D. A. (1991). Agreement between expert system and human raters' scores on complex constructed-response quantitative items. *Journal of Applied Psychology, 76*, 856-862.

Sebrechts, M. M., Enright, M., Bennett, R. E., & Martin, K. (1997). Using algebra word problems to assess quantitative ability: Attributes, strategies, and errors. *Cognition and Instruction, 14,* 285-343.

Sheehan, K. & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills*. (RR-94-14). Princeton, NJ: Educational Testing Service.

Shepard, L. A. (1992). Commentary: What policy makers who mandate tests should know about the new psychology of intellectual ability and learning. In B. R. Gifford & M. C. O'Connor (Eds.), *Changing assessments: Alternative views of aptitude, achievement and instruction* (pp. 301-328). Boston: Kluwer.

Sherman, H. (1995). The new world economy. In The Conference Board, (Ed.), *Business and economic outlook: How will the century close* (Report No. 1104-95-CH) (pp. 9-11). New York: The Conference Board.

Shute, V. J., & Glaser, R. (1990). A large-scale evaluation of an intelligent discovery world: Smithtown. *Interactive Learning Environments, 1*, 51-77.

Singley, M. K., & Bennett, R. E. (1995). *Toward computer-based performance assessment in mathematics* (RR-95-34). Princeton, NJ: Educational Testing Service.

Sternberg, R. J. (1991). Toward better intelligence tests. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 31-39). Englewood Cliffs, NJ: Prentice-Hall.

Sternberg, R. J. (1992). Ability tests, measurements, and markets. *Journal of Educational Psychology, 192*, 84, 134-140.

Treacy, M., & Wiersema, F. (1995). *The discipline of market leaders*. *Reading*, MA: Addison-Wesley.

van Eck, W. (1985). Electromagnetic radiation from video display units: An eavesdropping risk? *Computers & Security, 4*, 269-286.

Wenger, E. (1987). *Artificial intelligence and tutoring systems*. Los Altos, CA: Morgan Kaufmann.

Willingham, W. W., & Cole, N. S. (Eds.). (1997). *Gender and fair assessment*. Hillsdale, N. J.: Erlbaum.